ENSAE Doctoral Course – Lecture Notes

Insurance, Biases Discrimination & Fairness

Arthur Charpentier, July 2, 2023, v2

Contents

Chapter 1

Introduction

While the algorithms of machine learning methods have brought issues of discrimination and fairness back to the forefront, these topics have been the subject of an extensive literature over the past decades. But dealing with discrimination in insurance is fundamentally an ill-defined unsolvable problem. Nevertheless, we will try to connect the dots, to explain different perspectives, going back to the legal, philosophical and economic approaches to discrimination, before discussing the so-called concept of "actuarial fairness." We will offer some definitions, before introducing the book and the data used in the illustrative examples throughout the chapters.

1.1 A Brief Overview on Discrimination

1.1.1 Discrimination?

Definition 1.1.1 (Discrimination) *Merriam-Webster (2022). Discrimination is the act, practice, or an instance of separating or distinguishing categorically rather than individually.*

In this book, we will use this neutral definition of "discrimination". Nevertheless, Kroll et al. (2017) reminds us that the word "discrimination" carries a very different meaning in statistics and computer science, than it does in public policy. "*Among computer scientists, the word is a value-neutral synonym for differentiation or classification: a computer scientist might ask, for example, how well a facial recognition algorithm successfully discriminates between human faces and inanimate objects. But, for policymakers, "discrimination" is most often a term of art for invidious, unacceptable distinctions among people-distinctions that either are, or reasonably might be, morally or legally prohibited*." The word discrimination can then be used both in a purely descriptive sense (in the sense of making distinctions, as in this book), or in a normative manner, which implies that the differential treatment of certain groups is morally wrong, as shown by Alexander (1992), or more recently Loi and Christen (2021). To emphasize the second meaning, we can prefer the word "prejudice", that refers to an "*unjustifiable negative attitude*" (Dambrum et al. (2003) and Al Ramiah et al. (2010)) or an "*irrational attitude of hostility*" (Merriam-Webster (2022)) toward a group and its individual members. The definition of "discrimination" given in Correll et al. (2010) can be related to this one "*behaviour directed towards category members that is consequential for their outcomes and that*

is directed towards them not because of any particular deservingness or reciprocity, but simply because they happen to be members of that category." Here, the idea of "unjustified" difference is mentioned. But what if the difference can somehow be justified ? The notion of "merit" is key to the expression and experience of discrimination (we will discuss this in relation to ethics later). It is not an objectively defined criterion, but one rooted in historical and current societal norms and inequalities.

Avraham (2017) explained in one short paragraph the dilemma of considering the problem of discrimination in insurance. "*What is unique about insurance is that even statistical discrimination which by definition is absent of any malicious intentions, poses significant moral and legal challenges. Why? Because on the one hand, policy makers would like insurers to treat their insureds equally, without discriminating based on race, gender, age, or other characteristics, even if it makes statistical sense to discriminate* (...) *On the other hand, at the core of insurance business lies discrimination between risky and non-risky insureds. But riskiness often statistically correlates with the same characteristics policy makers would like to prohibit insurers from taking into account*." To illustrate this problem, and why writing about discrimination and insurance could be complicated, let us consider the example of "redlining" . Redlining has been an important issue (that we will discuss further in section 6.1.2), for the credit and the insurance industries, in the U.S., that started in the 30's.

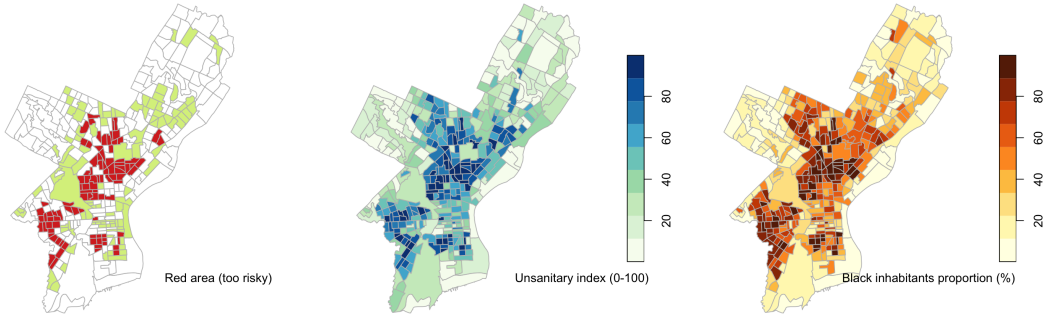


Figure 1.1: Map (freely) inspired by a Home Owners' Loan Corporation map from 1937, where red is used to identify neighborhoods where investment and lending were discouraged, on the left (see Crossney (2016) and Rhynhart (2020)). In the middle, some risk related variable (an fictitious "unsanitary index") per neighborhood of the city is presented, and on the right, a sensitive variable (the proportion of Black people in the neighborhood, again, freely created). Those maps are fictitious.

In 1935, the Federal Home Loan Bank Board (FHLBB) looked at more than 200 cities and created "*residential security maps*" to indicate the level of security for real-estate investments in each surveyed city. On the maps (see Figure 1.1 with a collection of fictitious maps), the newest areas—those considered desirable for lending purposes—were outlined in green and known as "Type A". "Type D" neighborhoods were outlined in red and were considered the most risky for mortgage support (on the left of Figure 1.1). Those areas were indeed those with a high proportion of dilapidated (or dis-repaired) buildings (as we can observe on the right of Figure 1.1). In the 70's, when looking at census data, sociologist noticed that red area, where insurers did not want to offer coverage, were also those with a high proportion of Black people, and following the work John McKnight and Andrew Gordon, "redlining" received more interest. On the map in the middle, we can observe information about the proportion of Black people. Thus, on the hand, it could be seen as "legitimate" to have a premium for household that could reflect somehow the general conditions

Box 1.1 **"*unfairly discriminatory*" insurance rates, according to U.S. legislation**

Arkansas law (23/3/67/2/23-67-208), 1987 *"a rate is not unfairly discriminatory in relation to another in the same class of business if it reflects equitably the differences in expected losses and expenses. Rates are not unfairly discriminatory because different premiums result for policyholders with like loss exposures but different expense factors, or with like expense factors but different loss exposures, if the rates reflect the differences with reasonable accuracy (...) A rate shall be deemed unfairly discriminatory as to a risk or group of risks if the application of premium discounts, credits, or surcharges among the risks does not bear areasonable relationship to the expected loss and expense experience among the various risks*."

Maine Insurance Code (24-A, 2303), 1969 *"Risks may be grouped by classifications for the establishment of rates and minimum premiums. Classification rates may be modified to produce rates for individual risks in accordance with rating plans that establish standards for measuring variations in hazards or expense provisions, or both. These standards may measure any differences among risks that may have a probable effect upon losses or expenses. No risk classification may be based upon race, creed, national origin or the religion of the insured (...) Nothing in this section shall be taken to prohibit as unreasonable or unfairly discriminatory the establishment of classifications or modifications of classifications or risks based upon size, expense, management, individual experience, purpose of insurance, location or dispersion of hazard, or any other reasonable considerations, provided such classifications and modifications apply to all risks under the same or substantially similar circumstances or conditions*."

of houses. On the other hand, it would be discriminatory to have a premium that is function of the ethnic's origin of the policyholder. The neighborhood, the "unsanitary index" and the proportion of Black people are here strongly correlated variables. Of course, there could be non-black people living in dilapidated houses outside of the red area, black people living in wealthy houses inside the red area, etc. If we work using aggregated data, it is difficult to disentangle information about sanitary conditions and racial information, to distinguish "legitimate" and "non-legitimate" discrimination, as discussed in Hellman (2011). Observe that in the case of "redlining" census and aggregated data are used, leading possibly to some "ecological fallacy" (as discussed in King et al. (2004) or Gelman (2009)). In the 2020's, we now have much more information (so called "big data era") and more complex models (machine learning literature), and we will see how to disentangle this complex problem, even if dealing with discrimination in insurance is probably still an ill-defined unsolvable problem, with strong identification issues. Nevertheless, as we will see, there are many way of looking at this problem, and we will try, here, to connect the dots, to explain different perspectives.

1.1.2 Legal Perspective on Discrimination

In Kansas, more than one hundred years ago, a law passed, allowing an insurance commissioner to review rates to ensure that they were not "*excessive, inadequate, or unfairly discriminatory with regards to individuals*", as mentioned in Powell (2020). Since them , the idea of "*unfairly discriminatory*" insurance rates has been discussed in many States, in the U.S. (see Box 1.1).

Unfortunately, as recalled in Vandenhole (2005), there is "*no universally accepted definition of discrimination*," and most legal documents usually provide (non-exhaustive) lists of the grounds on which discrimination is to be prohibited. For example, in the International Covenant on Civil and Political Rights, "*the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status*" (see Joseph and Castan (2013)). Such lists do not really

address the question of what discrimination really is. But looking for common features among those variables can be used to explain what discrimination is. For instance, discrimination is necessarily oriented toward some people based on their membership in a certain type of social group, with reference to a comparison group. Therefore, we should not discuss how well (or poorly) a person in a certain group is treated on some absolute scale, but rather how well she is treated relative to some other person who could be seen as "similar", but in the reference group. And again, this reference group is important since discrimination is not simply some differential treatment, that would be symmetrical: to have discrimination, there should be a favored and an non-favored group. As Altman (2011), wrote, "*as a reasonable first approximation, we can say that discrimination consists of acts, practices, or policies that impose a relative disadvantage on persons based on their membership in a salient social group.*"

1.1.3 Discrimination from a Philosophical Perspective

As mentioned already, we should not expect to have universal rules about discrimination. As mentioned in Hellman (2011), Supreme Court Justice Thurgood Marshall claimed once that "*a sign that says 'men only' looks very different on a bathroom door than on a courthouse door*." Nevertheless, philosophers have suggested definitions, starting with a distinction between "direct" and "indirect" discrimination. As mentioned in Lippert-Rasmussen (2014), it would be too simple to consider direct discrimination as intentional discrimination. A classical example would be a paternalistic employer who intend to help women by hiring them only for certain jobs, or for a promotion, as discussed in Jost et al. (2009). In that case, acts of direct discrimination can be unconscious in the sense that agents are unaware of the discriminatory motive behind decisions (related to the "*implicit bias*" discussed in Brownstein and Saul (2016a,b). Indirect discrimination corresponds to decisions with disproportionate effects, that might be seen as be discriminatory even if that is not the objective of the decision process mechanism. A standard example could be the one where the only way to enter a public building is by a set of stairs, that could be seen as a discrimination against people with disabilities who use wheelchairs, since they would be unable to enter the building; or if there would be a minimum height requirement for a job where height is not relevant, that could be seen as a discrimination against women, since they are generally shorter than men. On the one hand, for Young (1990), Cavanagh (2002) or Eidelson (2015), indirect discrimination should not be considered as discrimination, that should be strictly limited to "*intentional and explicitly formulated policies of exclusion or preference*." For Cavanagh (2002), in many cases, "*it is not discrimination they object to, but its effects; and these effects can equally be brought about by other causes*". On the other hand, Rawls (1971) considered structural indirect discrimination, that is when the rules and norms of society consistently produce disproportionately disadvantageous outcomes for the members of a certain group, relative to the other groups in society. Even it it is not intentional, it should be considered as discriminatory.

Let us get back to the moral grounds, to examine why discrimination is considered as wrong. According to Kahlenberg Richard (1996), racial discrimination should be considered "unfair" because it is associated with an immutable trait. Unfortunately, Boxill (1992) recalls that with such a definition, it would be also unfair to denying blind people a driver's license. And religion will challenge most definitions, since it is neither an immutable trait, nor a form of disability. Another approach is to claim that discrimination is wrong because it treats persons on the basis of inaccurate generalisations and stereotypes, as suggested by Schauer (2006). For Kekes (1995), treating a person only because she is a member of a certain social group is inherently unfair, since stereotyping treats people unequally "*without rational justification*." Thus, according to Flew (1993), racism is unfair because it treats individuals on the basis of traits that "*are strictly superficial and properly irrelevant to all, or almost all, questions of social status and employability.*" In other words, discrimination is wrong because it fails to treat individuals based on their merits. But in that case,

as Cavanagh (2002) observed, "*hiring on merit has more to do with efficiency than fairness*," that we will discuss further in the next section, on economic foundations of discrimination's. Finally, Lippert-Rasmussen (2006) and Arneson (1999, 2013) suggested to look at discrimination based on some consequentialist moral theory. In this approach, discrimination is wrong because it violates a rule that would be part of the social morality that maximizes overall moral value. Arneson (2013) writes that this view "*can possibly defend nondiscrimination and equal opportunity norms as part of the best consequentialist public morality.*"

A classical philosophical notion close to the idea of "non-discrimination" is the concept of "equality of opportunity". For Roemer and Trannoy (2016) "equality of opportunity" is a political ideal that is opposed to assigned-at-birth (caste) hierarchy, but not to hierarchy itself. To illustrate this point, consider the extreme case of caste hierarchy, where children acquire the social status of their parents. In contrast, "equality of opportunity" demands that the social hierarchy is determined by a form of equal competition among all members of the society. Rawls (1971) uses "equality of opportunity" to address the discrimination problem: everyone should be given a fair chance at success in a competition. This is also called "*substantive equality of opportunity*," and it is often implemented through metrics such as statistical parity and equalized odds, which assume that talent and motivation are equally distributed among sub-populations. This concept can be distinguished from the "substantive equality of opportunity", as defined in Segall (2013), where a person's outcome should be affected only by their choices, not their circumstances.

1.1.4 From Discrimination to Fairness

Humans have an innate sense of fairness and justice, with studies showing that even three-year-old children have demonstrated the ability to consider merit when sharing rewards, as shown by Kanngiesser and Warneken (2012), as well as chimpanzees and primates, Brosnan (2006), and many other animal species. And given that this trait is largely innate, it is difficult to define what is "fair," although many scientists have attempted to define notions of "fair" sharing, as Brams et al. (1996) recalls. On the one hand "fair" refers to legality (and to human justice, translated into a set of laws and regulations), and in a second sense, "fair" refers to an ethical or moral concept (and to an idea of natural justice). The second reading of the word "fairness" will be the most important here. According to one dictionary, fairness "*consists in attributing to each person what is due to him by reference to the principles of natural justice*." And being "just" raises questions related to ethics and morality (we will not differentiate here between ethics and morality).

This has to be related to a concept introduced in Feinberg (1970), called "*desert theory*," corresponding to the moral obligation that good actions must lead to better results. A student should deserve a good grade by virtue of having written a good paper, the victim of an industrial accident should deserve substantial compensation due to the negligence of his employer. For Leibniz or Kant, a person is supposed to deserve happiness in virtue for being morally good. In Feinberg (1970)'s approach, "*deserts*" are often seen as positive, but they are also sometimes negative, like fines, dishonor, sanctions, condemnations, etc (see Feldman (1995), Arneson (2007) or Haas (2013)). The concept of "*desert*" generally consists of a relationship between three elements: an agent, a deserved treatment or good, and the basis on which the agent is deserving.

We will evoke here the "ethics of models", or, as mentioned by Mittelstadt et al. (2016) or Tsamados et al. (2021), the "*ethics of algorithms*." A nuance exists with respect to the ethics of artificial intelligence, which deals with our behaviour or choices (as human beings) in relation to autonomous cars, for example, and which will attempt to answer questions such as "*should a technology be adopted if it is more efficient?*." The ethics of algorithms questions the choices made "by the machine" (even if they often reflect choices - or objectives - imposed by the person who programmed the algorithm).

Programming an algorithm in an ethical way must be done according to a certain number of standards. Two types of norms are generally considered by philosophers. The first is related to conventions, i.e. the rules

of the game (chess or Go), or the rules of the road (for autonomous cars). The second is made up of moral norms, which must be respected by everyone, and aim at the general interest. These norms must be universal, and therefore not favor any individual, or any group of individuals. This universality is fundamental for Singer (2011) who asks not to judge a situation with his own perspective, or that of a group to which one belongs to, but to take a neutral and "fair" point of view. Formally, in a typical way, in normative ethics, we would oppose consequentialism to deontologism.

As discussed previously, the ethical analysis of discrimination is related to the concept of "equality of opportunity," which holds that the social status of individuals depends solely on the service they can provide to society. As the second sentence of Article 1 of the 1789 Declaration of the Human Rights states "*les distinctions sociales ne peuvent être fondées que sur l'utilité commune*" (translated as[1] "*social distinctions may be founded only upon the general good*") or as Rawls (1971) points out, "*offhand it is not clear what is meant, but we might say that those with similar abilities and skills should have similar life chances. More specifically, assuming that there is a distribution of natural assets, those who are at the same level of talent and ability, and have the same willingness to use them, should have the same prospects of success regardless of their initial place in the social system, that is, irrespective of the income class into which they are born*." In the deontological approach, inspired by Emmanuel Kant, one forgets the utilities of each person, and simply imposes norms and duties. Here, regardless of the consequences (for the community as a whole), some things are not to be done. A distinction is typically made between egalitarian and proportionalist approaches. To go further, Roemer (1996, 1998) propose a philosophical approach, while Fleurbaey (1996) and Moulin (2004) consider an economic vision. And in a more computational context, Leben (2020) goes back to normative principles to assess the fairness of a model.

All ethics courses feature thought experiments, such as the popular "streetcar dilemma". In the original problem, stated in Foot (1967), a tram with no brakes is about to run over five people, and one of them has the opportunity to flip a switch that will cause the tram to swerve, but kill someone. What do we do? Or what should we do? Thomson (1976) suggested a different version, with a footbridge, where you can push a heavier person, who will crash into the track and die, but stop the tram. The latter version is often more disturbing because the action is indirect, and you start by murdering someone in order to save someone else. Some authors have used this thought experiment to distinguish between explanation (on scientific grounds, and based on causal arguments) and justification (based on moral precepts). This tramway experiment has been taken up in the moral psychology experiment, called, the *Moral Machine* project[2]. In this "game", one was virtually behind the wheel of a car, and choices were proposed: "*Do you run over one person or five people?*," "*Do you run over an elderly person or a child?*," "*Do you run over a man or a woman?*." Bonnefon (2019) revisits the experiment, and the series of moral dilemmas, where they obtained more than 40 million answers, from 130 countries. While naturally, numbers of victims were an important feature (we prefer to kill less people), age was also very important (priority to young people), and legal arguments seemed to emerge (we prefer to kill pedestrians who cross outside the dedicated crossings). These questions are important for self driving cars, as mentioned by Thornton et al. (2016).

For a philosopher, the question "*How fair is this model to this group?*" will always be followed by "*How fair by what normative principle?*." Measuring the overall effects on all those affected by the model (and not just the rights of a few) will lead to incorporating measures of fairness into an overall calculation of social costs and benefits. If we choose one approach, others will suffer. But this is the nature of moral choices, and the only responsible way to mitigate negative headlines is to develop a coherent response to these dilemmas, rather than ignore them. To speak of the ethics of models poses philosophical questions from which we cannot free ourselves, because, as we have said, a model aims to represent reality, "*what is*." To fight against

[1] See https://avalon.law.yale.edu/18th_century/rightsof.asp

[2] See https://www.moralmachine.net/

discrimination, or to invoke notions of fairness, is to talk about "*what should be*." We are once again faced with the famous opposition of Hume (1739). It is a well know properties of statistical models, as well as machine learning ones. As Chollet (2021) wrote it "*keep in mind that machine learning can only be used to memorize patterns that are present in your training data. You can only recognize what you've seen before. Using machine learning trained on past data to predict the future is making the assumption that the future will behave like the past*." For when we speak of "norm", it is important not to confuse the descriptive and the normative, or with other words, statistics (which tells us how things are) and ethics (which tells us how things should be). Statistical law is about "what is" because it has been observed to be so (e.g., *humans are bigger than dogs*). Human (divine, or judicial) law pertains to what is *is* because it has been decreed, and therefore *ought to be* (e.g. *humans are free and equal* or *humans are good*). One can see the "norm" as a regularity of cases, observed with the help of frequencies (or averages, as mentioned in the next chapter), for example, on the height of individuals, the length of sleep, in other words, data that makes up the description of individuals. Therefore, anthropometric data have made it possible to define, for example, an average height of individuals in a given population, according to their age; in relation to this average height, a deviation of 20% more or less determines gigantism or dwarfism. If we think of road accidents, it may be considered "abnormal" to have a road accident in a given year, at an individual (micro) level, because the majority of drivers do not have an accident. However, from the insurer's perspective (macro), the norm is that 10% of drivers have an accident. It would therefore be abnormal for no one to have an accident. This is the argument found in Durkheim (1897). From the singular act of suicide, if it is considered from the point of view of the individual who commits it, Durkheim tries to see it as a social act, therefore falling within a real norm, within a given society. From then on, suicide becomes, according to Durkheim, a "normal" phenomenon. Statistics then make it possible to quantify the tendency to commit suicide in a given society, as soon as one no longer observes the irregularity that appears in the singularity of an individual history, but a "social normality" of suicide. Abnormality is defined as "*contrary to the usual order of things*" (this might be considered an empirical, statistical notion), or "*contrary to the right order of things*" (this notion of right probably implies a normative definition), but also not conforming to the model. Defining a norm is not straightforward if we are only interested in the descriptive, empirical aspect, as actuaries do when they develop a model, but when a dimension of justice and ethics is also added, the complexity is bound to increase. We shall return in Chapter 4 to the (mathematical) properties that a "fair" or "equitable" model should be checked. Because if we ask a model to verify criteria not necessarily observed in the data, it is necessary to integrate a specific constraint into the model learning algorithm, with a penalty related to a fairness measure (just as we use "model complexity measure" to avoid overfit).

1.1.5 Economics Perspective on Efficient Discrimination

If jurists used the term "rational discrimination," economists used the term "efficient" or "statistical discrimination", such as in Phelps (1972) or Arrow (1973), following early work by Edgeworth (1922). Following Becker (1957) economists have tended to define discrimination as a situation where people who are "the same" are treated differently. Hence, a "discrimination" corresponds here to some "disparity", but we will use frequently the term "discrimination". More precisely, it is necessary to distinguish two standards. One standard corresponds to "disparate treatment," corresponding to "*any economic agent who applies different rules to people in protected groups is practicing discrimination*" as defined in Yinger (1998). The second discriminatory standard corresponds to "disparate impact". This corresponds to practices that seem to be neutral, but have the effect of disadvantaging one group more than others.

In labor economics, wages should be a function of productivity, which is unobservable when signing a contract, and therefore, as discussed in Riley (1975), Kohlleppel (1983) or Quinzii and Rochet (1985),

employers try to find signals. As claimed in Lippert-Rasmussen (2013), statistical discrimination occurs when "*there is statistical evidence which suggests that a certain group of people differs from other groups in a certain dimension, and its members are being treated disadvantageously on the basis of this information*." Those signals are observable variables that are correlated with productivity.

In the most common version of the model, employers use observable group membership as a proxy for unobservable skills, and rely on their beliefs about productivity correlates, in particular their estimates of average productivity differences between groups, as in Phelps (1972), Arrow (1973) or Bielby and Baron (1986). A variant of this theory is when there are no group differences in average productivity, but rather based on the belief that the variance in productivity is larger for some groups than for others, as in Aigner and Cain (1977) or Cornell and Welch (1996). In these cases, risk-adverse employers facing imperfect information might discriminate against groups with larger expected variances in productivity. According to England (1994), "statistical discrimination" might explain why there is still discrimination in competitive market. For Bertrand and Duflo (2017) "statistical discrimination" is a "*more disciplined explanation*" than the taste-based model initiated by Becker (1957), because the former "*does not involve an ad hoc (even if intuitive) addition to the utility function (animus toward certain groups) to help rationalize a puzzling behavior*."

Here, "statistical discrimination", rather than simply providing an explanation, can lead people to see social stereotypes as useful and acceptable, and therefore help to rationalise and justify discriminatory decisions. As suggested by Tilcsik (2021), economists, have theorised labour market discrimination constructing mathematically models that attribute discrimination to the deliberate actions of profit-maximising firms or utility-maximising individuals (as discussed in in Charles and Guryan (2011) or Small and Pager (2020)). And this view of discrimination has had influenced social science debates, legal decisions, corporate practices and public policy discussions, as mentioned in Ashenfelter and Oaxaca (1987), Dobbin (2001), Chassonnery-Zaïgouche (2020) or Rivera (2020). The most influential economic model of discrimination is probably the "statistical discrimination theory", discussed in the 70's, with Phelps (1972), Arrow (1973) and Aigner and Cain (1977). Applied to labour markets, this theory claims that employers have imperfect information about the future productivity of job applicants, which leads them to use easily observable signals, such as race or gender, to infer the expected productivity of applicants, as explained in Correll and Benard (2006). Employers who practice "statistical discrimination" rely on their beliefs about group statistics to evaluate individuals (corresponding to "discrimination" as defined in Definition 1.1.1). In this model, discrimination does not arise from a feeling of antipathy towards members of a group, it is seen as a rational solution to an information problem. Profit-maximising employers will use all the information available to them and, since individual-specific information is limited, they use group membership as a "proxy". Economists tend to view "statistical discrimination" as "*the optimal solution to an information extraction problem*" and sometimes describe it as "efficient" or "fair", as in Autor (2003), Norman (2003) and Bertrand and Duflo (2017). It should be stressed here that this approach, initiated in the 70's in the context of labor economics is essentially the same as the one underlying the concept of "actuarial fairness". Observe finally that the word "*statistical*" used here reinforces the image of discrimination as a rational, calculated decision, even though several models do not assume that employers' beliefs about group differences are based in statistical data, or any other type of systematic evidence. Employers' beliefs might be based on partial or idiosyncratic observations. As mentioned in Bohren et al. (2019) it is possible to have "*statistical discrimination with bad statistics*" here.

1.1.6 Algorithmic Injustice and Fairness of Predictive Models

While economists published extensively on discrimination in the job market in the 1970s, the subject has come back into the spotlight following a number of publications linked to predictive algorithms. *Correctional*

Offender Management Profiling for Alternative Sanctions, or `compas`, a tool widely used as a decision aid in the U.S. courts to assess a criminal's chance of re-offending, based on some risk scales for general and violent recidivism, and for pretrial misconduct. After several months of investigation, Angwin et al. (2016) looked back at the output of `compas` in a series of articles called "Machine Bias" (and subtitled "*Investigating Algorithmic Injustice*").

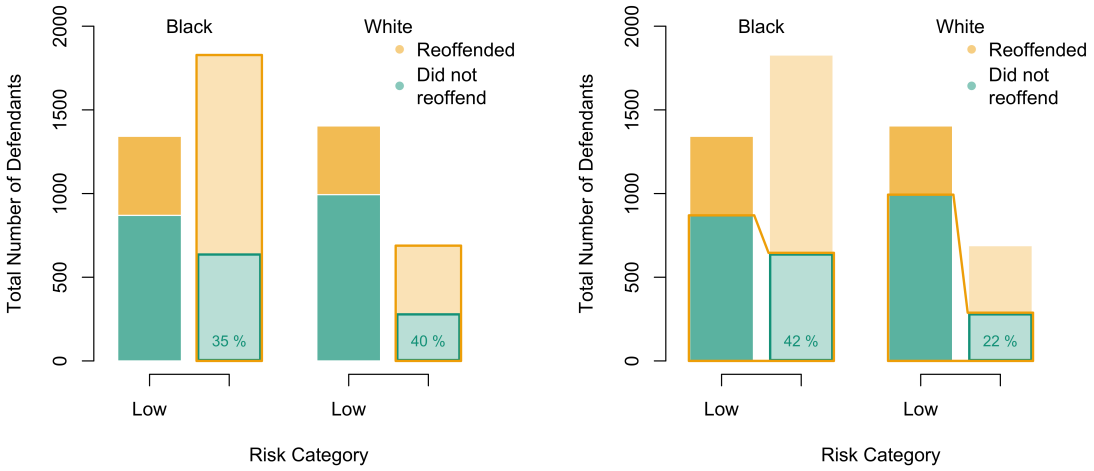


Figure 1.2: Two analysis of the same descriptive statistics of `compas` data, with the number of defendant (1) function of the race of defendant (black and white), (2) the risk category (binary, low and high) and (3) the indicator that the defendants reoffended, or not. On the left, the analysis of Dieterich et al. (2016) and on the right, the one of Feller et al. (2016).

As pointed out Feller et al. (2016), if we look at data from the `compas` dataset (from the `fairness` R package), on Figure 1.2, on the one hand (on the right of the figure)

- for White people, among those who did not re-offend, 78% were properly classified, since 22% did re-offend,
- for Black people, among those who did not re-offend, 58% were properly classified, since 42% did re-offend.

With standard terminology in classifiers and decision theory, the false negative rate is about two times higher for Black people (42% against 22%). As Larson et al. (2016) wrote it "*Black defendants were often predicted to be at a higher risk of recidivism than they actually were*. On the other hand (on the left of the figure) as Dieterich et al. (2016) observed

- for White people, among those who where classified as high risk, 40% did not re-offend,
- for Black people, among those who where classified as high risk, 35% did not re-offend.

Therefore, since the rate of recidivism is approximately equal at each risk score level, irrespective of race, it should not be claimed that the algorithm is racist. The initial approach will be called "*false positive rate parity*," while the second one is called "*predictive parity*." Obviously, there are reasonable arguments in favour of both contradictory positions. From this simple example, we see that having a valid and common definition of "fairness" or "parity" will be complicated.

Since then, many books and articles have addressed the issues highlighted in this article, namely the increasing power of these predictive decision-making tools, their ever-increasing opacity, the discrimination they replicate (or amplify), the 'biased' data used to train or calibrate these algorithms, and the sense of unfairness they produce. For instance, Kirkpatrick (2017) pointed out that "*the algorithm itself may not be biased, but the data used by predictive policing algorithms is colored by years of biased police practices*."

And justice is not the only area where such techniques are used. In the context of predictive health systems, Obermeyer et al. (2019) observed that a widely used health risk-prediction tool (predicting how sick individuals are likely to be, and the associated health-care cost), that is applied to roughly 200 million individuals in the U.S. per year, exhibited significant racial bias. More precisely, 17.7% of patients that the algorithm assigned to receive "extra care" were Black, and if the bias in the system was corrected for, as Ledford (2019) did, the percentage should increase to 46.5%.

Massive data, and machine learning techniques, have provided an opportunity to revisit a topic that has been explored by lawyers, economists, philosophers and statisticians for the past fifty years or longer. The aim here is to revisit these ideas, to shed new light on them, with a focus on insurance, and explore possible solutions. Lawyers, in particular, have discussed these predictive models, this "actuarial justice", as Thomas (2007), Harcourt (2011), Gautron and Dubourg (2015) or Rothschild-Elyassi et al. (2018) coined it.

The idea of bias and algorithmic discrimination is not a new one, as shown for instance by Pedreshi et al. (2008). However, over the past twenty years, the number of examples has continued to increase, with more and more interest in the media. "*AI biases caused 80% of black mortgage applicants to be rejected*" in Hale (2021), or "*How the use of AI risks recreating the inequity of the insurance industry of the previous century*" in Ito (2021). Pursuing David's 2015 analysis, McKinsey (2017) announced that artificial intelligence would disrupt the workplace (including the insurance and banking sectors, Mundubeltz-Gendron (2019)) particularly to replace lacklustre repetitive (human) work[3]. These replacements raise questions, and compel the market and the regulator to be cautious. For Reijns et al. (2021), "*the Dutch insurance sector makes it a mandate*", in an article on "*ethical artificial intelligence*," and in France, Défenseur des droits (2020) recalls that "*algorithmic biases must be able to be identified and then corrected*" because "*non-discrimination is not an option, but refers to a legal framework*." Bergstrom and West (2021) note, with a touch of irony, that there are people writing a bill of rights for robots, or devising ways to protect humanity from super-intelligent, Terminator-like machines, but that getting into the details of algorithmic auditing is often seen as boring, but necessary.

Live with blinders on, or close your eyes, rarely solves problems, although it has long been advocated as a solution to discrimination. As Budd et al. (2021) show, reverting to an Amazon experiment of removing names from CVs to eliminate gender discrimination does not work, since by hiding the candidate's name, the algorithm continued to preferentially choose men over women. Why did this happen? Simply because Amazon trained the algorithm from its existing resumes, with an over-representation of men, and there are elements of a resume (apart from the name) that can reveal a person's gender, such as a degree from a women's university, membership of a female professional organisation, or a hobby where the sexes are disproportionately represented. Proxies that are more or less correlated with the variable "protected" may

[3]Even if it seems exaggerated, because on the contrary, it is often humans who perform the repetitive tasks to help robots: "*in most cases, the task is repetitive and mechanical. One worker explained that he once had to listen to recordings to find those containing the name of singer Taylor Swift in order to teach the algorithm that it is a person*" as reported by Radio Canada in April 2019.

sustain a form of discrimination.

In this textbook, we will address to these issues, limiting ourselves to actuarial models in an insurance context, and almost exclusively, the pricing of insurance contracts. In Seligman (1983), the author asks the following basic question: "*If young women have fewer car accidents than young men - which is the case - why shouldn't women get a better rate? If industry experience shows - which it does - that women spend more time in hospital, why shouldn't women pay more?*." This type of question will be the starting point in our considerations in this textbook.

Paraphrasing Georges Clémenceau[4], who said (in 1887) that "*war is too serious a thing to be left to the military*," Worham (1985) argued that insurance segmentation was too important a task to be left to actuaries. Forty years later, we might wonder whether it is not worse to leave it to algorithms, and to clarify actuaries' role in these debates. In this introduction, we will begin by reviewing insurance segmentation and the foundations of actuarial pricing of insurance contracts. We will then review the various terms mentioned in the title, namely the notion of bias, discrimination and fairness, while proposing a typology of predictive models and data (in particular the so-called "sensitive" data, which may be linked to possible discrimination).

1.1.7 Discrimination Mitigation and Affirmative Action

Mitigating discrimination is usually seen as paradoxical, because in order to avoid discrimination, we must create another discrimination. More precisely, Supreme Court Justice Harry Blackmun stated, in 1978, "*in order to get beyond racism, we must first take account of race. There is no other way. And in order to treat some persons equally, we must treat them differently*." cited in Knowlton (1978), as mentioned in Lippert-Rasmussen (2020)). More formally, an argument in favour of affirmative action – called "*the present-oriented anti-discrimination argument* " – is simply that justice requires that we eliminate or at least mitigate (present) discrimination by the best morally permissible means of doing so, which corresponds to affirmative action. Freeman (2007) suggested a "*time-neutral anti-discrimination argument* ", in order to mitigate past, present, or future discrimination. But there are also arguments against affirmative action, corresponding to "*the reverse discrimination objection*," as defined in Goldman (1979): some might consider that there is an absolute ethical constraint against unfair discrimination (including affirmative action). To quote another Supreme Court Justice, in 2007, John G. Roberts of the US Supreme Court submits: "*The way to stop discrimination on the basis of race is to stop discriminating on the basis of race*" (Turner (2015) and Sabbagh (2007)). The arguments against affirmative action are usually based on two theoretical moral claims, according to Pojman (1998). The first denies that groups have moral status (or at least meaningful status). According to this view, individuals are only responsible for the acts they perform as specific individuals and, as a corollary, we should only compensate individuals for the harms they have specifically suffered. The second asserts that a society should distribute its goods according to merit.

1.2 From Words and Concepts to Mathematical Formalism

1.2.1 Mathematical Formalism

The starting point of any statistical or actuarial model is to suppose that observations are realisations of random variables, in some probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$ (see Rolski et al. (2009) for example, or any actuarial textbook). Therefore, let $\mathbb{P}$ denote the "true" probability measure, associated with random variables

[4]Member of the Chamber of Deputies from 1885 and 1893 and then Prime Minister of France from 1906 to 1909 and again from 1917 until 1920.

$(\boldsymbol{Z}, Y) = (S, \boldsymbol{X}, Y)$. Here, features $\boldsymbol{Z}$ can be splitted into a couple $(S, \boldsymbol{X})$, where $\boldsymbol{X}$ is the non-sensitive information while S is the sensitive attribute[5]. Y is the outcome we want to model, which would correspond to the annual loss of a given insurance policy (insurance pricing), the indicator of a false claim (fraud detection), the number of visits to the dentist (partial information for insurance pricing), the occurrence of a natural catastrophe (claims management), the indicator that the policyholder will purchase insurance to a competitor (churn model), etc. Thus, here, we have a triplet $(S, \boldsymbol{X}, Y)$, defined on $\mathcal{S} \times \mathcal{X} \times \mathcal{Y}$, following some unknown distribution $\mathbb{P}$. $\mathbb{P}_n$ will denote the empirical probabilities from the insurer portfolio, associated with data $\mathcal{D}_n = \{(\boldsymbol{z}_i, y_i)\} = \{(s_i, \boldsymbol{x}_i, y_i)\}$, where $i = 1, 2, \cdots, n$.

Observe that it is always assumed that $\mathcal{S}$ is somehow fixed in advance, and will not be learnt: gender will be considered as a binary categorical variable, sensitive and protected. In most cases, s will be a categorical variable, and in order to avoid heavy notations, we will simple consider a binary sensitive attribute (denoted $s \in \{\texttt{A}, \texttt{B}\}$ to remain quite general, and avoid $\{0, 1\}$ not to get confusions with values taken by y in a classification problem). $\mathcal{Y}$ will depend on the model considered: in a classification problem, $\mathcal{Y}$ will usually correspond to $\{0, 1\}$, while in a regression problem, $\mathcal{Y}$ will correspond to the real line $\mathbb{R}$. We can also consider counts, when $y \in \mathbb{N}$ (i.e. $\{0, 1, 2, \cdots\}$).

Through the book, we will consider models, that are formally function $m : \mathcal{S} \times \mathcal{X} \to \mathcal{Y}$, that we will estimated from some training dataset $\mathcal{D}_n$. Considering models $m : \mathcal{X} \to \mathcal{Y}$ ("*gender-blind*" if s denote the gender, or "*color-blind*" if s denote the race, etc) is supposed to create a more "fair" model, unfortunately, in a very weak sense (since many variables in $\boldsymbol{x}$ might be strongly correlated with s). After estimating a model, we can use it to obtain predictions, denoted $\widehat{y}$.

1.2.2 Legitimate Segmentation and Unfair Discrimination

In the previous section, we have tried to explain that there could be "legitimate" and "illegitimate" discrimination, "fair" and "unfair". We will consider here a first attempt to illustrate that issue, with very simple dataset (with simulated data). Consider a risk, and let y denote the occurrence of that risk (hence, y is binary). As we will discuss in Chapter 2, it is legitimate to ask policyholders to pay a premium that is proportional to $\mathbb{P}[Y = 1]$, the probability that the risk occurs (that will be the idea of "actuarial fairness"). Assume now that this occurrence is related to a single feature x : the larger x, the more likely the risk will occur. A classical example could be the occurrence of the death of a person, where x is the age of that person. Here, the correlation between y and x will be coming from a common (unobserved) factor, z. In a small dataset, `toydata1` (divided into a training dataset, `toydata1_train`, and a validation dataset, `toydata1_validation`), we have simulated values, where the confounding variable Z (that will not be observed, and used, in the modeling process) is a Gaussian variable, $Z \sim \mathcal{N}(0, 1)$, and then

$$\begin{cases} X = Z + \epsilon, \ \epsilon \sim \mathcal{N}(0, 1/2^2), \\ S = \mathbf{1}(Z + \eta > 0), \ \eta \sim \mathcal{N}(0, 1/2^2), \\ Y = \mathbf{1}(Z + \nu > 0), \ \ \nu \sim \mathcal{N}(0, 1/2^2). \end{cases}$$

The sensitive attributed, that will take values 0 (or A) and 1 (or B), does not influence y, and therefore it might not be legitimate to use it (it could be seen as an "illegitimate discrimination"). Note that z influences all variables, x, s and y (with a probit model for the last two), and because of that unobserved confounding variable z, all variables are here (strongly) correlated. On Figure 1.3, we can visualize the dependence between x and y (via boxplots of x given y) on the left, and between x and s (via boxplots of x given s) on the right. For example, if $x \sim -1$, then y takes values in $\{0, 1\}$ respectively with 25% and 75% chance. It

[5]For simplicity, in most of the book, we will discuss the case where S is a single sensitive attribute.

will be 75% and 25% chance if $x \sim +1$. Similarly, when $x \sim -1$, s is four times more likely to be in group A than in group B.

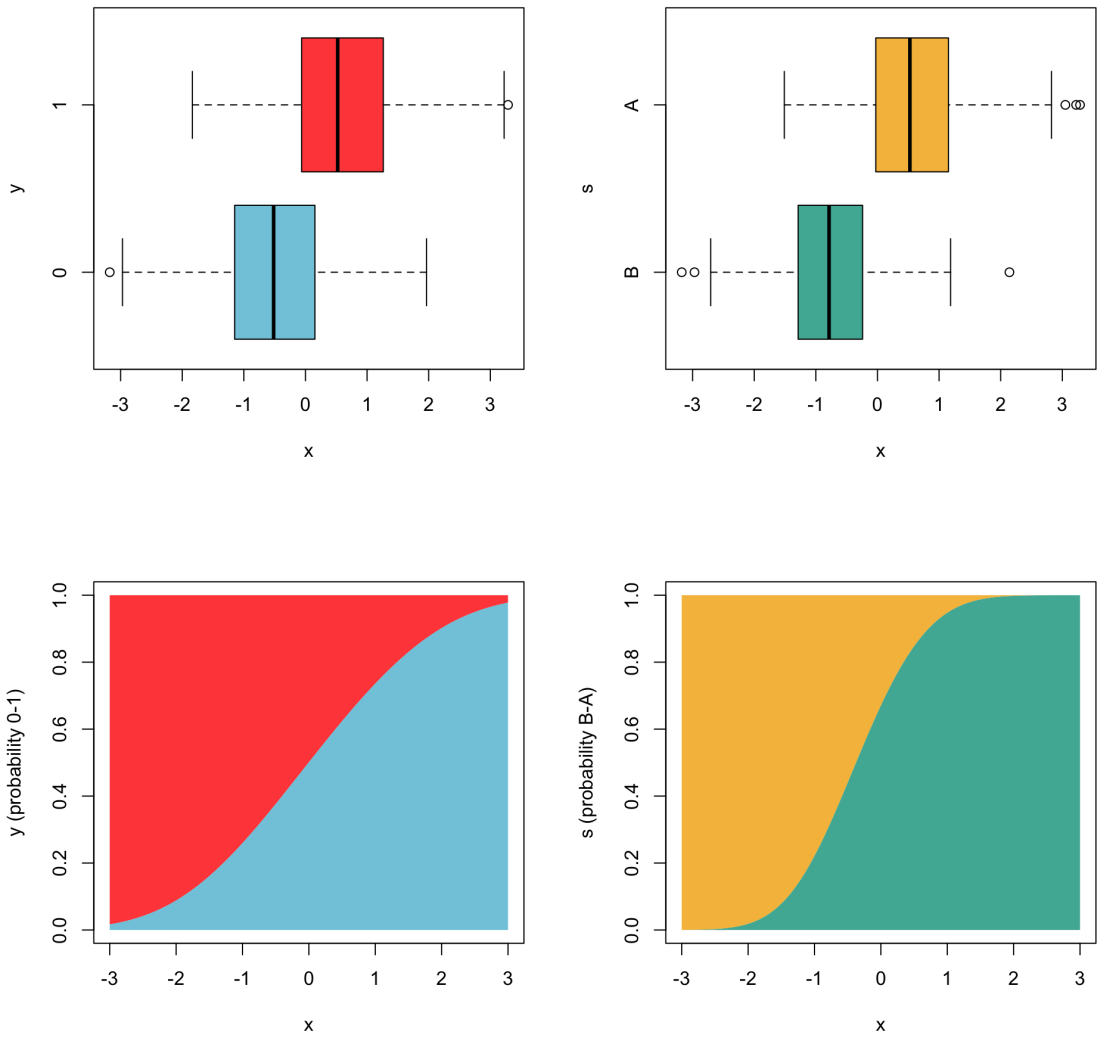


Figure 1.3: On top, boxplot of x conditional on y, with $y \in \{0, 1\}$ on the left, and conditional on s, with $s \in \{\texttt{A}, \texttt{B}\}$ on the right, from the `toydata1` dataset. Below, the curve on the left is $x \mapsto \mathbb{P}[Y = 1|X = x]$ while the curve on the right is $x \mapsto \mathbb{P}[S = \texttt{A}|X = x]$. Hence, when $x = +1$, $\mathbb{P}[Y = 1|X = x] \sim 75\%$, and therefore $\mathbb{P}[Y = 0|X = x] \sim 25\%$ (on the left), while when $x = +1$, $\mathbb{P}[S = \texttt{A}|X = x] \sim 95\%$, and therefore $\mathbb{P}[S = \texttt{B}|X = x] \sim 5\%$ (on the right).

When fitting a logistic regression to predict y based both on x and s, from `toydata1_train`, observe

that variable x is clearly significant, but not s (using `glm` in R, see Section 3.3 for more details about standard classifiers, starting with the logistic regression):

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2983     0.2083  -1.432    0.152
x             1.0566     0.1564   6.756 1.41e-11 ***
s == A        0.2584     0.2804   0.922    0.357
```

Without the sensitive variable s, we obtain a logistic regression on x only, that could be seen as "*fair through unawareness*." The estimation yields

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.1390     0.1147  -1.212    0.226
x             1.1344     0.1333   8.507   <2e-16 ***
```

Here, $\widehat{m}(x)$, that estimates $\mathbb{E}[Y|X = x]$, is equal to

$$\widehat{m}(x) = \frac{\exp[-0.1390 + 1.1344\, x]}{1 + \exp[-0.1390 + 1.1344\, x]}.$$

But it does not mean that this model is perceived as "fair" by everyone. On Figure 1.4, we can visualize the probability that scores exceed a given threshold t, here 50%. Even without using s as a feature in the model, $\mathbb{P}[\widehat{m}(X) > t|S = s]$ does depend on s, whatever the threshold t. And if $\mathbb{E}[\widehat{m}(X)] \sim 50\%$, observe that $\mathbb{E}[\widehat{m}(X)|S = \texttt{A}] \sim 65\%$ while $\mathbb{E}[\widehat{m}(X)|S = \texttt{B}] \sim 25\%$. With our premium interpretation, it means that, on average, people that belong in group `A` will pay a premium at least twice the one paid by people in group `B`. Of course, *ceteris paribus* it is not the case, since individuals with the same x will have the same prediction, whatever s, but overall, we observe a clear difference. One can easily transfer this simple example to many real-life applications.

Through this book, we will provide examples of such situations, then formalize some measures of fairness, and finally discuss methods used to mitigate a possible discrimination in a predictive model $\widehat{m}$, even if $\widehat{m}$ is not a function of the sensitive attribute (fairness through unawareness).

1.3 Structure of the Book

In part I we will get back to **insurance and predictive modeling**. In Chapter 2, we will present applications of predictive modeling in insurance, emphasizing insurance ratemaking and premium calculations, first in the context of homogeneous policyholders, and then of heterogeneous policyholders. We will discuss "segmentation" from a general perspective, the statistical approach being discussed in Chapter 3. In that chapter, we will present standard supervised models, with GLMs, penalized versions, neural nets, trees, and ensemble approaches. In Chapter 4, we will then address the questions of interpretation and explanation of predictive models, as well as accuracy and calibration.

In part II, we will discuss further **segmentation and discrimination** and sensitive attribute in the context of insurance modeling. In Chapter 5, we will provided a classification and a typology of pricing variables. In Chapter 6, we will discuss direct discrimination (with race, gender, age, and genetic discrimination), as well as indirect direction We will discuss also **biases and data**, in Chapter 7, with a discussion about data,

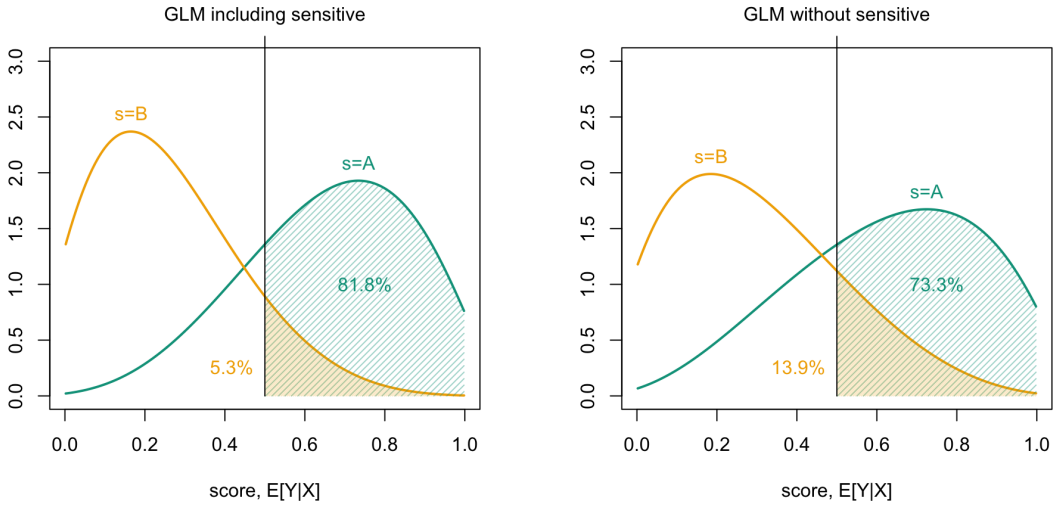


Figure 1.4: Distribution of the score $m(\boldsymbol{X}, S)$, conditional on A and B, on the left, and distribution of the score $m(\boldsymbol{X})$ without the sensitive variable, conditional on A and B, on the right (fictitious example). In both cases, logistic regressions are considered. From this score, we can get a classifier $\widehat{y} = \mathbf{1}(m(\boldsymbol{z}) > t)$ (where $\boldsymbol{z}$ is either $(\boldsymbol{x}, s)$, on the left, or simply $\boldsymbol{x}$, on the right). Here, we consider cut-off $t = 50\%$. Areas on the right of the vertical line (at $t = 50\%$) correspond to the proportion of individuals classified as $\widehat{y} = 1$, in both groups, A and B.

observations and experiments. We will get back on how data are collected before getting back to the popular adage "*correlation is not causation*," and start discussing causal inference and counterfactuals.

In part III, we present various approaches to **quantify fairness**, with a focus in Chapter 8, on "*group discrimination*" concepts, while "*individual fairness*" will be presented in Chapter 9. And finally, in part IV, we will discuss **mitigation of discrimination**, using three approaches: The pre-processing approach, in Chapter 10, in-processing approach in Chapter 11 and post-processing approach in Chapter 12.

1.4 Datasets and Case Studies

In the following chapters, and more specifically in parts III and IV, we will use both generated data, and publicly available real datasets to illustrate various techniques, either to quantify a potential discrimination (in part III) or to mitigate it (in part IV). All the datasets are available from the Github repository, in[6] R,

```
> library(devtools)
> devtools::install_github("freakonometrics/InsurFair")
> library(InsurFair)
```

[6]See Charpentier (2014) for a general overview on the use of R in actuarial science. Note that some packages mentioned here also exist in Python, in `scikit-learn`, as well as packages dedicated to fairness, such as `fairlearn`, or `aif360`).

The first toy dataset is the one discussed previously, in Section 1.2.2, with `toydata1_train` and `toydata1_valid`, with (only) three variables `y` (binary outcome), `s` (binary sensitive attribute) and `x` (drawn from a Gaussian variable).

```
> str(toydata1_train)
'data.frame': 600 obs. of  3 variables:
 $ x  : num  0.7939 0.5735 0.9569 0.1299 -0.0606 ...
 $ s  : Factor w/ 2 levels "B","A": 1 1 2 1 2 2 2 1 1 1 ...
 $ y  : Factor w/ 2 levels "0","1": 1 1 2 2 2 2 2 1 2 1 ...
```

As discussed, the three variables are correlated, since they are all based on an unobserved common variable z.

The `toydata2` dataset consists in two generated data, $n = 5000$ are used as a training sample, and $n = 1000$ are used for validation. The process used to generate the data is the following:

- the binary sensitive attribute, $s \in \{\texttt{A}, \texttt{B}\}$, is drawn, with respectively 60% and 40% individuals in each group
- $(x_1, x_3) \sim \mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$, with some correlation of 0.4 when $s = \texttt{A}$ and 0.7 when $s = \texttt{B}$
- $x_2 \sim \mathcal{U}([0, 10])$, independent of x_1 and x_3
- $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 \mathbf{1}_{\texttt{B}}(s)$, that does not depend on x_3
- $y \sim \mathcal{B}(p)$ where $p = \exp(\eta)/[1 + \exp(\eta)] = \mu(x_1, x_2, s)$.

On Figure 1.5, we can visualize scatter plots with x_1 on the x-axis and x_2 on the y-axis, with on the left, colors depending on y ($y \in \{\texttt{GOOD}, \texttt{BAD}\}$, or $y \in \{0, 1\}$) and depending on s ($s \in \{\texttt{A}, \texttt{B}\}$) on the right. On Figure 1.6, we can visualize level curves of $(x_1, x_2) \mapsto \mu(x_1, x_2, \texttt{A})$ on the left and $(x_1, x_2) \mapsto \mu(x_1, x_2, \texttt{B})$ on the right, where $\mu(x_1, x_2, s)$ are the true probabilities used to generate the dataset. Colors reflect the value of the probability (on the right part) and are coherent with $\{\texttt{GOOD}, \texttt{BAD}\}$.

Then, there will be real data. The `germancredit` dataset, collected in Hofmann (1990) and used in the `CASdataset` package, from Charpentier (2014), contains 1,000 observations and 23 attributes. The variable of interest y is a binary variable indicating whether a person experienced a default of payment. There are 70% on 0's ("good" risks), 30% of 1's ("bad" risks). The sensitive attribute is the gender of the person (binary, with 69% women (B) and 31% men (A), but we can also use the age, categorized.

The `frenchmotor` datasets, from Charpentier (2014), in personal motor insurance, with underwriting data, and information about claim occurrence (here considered as binary). It is obtained as the aggregation of `freMPL1`, `freMPL2`, `freMPL3` and `freMPL4`, while keeping only observations with `exposure` exceeding 0.9. Here the sensitive attribute is $s =$ `Gender`, which is a binary feature, and the goal will be to create a score that reflects the probability to claim a loss (during the year). The entire dataset contains $n = 12,437$ policyholders and 18 variables. A subset with 70% of the observations is used for training, and 30% are used for observation. Note that variable `SocioCateg` contains here 9 categories (only the first digit in the categories is considered). In numerical applications, two specific individuals (named `Andrew` and `Barbara`) are considered, to illustrate various points.

The `telematic` dataset is an original dataset, 1177 insurance contracts, observed during two years. We have claims data for 2019 (here `claim` is binary, `no` or `yes`, 13% of the policyholders did claim a loss), the age and the gender (`gender`) of the driver, and some telematic data for 2018 (including `Total_Distance`, `Total_Time`, as well as `Drive_Score`, `Style_Score`, `Corner_Score`, `Acceleration_Score` or `Braking_Score`, including also some binary scores related to "heavy" acceleration or braking).

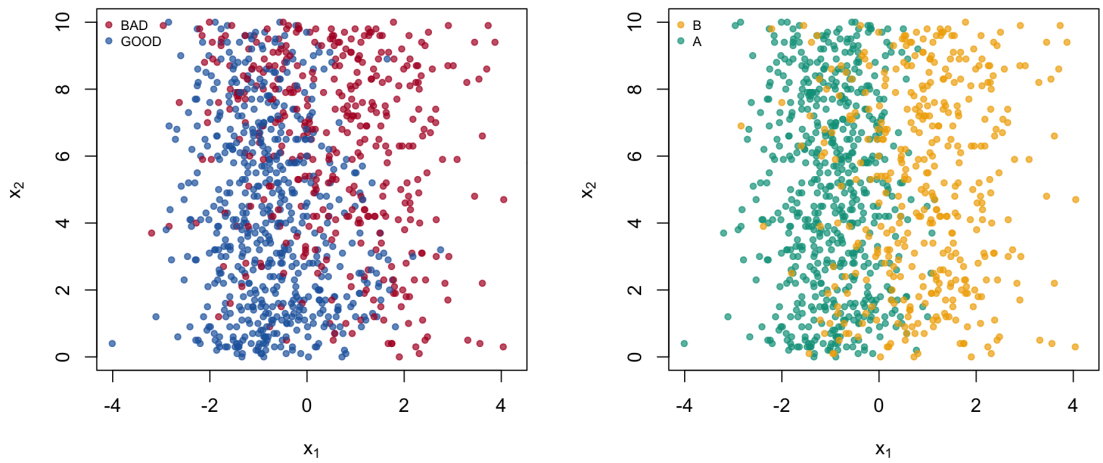


Figure 1.5: Scatter plot on `toydata2`, with x_1 on the x-axis and x_2 on the y-axis, with on the left, colors depending on the outcome y ($y \in \{$`GOOD` , `BAD`$\}$, or $y \in \{0\,,\,1\}$) and depending on the sensitive attribute s ($s \in \{$`A` , `B`$\}$) on the right.

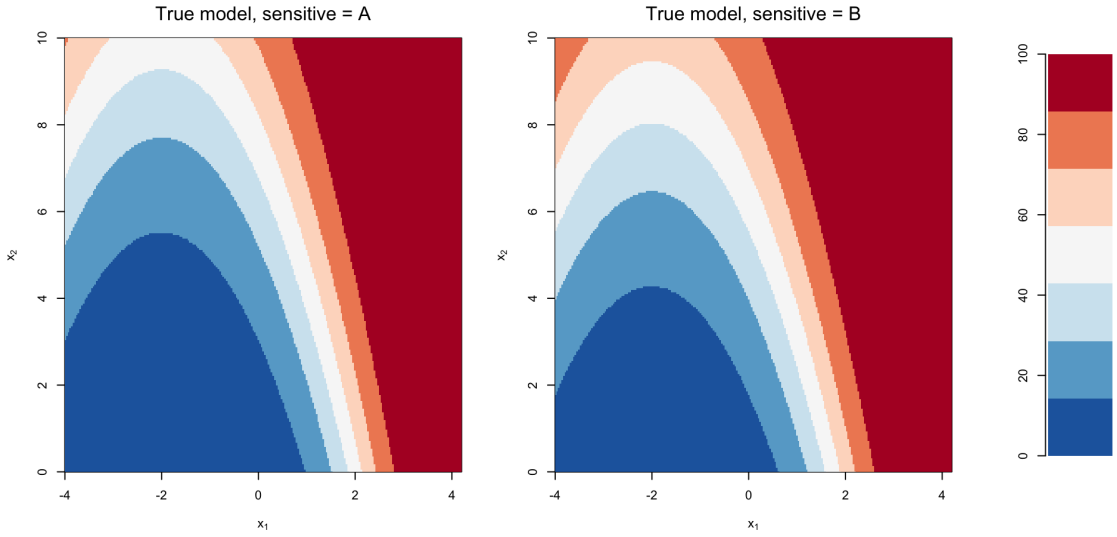


Figure 1.6: Level curves of $(x_1, x_2) \mapsto \mu(x_1, x_2,$ `A`$)$ on the left and $(x_1, x_2) \mapsto \mu(x_1, x_2,$ `B`$)$ on the right, the true probabilities used to generate the `toydata2` dataset. The blue area in the lower left corner corresponds to $\widehat{y}$ close to 0 (or `GOOD` risk), while the red area in the upper right corner corresponds to $\widehat{y}$ close to 1 (or `BAD` risk).

Part I

Insurance and Predictive Modeling

"*Predictive modeling involves the use of data to forecast future events. It relies on capturing relationships between explanatory variables and the predicted variables from past occurrences and exploiting these relationships to predict future outcomes. Forecasting future financial events is a core actuarial skill – actuaries routinely apply predictive modeling techniques in insurance and other risk management application,*" Frees et al. (2014a).

"*The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work-that is, correctly to describe phenomena from a reasonably wide area,*" Von Neumann (1955)

"*In economic theory, as in Harry Potter, the Emperor's New Clothes or the tales of King Solomon, we amuse ourselves in imaginary worlds. Economic theory spins tales and calls them models. An economic model is also somewhere between fantasy and reality. Models can be denounced for being simplistic and unrealistic, but modeling is essential because it is the only method we have of clarifying concepts, evaluating assumptions, verifying conclusions and acquiring insights that will serve us when we return from the model to real life. In modern economics, the tales are expressed formally: words are represented by letters. Economic concepts are housed within mathematical structures,*" Rubinstein (2012)

Chapter 2

Fundamentals of Actuarial Pricing

'*Insurance is the contribution of the few to the misfortune of the many*" is a simple way to describe what insurance is. But it doesn't say what the "*contribution*" should, to be fair.In this chapter, we will return to the fundamentals of pricing and risk sharing, and at the end we will mention other models used in insurance (to predict future payments to be provisioned, to create a fraud score, etc.)

Even though insurers will not be able to predict which of their clients will suffer a loss, they should be capable of estimating probabilities to claim a loss, and possibly the distribution of their aggregate losses, with an acceptable margin of error, and budgeting accordingly. The role of actuaries is to run statistical analysis to measure individual risk and price it.

2.1 Insurance

The insurance business is characterised by an inverted production cycle. In return for a premium - the amount of which is known when the contract is taken out - the insurer undertakes to cover a risk, the unknown date and amount, according to the definition of "actuarial pricing." In order to do this, the insurer will pool the risks within a mutuality. Insurance's universal secret is therefore the pooling of a large number of insurance contracts within a mutuality, in order to allow compensation to be made between the risks that have been damaged and those for which the insurer has collected premiums without having had to pay out any benefits, as Petauton (1998) argues. To use Chaufton (1886)'s formulation, insurance is the "*compensation of the effects of chance by mutuality organised according to the laws of statistics*." The first important concept is mutualization,

Definition 2.1.1 (Mutuality) *Wilkie (1997). Mutuality is considered as the normal form of commercial private insurance, where participants contribute to the risk pool through a premium that relates to their particular risk at the time of the application, i.e. the higher the risk that they bring to the pool, the higher the premium required.*

Through effective underwriting, Wilkie (1997) claims that "*the risk is evaluated by the insurer as thoroughly as possible, based on all the facts that are relevant and available*." Participation in mutual

insurance schemes is voluntary and the amount of cover that the individual purchases is discretionary. An essential feature of mutual insurance is segmentation, or discrimination in underwriting, leading to significant differences in premium rates for the same amount of life cover for different participants. Viswanathan (2006) gives several examples. The second concept it solidarity,

Definition 2.1.2 (Solidarity) *Wilkie (1997). Solidarity is the basis of most national or social insurance schemes. Participation in such state-run schemes is generally compulsory and individuals have no discretion over their level of cover. All participants normally have the same level of cover. In solidarity schemes the contributions are not based on the expected risk of each participant.*

In those state-run schemes, contributions are often just equal for all, or it can be according to the individual ability to pay (such as percentage of income). Since everybody pays the same contribution rate, the low-risk participants are effectively subsidising the high-risk participants. With an insurance economics perspective, agents make decisions individually, forgetting that the decisions they make often go beyond their narrow self-interest, reflecting instead broader community and social interests, even in situations where they are not known to each other. This is not altruism, *per se*, but rather a notion of strong reciprocity, the "*predisposition to cooperate even when there is no apparent benefit in doing so*," as formalized in Gintis (2000) and Bowles and Gintis (2004).

Solidarity is important in insurance. In most countries, employer-based health insurance include maternity benefits for everyone. In the United States, a federal law that says it's discriminatory not to do so. "*Yes, men should pay for pregnancy coverage, and here's why*, said Hiltzik (2013), *it takes two to tango*." No man has ever given birth to a baby, but it's also true that no baby has ever been born without a man being involved somewhere along the line. "*Society has a vested interest in healthy babies and mothers*" and "*universal coverage is the only way to make maternity coverage affordable*," therefore, solidarity is imposed, and men should pay for pregnancy coverage.

One should probably stress here that insurance does not eliminate the risk, but to transfer it, and this transfer is done according to a social philosophy chosen by the insurer. In "public insurance", as Ewald (1986) reminds us, aim to transfer risk from individuals to a wider social group, by "socialising", or redistributing risk "more fairly" within the population. Thus, low-risk individuals pay insurance premiums at a higher rate than their risk profile would suggest, even if this seems "inefficient" from an economic point of view. Social insurance, organised according to principles of solidarity, where access and coverage are independent of risk status, and sometimes of ability to pay (as noted by Mittra (2007)), although in many cases the premium is proportional to the income of the policyholder, is usually provided by public rather than private entities. For some social goods, such as health care, long-term care and perhaps even basic mortgage life insurance, it may simply be inappropriate to provide such products through a mutuality-based model that inevitably excludes some individuals, as "*primary social goods, because they are defined as something to which everyone has an inalienable right, cannot be distributed through a system that excludes individuals based on their risk status or ability to pay*." Mutual insurance companies are often seen as an intermediary between such public insurance and for-profit insurance companies.

And as Lasry (2015) points out, "*insurance has long been faced with a dilemma: on the one hand, better knowledge of a risk allows for better pricing; better knowledge of risk factors can also encourage prevention; on the other hand, mutualisation, which is the basis of insurance, can only subsist in most cases in a situation of relative ignorance (or even a legal obligation of ignorance).*" Actuaries will then seek to classify or segment risks, all based on the idea of mutualisation. We shall return to the mathematical formalism of this dilemma. De Pril and Dhaene (1996) point out that segmentation is a technique that the insurer uses to differentiate the premium and possibly the cover, according to a certain number of specific characteristics of the risk to be policyholder (hereinafter referred to as segmentation criteria), with the aim of achieving

Box 2.1 **Insurance & underwriting** (in French law), by Rodolphe Bigot[1]

The insurance transaction and the underlying mutualisation are based on so-called risk selection. Apart from most group insurances, which consist of a kind of mutualisation within a mutualisation, the insurer refuses or accepts that each applicant for insurance enters the mutualisation constituted by the group of policyholders. This selection of risks "*confines the mutualisation to the policyholders accepted by the insurer, who is always considered, in insurance contract law, as the one who accepts the contract proposed to him by the applicant for insurance*," wrote Monnet (2017), p. 13 and following). In this respect, it should be recalled that the economics of the insurance transaction require that the insurance company be given a great deal of freedom to accept or refuse the risk proposed to it. Proof of this freedom in the area of personal insurance is the provisions of Article 225-3 of the Criminal Code, which exclude from the scope of application of the criminal repression of discrimination in the supply of goods and services provided for in Article 225-2 "*discrimination based on the state of health, when it consists of operations whose purpose is the prevention and coverage of risks of death, risks of harm to the physical integrity of the person, or risks of incapacity for work or disability*." However, not to admit a limit to this freedom of the insurer would lead to the evacuation of important social considerations and to the exclusion not only of insurance but also of the goods and services linked to it (such as borrowing, and therefore access to property) of the most exposed persons (Bigot and Cayol (2020), p. 540). The question of the right to insurance arises here (Pichard (2006)). To this end, "*having access to insurance means not only the very possibility of taking out a contract for coverage, but perhaps also at a reasonable economic cost, not prohibitive, not dissuasive. In societies where the need for security, or even comfort, is a leitmotif, the question is very relevant*." (Noguéro (2010), p. 633)

a better match between the estimated cost of the claim and the burdens that a given person places on the community of policyholders and the premium that this person has to pay for the cover offered. In Box 2.1, Rodolphe Bigot gets back on general legal considerations regarding risk selection in insurance (in France, but most principles can be observed elsewhere). Underwriting is the term used to describe the decision-making process by which insurers determine whether to offer, or refuse, an insurance policy to an individual based on the available information (and the requested amount). Gandy (2016) asserts that "right to underwrite" is basically a right to discriminate. Hence, "higher premium" corresponds to a rating decision, "exclusion waiver" is a coverage decision while "denial" is an underwriting decision.

2.2 Premiums and Benefits

Comparing policyholders is always tricky, as not only do they potentially have different risks, but policyholders may also have different preferences (and therefore choose different policies). First of all, it is important to distinguish between coverages. In a car insurance policy, the "third party liability" cover is the compulsory component, covering exclusively the damage that the insured car might cause to a third party. But some policyholders may want (or need) more extensive protection. Other standard types of cover include "comprehensive" cover, which covers all damage to the vehicle (regardless of the circumstances of the accident or the driver's responsibility), "collision" cover, which reimburses the owner for damage caused to the vehicle in the event of a collision with a third party, and "fire and theft" cover, which compensates the owner of the vehicle if it is damaged or destroyed by fire, or if it is stolen. Some insurers also offer "mechanical breakdown" cover, which allows the insurance to compensate for the cost of repairs related to a breakdown, or "vehicle contents" cover, which offers compensation in the event of damage to or disappearance of items inside the insured vehicle. There may also be "assistance" cover, which provides services in the event of a breakdown, such as breakdown assistance, towing, repatriation, etc. Another possible source of difference is

[1]Lecturer in private law, UFR of Law, Le Mans University, member of Thémis-UM and Ceprisca

the indemnity, which may vary according to the choice of the deductible level, Buchanan and Priest (2006). As a reminder, the deductible is the amount that remains payable by the policyholder after the insurer has compensated for a loss. The absolute (or fixed) excess is the most common in car insurance: in a policy with an excess of €150, if the repair costs amount to €250, the insurance company will pay €100 and the remaining €150 will be paid by the policyholder. Many insurers now offer “mileage excesses", defining a perimeter around the vehicle’s usual parking place: within this perimeter, the assistance guarantee will not work. However, if a breakdown occurs outside this perimeter, the assistance guarantee can be called upon.

Also, it is difficult to compare the auto insurance premium paid by different people. In Figure 2.1, we can see that the choice of auto insurance coverage is strongly dependent on age, with young drivers opting overwhelmingly for compulsory coverage (one-third of drivers between 20 and 25 years of age), and older drivers taking out more “comprehensive" insurance (90% of drivers between 70 and 80 years of age). Choosing different coverage inevitably translates into higher bills, as older people may have a more expensive policy simply because they require more coverage.

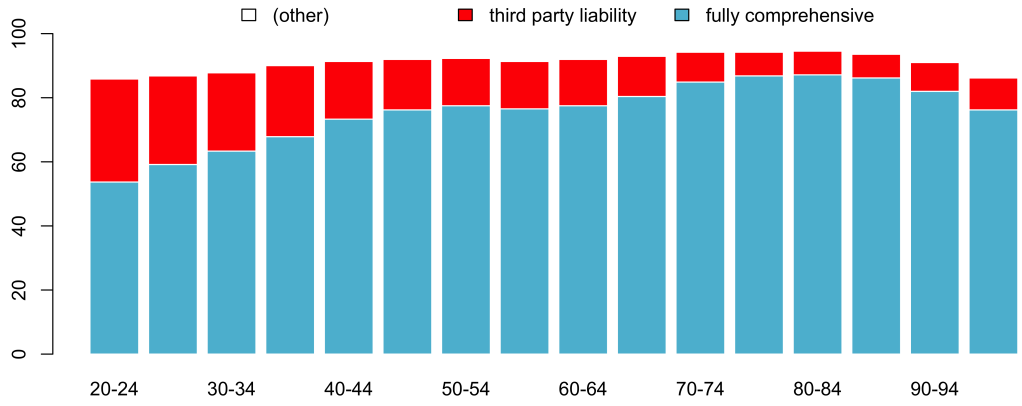


Figure 2.1: Coverage selected by auto insurance policyholders based on age, with the basic mandatory coverage, “third party insurance" and the broadest coverage nown as “fully comprehensive."
(source: personal communication, real data from an insurance company in France)

As mentioned earlier, a natural idea is that each policyholder should be offered a premium that is proportional to the risk he or she represents, to avoid that another company will attract this customer with a more attractive contract. This principle of “personalization" could be seen to have the virtues of fairness (since each individual pays according to the risk he or she passes on to the community) and can even be reconciled with the principle of mutualization: all that’s needed (provided the market is large enough) is to group individuals into mutuals that are homogeneous from the point of view of risk. This very general principal does not say anything about how to build a fair tariff. A difficult task lies in the fact that insurers have incomplete information about their customers. It is well known that the observable characteristics of policyholders (which will can be used in pricing) explain only a small proportion of the risks they represent. The only remedy for this imperfection is to self-select policyholders by differentiating the cover offered to

them, i.e. a non-linear scale linking the premium to be paid to the amount of the deductible accepted. As mentioned in the previous chapter, observe that there is a close analogy between this concept of "fair tariff", or "actuarial fairness" and that of "*equilibrium with signal*" proposed by Spence (1974, 1976) to describe the functioning of certain labor markets. Riley (1975) proposed a more general model that could be applied to insurance markets, among others. Cresta and Laffont (1982) proved the existence of fair insurance rates for a single risk. While the structure of equilibria with signal is now well understood in the case of a one-dimensional parameter, the same cannot be said for cases where several parameters are involved. Kohlleppel (1983) gave an example of the non-existence of such an equilibrium in a model satisfying the natural extension of Spence's hypotheses. Since insurance is generally a highly competitive and regulated market, the insurer must use all the statistical tools and data at its disposal to build the best possible rates. At the same time, its premiums must be aligned with the company's strategy and take into account competition.Because of the important role played by insurance in society, premiums are also scrutinized by regulators. They must be transparent, explainable and ethical. Thus, pricing is not only statistical, it also carries strategic and societal issues. These different issues can push the insurer to offer fairer premiums in relation to a given variable. For example, the regulations require insurers to present fair premiums according to the gender of the policy holder given their strategies and to offer fair premiums according to age. Regardless of the reason why an insurance player must present fairer pricing in relation to a variable, it must be able to define, measure and then mitigate the ethical bias of its pricing while preserving its consistency and performance.

2.3 Premium and Fair Technical Price

2.3.1 Case of a Homogeneous Population

Before talking about segmentation, let us examine the case of a homogeneous portfolio, where the policyholders are faced with the same probability of occurrence of a claim, the same extent of damage, the same sum insured, etc., from ομογενης (homogenes), "*of the same race, family or kind*," (from ομος, homos, "same", and γενος, genos, "kinds"). In non-life insurance, the insurer undertakes (in exchange for the payment of a premium, the amount of which is decided when the contract is signed) to cover the claims that will occur during the year. If y is the annual cost of a randomly selected policyholder, then we define[2] the "pure premium" as $\mathbb{E}[Y]$.

Definition 2.3.1 (Pure premium (homogeneous risks)) *Let Y be the non-negative random variable corresponding to the total annual loss associated with a given policy, then the pure premium is $\mathbb{E}[Y]$.*

If we consider the risk of losing 100 with probability p (and nothing with probability $1-p$), the pure premium for this risk (economists would call it a lottery) is $100\,p$. The premium for an insurance contract is then proportional to the probability of having a claim. In the following, our examples will often be limited to the estimation of this probability. In a personal insurance contract, the period of cover is longer, and it is necessary to discount future payments[3]. For example, in a personal insurance policy, the policyholder would pay 100 upon the death of a person (to his beneficiaries): if T is the (random) remaining life of a

[2]Denuit and Charpentier (2004) discuss the mathematical formalism that allows such a writing. In particular, the expectation is calculated according to a probability $\mathbb{P}$ corresponding to the "historical" probability (there is no law of one price in insurance, contrary to the typical approach in market finance, in the sense of Froot et al. (1995). This will be discussed further in Section 3.1.1.). Insurance, like econometrics, is formalized in a probabilistic world, perhaps unlike many machine learning algorithms which are derived without a probabilistic model, as discussed in Charpentier et al. (2018).

[3]Without discounting, since death is (at an infinite time horizon) certain, the pure premium would be exactly the amount of the capital paid to the beneficiaries.

Box 2.2 **Moral Hazard**, Michelbacher (1926)

"Moral hazard is the Bogey Man who will catch the unwary insurance official who does not watch out. When insurance is under consideration he is always present in one guise or another, sometimes standing out in bold relief, but more often lurking in the background where he employs every expedient to avoid detection. In all the ramifications of insurance procedure, from the binding of the risk until the last moment of policy coverage has expired, his insidious influence may manifest itself, usually where it is least expected. In the other case his ignorance, carelessness, inattention or recklessness may involve the carrier in claims which the ordinarily prudent policyholder would avoid. The unsafe automobile driver; the employer whose attitude toward safety is not proper; the careless person who loves display and is notoriously lax in the protection of his jewelry: these and many others are "bad risks" for the insurance carrier because they prevent the proper functioning of the law of averages and introduce the certainty of loss into the insurance transaction. It will be noted that the term "moral hazard" as employed in this discussion is used in a much broader sense than the following definition, which is typical of common usage, would imply: "The hazard is the deflection or variation from the accepted standard of what is right in one's conduct. Moral Hazard is that risk or chance due to the failure of the positive moral qualities of a person whose interests are affected by a policy of insurance."

policyholder, at the time the policy is taken out, then the pure premium is the probable present value of the future flows, i.e.

$$a = \mathbb{E}\left[\frac{100}{(1+r)^T}\right] = \sum_{t=0}^{\infty} \frac{100}{(1+r)^t} \cdot \mathbb{P}[T=t],$$

for some discount rate r. But this assumption of homogeneous risks is clearly too strong in many insurance applications. For instance in the case of death insurance, because the law of T should depend e.g. on the age of the policyholder at the time of the contract, as we will discuss in Section 2.3.3. But before, let us get back to classical concepts about economic decisions, when facing uncertain events.

2.3.2 The Fear of Moral Hazard and Adverse-Selection

In the context of insurance, moral hazard refers to the impact of insurance on incentives to reduce risks. An individual facing an accident risk such as of the loss of a home, car or the risk of medical expenses, can generally take actions to reduce the risk. Without insurance, the costs and benefits of accident avoidance, or precaution, are internal to the individual and the incentives for avoidance are optimal. With insurance, some of the accident costs are borne by the insurer, as recalled in Winter (2000).

Definition 2.3.2 (Adverse selection) *Laffont and Martimort (2002). Adverse selection is a market situation where buyers and sellers have different information. "Adverse selection" characterizes principal-agent models in which an agent has private information before a contract is written.*

Definition 2.3.3 (Moral hazard) *Arrow (1963). In economics, a moral hazard is a situation where an economic actor has an incentive to increase its exposure to risk because it does not bear the full costs of that risk.*

There has been a lot of publications about adverse selection and moral hazard in life insurance, that creates a demand for insurance which is positively correlated with the insured's risk of loss, and could be seen as immoral, or unethical. In Box 2.2, one of the oldest discussion about moral hazard, in Michelbacher (1926), is reproduced.

All actuaries have been lulled by Akerlof's 1970 fable of "*lemons*". The insurance market is characterized by information asymmetries. From the insurer's point of view, these asymmetries mainly concern the need to find adequate information on the customer's risk profile. A decisive factor in the success of an insurance business model is the insurer's ability to estimate the cost of risk as accurately as possible. While in the case of some simple product lines, such as motor insurance, the estimation of the cost of risk can be largely or fully automated and managed in-house, in areas with complex risks, the assistance of an expert third party can mitigate this type of information asymmetry. With Akerlof's terminology, some insurance buyers are considered low risk peaches, while others are high risk lemons. In some cases, insurance buyers know (to some extent) whether they are lemons or peaches. If the insurance company could tell the difference between lemons and peaches, it would have to charge peaches a premium related to the risk of the peaches and lemons a premium related to the risk of the lemons, according to a concept of actuarial fairness, as Baker (2011) reminds us. But if actuaries are not able to differentiate between lemons and peaches, then they will have to charge the same price for an insurance contract. The main difference between the market described by Akerlof (1970) (in the original fable it was a market for used cars) and an insurance market is that the information asymmetry was initially (in the car example) in favour of the seller of an asset. In the field of insurance, the situation is often more complex. In the field of car insurance, Dalziel and Job (1997) pointed out the optimism bias of most drivers who all think they are "good risks". The same bias will be found in many other examples, as mentioned by Royal and Walls (2019), but excluding health insurance, where the policyholder may indeed have more information than the insurer.

To use the description given by Chassagnon (1996), let us suppose that an insurer covers a large number of agents who are heterogeneous in their probability of suffering a loss. The insurer proposes a single price that reflects the average probability of loss of the agent representative of this economy, and it becomes unattractive for agents whose probability of suffering an accident is low to insure themselves. A phenomenon of selection by price therefore occurs and it is said to be unfavourable because it is the bad agents who remain. To guard against this phenomenon of anti-selection, risk selection and premium segmentation are necessary. "*Adverse-selection disappears when risk analysis becomes sufficiently effective for markets to be segmented efficiently*," says Picard (2003), "*doesn't the difficulty econometricians have in highlighting real anti-selection situations in the car insurance market reflect the increasingly precise evaluation of the risks underwritten by insurers* ?"

2.3.3 Case of a Heterogeneous Population

It is important to have model that can capture this heterogeneity (from ετερογενης, heterogenes, "*of different kinds*," from ετερος, heteros, "other, another, different" and γενος, genos, "kinds"). To get back to our introductory example, if T_x is the age at (random) death of the policyholder of age x at the time the contract was taken out (so that $T_x - x$ is the residual life span), then the pure premium corresponds to the expected present value of future flows, i.e.

$$a_x = \mathbb{E}\left[\frac{100}{(1+r)^{T_x-x}}\right] = \sum_{t=0}^{\infty} \frac{100}{(1+r)^t} \cdot \mathbb{P}[T_x = x+t],$$

for a discount rate r, which is written, using a more statistical terminology

$$a_x = \sum_{t=0}^{\infty} \frac{100}{(1+r)^t} \cdot \frac{L_{x+t-1} - L_{x+t}}{L_{x+t-1}},$$

where L_t is the number of people alive at the end of t years in a cohort that we would follow, so that $L_{x+t-1} - L_{x+t}$ is the number of people alive at the end of $x+t-1$ years but not $x+t$ years (and therefore

dead in their t-th year). It is De Witt (1671) who first proposed this premium for a death insurance, where discriminating according to age seems legitimate.

But we can go further, because $\mathbb{P}[T_x = t]$, the probability that the policyholder of age x at the time of subscription will die in t years, could also depend on his gender, his health history, and probably on other variables that the insurer might know. And in this case, it is appropriate to calculate conditional probabilities, $\mathbb{P}[T_x = t|\text{woman}]$ or $\mathbb{P}[T_x = t|\text{man smoker}]$.

2.4 Mortality Tables and Life Insurance

To illustrate heterogeneity, let us continue with mortality tables, since many tables are public, and openly available. The first modern mortality table was constructed in 1662 by John Graunt in London, and the first scientific mortality table was presented to the Royal Academy by Edmund Halley, in 1693, "*Estimate of the Degree of Mortality of Mankind, drawn from curious tables of the births and funerals at the city of Breslau*". At first, tables were constructed on data obtained from general population statistics, namely the Northampton (also called "*Richard Price*" life table) and Carlisle tables, Milne (1815) or Gompertz (1825, 1833). In order to compute adequate premium rates, insurance companies began to keep accurate and reliable records of their own mortality experience. The first life table constructed on the basis of insurance data was completed in 1834 by actuaries of the Equitable Assurance of London, as discussed in Sutton (1874) and Nathan (1925). Later one, American life insurance companies had the benefit of the English experience. As mentioned in Cassedy (2013), English mortality tables tended to overestimate death rates (both in the U.S. and in England, contributing to the prosperity of life insurance companies), and according to Zelizer (2018), The Presbyterian and Episcopalian Funds relied on the Scottish mortality experience (the first life table was constructed in 1775, as mentioned in Houston (1992)), while the Pennsylvania Company and the Massachusetts Hospital Life Insurance Company used the Northampton table. From the 1830s to the 1860s, American companies based their premiums on the Carlisle table. In 1868, Sheppard Homans (actuary of the Mutual Life Insurance Company) and George Phillips (Equitable's actuary) produce the first comprehensive table of American mortality in Homans and Phillips (1868), named the "*American Experience*" table in Ransom and Sutch (1987).

2.4.1 Gender Heterogeneity

As surprising as it may seem, Pradier (2011) noted that before the end of the XVIII-th century, in the U.K. and in France, the price of life annuities hardly ever depended on the sex of the subscriber. However, the first separate mortality tables, between men and women, constituted as early as 1740 by Nicolas Struyck (published in appendices of a geography article, Struyck (1740)) showed that women generally lived longer than men, in Table 2.1. Struyck (1740) (translated in Struyck (1912)) shows that at age 20, life expectancy (residual) is 30 years $^3/_4$ for men and 35 years $^1/_2$ for women. It also provides life annuity tables by gender. For a 50-year-old woman, a life annuity was worth 969 florins, compared to 809 florins for a man of the same age. This substantial difference seemed to legitimize a differentiation of premiums. Here, 424 men (L_x) and 468 women (out of one thousand respective births) had reached 40 years of age ($x = 40$). And among those who have reached 40 years of age, 12.5% of men and 9.6% of women will die within 5 years (mathematically denoted ${}_5p_x = \mathbb{P}[T \leq x + 5|T > x]$).

According to Pradier (2012), it was not until the Duchy of Calenberg's widows' fund went bankrupt in 1779 that the age and sex of subscribers were used in conjunction to calculate annuity prices. In France, in 1984, the regulatory authorities of the insurance markets decided to use regulatory tables established for the general population by INSEE, based on the population observed over 4 years, namely the PM 73-77 table for

men						women					
x	L_x	${}_5p_x$	x	L_x	${}_5p_x$	x	L_x	${}_5p_x$	x	L_x	${}_5p_x$
0	1000	29.0%	45	371	16.6%	0	1000	28.9%	45	423	11.8%
5	710	5.6%	50	313	19.2%	5	711	5.2%	50	373	14.7%
10	670	4.2%	55	253	22.9%	10	674	3.3%	55	318	18.2%
15	642	5.5%	60	195	27.2%	15	652	4.3%	60	260	21.2%
20	607	6.6%	65	142	31.7%	20	624	5.8%	65	205	26.8%
25	567	7.9%	70	97	37.1%	25	588	6.8%	70	150	33.3%
30	522	9.2%	75	61	45.9%	30	548	7.3%	75	100	45.0%
35	474	10.5%	80	33	51.5%	35	508	7.9%	80	55	56.4%
40	424	12.5%	85	16		40	468	9.6%	85	24	

Table 2.1: Excerpt from the Men and Women life tables in 1720 (source: Struyck (1912), page 231), for pseudo-cohorts of one thousand people ($L_0 = 1000$).

men and the PF 73-77 table for women, renamed TD and TV 73-77 tables, respectively (with an analytical extension beyond 99 years). While the primary factor in mortality is age, gender is also an important factor, as shown in the TD-TV Table. For more than a century, the mortality rate for men has been higher than that of women in France.

In practice, however, the actuarial pricing of life insurance policies has continued to be established without taking into account the gender of the policyholder. In fact, the reason two tables were used was that the male table was the regulatory table for life insurance (PM became TD, for "*table de décès*," or "death table"), and the female table became the table for life insurance (PF became TV, for "*table de vie*," or "life table"). In 1993, the TD and TV 88-90 tables replaced the two previous tables, with the same principle, i.e., the use of a table built on a male population for death insurance, and a table built on a female population for life insurance. From a prudential point of view, the female table models a population that has, on average, a lower mortality rate, and therefore lives longer.

TD 73-77		TV 73-77		TD 88-90		TV 88-90		INED men		INED women	
0	100000	0	100000	0	100000	0	100000	0	100000	0	100000
10	97961	10	98447	10	98835	10	99129	10	99486	10	99578
20	97105	20	98055	20	98277	20	98869	20	99281	20	99471
30	95559	30	97439	30	96759	30	98371	30	98656	30	99247
40	93516	40	96419	40	94746	40	97534	40	97661	40	98810
50	88380	50	94056	50	90778	50	95752	50	95497	50	97645
60	77772	60	89106	60	81884	60	92050	60	90104	60	94777
70	57981	70	78659	70	65649	70	84440	70	78947	70	89145
80	28364	80	52974	80	39041	80	65043	80	59879	80	77161
90	4986	90	14743	90	9389	90	24739	90	25123	90	44236
100	103	100	531	100	263	100	1479	100	1412	100	4874
110	0	110	0	110	0	110	2				

Table 2.2: Excerpt from French tables, with TD and TV 73-77 on the left, TD and TV 88-90 in the center, and INED 2017-2019 on the right.

In 2005, the TH and TF 00-02 tables were used as regulatory tables, still with tables founded on different populations, namely men and women respectively. But this time, the term men (H, for *hommes*) and women (F, for *femmes*) is maintained, as regulations allowed for the possibility of different pricing for men and women. A ruling by the Court of Justice of the European Union on March 1, 2011, however, made gender-differentiated pricing impossible (as of December 21, 2012), on the grounds that they would discriminate. In comparison, recent (French) INED tables are also mentioned in the Table 2.2, on the right.

2.4.2 Health and Mortality

Beyond gender, all sorts of "discriminating variables" have been studied, in order to build, for example, mortality tables depending on whether the person is a smoker or not, as in Benjamin and Michaelson (1988), in Table 2.3. Indeed, since Hoffman (1931), or Johnston (1945), actuaries had observed that exposure to tobacco, and smoking, had an important impact on the policyholder's health. As Miller and Gerstein (1983) wrote, "*it is clear that smoking is an important cause of mortality*."

	Men			Women	
	non-smoker	smoker		non-smoker	smoker
25	48.4	42.8	25	52.8	49.8
35	38.7	33.3	35	43.0	40.1
45	29.2	24.2	45	33.5	31.0
55	20.3	16.5	55	24.5	22.6
65	12.8	10.4	65	16.2	15.1

Table 2.3: Residual life expectancy (in years) by age (25-65 years) for smokers and nonsmokers (source: Benjamin and Michaelson (1988), for 1970-1975 data in the United States).

There are also mortality tables (or calculations of residual life expectancy) by level of body mass index (BMI, introduced by Adolphe Quetelet in the mid-XIX-th century), as calculated by Steensma et al. (2013) in Canada. A "normal" index refers to people with an index between 18.5 and 25kg/m^2; "overweight" refers to an index between 25 and 30kg/m^2; obesity level I refers to an index between 30 and 35kg/m^2, and obesity level II refers to an index exceeding 35kg/m^2. The Table shows some of the elements. These orders of magnitude are comparable with Fontaine et al. (2003) among the pioneer studies, Finkelstein et al. (2010), or more recently Stenholm et al. (2017).

	Men				Women				
	normal	overweight	obese I	obese II		normal	overweight	obese I	obese II
20	57.2	61.0	59.1	53.5	20	62.8	66.5	64.6	59.3
30	47.6	51.4	49.4	44.1	30	53.0	56.7	54.8	49.5
40	38.1	41.7	39.9	34.7	40	43.3	46.9	45.0	39.9
50	28.9	32.4	30.6	25.8	50	33.8	37.3	35.5	30.6
60	20.4	23.6	21.9	17.6	60	24.9	28.1	26.4	21.9
70	13.2	15.8	14.4	10.9	70	16.8	19.7	18.2	14.3

Table 2.4: Residual life expectancy (in years), as a function of age (between 20 and 70 years) as a function of BMI level (source: Steensma et al. (2013)).

2.4.3 Wealth and Mortality

Higher incomes are associated with longer life expectancy, as mentioned already in Kitagawa and Hauser (1973) with probably the first documented analysis. But despite the importance of socioeconomic status on mortality and survival, Yang et al. (2012), Chetty et al. (2016) and Demakakos et al. (2016) stressed that wealth has been underinvestigated as a predictor of mortality. Duggan et al. (2008) and Waldron (2013) used social security data in the United States. In France, disparities of life expectancy by social categories are well known. Recently, Blanpain (2018) created some life tables per wealth quantiles. An excerpt can be visualized on Table 2.5, with men on the left, women on the right, and fictional cohorts, with the poorest 5% on the population on the left ("0-5%") and the richest 5% on the right ("95-100%"). Force of mortality, as a function of the age, the gender, and the wealth quantile, can be visualized on Figure 2.2.

	Men				Women		
	0-5%	45-50%	95-100%		0-5%	45-50%	95-100%
0	100000	100000	100000	0	100000	100000	100000
10	99299	99566	99619	10	99385	99608	99623
20	99024	99396	99469	20	99227	99506	99526
30	97930	98878	99094	30	98814	99302	99340
40	95595	98058	98627	40	97893	98960	99074
50	90031	96172	97757	50	95021	97959	98472
60	77943	91050	95649	60	88786	95543	97192
70	59824	79805	90399	70	79037	90408	94146
80	38548	59103	76115	80	63224	79117	85825
90	13337	23526	38837	90	31190	45750	55918
100	530	1308	3231	100	2935	5433	8717

Table 2.5: Excerpt of life tables per wealth quantile, and gender, in France, from Blanpain (2018).

2.5 Modeling Uncertainty and Capturing Heterogeneity

2.5.1 Groups of Predictive Factors

A multitude of criteria can be used to create rate classes, as we've seen in the context of mortality. To get a good predictive model, as in standard regression models, we simply look for variables that are significantly correlated with the variable of interest, as mentioned by Wolthuis (2004). For instance, in the case of car insurance, the following information was proposed in Bailey and Simon (1959): use (leisure - pleasure - or professional - business), age (under 25 or not), gender and marital status (married or not). Specifically, five risk classes are considered, with rate surcharges relative to the first class (which is used here as a reference)

- "*pleasure, no male operator under 25*," (reference),
- "*pleasure, nonprincipal male operator under 25*," +65%,
- "*business use*," +65%,
- "*married owner or principal operator under 25*," +65%,

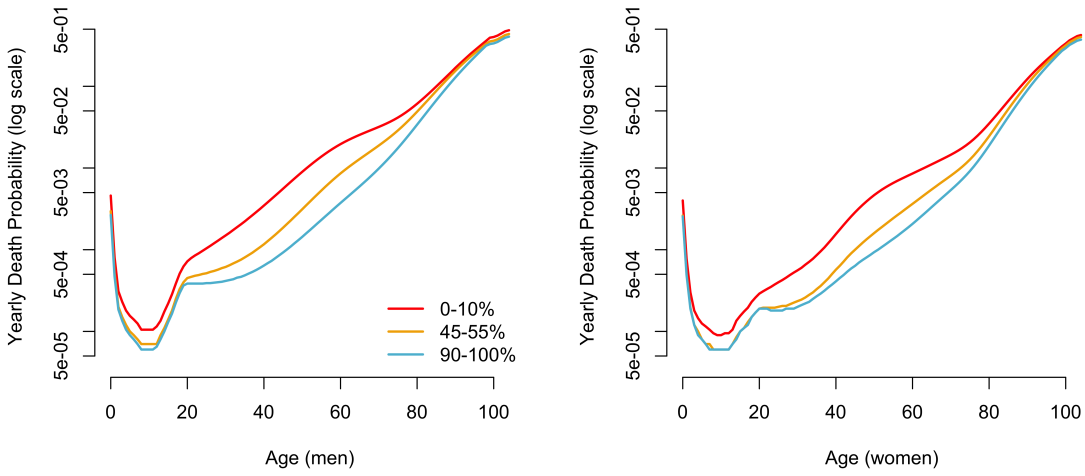


Figure 2.2: Force of mortality (log scale) for men, on the left and women, on the right, for various income quantile (bottom, medium, and upper 10%), in France.
(data from Blanpain (2018))

- "*unmarried owner or principal operator under 25,*" +140%.

In the 1960s, the rate classes resembled those that would be produced by classification (or regression) trees such as those introduced by Breiman et al. (1984). But using more advanced algorithms, Davenport (2006) points out that when an actuary creates risk classes and rate groups, and in most cases, these 'groups' are not self-aware, they are not conscious (at most, the actuary will try to describe them by looking at the averages of the different variables). These groups, or risk classes, are built on the basis of available data, and exist primarily as the product of actuarial models. And as Gandy (2016) points out, there is no physical basis for group members to identify other members of "their group". As discussed in Section 3.2, these risk groups, developed at a particular point in time, create a transient collusion between policyholders, who are likely to change groups as they move, change cars, or even simply grow older.

2.5.2 Probabilistic Models

Consider here a probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is a set of "events" on Ω ($\mathcal{A} \in \mathcal{F}$ is an "event"). Recall briefly that $\mathbb{P}$ is a function $\mathcal{F} \to [0,1]$ satisfying some properties, such as $\mathbb{P}(\Omega) = 1$; for disjoint events, an "additivity property": $\mathbb{P}(\mathcal{A} \cup \mathcal{B}) = \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B})$; a "subset property" (or "inclusion property"), if $\mathcal{A} \subset \mathcal{B}$, $\mathbb{P}(\mathcal{A}) \leq \mathbb{P}(\mathcal{B})$, as in Cardano (1564) or Bernoulli (1713), or for multiple (possibly infinite) disjoint events as in Kolmogorov (1933), $\mathcal{A}_1, \cdots, \mathcal{A}_n, \cdots$,

$$\mathbb{P}(\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_n \cup \cdots) = \mathbb{P}(\mathcal{A}_1) + \cdots + \mathbb{P}(\mathcal{A}_n) + \cdots$$

inspired by Lebesgue (1918), etc. In Section 3.1.1 we will get back on probability measures, since they are extremely important to assess the well-calibration of a model, as well as its fairness. But in this section, we need to recall two important properties that are crucial to model heterogeneity.

Propoition 2.5.1 (Total probability) *If $(\mathcal{B}_i)_{i\in I}$ is a partition of Ω (an exhaustive (finite or countable) set of disjoint events),*

$$\mathbb{P}(\mathcal{A}) = \sum_{i\in I} \mathbb{P}(\mathcal{A}\cap\mathcal{B}_i) = \sum_{i\in I} \mathbb{P}(\mathcal{A}|\mathcal{B}_i)\cdot\mathbb{P}(\mathcal{B}_i),$$

where, by definition, $\mathbb{P}(\mathcal{A}|\mathcal{B}_i)$ denotes the conditional probability of the occurrence of $\mathcal{A}$, given that $\mathcal{B}_i$ occurred.

See Feller (1957) or Ross (1972).

An immediate consequence is the law of total expectations,

Propoition 2.5.2 (Total expectations) *For any measurable random variable Y with finite expectation,*

$$\mathbb{E}(Y) = \sum_{i\in I} \mathbb{E}(Y|\mathcal{B}_i)\cdot\mathbb{P}(\mathcal{B}_i).$$

See Feller (1957) or Ross (1972).

This formula can be written simply in the case where two sets, two subgroups, are considered, for example related to the gender of the individual,

$$\mathbb{E}(Y) = \mathbb{E}(Y|\text{woman})\cdot\mathbb{P}(\text{woman}) + \mathbb{E}(Y|\text{man})\cdot\mathbb{P}(\text{man}).$$

If Y denotes the life expectancy at birth of an individual, the literal translation of the previous expression is that the life expectancy at birth of a randomly selected individual (on the left) is a weighted average of the life expectancies at birth of females and males, the weights being the respective proportions of males and females in the population. And since $\mathbb{E}(Y)$ is an average of the two,

$$\min\{\mathbb{E}(Y|\text{woman}),\mathbb{E}(Y|\text{man})\} \leq \mathbb{E}(Y) \leq \max\{\mathbb{E}(Y|\text{woman}),\mathbb{E}(Y|\text{man})\},$$

in other words, treating the population as homogeneous, when it is not, mean that one group is subsidized by the other, which is called “actuarial unfairness", as discussed by Landes (2015), Frezal and Barry (2019) or Heras et al. (2020). The greater the difference between the two conditional expectations, the greater the unfairness. This “unfairness" is also called “cross-financing" since one group will subsidize the other one.

Definition 2.5.1 (Pure premium (heterogeneous risks)) *Let Y be the non-negative random variable corresponding to the total annual loss associated with a given policy, with covariates $\boldsymbol{x}$, then the pure premium is $\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}]$.*

We will use notation $\mu(\boldsymbol{x})$, named also “regression function" (see Definition 3.3.1). We will also use notations $\mathbb{E}_Y[Y]$ (for $\mathbb{E}[Y]$) and $\mathbb{E}_{Y|\boldsymbol{X}}[Y|\boldsymbol{X}=\boldsymbol{x}]$ (for $\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}]$) to emphasize the measure used to compute the expected value (and to avoid confusions). For example, we can write

$$\mathbb{E}_Y[Y] = \int_{\mathbb{R}} y f_Y(Y)\mathrm{d}y \text{ and } \mathbb{E}_{Y|\boldsymbol{X}}[Y|\boldsymbol{X}=\boldsymbol{x}] = \int_{\mathbb{R}} y f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})\mathrm{d}y = \int_{\mathbb{R}} y\frac{f_{Y,\boldsymbol{X}}(y,\boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})}\mathrm{d}y.$$

The law of total expectations (Proposition 2.5.2) can be written, with that notation

$$\mathbb{E}_Y[Y] = \mathbb{E}_{\boldsymbol{X}}\big[\mathbb{E}_{Y|\boldsymbol{X}}[Y|\boldsymbol{X}]\big].$$

An alternative is to write, with synthetic notations $\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y|\boldsymbol{X}]\big]$, where the same notation – $\mathbb{E}$ – is used indifferently to describe the same operator on different probability measures.

The law of total expectations can be written

$$\mathbb{E}_Y[Y] = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{Y|\boldsymbol{X}}[Y|\boldsymbol{X}]\right] = \mathbb{E}_{\boldsymbol{X}}\left[\mu(\boldsymbol{X})\right],$$

which is a desirable property we want to have on any pricing function m (also called "globally unbiased", see Definition 4.3.6):

Definition 2.5.2 (Balance Property) *A pricing function m satisfies the balance property if $\mathbb{E}_{\boldsymbol{X}}[m(\boldsymbol{X})] = \mathbb{E}_Y[Y]$.*

The name "balance property" comes from accounting, since we want assets (what comes in, or premiums, $m(\boldsymbol{x})$) to equal liabilities (what goes out, or losses y) on average. This concept , as it appears in economics in Borch (1962), corresponds to "actuarial fairness," and is based on a match between the total value of collected premiums and the total amount of legitimate claims made by the policyholder. Since it is impossible for the insurer to know what future claims will actually be like, it is considered actuarially fair to set the level of premiums on the basis of the historical claims record of people in the same (assumed) risk class. It is on this basis that discrimination is considered "fair" in distributional terms, as explained in Meyers and Van Hoyweghen (2018). Otherwise, the redistribution would be considered "unfair", with forced solidarity from the low risk group to the high risk group. This "fairness" was undermined in the 1980s, when private insurers limited access to insurance for people with AIDS, or at risk of developing it, as Daniels (1990) recalls. Feiring (2009) goes further in the context of genetic information, "*since the individual has no choice in selecting his genotype or its expression, it is unfair to hold him responsible for the consequences of the genes he inherits - just as it is unfair to hold him responsible for the consequences of any distribution of factors that are the result of a natural lottery*." In the late 1970s (see Boonekamp and Donaldson (1979), Kimball (1979) or Maynard (1979)), the idea that the proportionality between the premium and the risk incurred would guarantee fairness between policyholders began to be translated into conditional expectation (conditional on the risk factors retained).

As discussed in Meyers and Van Hoyweghen (2018), who trace the emergence of actuarial fairness from its conceptual origins in the early 1960s to its position at the heart of insurance thinking in the 1980s, the concept of "actuarial fairness" appeared while as more and more countries adopted anti-discrimination legislation. At that time, insurers positioned "actuarial fairness" as a fundamental principle that would be jeopardized if the industry did not benefit from exemptions to such legislation. For instance, according to the Equality Act 2010 in the U.K. "*it is not a contravention (...) to do anything in connection with insurance business if (a) that thing is done by reference to information that is both relevant to the assessment of the risk to be insured and from a source on which it is reasonable to rely, and (b) it is reasonable to do that thing* ," as Thomas (2017) wrote it.

In most applications, there is a strong heterogeneity within the population, with respect to risk occurrence and risk costs. For example, when modeling mortality, the probability to die within a given year can be above 50% for very old and sick people, and less than 0.001% for pre-teenagers. Formally, the heterogeneity will be modelled by a latent factor Θ. If y designates the occurrence (or not) of an accident, y is seen as the realization of a random variable Y which follows a Bernoulli distribution, $\mathcal{B}(\Theta)$, where Θ is a non-observable latent variable (as in Gourieroux (1999) or Gourieroux and Jasiak (2007)). If y denotes the number of accidents occurring during the year, Y follows a Poisson distribution, $\mathcal{P}(\Theta)$ (or a binomial-negative model, or a parametric model with inflation of zeros, etc., as in Denuit et al. (2007)). If y notes the total cost, Y follows a Tweedie distribution, or more generally a compound Poisson distribution, which we will denote by $\mathcal{L}(\Theta, \varphi)$, where $\mathcal{L}$ denotes a distribution with mean Θ, and where φ is a dispersion parameter (see Definition 3.3.13 for more details). The goal of the segmentation is to constitute ratemaking classes

(denoted B_i previously) in an optimal way, i.e. by ensuring that one class does not subsidize the other, from observable characteristics, noted $\boldsymbol{x} = (x_1, x_2, \cdots, x_k)$. Crocker and Snow (2013) speaks of "*categorization based on immutable characteristics*." For Gouriéroux (1999), it is the "*static partition*" used to constitute sub-groups of homogeneous risks ("*in a given class, the individual risks are independent, with identical distributions*"). This is what a classification or regression tree does, the B_i's being the leaves of the tree, with the previous probabilistic notations. If y designates the occurrence of an accident, or the annual (random) load, the actuary will try to approximate $\mathbb{E}[Y|\boldsymbol{X}]$, from training data. In an econometric approach, if y designates the occurrence (or not) of an accident, and if $\boldsymbol{x}$ designates the set of observable characteristics of the policyholder, $Y|\boldsymbol{X} = \boldsymbol{x}$ follows a Bernoulli distribution, $\mathcal{B}(p_{\boldsymbol{x}})$, for example

$$p_{\boldsymbol{x}} = \frac{\exp(\boldsymbol{x}^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}^\top \boldsymbol{\beta})} \text{ or } p_{\boldsymbol{x}} = \Phi(\boldsymbol{x}^\top \boldsymbol{\beta})$$

for a logistic or probit regression, respectively [4]. If Y designates the number of accidents that occurred during the year, $Y|\boldsymbol{X} = \boldsymbol{x}$ follows a Poisson distribution, $\mathcal{P}(\lambda_{\boldsymbol{x}})$, with typically $\lambda_{\boldsymbol{x}} = \exp(\boldsymbol{x}^\top \boldsymbol{\beta})$. If Y denotes the annual cost, $Y|\boldsymbol{X} = \boldsymbol{x}$ follows a Tweedie distribution, or more generally a compound Poisson distribution, $\mathcal{L}(\mu_{\boldsymbol{x}}, \varphi)$, where $\mathcal{L}$ denotes a distribution of mean μ, with $\mu_{\boldsymbol{x}} = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ (for more details, Denuit and Charpentier (2004, 2005)).

To return to the analysis of Denuit and Charpentier (2004), if we assume that the risks are homogeneous, the pure premium will be $\mathbb{E}[Y]$, and we have the risk sharing table, Table 2.6. Without purchasing insurance, policyholder face a random loss Y. With insurance, policyholder face a fixed loss $\mathbb{E}[Y]$. The risk is transferred to the insurance company, that faces random loss $Y - \mathbb{E}[Y]$. On average, the loss for the insurance company is null, and all the risk is carried by the insurer.

	Policyholder	Insurer
Loss	$\mathbb{E}[Y]$	$Y - \mathbb{E}[Y]$
Average loss	$\mathbb{E}[Y]$	0
Variance	0	$\text{Var}[Y]$

Table 2.6: Individual loss, its expected value and its variables, for the policyholder on the left, and the insurer on the right. $\mathbb{E}[Y]$ is the premium paid, and Y the total loss.

At the other extreme, if the latent risk factor Θ were observable, the requested pure premium would be $\mathbb{E}[Y|\Theta]$, and we would have the split of Table 2.7.

	Policyholder	Insurer
Loss	$\mathbb{E}[Y\|\Theta]$	$Y - \mathbb{E}[Y\|\Theta]$
Average loss	$\mathbb{E}[Y]$	0
Variance	$\text{Var}[\mathbb{E}[Y\|\Theta]]$	$\text{Var}[Y - \mathbb{E}[Y\|\Theta]]$

Table 2.7: Individual loss, its expected value and its variables, for the policyholder on the left, and the insurer on the right. $\mathbb{E}[Y|\Theta]$ is the premium paid, and Y the total loss.

[4] Φ is here the distribution function of the centred and reduced normal distribution, $\mathcal{N}(0, 1)$.

Propoition 2.5.3 (Variance decomposition (1)) *For any measurable random variable Y with finite variance*

$$\text{Var}[Y] = \underbrace{\mathbb{E}[\text{Var}[Y|\Theta]]}_{\rightarrow \textit{ insurer}} + \underbrace{\text{Var}[\mathbb{E}[Y|\Theta]]}_{\rightarrow \textit{ policyholder}}.$$

See Denuit and Charpentier (2004).

Finally using only observable features, denoted $\boldsymbol{x} = (x_1, x_2, \cdots, x_k)$, we would have the decomposition of Table 2.8.

	Policyholder	Insurer
Loss	$\mathbb{E}[Y\|\boldsymbol{X}]$	$Y - \mathbb{E}[Y\|\boldsymbol{X}]$
Average loss	$\mathbb{E}[Y]$	0
Variance	$\text{Var}[\mathbb{E}[Y\|\boldsymbol{X}]]$	$\mathbb{E}[\text{Var}[Y\|\boldsymbol{X}]]$

Table 2.8: Individual loss, its expected value and its variables, for the policyholder on the left, and the insurer on the right. $\mathbb{E}[Y|\boldsymbol{X}]$ is the premium paid, and Y the total loss.

Propoition 2.5.4 (Variance decomposition (2)) *For any measurable random variable Y with finite variance*

$$\text{Var}[Y] = \underbrace{\mathbb{E}[\text{Var}[Y|\boldsymbol{X}]]}_{\rightarrow \textit{ insurer}} + \underbrace{\text{Var}[\mathbb{E}[Y|\boldsymbol{X}]]}_{\rightarrow \textit{ policyholder}},$$

where

$$\begin{aligned}\mathbb{E}[\text{Var}[Y|\boldsymbol{X}]] &= \mathbb{E}[\mathbb{E}[\text{Var}[Y|\Theta]|\boldsymbol{X}]] + \mathbb{E}[\text{Var}[\mathbb{E}[Y|\Theta]|\boldsymbol{X}]] \\ &= \underbrace{\mathbb{E}[\text{Var}[Y|\Theta]]}_{\textit{perfect ratemaking}} + \underbrace{\mathbb{E}\{\text{Var}[\mathbb{E}[Y|\Theta]|\boldsymbol{X}]\}}_{\textit{misclassification}}.\end{aligned}$$

See Denuit and Charpentier (2004).

This "misclassification" term (on the right) is called "*subsidierende solidariteit*" in De Pril and Dhaene (1996), or "*subsidiary solidarity*", as opposed to "*kanssolidariteit*" or "*random solidarity*" term (on the left). As certainty replaces uncertainty, it will never disappear, at least as long as it is a matter of predicting what may happen during a year at the time of subscription. For Corlier (1998), segmentation "*decreases the solidarity of risks belonging to different segments*." And for Löffler et al. (2016), citing a McKinsey report on the future of the insurance industry, mentioned that massive data will inevitably lead to de-mutualization and an increased focus on prediction. Nevertheless Lemaire et al. (2016) suggested some practical limits, since "*this process of segmentation, the sub-division of a portfolio of drivers into a large number of homogeneous rating cells, only ends when the cost of including more risk factors exceeds the profit that the additional classification would create, or when regulators rule out new variables*."

The link with solidarity is discussed in Gollier (2002), which reminds us that solidarity is fundamentally about making transfers in favor of disadvantaged people, compared to advantaged people, as discussed in the introduction of this chapter. But a very limited version of solidarity is taken into account in the context of insurance: "*solidarity in insurance means deciding not to segment the corresponding risk market on the basis of the observable characteristics of individuals' risks*," as in health insurance or unemployment insurance. It should be noted that while historically, the $\boldsymbol{x}$ variables were discretized in order to make "tariff classes", it is

now conventional to consider continuous variables as such, or even to transform them, while maintaining a relative regularity. According to De Wit and Van Eeghen (1984), in the past, it used to be very difficult to discover risk factors both in a qualitative and in a quantitative sense. "*Solidarity was therefore, unavoidably, considerable. But recent developments have changed this situation: with the help of computers it has become possible to make thorough risk analyses, and consequently to arrive at further premium differentiation*."

Again, the difficulty with pricing is that this underlying risk factor Θ is not observable. Not capturing it would lead to unfairness, as it would unduly subsidize the "riskier" (likely to have more expensive claims) individuals by the "less risky". Baker and Simon (2002) went further, arguing that the reason some people are classified as "low risk" and others as "high risk" is irrelevant. Speaking of automating accountability, Baker and Simon (2002) argued that it was important to make people accountable for the risk they bring to mutuality, especially the riskiest policyholders, in order for the least risky policyholder to "*feel morally comfortable*" (as Stone (1993) put it). The danger is that, in this way, the allocation of each person's contributions to mutuality would be the result of an actuarial calculation, as Stone (1993) put it. Porter (2020) said that this process was "*a way of making decisions without seeming to decide*." We will review this point when we discuss exclusions and the interpretability of models. The insurer will then use proxies to capture this heterogeneity, as we have just seen. A proxy (one might call it a "proxy variable") is a variable that is not significant in its own right, but which replaces a useful but un-observable, or un-measurable, variable according to Upton and Cook (2014).

Most of our discussion will focus on tariff discrimination, and more precisely on the "technical" tariff. As mentioned in the introduction, from the point of view of the policyholder, this is not the most relevant variable. Indeed, in addition to the actuarial premium (the pure premium mentioned earlier), there is a commercial component, as an insurance agent may decide to offer a discount to one policyholder or another, taking into account a different risk aversion or a greater or lesser price elasticity, see Meilijson (2006). But an important underlying question is "*is the provided service the same?*." Ingold and Soper (2016) review the example of Amazon not offering the same services to all its customers, in particular same-day-delivery offers, offered in certain neighborhoods, chosen by an algorithm that ultimately reinforced racial bias (by never offering same-day delivery in neighborhoods composed mainly of minority groups). A naive reading of prices on Amazon would be biased because of this important bias in the data, which should be taken into account. As Calders and Žliobaite (2013) reminds us, "*unbiased computational processes can lead to discriminative decision procedures*." In insurance, one could imagine that a claims manager does not offer the same compensation to people with different profiles – some people being less likely to dispute than others. It is important to better understand the relationship between the different concepts.

2.5.3 Interpreting and Explaining Models

A large part of the actuary's job is to motivate, and explain, a segmentation. Some authors, such as Pasquale (2015), Castelvecchi (2016) or Kitchin (2017), have pointed out that machine learning algorithms are characterized by their opacity and their "incomprehensibility", sometimes called "black box" properties. And it is essential to explain them, to tell a story. For Rubinstein (2012), models are "fables": "*In economic theory, as in Harry Potter, the Emperor's New Clothes or the tales of King Solomon, we entertain ourselves in imaginary worlds. Economic theory spins tales and calls them models. An economic model is also somewhere between fantasy and reality* (...) *the word model sounds more scientific than the word fable or tale, but I think we are talking about the same thing*." In the same way, the actuary will have to tell the story of his model, before convincing the underwriting and insurance agents to adopt it. But this narrative is necessarily imprecise. As Saint Augustine said, "*What is time? If no one asks me, I know. But if someone asks me and I want to explain it, then I don't know anymore.*"

One can hear that age must be involved in the prediction of the frequency of claims in car insurance, and indeed, as we see in Figure 2.3, the prediction will not be the same at 18, 25 or 55 years of age. Quite naturally, a premium surcharge for young drivers can be legitimised, because of their limited driving experience, coupled with unlearned reflexes. But this story does not tell us what order of magnitude of this surcharge would seem legitimate. Going further, the choice of model is far from neutral on the prediction: for a 22-year-old policyholder, relatively simple models propose an extra premium of +27%, +73%, +82% or +110% (compared to the average premium for the entire population). While age discrimination may seem logical, how much difference can be allowed here, and would be perceived as "quantitatively legitimate"? In Section 4.1, we will present standard approaches used to interpret actuarial predictive models, and explain predicted outcomes.

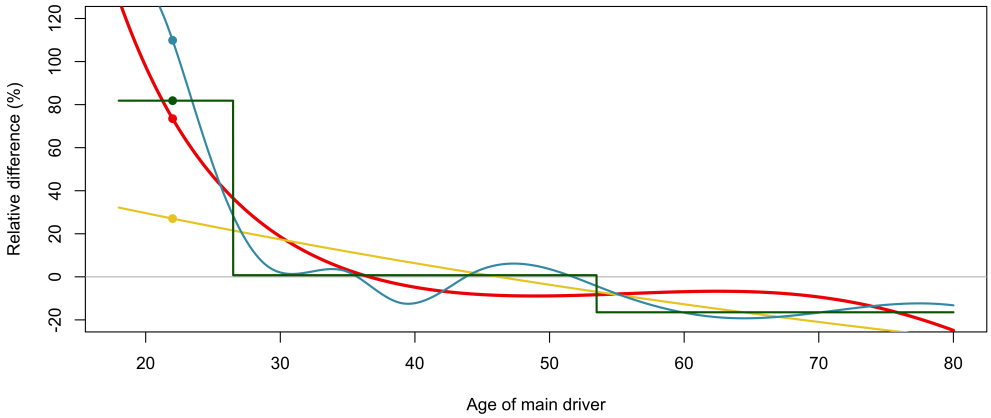


Figure 2.3: The evolution of auto insurance claim frequency as a function of primary driver age, relative to overall annual frequency, with a Poisson regression in yellow, a smoothed regression in red, a smoothed regression with a *too* small smoothing bandwidth in blue, and with a regression tree in green. Dots on the left are the predictions for a 22-year-old driver.
(data from `CASdataset` R package, see Charpentier (2014))

2.6 From Technical to Commercial Premiums

So far, we have discussed heterogeneity in the technical pure premiums, but in real life applications, some additional heterogeneity can yield additional sources for discrimination.

2.6.1 Homogeneous Policyholders

Technical or actuarial premiums are purely based on risk characteristics, while commercial premiums are based on economic considerations. In classical textbooks in economics of insurance (see Dionne and Harrington (1992); Dionne (2000, 2013) or Eisen and Eckles (2011)), homogeneous agents are considered, perfectly informed (not in the sense that there is no randomness, but they perfectly know the odds of

unfortunate events), having a utility function u (perfectly known also), and a wealth w, will agree to transfer the risk (or parts of the risk) against the payment of a premium π if it satisfies

$$u(w-\pi) \geq \mathbb{E}\big[u(w-Y)\big].$$

Thus, an insurer, also with a perfect knowledge of the wealth and the utility of the agent (or his or her risk aversion), could ask the following premium.

Definition 2.6.1 (Indifference utility principle) *Let Y be the non-negative random variable corresponding to the total annual loss associated with a given policy, for a policyholder with utility u and wealth w, the indifference premium is*

$$\pi = w - u^{-1}\left(\mathbb{E}\big[u(w-Y)\big]\right).$$

If u is the identity function, $\pi = \mathbb{E}\big[Y\big]$, corresponding to the technical, actuarial, pure premium. And if the agent is risk adverse, u is concave and $\pi \geq \mathbb{E}\big[Y\big]$.

On Figure 2.4, we can visualize both π and $\mathbb{E}(Y)$. Consider here a simple case, where the policyholder, with wealth w, and with a concave utility function u facing a possible random loss, where, with probability p the agent losses her wealth. On Figure 2.4, we use $p = 2/5$ just to get a visual perspective. Here, the random loss Y will take two values,

$$Y = \begin{cases} y_2 = w \text{ with probability } p = 2/5 \\ y_1 = 0 \text{ with probability } 1-p = 3/5, \end{cases}$$

or equivalently, the wealth will be

$$w - Y = \begin{cases} w - y_2 = 0 \text{ with probability } p = 2/5 \\ w - y_1 = w \text{ with probability } 1-p = 3/5. \end{cases}$$

The technical pure premium is here $\pi_0 = \mathbb{E}(Y) = py_2 + (1-p)y_1 = pw$, and when paying that premium, the wealth would be $w - \pi_0 = (1-p)w$.

If the agent is risk adverse (strictly), $u(w-\pi_0) > \mathbb{E}[u(w-Y)]$, in the sense that the insurance company can ask for a higher premium than the pure premium

$$\begin{cases} \pi_0 = \mathbb{E}[Y] > 0 & : \text{ actuarial (pure) premium} \\ \pi - \pi_0 = w - \mathbb{E}[Y] - u^{-1}\left(\mathbb{E}\big[u(w-Y)\big]\right) \geq 0 & : \text{ commercial loading.} \end{cases}$$

We will come back on the practice of price optimization in Section 2.6.3.

2.6.2 Heterogeneous Policyholders

"Actuarial fairness" is about a "legitimate" discrimination when it is based on a risk factor. Thomas (2012) mentioned that some laws prohibit discrimination with respect to the age, but often contains exceptions for insurance underwriting. However, these exemptions usually apply only to differences which are justified by differences in the underlying risk, and the technical premium. In the previous section, we have discussed economic models, based on individual wealth w and utility functions u, that can actually be heterogeneous. As mentioned in Boonen and Liu (2022), with information about personal characteristics, the insurer can customize insurance coverage and premium for each individual in order to optimize her objective function.

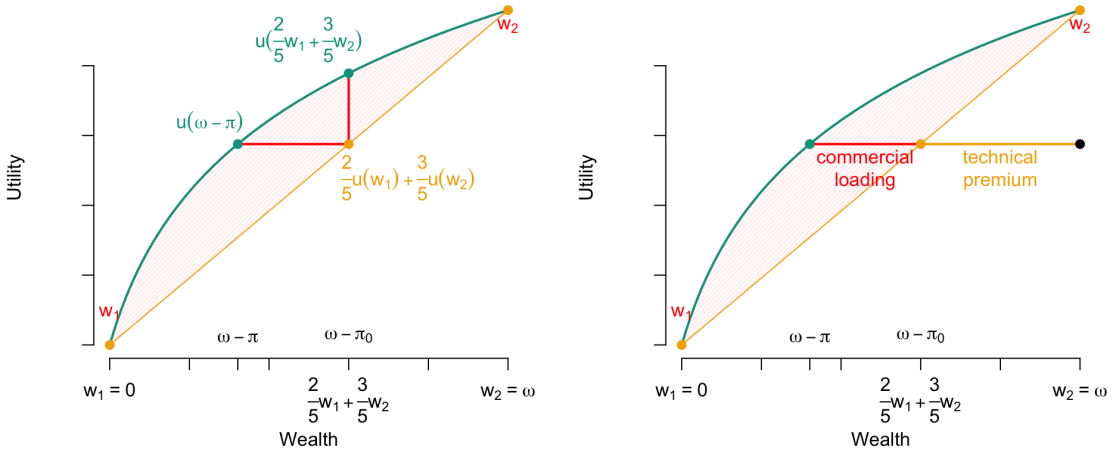


Figure 2.4: Utility and (ex-post) wealth, with an increasing concave utility function u, while the straight line correspond to a linear utility u_0 (risk neutral). Starting with initial wealth ω, agent will have random wealth W after one year, with two possible states: either w_1 (complete loss, on the left part of the x-axis) or $w_2 = \omega$ (no loss, on the right part of the x-axis). Complete loss occurs with 40% chance ($^2/_5$). π_0 is the pure premium (corresponding to a linear utility) while π is the commercial premium. All premiums in the colored area are high enough for the insurance company, and low enough for the policyholder.

Commercial insurance premiums will then depend on observable variables that are correlated with the individual's risk-aversion parameter, such as the gender, the age or even the race (as considered in Pope and Sydnor (2011)). Such discrimination might violate insurance regulations with respect to discrimination.

In the context of heterogeneity on the underlying risk only, consider the case where heterogeneity is captured through covariates $\boldsymbol{X}$ and that agents have the same wealth w and the same utility u,

$$\begin{cases} \pi_0(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}] > 0 & \text{is the actuarial premium} \\ \pi - \pi_0 = w - \mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}] - u^{-1}\left(\mathbb{E}\left[u(w-Y)\middle|\boldsymbol{X}=\boldsymbol{x}\right]\right) \geq 0 & \text{is the commercial loading.} \end{cases}$$

For example, on Figure 2.5, we have on the left, the same example as Figure 2.4, corresponding to some "good" risk,

$$Y = \begin{cases} y_2 = w \text{ with probability } p = {}^2/_5 \\ y_1 = 0 \text{ with probability } 1-p = {}^3/_5. \end{cases}$$

On the right, we have some "bad risk", where the value of the loss is unchanged, but the probability to claim a loss is higher ($p' > p$). On Figure 2.5

$$Y = \begin{cases} y_2 = w \text{ with probability } p' = {}^3/_5 > {}^2/_5 \\ y_1 = 0 \text{ with probability } 1-p' = {}^2/_5. \end{cases}$$

In that case, it could be seen as legitimate, and fair, to ask a higher technical premium, and possibly to add then the appropriate loading.

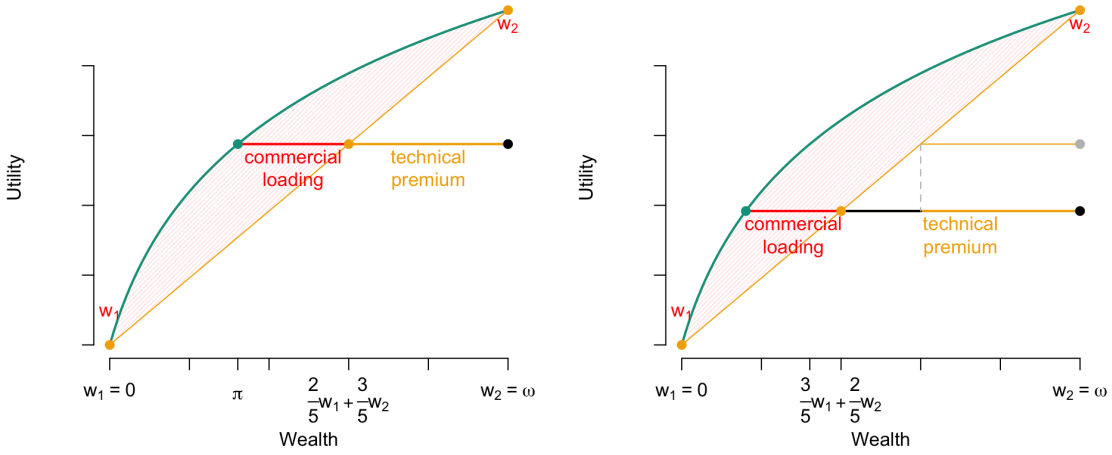


Figure 2.5: On the left, same graph as on Figure 2.4, with utility and (ex-post) wealth, with an increasing concave utility function u, and random wealth W after one year, with two possible states: either w_1 (complete loss) or $w_2 = \omega$(no loss). Complete loss occurs with 40% chance ($^2/_5$) on the left, while complete loss occurs with 60% chance ($^3/_5$) on the right. Agents have identical risk aversion and wealth, on both graphs, technical premium is higher when risk is higher (additional **black** part), commercial loading is here the same.

If heterogeneity is no longer on the underlying risk, but on the risk aversion (or possibly the wealth), if u is now function of some covariates, $u_{\boldsymbol{x}}$ we should write

$$\begin{cases} \pi_0(\boldsymbol{x}) = \mu(\boldsymbol{x}) = \mathbb{E}[Y] > 0 & : \text{ actuarial premium} \\ \pi - \pi_0 = w - \mathbb{E}[Y] - u_{\boldsymbol{x}}^{-1}\left(\mathbb{E}\left[u_{\boldsymbol{x}}(w-Y)\right]\right) \geq 0 & : \text{ commercial loading.} \end{cases}$$

Here, we used he expected utility approach, from Von Neumann and Morgenstern (1953), to illustrate, but alternatives could be considered.

The approach described previously is also named “differential pricing", where customers with a similar risk are charged different premiums (for reasons other than risk occurrence and magnitude). In this line, Central Bank of Ireland (2021) considered “price walking" as discriminatory. “Price walking" corresponds to the case where longstanding, loyal policyholders are charged higher prices for the same services compared to customers that have just switched to that provider. This is a well-documented practice in the telecommunications industry, that can be observed also in insurance (see Guelman et al. (2012) or He et al. (2020) who model attrition rate or “customer churn"). According to Central Bank of Ireland (2021), the practice of price walking is “unfair" and could result in unfair outcomes for some groups of consumers, both in the private motor insurance and household insurance markets. For example, long-term customers (those who stayed with the same insurer for nine years or more) pay, on average, 14% more on private car insurance and 32% more on home insurance than the equivalent customer renewing for the first time.

Box 2.2 **Price Optimization in the United States**

- **Alaska**, Wing-Heir (2015)

The practice of adjusting either the otherwise applicable manual rates or premiums or the actuarially indicated rates or premiums based on any of the following is considered inconsistent with the statutory requirement that "rates shall not be (...) unfairly discriminatory", whether or not such adjustment is included within the insurer's rating plan:

a. *Price elasticity of demand;*
b. *Propensity to shop for insurance;*
c. *Retention adjustment at an individual level; and*
d. *A policyholder's propensity to ask questions or file complaints.*

- **California**, Volkmer (2015)

Price Optimization does not seek to arrive at an actuarially sound estimate of the risk of loss and other future costs of a risk transfer. Therefore, any use of Price Optimization in the ratemaking/pricing process or in a rating plan is unfairly discriminatory in violation of California law.

- **district of Columbia**, Taylor (2015)

Price optimization refers to an insurer's practice of charging the maximum premium that it expects an individual or class of individuals to bear, based upon factors that are neither risk of loss related nor estimated expense related. For example, an insurer may charge a non-price sensitive individual a higher premium than it would charge a price sensitive individual; despite their risk characteristics being equal. This practice is discriminatory and it violates the District's anti-discrimination insurance laws codified at D.C. Official Code §31-2231.13(c), 31-2703(a) and 31-2703(b).

- **Pennsylvania**, Miller (2015b)

With the advent of sophisticated pricing tools, including computer software and rating models referred to as price optimization, insurers, rating organizations and advisory organizations are reminded that policyholders and applicants with identical risk classification profiles—that is, risks of the same class and essentially the same hazard—must be charged the same premium. Rates that fail to reflect differences in expected losses and expenses with reasonable accuracy are unfairly discriminatory under Commonwealth law and will not be approved by the Department.

2.6.3 Price Optimization and Discrimination

"*We define price optimization in P&C insurance as the supplementation of traditional supply-side actuarial models with quantitative customer demand models*," explained Bugbee et al. (2014). Duncan and McPhail (2013), Guven and McPhail (2013) and Spedicato et al. (2018) mention that such practices are intensively discussed by practitioners, even if they did not get much attention in academic journal. Notable exceptions would be Morel et al. (2003) that introduced myopic pricing, while more realistic approaches, named "semi-myopic pricing strategies, were discussed in Krikler et al. (2004) or more recently Ban and Keskin (2021).

Many regulators believe that price optimization is unfairly discriminatory (as shown in Box 2.2, with some regulations in some states, in the U.S.). Is it a legitimate discrimination to have premiums function on willingness or ability to pay, and risk aversion? According to the Code of Professional Conduct[5] (Precept 1, on "Professional Integrity"), "*an actuary shall act honestly (...) to fulfill the profession's responsibility to the public and to uphold the reputation of the actuarial profession.*"

[5]See https://www.soa.org/about/governance/about-code-of-professional-conduct/

2.7 Other Models in Insurance

So far, we have discussed only premium principles, but predictive models are used almost everywhere in insurance.

2.7.1 Claims Reserving and IBNR

Another interesting application is reserving (see Hesselager and Verrall (2006) or Wüthrich and Merz (2008) for more details). Loss reserves are a major item in the financial statement of an insurance company and in terms of how it is valued from the perspective of possible investors. The development and the release of reserves are furthermore important input variables to calculate the MCEV (market consistent embedded value), which "*provides a means of measuring the value of such business at any point in time and of assessing the financial performance of the business over time*," as explained in American Academy of Actuaries (2011). Hence, the estimates of unpaid losses give management important input for their strategy, pricing and underwriting. A reliable estimate of the expected losses is therefore crucial. Traditional models of reserving for future claims are mainly based on claims triangles (e.g., Chain Ladder or Bornhütter-Ferguson as distribution-free methods, as described in De Alba (2004)) or distribution-based (stochastic) models with aggregated data on the level of the gross insurance portfolio or on the level of a subportfolio as the methodology requires the use of portfolio-based parameters, e.g., reported or paid losses, prior expected parameters like losses or premiums. The reserving amount can be influenced by many factors, for example the composition of the claim, medical advancement, life expectancy, legal changes, etc. The consequence is a loss of potentially valuable information on the level of the single contract as the determining drivers are entirely disregarded.

2.7.2 Fraud Detection

Fraud is not self-revealed, and therefore, it must be investigated said Guillen (2006) and Guillen and Ayuso (2008). Tools for detecting fraud span all kind of actions undertaken by insurers. They may involve human resources, data mining (see Neural Networks), external advisors, statistical analysis, and monitoring. The currently available methods to detect fraudulent or suspicious claims based on human resources rely on video or audiotape surveillance, manual indicator cards, internal audits, and information collected from agents or informants. Methods based on data analysis seek external and internal data information. Automated methods use various machine learning techniques, such as selecting fuzzy set clustering in Derrig and Ostaszewski (1995) simple regression models in Derrig and Weisberg (1998), or GLMS, with a logistic regression in Artıs et al. (1999); Artís et al. (2002) and a probit model in Belhadji et al. (2000) or neural networks, in Brockett et al. (1998) or Viaene et al. (2005).

2.7.3 Mortality

The modeling and forecasting of mortality rates have been subject to extensive research in the past. The most widely used approach is the "Lee-Carter Model", from Lee and Carter (1992) with its numerous extensions. More recent approaches involve non-linear regression and generalized linear models (GLM). But recently, many machine learning algorithms have been used to detect (unknown) patterns, such as Levantesi and Pizzorusso (2019), with decision trees, random forests, and gradient boosting. Perla et al. (2021) generalized the Lee-Carter Model with a simple convolutional network.

2.7.4 Parametric Insurance

Parametric insurance is also an area where predictive models are important. Here, we consider guaranted payment of a predetermined amount of an insurance claim upon the occurrence of a stipulated triggering event, which must be some predefined parameter or metric specifically related to the insured's particular risk exposure, as explained in Hillier (2022) or Jerry (2023)

2.7.5 Data and Models to Understand the Risks

If most previous section where about making money using data and models, from the insurance perspective, it should be also highlighted that insurance companies have helped improve the quality of life in many countries, using data they collected. A classic example is the development of epidemiology. Indeed, as early as the XIX-th century, insurance doctors initiated an approach that prefigures systematic medical examinations, developing our contemporary medicine, based on prevention, or more and more oriented towards patients who ignore themselves. As early as 1905, John Welton Fischer, medical director of the Northwestern Mutual Life Insurance Company and a member of the Association of Life Insurance Medical Directors of America, became interested in the routine measurement of blood pressure in the examination of life insurance applicants. He was the first to do so at a time when the blood pressure monitor, which had just been invented, had not really proved its worth and remained He was the first to do so at a time when the newly invented tensiometer had not really proved its worth and was still confined to experimental use. At the beginning of 1907, Fischer began to measure the systolic blood pressure of applicants aged between 40 and 60 years. He then instructed his company's physicians to perform this measurement in cities with more than 100,000 inhabitants. By 1913, 85% of his company's applicants had had their blood pressure measured, a reminder of his diagnosis. Although Fischer's conclusions are clear, he does not explain how he foresaw the importance of this measurement as a risk factor. When Fischer proposed the introduction of blood pressure measurement for the newly insured, there was no information on the prognosis associated with elevated blood pressure, nor was there a clear definition of what "normal" pressure should be, as Kotchen (2011) recalls. The relationship between blood pressure and cardiovascular morbidity was still completely unknown, despite some work by clinicians. Insurance companies produced the first prospective statistics for hypertension, a term that did not then refer to any well-defined disease or concept. In 1911, Fischer wrote a letter to the Medical Directors Association explaining to his peers that "*the sphygmomanometer is indispensable in life insurance examinations, and the time is not far distant when all progressive life insurance companies will require its use in all examinations of applicants for life insurance*."

In 1915, the Prudential Life Insurance Company had already measured the blood pressure of 18,637 applicants, the New York Life Insurance Company of 62,000 applicants for insurance and, in 1922, the New York experiment of the Metropolitan Life Insurance Company totalled 500,000 examinations in more than 8,000 insureds, recalls Dingman (1927). No private practitioner, no hospital doctor, no organization, until then, had been able to compile such statistics. In a series of reports that began with Dublin (1925), the Actuarial Society of America described the population-based distribution of blood pressure, the age-related increases of blood pressure, and the relationships of blood pressure to both body size and mortality. This report studied a cohort of 20,000 insured persons, aged 38 to 42 years, with measurements of systolic and diastolic blood pressure. The report showed an increase in systolic and diastolic blood pressures with age. At younger ages, systolic and diastolic blood pressures were lower in women than in men. Blood pressure also increased progressively with age in both men and women. The report also showed that systolic and diastolic blood pressures increased with height in men, defined in terms of "build groups" (average weight for each inch of height) in different age groups of men. He eventually noted that changes in diastolic blood pressure were more important than changes in systolic blood pressure in predicting mortality. For insurers,

this information, although measured on an ad-hoc basis, was sufficient to exclude certain individuals or to increase their insurance premiums. The designation of hypertension as a risk factor for reduced life expectancy was not based on research into the risk factors of hypertension, but on a simple measure of correlation and risk analysis. And the existence of a correlation did not necessarily indicate a causal link, but this was not the concern of the many physicians working for insurers. Medical research was then able to work on a better understanding of these phenomena, observed by the insurers, who had access to these data (because they had the good idea to collect them).

Chapter 3

Models: Overview on Predictive Models

In this chapter, we will give an overview on predictive modeling, used by actuaries. Historically, we moved from relatively homogeneous portfolios to tariff classes, and then to modern insurance the concept of “premium personalization". Modern modelling techniques will be presented, starting with econometric approaches before presenting machine learning techniques.

As we’ve seen in the previous chapter, insurance is deeply related to predictive modeling. But contrary to popular opinion that models and algorithms are purely objective, O’Neil (2016) explains in her book that “*models are opinions embedded in mathematics (...) A model’s blind spots reflect the judgments and priorities of its creators*." In this chapter (and the next one), we will get back to general ideas about actuarial modeling.

3.1 Predictive Model, Algorithms and “Artificial Intelligence"

3.1.1 Probabilities and Predictions

We will not start a philosophical discussion about risk and uncertainty, here, but in actuarial science, all stories begin with a probabilistic model. “*Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means*" said Bertrand Russell in a conference, back in the early 30’s, quoted in Bell (1945) . Very often, the “physical” probabilities receive an objective value, on the basis of the law of large numbers, since the empirical frequency converge towards “the probability" (frequentist theory of probabilities),

Propoition 3.1.1 (Law of Large Numbers (1)) *Consider an infinite collection of i.i.d. random variables* $X, X_1, X_2, \cdots, X_n, \cdots$ *in a probabilistic space* $(\Omega, \mathcal{F}, \mathbb{P})$*, then*

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i \in \mathcal{A})}_{\text{(empirical) frequency}} \overset{a.s.}{\to} \underbrace{\mathbb{P}(X \in \mathcal{A})}_{\text{probability}} \text{ as } n \to \infty.$$

Strong law of large numbers (also called Kolmogorov's law), see Loève (1977), or any probability textbook.

This is a so-called "physical" probability, or "objective". Of course, there is no need to repeat throws of dice to affirm that (with a perfectly balanced die) the probability of obtaining 6 at the time of a throw is equal to $1/6$ (by symmetry of the cube). But if we repeat the experience of throwing a dice millions of time, $1/6$ should be close to the frequency of appearance of 6. Almost 200 years ago, Cournot (1843) already distinguished a "objective meaning" of the probability (as measure of the physical possibility of realization of a random event) and a "subjective meaning" (the probability being a judgement made on an event, this judgement being linked to the ignorance of judgment being linked to the ignorance of the conditions of the realization of the event).

But if we use that definition, we are unable to make sense of the probability of a "*single singular event*," as noted by von Mises (1928, 1939): "*When we speak of the 'probability of death', the exact meaning of this expression can be defined in the following way only. We must not think of an individual, but of a certain class as a whole, e.g., 'all insured men forty-one years old living in a given country and not engaged in certain dangerous occupations'. A probability of death is attached to the class of men or to another class that can be defined in a similar way. We can say nothing about the probability of death of an individual even if we know his condition of life and health in detail. The phrase 'probability of death', when it refers to a single person, has no meaning for us at all.*" And there are even deeper paradoxes, that can be related to latent risk factors discussed in the previous chapter, and the "true underlying probability" (to claim a loss, or to die). In a legal context, Fenton and Neil (2018) quoted a judge, who was told that a person what less than 50% to be guilty: "*look, the guy either did it or he didn't do it. If he did then he is 100% guilty and if he didn't then he is 0% guilty; so giving the chances of guilt as a probability somewhere in between makes no sense and has no place in the law*." The main difference with actuarial pricing, is that we should estimate probabilities associated with future event. But still, one can wonder if "the true probability" is a concept that makes sense when signing a contract. Thus, the goal here will be to train a model that will compute a score, that might be interpreted as a "probability" (this will raise the question of " calibration" of a model, the connection between that score and the "observed frequencies" (interpreted as probabilities), as discussed in Section 4.3.3).

A classical convention in actuarial literature is to suppose that random variables live in a probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$, without much discussions about the probability measure $\mathbb{P}$. In most sections of this book, $\mathbb{P}$ is the (unobservable) probability measure associated with the portfolio of the insurer (or associated with the training dataset). For instance, using the validation dataset from `germancredit`, let X denote the age of the person with a load, $\widehat{Y}$ the probability to have a default (with a logistic regression), and the gender is the sensitive attribute $S \in \{\texttt{A}, \texttt{B}\}$, and

$$\begin{cases} \mathbb{P}_n[X \in [18;25] | S = \texttt{A}] = 32\% \text{ and } \mathbb{P}_n[\widehat{Y} > 50\% | S = \texttt{A}] = 25\% \\ \mathbb{P}_n[X \in [18;25] | S = \texttt{B}] = 10\% \text{ and } \mathbb{P}_n[\widehat{Y} > 50\% | S = \texttt{B}] = 21\%. \end{cases}$$

But because there is competition in the market, $\mathbb{P}_n$ (and probably $\mathbb{P}$) can be different than $\mathbb{P}_0$, the probability measure for the entire population

$$\begin{cases} \mathbb{P}_0[X \in [18;25] | S = \texttt{A}] = 20\% \text{ and } \mathbb{P}_0[\widehat{Y} > 50\% | S = \texttt{A}] = 20\% \\ \mathbb{P}_0[X \in [18;25] | S = \texttt{B}] = 15\% \text{ and } \mathbb{P}_0[\widehat{Y} > 50\% | S = \texttt{B}] = 15\%. \end{cases}$$

There could be also some target probability measure $\mathbb{P}_\star$ since underwriters can be willing to target some specific segments of the population, as discussed in chapters and 7 and 12 (on change of measures),

$$\begin{cases} \mathbb{P}_\star[X \in [18;25] | S = \texttt{A}] = 25\% \text{ and } \mathbb{P}_\star[\widehat{Y} > 50\% | S = \texttt{A}] = 25\% \\ \mathbb{P}_\star[X \in [18;25] | S = \texttt{B}] = 20\% \text{ and } \mathbb{P}_\star[\widehat{Y} > 50\% | S = \texttt{B}] = 20\%. \end{cases}$$

or possibly some "fair" measure $\mathbb{Q}$, as we will discuss in chapters 8 (on quantifying group fairness) and 12 (when mitigating discrimination), that will satisfy some independence properties,

$$\begin{cases}\mathbb{Q}[X \in [18;25]|S=\texttt{A}] = 39\% \text{ and } \mathbb{Q}[\widehat{Y}>50\%|S=\texttt{A}] = 23\% \\ \mathbb{Q}[X \in [18;25]|S=\texttt{B}] = 15\% \text{ and } \mathbb{Q}[\widehat{Y}>50\%|S=\texttt{B}] = 23\%.\end{cases}$$

It is also possible to mention here the fact that the model is fitted on past data, associated with probability measure, $\mathbb{P}_n$ but because of the competition on the market, or because of the general economic context, the structure of the portfolio might change. The probability measure for next year will then be $\mathbb{P}'_{n'}$ (with

$$\begin{cases}\mathbb{P}'_{n'}[X \in [18;25]|S=\texttt{A}] = 35\% \text{ and } \mathbb{P}'_{n'}[\widehat{Y}>50\%|S=\texttt{A}] = 27\% \\ \mathbb{P}'_{n'}[X \in [18;25]|S=\texttt{B}] = 20\% \text{ and } \mathbb{P}'_{n'}[\widehat{Y}>50\%|S=\texttt{B}] = 27\%,\end{cases}$$

if our score, used to assess whether we give a loan to some clients attracts more young (and more risky) people. We will not discuss this issue here, but the "generalization" property should be with respect to new un-observable and hard to predict probability measure $\mathbb{P}'_{n'}$ (and not $\mathbb{P}$ as usually considered in machine learning, as discussed in the next sections).

Just a short comment before going further: in the equations above, the vertical bar corresponds to a "conditional probability". $\mathbb{P}[\mathcal{A}|\mathcal{B}]$ is the conditional probability probability that event A occurs given the information that event $\mathcal{B}$ occurred. It is the ratio of the probability that both $\mathcal{A}$ and $\mathcal{B}$ occurred (corresponding to $\mathbb{P}[\mathcal{A}\cap\mathcal{B}]$) over the probability that $\mathcal{B}$ occurred. Bayes on that definition, we can derive Bayes formula,

Propoition 3.1.2 (Bayes formula) *Given two events $\mathcal{A}$ and $\mathcal{B}$ such that $\mathbb{P}[\mathcal{B}]\neq 0$,*

$$\mathbb{P}[\mathcal{A}|\mathcal{B}] = \frac{\mathbb{P}[\mathcal{B}|\mathcal{A}]\cdot\mathbb{P}[\mathcal{A}]}{\mathbb{P}[\mathcal{B}]} \propto \mathbb{P}[\mathcal{B}|\mathcal{A}]\cdot\mathbb{P}[\mathcal{A}].$$

Bayes (1763) and Laplace (1774). Besides the mathematical expression, that formula has two possible interpretations. The first one corresponds to an "update of beliefs", from a *prior* distribution $\mathbb{P}[\mathcal{A}]$ to a *posterior* distribution $\mathbb{P}[\mathcal{A}|\mathcal{B}]$, given some additional information B. The second one is related to a "inverse problem," where we try to determine the causes of a phenomenon from the experimental observation of its effects. A classical example could be the one where $\mathcal{A}$ is a disease and $\mathcal{B}$ is a symptom (or a set of symptoms), and with Bayes' rule (see Spiegelhalter et al. (1993) for more details, with multiple diseases and multiple symptoms)

$$\mathbb{P}[\text{disease}|\text{symptom}] \propto \mathbb{P}[\text{symptom}|\text{disease}]\cdot\mathbb{P}[\text{disease}].$$

Another close example would be the one where $\mathcal{B}$ is the result of a test, and

$$\mathbb{P}[\text{cancer}|\text{test positive}] \propto \mathbb{P}[\text{test positive}|\text{cancer}]\cdot\mathbb{P}[\text{cancer}].$$

3.1.2 Models

Predictive models are used to capture a relationship between a response variable y (typically a claim occurrence, a claim frequency, a claim severity, or an annual cost) and a collection of predictors, denoted[1] $\boldsymbol{x}$, as explained in Schmidt (2006). If y is binary, a classical model will be the logistic regression for example

[1]As discussed previously, notation $\boldsymbol{z}$ will also be used later on, and we will distinguish admissible predictors $\boldsymbol{x}$, and sensitive ones $\boldsymbol{s}$. In this chapter, we will mainly use $\boldsymbol{x}$, as in most textbooks.

(see Dierckx (2006)), and more generally, actuaries have used intensively Generalized Linear Models (GLM, see Frees (2006) or Denuit et al. (2019a)) where a prediction of the outcome y is obtained by transforming a linear combination of predictors. Econometric models (and GLM) are popular since they rely strongly on probabilistic models, and the insurance business is based on randomness of events.

For Ekeland (1995), modelling is the (intellectual) construction of a mathematical model, i.e., a network of equations supposed to describe reality. Very often, a model is also, above all, a simplification of this reality. A model that is too complex is not a good model. This is the idea of over-learning (or "overfit") that is found in statistics, or the concept of parsimony, sometimes called "Ockham's razor" (as in Figure 3.1), which is typical in econometrics and discussed by William of Ockham (in the XIV-th century). As Milanković (1920) stated, "*in order to be able to translate the phenomena of nature into mathematical language, it is always necessary to admit simplifications and to simplify certain influences and irregularities*." The model is a simplification of the world, or, as Korzybski (1958) said in a geography context, "*a map is not the territory it represents, but, if correct, it has a similar structure to the territory, which accounts for its usefulness*." The map is not the territory : the map reflects our representation of the world, whereas the territory is the world as it really is. We naturally think of Borges (1946) (or Umberto Ecco's pastiche *of the impossibility of constructing the map 1: 1 of the Empire*, in Eco (1992))[2], "*en aquel Imperio, el Arte de la Cartografía logró tal Perfección que el mapa de una sola Provincia ocupaba toda una Ciudad, y el mapa del Imperio, toda una Provincia. Con el tiempo, estos Mapas Desmesurados no satisficieron y los Colegios de Cartógrafos levantaron un Mapa del Imperio, que tenía el tamaño del Imperio y coincidía puntualmente con él.*"

The notion of model will increasingly be replaced by the term "algorithm", or even "artificial intelligence", or shortly A.I. (notably in the press, see Milmo (2021), Swauger (2021) or Smith (2021) among many others). For Zafar et al. (2019), "algorithm" means predictive models (decision rules) calibrated from historical data through data mining techniques. To understand the difference, Cardon (2019) gives an example to explain what machine learning is. It is quite simple to write a program that converts a temperature in degrees Fahrenheit to a temperature given in degrees Celsius. To do this, there is a simple rule: subtract 32 from the temperature in degrees Fahrenheit x and multiply the result by $5/9$ (or divide by 1.8), i.e.

$$y \leftarrow \frac{5}{9}(x - 32).$$

A machine learning (or artificial intelligence) approach offers a very different solution. Instead of coding the rule into the machine (what computer scientists might call "*Good Old Fashioned Artificial Intelligence*," as Haugeland (1989)), we simply give to the machine several examples of matches between temperatures in Celsius and Fahrenheit (x_i, y_i). We enter the data into a training database, and the algorithm will learn a conversion rule by itself, looking for the closest candidate function to the data. We can then find an example like the one in Figure 3.1, with some data and different models (one simple (linear) and one more complex).

It is worth noting that the "complexity" of certain algorithms, or their "opacity" (which leads to the term "black box"), has nothing to do with the optimisation algorithm used (in deep learning, back-propagation is simply an iterative mechanism for optimising a clearly described objective). It is mainly that the model obtained may seem complex, impenetrable, to take into account the possible interactions between the predictor variables for example. For the sake of completeness, a distinction should be made between classical supervised machine learning algorithms and reinforcement learning techniques. The latter case describes sequential (or iterative) learning methods, where the algorithm learns by experimenting, as described in Charpentier et al. (2021). We will find these algorithms in automatic driving for example, or if we wanted to

[2] "*In that Empire, the Art of Cartography achieved such Perfection that the map of a single Province occupied an entire City, and the map of the Empire, an entire Province. In time, these inordinate maps were not satisfactory and the Colleges of Cartographers drew up a map of the Empire, which was the size of the Empire and coincided exactly with it*" [personal translation].

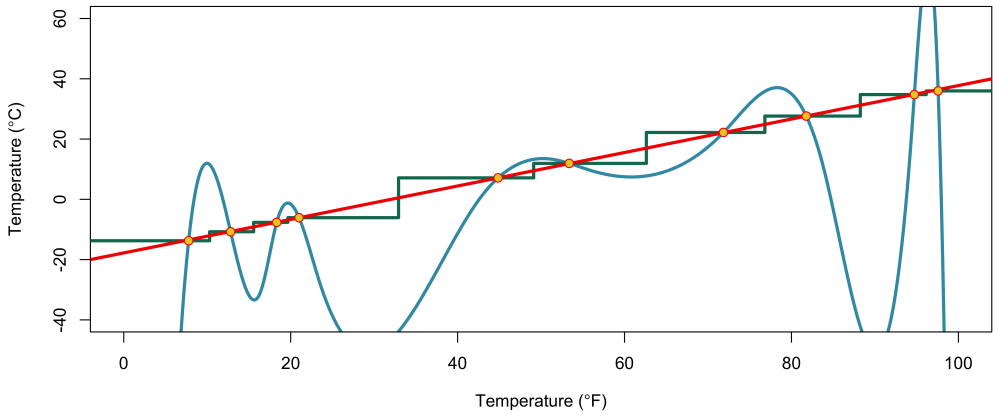


Figure 3.1: A simple (linear) model, a piecewise constant model, or a complex (non-linear but continuous) model, from about ten observations (x_i, y_i) where x is a temperature in degrees Fahrenheit and y is the temperature in degrees Celsius, at the same location i.

model correctly the links between the data, the constructed model, the new connected data, the update of the model, etc. But we will not insist more on this class of models here.

To conclude this introduction, let us stress that in insurance models, the goal is not to predict "who" will die, and get involved in a car accident. Actuaries create scores, that will be interpreted as the probability to die, or the probability to get a bodily injury claims, in order to compute "fair" premiums. To use a classic statistical example, let y denote the face of a dice, potentially lumpy. If p is the (true) probability of falling on 6 (say 14.5752%), we say at first that we are able to acquire information about the way the dice was made, about its roughness, its imperfections, that will allow us to refine our knowledge on this probability, but also that we have a model able to link the information in an adequate way. Knowing better the probability of falling on 6 does not guarantee that the dice will fall on a 6, the random component does not disappear, and will never disappear. Translated into the insurance problem, p might denote the "true" probability that a specific policyholder will get involved in a car accident. Based on external information $\boldsymbol{x}$, some model will predict that the probability to get an accident is $\widehat{p}_{\boldsymbol{x}}$ (say 11.1245%). As mentioned by French statistician Alfred Sauvy, "*dans toute statistique, l'inexactitude du nombre est compensée par la précision des décimales*" (or "*in all statistics, the inaccuracy of the number is compensated by the precision of the decimals*," infinite precision we might add). The goal is not to find a model returns either 0% or 100% (this will happen with several machine learning algorithms), simply to assess with confidence a valid probability, used to compute for a "fair" premium. And an easy way is perhaps to use simple categories: probability to get a 6 less than 15% (less than a fair dice), between 15% and 18.5% (close to a fair dice) and more than 18.5% (more than a fair dice).

3.2 From Categorical to Continuous Models

“Risk classification" is an old and natural way to get insurance premiums, as explained in Section 2.3.3. Not only higher-rated insureds are less likely to engage in the risky activity, risk classification provides incentives for risk reduction (merit rating in auto insurance encourages adherence to traffic regulations; experience-rating in workers' compensation encourages employers to eliminate workplace hazards, etc). And as suggested by Feldblum (2006), risk classification also promotes social welfare by making insurance coverage more widely available.

3.2.1 Historical Perspective, Insurers as Clubs

In ancient Rome, a *collegium* (plural *collegia*) was an association. They functioned as social clubs, or religious collectives, whose members worked towards their shared interests, as explained in Verboven (2011). During Republican Rome (that began with the overthrow of the Roman Kingdom, 509 before the common era, and ended with the establishment of the Roman Empire, 27 before our era), military *collegia* were created. As explained in Ginsburg (1940), upon the completion of his service a veteran had the right to join one of the many *collegia veteranorum*. The Government established in each legion special savings banks. Half of the cash bonuses, *donativa*, which the emperors made to the soldiers on various occasions, was not handed over to the beneficiaries in cash but was deposited to the account of each soldier in his legion's savings ban. This could be seen as an insurance scheme, and risks against which a member was insured were diverse. In case of retirement, upon the completion of his term of service, he received a premium which helped him somewhat to arrange the rest of his life. The membership in a *collegium* gave him a mutual insurance against “*unforeseen risks*." These *collegia*, besides being cooperative insurance companies, had other functions. And because of the structure of those *collegia* based on corporatism, members were quite homogeneous.

Sometime in the early 1660th, the *Pirate's Code* was supposedly written by the Portuguese buccaneer Bartolomeu Português. An interestingly, a section is explicitly dedicated to insurance and benefits: “*a standard compensation is provided for maimed and mutilated buccaneers. Thus they order for the loss of a right arm six hundred pieces of eight, or six slaves; for the loss of a left arm five hundred pieces of eight, or five slaves; for a right leg five hundred pieces of eight, or five slaves; for the left leg four hundred pieces of eight, or four slaves; for an eye one hundred pieces of eight, or one slave; for a finger of the hand the same reward as for the eye*," see Barbour (1911) (or more recently Leeson (2009) and Fox (2013) about this piratical schemes).

In the XIX-th century, in Europe, mutual aid societies involved a group of individuals who made regular payments into a common fund in order to provide for themselves in later, unforeseeable moments of financial hardship or of old age. As mentioned by Garrioch (2011), in 1848, there were in Paris 280 mutual aid societies with well over 20,000 members. For example, the *Société des Arts Graphiques*, was created in 1808. It admitted only men over twenty and under fifty, and it charged much higher admission and annual fees for those who joined at a more advanced age. In return, they received benefits if they were unable to work, reducing over a period of time, but in case of serious illness the Society would pay the admission fee for a hospice. In England, there were “friendly societies", as described in Ismay (2018). In France, after the 1848 revolution and Louis Napoléon Bonaparte's coup d'état in 1851, mutual funds were seen as a means of harmonizing social classes. The money collected through contributions came to the rescue of unfortunate workers, who would no longer have any reason to radicalize. It was proposed that insurance should become compulsory (Bismark proposed this in Germany in 1883), but the idea was rejected in favor of giving workers the freedom to contribute, as the only way to moralize the working classes, as Da Silva (2023) explains. In 1852, of the 236 mutual funds created, 21 were on a professional basis, while the other 215 were on a territorial basis.

And from 1870 onwards, mutual funds diversified the professional profile of contributors beyond blue-collar workers, and expanded to include employees, civil servants, the self-employed and artists. But the amount of the premium is not linked to the risk. As Da Silva (2023) puts it,"*mutual insurers see in the actuarial figure the programmed end of solidarity.*" For mutual funds, solidarity is essential, with everyone contributing according to their means and receiving according to their needs. Around the same time, in France, the first insurance companies appeared, based on risk selection, and the first mathematical approaches to calculating premiums. Hubbard (1852) advocates the introduction of an "*English-style scientific organization*" in their management . For its members, they must, like insurance companies, be able to know "*the probable average of the claims*" that they might have to cover. The development of tables should lead insurers to adopt the principle of contributions varying according to the age of entry and the specialization of contributions and funds (health/retirement). It is with this in mind that they draw up tables. For Stone (1993) and Gowri (2014) the defining feature of "modern insurance" is its reliance on segmenting the risk pool into distinct categories, each receiving a price corresponding to the particular risk that the individuals assigned to that category are expected to represent (as accurately as can be estimated by actuaries).

3.2.2 "Modern Insurance" and Categorization

Once heterogeneity with respect to the risk was observed in portfolios, insurers have operated by categorizing individuals into risk classes and assigning corresponding tariffs. This ongoing process of categorization ensures that the sums collected, on average, are sufficient to address the realized risks within specific groups. The aim of risk classification, as explained in Wortham (1986), is to identify the specific characteristics that are supposed to determine an individual's propensity to suffer an adverse event, forming groups within which the risk is (approximately) equally shared. The problem, of course, is that the characteristics associated with various types of risk are almost infinite; as they cannot all be identified and priced in every risk classification system, there will necessarily be unpriced sources of heterogeneity between individuals in a given risk class.

In 1915, as mentioned in Rothstein (2003), the president of the Association of Life Insurance Medical Directors of America noted that the question asked almost universally of the Medical Examiner was "*What is your opinion of the risk? Good, bad, first-class, second-class, or not acceptable?* ." Historically, insurance prices were a (finite) collection of prices (maybe more than than the two classes mentioned, "*first-class* and "*second-class*"). In Box 3.1, in the early 1920's, Albert Henry Mowbray, who worked for New York Life Insurance Company and later Liberty Mutual, who was also an actuary for state-level insurance commissions in New Carolina and California, and the National Council on Workmen's Insurance, gives his perspective on insurance rate making.

Several articles and textbooks in sociology tried to understand how classification mechanisms establish symbolic boundaries that reinforce group identities, such as Bourdieu (2018), Lamont and Molnár (2002), Massey (2007), Ridgeway (2011), Fourcade and Healy (2013) or Brubaker (2015). But here, those "groups" or "classes" do not share any identity, and Simon (1988) or Harcourt (2015b) use the term "actuarial classification" (where "actuarial" designates any decision-making technique that relies on predictive statistical methods, replacing more holistic or subjective forms of judgment). In those class-based systems, based on insurance rating table (or grid), results are determined by assigning individuals to a group in which each person is positioned as "average" or "typical". "*[Most] actuaries cannot think of individuals except as members of groups*" claimed Brilmayer et al. (1979). Each individual is assigned the same value as all other members of the group to which it is assigned (as opposed to models discussed in section 3.3, where model give to each individual its own unique value or score, as close as possible, as explained in Fourcade (2016)). Simon (1987, 1988), and then Feeley and Simon (1992), defined "actuarialism", that designate the use of statistics to guide "*class-based decision-making*," used to price pensions and insurance. As explained in Harcourt

Box 3.1 **Historical perspective**, Albert Henry Mowbray (1921)

"*Classification of risks in some manner forms the basis of rate making in practically all branches of insurance. It would appear therefore that there should be some fundamental principle to which a correct system of classification in any branch of insurance should conform* (...) *As long ago as the days of ancient Greece and Rome the gradual transition of natural phenomena was observed and set down in the Latin maxim,* 'natura non agit per altum'*. If each risk, therefore is to be precisely rated, it would be necessary to recognize very minute differences and precisely measure them.* (...) *Since we are not capable of covering a large field fully and at the same time recognizing small differences in all parts of the field, it is natural that we resort to subdivision of the field b y means of classification, thereby concentrating our attention on a smaller interval which may again be subdivided by further classification, and the system so carried on to the limit to which we find it necessary or desirable to go. But however far we may go in any system of classification, whether in the field of pure or applied science including the business or insurance, we shall always find difficulties presented b y the borderline case, difficulties which arise from the continuous character of natural phenomena which we are attempting to place in more or less arbitrary divisions. While thus acknowledging that classification will never completely solve the problem of recognizing differences between individuals, nevertheless classification seems to be necessary at least as a preliminary step toward such recognition in any field of study. The fact that a complete and final solution cannot be made is, therefore, no justification for completely discarding classification as a method of approach. Since it is insurance hazards that we undertake to measure and classify, the preliminary step in studying classification theory may well be to ask what is an insurance hazard and how it may be determined. It must be evident to the members of this Society that an insurance hazard is what is termed "a mathematical expectation", that is a product of a sum at risk and the probability of loss from the conditions insured against, e.g., the destruction of a piece of property by fire, the death of an individual, etc. If the net premiums collected are so determined on the basis of the true natural probability a n d there is a sufficient spread then the sums collected will just cover the losses and this is what should be*."
"*1. The classification should bring together risks which have inherent in their operation the same causes of loss. 2. The variation from risk to risk in the strength of each cause or at least of the more important should not be greater than can be handled by the formula b y which the classification is subdivided, i.e., the Schedule and / or Experience Rating Plan used. 3. The classification should not cover risks which include, as important elements of their hazard, causes which are not common to all. 4. The classification system and the formula for its extension (Schedule and / or Experience Rating Plans) should be harmonious. 5. The basis throughout should be the outward, recognizable indicia of the presence and potency of the several inherent causes of loss including extent as well as occurrence of loss.*"

(2015b), this "actuarial classification" is the constitution of groups with no experienced social significance for the participants. A person classified as a particular risk by an insurance company shares nothing with the other people so classified, apart from a series of formal characteristics (e.g. age, sex, marital status, etc.). As we will see in Section 4.1 on interpretability, actuaries try *ex-post* to give social representations to those groups. For Austin (1983) and Simon (1988), categories used by the insurance company when grouping risks are "*singularly sterile*," resulting in inert, immobile and deactivated communities, corresponding to "*artificial*" groups. These are not groups organized around a shared history, common experiences or active commitment, forming some "*aggregates*" - living only in the imagination of the actuary who calculates and tabulates, not in any lived form of human association. If Hacking (1990) observed that standard classes creates coherent group identities (causing possible stereotypes and discrimmination, as we will discuss in Part III), Simon (1988), provocatively suggests that actuarial classifications can in turn "*undo*" people's identity. As mentioned in Abraham (1986), the goal for actuaries is to create groups, or "classes" made up of individuals who share a series of common characteristics and are therefore presumed to represent the same risk. Following François (2022), we could claim that actuarial techniques reduce individuals to a series of formal roles that have no "*moral density*" and therefore do not grant an "*identity*" that organizes a coherent sense of self. And the inclusion of nominally "*demoralized categories*," such as gender, in class-based rating systems makes their total demoralization difficult to achieve - and is in itself an issue of struggle. Heimer (1985) used the term "*community of fate*." These "*communities*" created artificially by statisticians are, in

Box 3.2 **Historical perspective**, Harry S. Havens (1979)

"The price which a person pays for automobile insurance depends on age, sex, marital status, place of residence and other factors. This risk classification system produces widely differing prices for the same coverage for different people. Questions have been raised about the fairness of this system, and especially about its reliability as a predictor of risk for a particular individual. While we have not tried to judge the propriety of these groupings, and the resulting price differences, we believe that the questions about them warrant careful consideration by the State insurance departments. In most States the authority to examine classification plans is based on the requirement that insurance rates are neither inadequate, excessive, nor unfairly discriminatory. The only criterion for approving classifications in most States is that the classifications be statistically justified – that is, that they reasonably reflect loss experience. Relative rates with respect to age, sex, and marital status are based on the analysis of national data. A youthful male driver, for example, is charged twice as .much as an older driver all over the country (...) It has also been claimed that insurance companies engage in redlining –the arbitrary denial of insurance to everyone living in a particular neighborhood. Community groups and others have complained that State regulators have not been diligent in preventing redlining and other forms of improper discrimination that make insurance unavailable in certain areas. In addition to outright refusals to insure, geographic discrimination can include such practices as: selective placement of agents to reduce business in some areas, terminating agents and not renewing their book of business, pricing insurance at un-affordable levels, and instructing a.gents to avoid certain areas. We reviewed what the State insurance departments were doing in response to these problem. To determine if redlining exists, it is necessary to collect data on a geographic oasis. Such data should include current insurance policies, new policies being written, cancellations, and non-renewals. It is also important to examine data on losses by neighborhoods within existing rating territories because marked discrepancies within territories would cast doubt on the validity of territorial boundaries. let, not even a fifth of the States collect anything other than loss data, and that data is gathered on a territory-wide basis."

that sense, very different from the communities of workers, neighbors and co-religionists that characterized the traditional mutual organizations displaced by modern forms of insurance, as explained in Gosden (1961), Clark and Clark (1999), Levy (2012) or Zelizer (2017). Furthermore, Rouvroy et al. (2013) and Cheney-Lippold (2017) point out that scoring technologies are continually swapping predictors, "*shuffling the cards*," so that there is no stable basis for constructing group memberships, or a coherent sense.

Harry S. Havens in the late 1970's gave the desription mentioned in Box 3.2.

In Box 3.3, a paragraph from Casey et al. (1976) provides some historical perspective, by Barbara Casey, Jacques Pezier and Carl Spetzler.

3.2.3 Mathematics of Rating Classes

As mentioned in Section 2.5, an important theorem when modeling heterogeneity is the variance decomposition property, or "law of total variance" (corresponding to Pythagorean Theorem, see Proposition 2.5.3),

$$\text{Var}[Y] = \underbrace{\mathbb{E}[\text{Var}[Y|\Theta]]}_{\text{within}} + \underbrace{\text{Var}[\mathbb{E}[Y|\Theta]]}_{\text{between}}.$$

Here the variance of the outcome Y is decomposed in two parts, one representing the variance due to the variability of the underlying risk factor Θ, and one reflecting the inherent variability of Y if Θ did not vary (the homogeneous case). One can recognize that a similar idea is the basis for ANOVA models ("Analysis of Variance", as formalized in Fisher (1921) and Fisher and Mackenzie (1923)) where the total variability is split into the "within groups" and the "between groups". The "one-way analysis of variance", is a technique that can be used to compare whether two (or more) sample's means are significantly different or not. If the

Box 3.3 **Historical perspective**, Casey et al. (1976)

"The opinion that distinctions based on sex, or any other group variable, necessarily violate individual rights reflect ignorance of the basic rules of logical inference in that it would arbitrarily forbid the use of relevant information. It would he equally fallacious to reject a classification system based on socially acceptable variables because the results appear discriminatory. For example, a classification system may be built on use of car, mileage. merit rating, and other variables, excluding sex. However, when verifying the average rates according to sex one may discover significant differences between males and females. Refusing to allow such differences would he attempting to distort reality by choosing to be selectively blind. Indeed, the rationale that proscribing the use of certain rating variables is in the public interest because.under imperfect risk assessment systems, actuarial fairness is not achieved for some – albeit unidentifiable – individuals is fundamentally contradictory. It promotes a remedy for unfairness to some that increases the unfairness overall (by the same actuarial yardstick) and redistributes it. On the other hand, the opinion that distinction based on sex, or any other group variable. necessarily violate individual right reflects ignorance of the basic rules of logical inference in that it would arbitrarily> forbid the use of relevant information. It would be equally fallacious to reject a classification system based on socially acceptable variables because the results appear discriminatory. For example, a classification system may be built on use of car, mileage, merit rating, and other variables, excluding sex. However, when verifying the average rates according to sex one may discover significant differences between males and females. Refusing to allow such differences would be attempting to distort reality by choosing to be selectively blind. "The use of rating territories is a case in point. Geographical divisions, however designed, are often correlated with socio-demographic factors such as income level and race because of natural aggregation or forced segregation according to these factors. Again we conclude that insurance companies should be free to delineate territories and assess territorial differences as well as they can. At the same time, insurance companies should recognize that it is in their best interest to be objective and use clearly relevant factors to define territories lest they be accused of invidious discrimination by the public. One possible standard does exist for exception to the counsel that particular rating variables should not be proscribed. What we have called 'qua1 treatment' standard of fairness may precipitate a societal decision that the process of differentiating among individuals on the basis of certain variables is discriminatory and intolerable. This type of decision should be made on a specific. statutory basis. Once taken. it must be adhered to in private and public trans- actions alike and enforced by the insurance regulator. This is. in effect, a standard for conduct that by design transcends and preempts economic considerations. Because it is not applied without economic cost. however. insurance regulators and the industry should participate in and inform legislative deliberations that would ban the use of particular rating variables as discriminatory."

outcome y is continuous (extensions can be obtained for binary variables, or counts), suppose that

$$y_{i,j} = \mu_j + \varepsilon_{i,j},$$

where i is the index over individuals, and j the index over groups (with $j = 1, 2, \cdots, J$). μ_j is the mean of the observations for group j, and errors $\varepsilon_{i,j}$ are supposed to be zero-mean (normally distributed is a classical assumption). One could also write

$$y_{i,j} = \mu + \alpha_j + \varepsilon_{i,j}, \text{ where } \alpha_1 + \alpha_2 + \cdots + \alpha_J = 0,$$

where μ is the overall mean, while α_j is the deviation from the overall mean, for group j. Of course, one can generalize that model to multiple factors. In the "two-way analysis of variance", two types of groups are considered

$$y_{i,j,k} = \mu_{j,k} + \varepsilon_{i,j,k},$$

where j the index over groups according to the first factor, while k is the index over groups according to the second factor. $\mu_{j,k}$ is the mean of the observations for groups j and k, and errors $\varepsilon_{i,j,k}$ are supposed to be zero-mean. We can write the mean as a linear combination of factors, in the sense that

$$y_{i,j,k} = \underbrace{\mu + \alpha_j + \beta_k + \gamma_{j,k}}_{=\mu_{j,k}} + \varepsilon_{i,j,k},$$

where μ is still the overall mean, while α_j and β_j correspond to the deviation from the overall mean, while $\gamma_{i,k}$ is the non-additive interaction effect. In order to have identifiability of the model, some "sum-to-zero" constraints are added, as previously,

$$\sum_{j=1}^{J}\alpha_j=\sum_{j=1}^{J}\gamma_{j,k}=0 \text{ and } \sum_{k=1}^{K}\beta_k=\sum_{k=1}^{K}\gamma_{j,k}=0.$$

A more modern way to consider those models is to use linear models. For example, for the "one-way analysis of variance", we can write $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ where

$$\begin{cases}\boldsymbol{y}=(y_{1,1},\cdots,y_{n_1,1},y_{1,2},\cdots,y_{n_2,2},\cdots,y_{1,J},\cdots,y_{n_J,J})\\ \boldsymbol{\varepsilon}=(\varepsilon_{1,1},\cdots,\varepsilon_{n_1,1},\varepsilon_{1,2},\cdots,y_{n_2,2},\cdots,\varepsilon_{1,J},\cdots,\varepsilon_{n_J,J})\end{cases}$$

$\boldsymbol{\beta}=(\beta_0,\beta_1,\cdots,\beta_J)^\top$ and

$$\boldsymbol{X}=[\mathbf{1}_n,\boldsymbol{A}] \text{ where } \boldsymbol{A}=\begin{pmatrix}\mathbf{1}_{n_1}&\mathbf{0}&\cdots&\mathbf{0}\\ \mathbf{0}&\mathbf{1}_{n_2}&\cdots&\mathbf{0}\\ \vdots&\vdots&&\vdots\\ \mathbf{0}&\mathbf{0}&\cdots&\mathbf{1}_{n_J}\end{pmatrix}, \text{ and } \boldsymbol{X}=\begin{pmatrix}\mathbf{1}_{n_1}&\mathbf{1}_{n_1}&\mathbf{0}&\cdots&\mathbf{0}\\ \mathbf{1}_{n_2}&\mathbf{0}&\mathbf{1}_{n_2}&\cdots&\mathbf{0}\\ \vdots&\vdots&\vdots&&\vdots\\ \mathbf{1}_{n_J}&\mathbf{0}&\mathbf{0}&\cdots&\mathbf{1}_{n_J}\end{pmatrix}$$

are respectively $n\times J$ and $n\times(J+1)$ matrices. In the first approach, $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, and the ordinary least square estimate is

$$\widehat{\boldsymbol{\beta}}=(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{y}=(\overline{y}_{\cdot 1},\cdots,\overline{y}_{\cdot J})\in\mathbb{R}^J, \text{ where } \overline{y}_{\cdot j}=\frac{1}{n_j}\sum_{i=1}^{n_j}y_{i,j},$$

so that $\widehat{\mu}_j=\overline{y}_{\cdot j}$ is simply the average within group j. In the second case, if $y_{i,j}=\mu+\alpha_j+\varepsilon_{i,j}$, where $\alpha_1+\alpha_2+\cdots+\alpha_J=0$, we can prove that

$$\widehat{\mu}=\widetilde{y}=\frac{1}{J}\sum_{j=1}^{J}\overline{y}_{\cdot j} \text{ and } \widehat{\alpha}_j=\overline{y}_{\cdot j}-\widetilde{y},$$

where the estimator of μ is the average of group averages. An alternative is to change slightly the constraint, so that $n_1\alpha_1+n_2\alpha_2+\cdots+n_J\alpha_J=0$, and in that case

$$\widehat{\mu}=\overline{y}=\frac{1}{n}\sum_{j=1}^{J}n_j\overline{y}_{\cdot j} \text{ and } \widehat{\beta}_j=\overline{y}_{\cdot j}-\overline{y}.$$

Let $j\in\{1,2,\cdots,J\}$ and $k\in\{1,2,\cdots,K\}$, and let $n_{j,k}$ denote the number of observations in group j for the first factor, and k for the second. Define averages

$$\overline{y}_{\cdot jk}=\frac{1}{n_{jk}}\sum_{i=1}^{n_{jk}}y_{ijk},\ \overline{y}_{\cdot j\cdot}=\frac{1}{n_{j\cdot}}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}y_{ijk}, \text{ and } \overline{y}_{\cdot\cdot k}=\frac{1}{n_{\cdot k}}\sum_{j=1}^{J}\sum_{i=1}^{n_{jk}}y_{ijk}.$$

The model is here

$$y_{ijk}=\mu+\alpha_j+\beta_k+\gamma_{jk}+\varepsilon_{ijk}$$

that we can write, using $(J + K + JK)$ indicators (vectors in dimension n, with respectively, in each block, $n_{j\cdot}$, $n_{\cdot k}$ and n_{jk} one's), as a classical regression problem. As previously, under identifiability assumptions, it is possible to have interpretable estimates to those quantities.

$$\widehat{\mu} = \overline{y},\ \widehat{\alpha}_j = \overline{y}_{j\cdot} - \overline{y},\ \widehat{\beta}_k = \overline{y}_{\cdot,k} - \overline{y}$$

$$\widehat{\gamma}_{jk} = \overline{y}_{jk} - \overline{y}_{j\cdot} - \overline{y}_{\cdot k} + \overline{y}$$

Without the non-additive interaction effect, the model becomes

$$y_{ijk} = \mu + \alpha_j + \beta_k + \varepsilon_{ijk}.$$

Such models where used historically in claims reserving (see Kremer (1982) for instance, for a formal connection), and, of course, in ratemaking. As explained in Bennett (1978), "*in a rating structure used in motor insurance there may typically be about eight factors, each having a number of different levels into which risks may be classified and then be charged different rates of premium* ", with either an "additive model" or a "multiplicative model" for the premium μ (with notations of Bennett (1978)),

$$\begin{cases} \widehat{\mu}_{jk\cdots} = \widehat{\mu} + \widehat{\alpha}_j + \widehat{\beta}_k + \cdots & \text{additive,} \\ \widehat{\mu}_{jk\cdots} = \widehat{\mu} \cdot \widehat{\alpha}_j \cdot \widehat{\beta}_k \cdots & \text{multiplicative,} \end{cases}$$

where α_j is a parameter value for the i-th level of first risk factor, etc., and μ is a constant corresponding to some "overall average level".

Historically, classification relativities were determined one dimension at a time (see Feldblum and Brosius (2003), and the appendices to McClenahan (2006) and Finger (2006) for some illustration of the procedure). Then Bailey and Simon (1959) and Bailey and Simon (1960) introduced the "minimum bias procedure"

On Figure 3.2, we can visualize a dozen classes associated with credit risk (on the `germancredit` database), with on the x-axis, predictions given by two models, and the empirical default probability on the y-axis (that will correspond to a discrete version of the calibration plot described in Section 4.3.3).

As discussed in Agresti (2012, 2015), there are strong connections between those approaches based on groups and linear models, and actuarial research started to move towards "continuous" models. Nevertheless, the use of categories has been popular in the industry for several decades. For example, Siddiqi (2012) recommends to cut all continuous variables into bins, using a so-called "weight-of-evidence binning" technique, usually seen as an "optimal binning" for numerical and categorical variables using methods including tree-like segmentation, or chi-square merge. In R, it can be performed using the `woebin` function of the `scorecard` package. For example, on the `germancredit` dataset, three continuous variables are divided into bins, as on Figure 3.3. For the duration (in months), bins are A=$[0, 8)$, B=$[8, 16)$, C=$[16, 34)$, D=$[34, 44)$ and E=$[44, 80)$; for the credit amount, bins are A=$[0, 1400)$, B=$[1400, 1800)$, C=$[1800, 4000)$, D=$[4000, 9200)$, B=$[9200, 20000)$; and for the age of the applicant, A=$[19, 26)$, B=$[26, 28)$, C=$[28, 35)$, D=$[35, 37)$, and E=$[37, 80)$. The use of categorical features, to create ratemaking classes is now obsolete, we more and more actuaries consider "individual" pricing models.

3.2.4 From Classes to Score

Instead of considering risk classes, the measurement of risk can take a very different form, which we could be call "individualization". In many respects, the latter is a kind of asymptotic limit of the asymptotic limit of the first one, when the number of classes increases. By significantly reducing the population through the assignment of individuals to exclusive categories, and ensuring that each category consists of

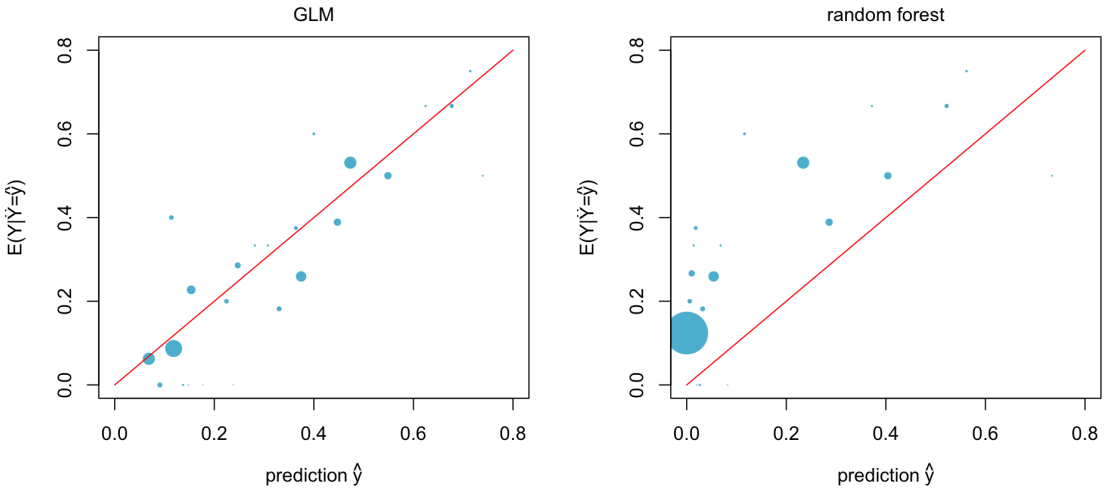


Figure 3.2: Scatterplot with predictions $\widehat{y}$ on various groups, and average outcomes $\overline{y}$, on the database `germancredit`, with the logistic regression (GLM) on the left and the random forest (RF) on the right. Size of circles are proportional to size of groups.

a single individual, the processes of "categorization" and "individualization" begin to converge. Besides computational aspects (discussed in the next section), this approach is fundamentally altering the logical properties of reasoning, as discussed in François (2022) and Krippner (2023). Individuals are given a very precise "score" (which of course can be shared with others). These scores are not discrete, discontinuous, qualitative categories, but numbers that we can consequently engage in calculations (as explained in the previous chapter). These numbers have moreover cardinal value, in the sense that they do not only allow us to classify risks in relation to each other risks in relation to each other, but where they designate a quantity (of risk) on which it is possible to. Those score make the world immediately ordered and readable: scores simplify the world, they order and hierarchize it, as François (2022) claim. Thanks to those scores, individual policyholders are directly comparable. As Friedler et al. (2016) explained, "*the notion of the group ceases to be a stable analytical category and becomes a speculative ensemble assembled to inform a decision and enable a course of action* (...) *Ordered for a different purpose, the groups scatter and reassemble differently.*" In the next section, we will present techniques used by actuaries to model risks.

3.3 Supervised Models and "Individual" Pricing

If we are going to talk here mainly about insurance pricing models, i.e. supervised models where the variable of interest y will be the occurrence of a claim in the coming year, the number of claims, or the total charge, it is worth keeping in mind that the input data ($\boldsymbol{x}$) can be the predictions of a model. For example x_1 could be an observed acceleration score from the previous year (computed by an external provider who had access to the raw telematics data), x_2 could be the distance to the nearest fire station (extrapolated from distance-to-address software), x_3 can the age of the policyholder, x_4 could be a prediction of the number of kilometres driven,

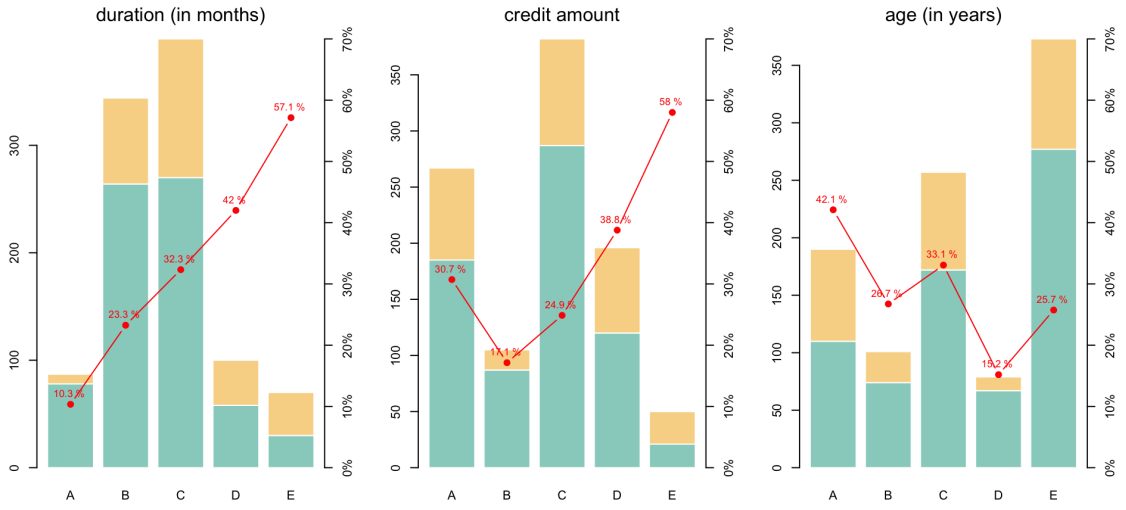


Figure 3.3: From continuous variables to categories (five categories $\{A, B, C, D, E\}$), for three continuous variables of the `germancredit` dataset: duration of the credit, amount of the credit, and age of the applicant. Bars in the background are the number of applicants in each bin (y-axis on the left), and the line if the probability of having a default (y-axis on the right).

etc (in Chapter 5, we will discuss more predictive variables used by actuaries). In the learning database, the "past observations" y_i can also be predictions, especially we want wishes to keep recent claims, still pending, but where the claims manager can give an estimate, based on human experts, but also on "black box" algorithms. We can think of those applications that give a cost estimate of a car damage claim based on a photo of the vehicle sent by the policyholder, or the use of compensation scales for claims not yet settled.

As we will see in this section, a natural "model" or "predictor" for a variable y is related to the conditional expected value. If y corresponds to the total annual loss associated with a given policy, we have seen in the previous chapter that $\mathbb{E}[Y]$ was the "homogeneous pure premium" (see Definition 2.3.1) and $\mathbb{E}[Y|\boldsymbol{X}]$ corresponds to the "heterogeneous pure premium" (see Definition 2.5.1). In the classical collective risk model, $Y = Z_1 + \cdots + Z_N$ is a compound sum, a random sum of random individual losses, and under standard assumptions (see Charpentier and Denuit (2004) or Denuit et al. (2007)), $\mathbb{E}[Y|\boldsymbol{X}] = \mathbb{E}[N|\boldsymbol{X}] \cdot \mathbb{E}[Z|\boldsymbol{X}]$, where the first term $\mathbb{E}[N|\boldsymbol{X}]$ is the expected annual claim frequency for a policyholder with characteristics $\boldsymbol{X}$, while $\mathbb{E}[Z|\boldsymbol{X}]$ is the average cost of a single claim. In this chapter, quite generally, y is one of those variables of interest used to calculate a premium.

If x and y are discrete variables, recall that

$$\mathbb{E}[Y|X = x] = \sum_{y} y \, \mathbb{P}[Y = y|X = x].$$

Quite naturally, in the absolutely continuous case, we would write

$$\mathbb{E}[Y|X = x] = \int y \, f(y|x) \mathrm{d}y = \int y \, \frac{f(x, y)}{f(x)} \mathrm{d}y,$$

with standard notation. Those functions are interesting since we have the following decomposition

$$Y = \mathbb{E}[Y|X = x] + \underbrace{\big(Y - \mathbb{E}[Y|X = x]\big)}_{=\varepsilon},$$

where $\mathbb{E}[\varepsilon|X = x] = 0$. It should be stressed that the extension to the case where X is absolutely continuous is formally slightly complicated since $\{X = x\}$ is an event with probability 0, and then $\mathbb{P}[Y \in \mathcal{A}|X = x]$ is not properly defined (in Bayes formula, in Proposition 3.1.2). As stated in Kolmogorov (1933)[3], "*der Begriff der bedingten Wahrscheinlichkeit in bezug auf eine isoliert gegebene Hypothese, deren Wahrscheinlichkeit gleich Null ist, unzulässig*." Billingsley (2008), Rosenthal (2006) or Resnick (2019) provide theoretical functions for the notation "$\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$",

Definition 3.3.1 (Regression function μ) *Let Y be the non-negative random variable of interest, observed with covariates $\boldsymbol{X}$, the regression function is $\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$.*

Without getting too much into details (based on measure theory), we will invoke here the "*law of the unconscious statistician*" (as coined in Ross (1972) and Casella and Berger (1990)), and write

$$\mathbb{E}\left[\varphi(Y)\right] = \int_{\Omega} \varphi\big(Y(\omega)\big)\mathbb{P}(\mathrm{d}\omega) = \int_{\mathbb{R}} \varphi(y)\mathbb{P}_Y(\mathrm{d}y),$$

for some random variable $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$, with law $\mathbb{P}_Y$. And we will take even more freedom when conditioning. As discussed in Proschan and Presnell (1998), "*statisticians make liberal use of conditioning arguments to shorten what would otherwise be long proofs*," and we will do the same here. Heuristically (the prof can be found in Pfanzagl (1979) and Proschan and Presnell (1998)), a version of $\mathbb{P}(Y \in \mathcal{A}|X = x)$ can be obtained as a limit of conditional probabilities given that X lies in small neighborhoods of x, the limit being taken as the size of the neighborhood tends to 0,

$$\mathbb{P}(Y \in \mathcal{A}|X = x) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(\{Y \in \mathcal{A}\} \cap \{|X - x| \leq \epsilon\})}{\mathbb{P}(\{|X - x| \leq \epsilon\})} = \lim_{\epsilon \to 0} \mathbb{P}(Y \in \mathcal{A}||X - x| \leq \epsilon),$$

that can be extended in higher dimension, using some distance between $\boldsymbol{X}$ and $\boldsymbol{x}$, and then use that approach to define[4] "$\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$". In section 4.1, we will have a brief discussion about a related problem, which is the distinction between $\mathbb{E}[\varphi(x_1, X_2)]$ and $\mathbb{E}[\varphi(X_1, X_2)|X_1 = x_1]$.

3.3.1 Machine Learning Terminology

Suppose that random variables $(\boldsymbol{X}, Y)$ are defined on a probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$, and we observe a finite sample $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)$. Based on that sample, we want to estimate, or learn, a model m that will be a good approximation of the unobservable regression function μ, where $\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$.

In the specific case where y is a categorical variable, for example a binary variable (taking here values in $\{0, 1\}$), there is strong interest in the machine learning literature not to estimate the regression function μ, but to construct a "classifier", which predicts the class. For example, in the logistic regression (see Section 3.3.2), we suppose that $(Y|\boldsymbol{X} = \boldsymbol{x}) \sim \mathcal{B}(\mu(\boldsymbol{x}))$ where $\text{logit}[\mu(\boldsymbol{x})] = \boldsymbol{x}^\top\boldsymbol{\beta}$, and $\mu(\boldsymbol{x})$ has two interpretations, since $\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ and $\mu(\boldsymbol{x}) = \mathbb{P}[Y = 1|\boldsymbol{X} = \boldsymbol{x}]$. From this regression function, one can easily construct a "classifier" by considering $m_t(\boldsymbol{x}) = \mathbf{1}(m(\boldsymbol{x}) > t)$, taking values in $\{0, 1\}$ (like y), for some appropriate cutoff threshold $t \in [0, 1]$.

[3] "*the notion of conditional probability is inadmissible in relation to a hypothesis given in isolation whose probability is zero*" [personal translation].

[4] Which corresponds to a very standard idea in non-parametric statistics, see Tukey (1961), Nadaraya (1964) or Watson (1964).

Definition 3.3.2 (Loss ℓ) *A loss function ℓ is a function defined on $\mathcal{Y}\times\mathcal{Y}$ such that $\ell(y,y')\geq 0$ et $\ell(y,y)=0$.*

A loss is not necessarily a distance (between y and y') since symmetry is not required, neither is the triangle inequality. Some losses are simply a function (called "cost") of some distance between y and y'.

Definition 3.3.3 (Risk $\mathcal{R}$) *For a fitted model $\widehat{m}$, its risk is*

$$\mathcal{R}(\widehat{m})=\mathbb{E}\Big[\ell\left(Y,\widehat{m}(\boldsymbol{X})\right)\Big].$$

For instance, in a regression problem, a quadratic loss function ℓ_2 is used

$$\ell_2(y,\widehat{y})=(y-\widehat{y})^2,$$

and the risk (names "quadratic risk") is then

$$\mathcal{R}_2(\widehat{m})=\mathbb{E}\Big[\left(Y-\widehat{m}(\boldsymbol{X})\right)^2\Big],$$

where $\widehat{m}(\boldsymbol{x})$ is some prediction. Observe that

$$\mathbb{E}[Y]=\underset{m\in\mathbb{R}}{\operatorname{argmin}}\big\{\mathcal{R}_2(m)\big\}=\underset{m\in\mathbb{R}}{\operatorname{argmin}}\Big\{\mathbb{E}\Big[\ell_2\left(Y,m\right)\Big]\Big\}.$$

The fact that the expected value minimizes the expected loss for some loss function (here ℓ_2) is named "elicitable" in Gneiting et al. (2007). From this property, we can undestand why the expected value is also called "best estimate" (see also the connection to Bregman distance, in Definition 3.3.12). As discussed in Huttegger (2013), the use of a quadratic loss function give rise to a rich geometric structure, for variables that are squared integrable, which is essentially very close to the geometry of Euclidean spaces (L^2 being an Hilbert space, with an inner product, and a projection operator, we will come back to this point in Chapter 10, in "pre-processing" approaches). Up to a monotonic transformation (the square root function), the distance here is the expectation of the quadratic loss function.

The quantile loss $\ell_{q,\alpha}$, for some $\alpha\in(0,1)$ is

$$\ell_{q,\alpha}(y,\widehat{y})=\max\left\{\alpha(y-\widehat{y}),(1-\alpha)(\widehat{y}-y)\right\}=(y-\widehat{y})\big(\alpha-\mathbf{1}_{(y<\widehat{y})}\big).$$

For example, Kudryavtsev (2009) used a quantile loss function in the context of ratemaking. It is called "quantile" loss since

$$Q(\alpha)=F^{-1}(\alpha)\in\underset{q\in\mathbb{R}}{\operatorname{argmin}}\big\{\mathcal{R}_{q,\alpha}(q)\big\}=\underset{q\in\mathbb{R}}{\operatorname{argmin}}\Big\{\mathbb{E}\Big[\ell_{q,\alpha}\left(Y,q\right)\Big]\Big\}$$

Indeed,

$$\underset{m}{\operatorname{argmin}}\big\{\mathcal{R}_{q,\alpha}(q)\big\}=\underset{m}{\operatorname{argmin}}\left\{(\alpha-1)\int_{-\infty}^{q}(y-q)dF_Y(y)+\alpha\int_{q}^{\infty}(y-q)dF_Y(y)\right\},$$

and by computing the derivative of the expected loss via an application of the Leibniz integral rule,

$$0=(1-\alpha)\int_{-\infty}^{q^\star}dF_Y(y)-\alpha\int_{q^\star}^{\infty}dF_Y(y)$$

so that $0=F_Y(q^\star)-\alpha$. Thus, quantiles are also "ellicitable" functional. When $\alpha=1/2$ (the median), we recognize the least absolute deviation loss ℓ_1, $\ell_1(y,\widehat{y})=|y-\widehat{y}|$.

Definition 3.3.4 (Empirical risk $\widehat{\mathcal{R}}_n$) *Given a sample $\{(y_i, \boldsymbol{x}_i), i = 1, \cdots, n\}$, define the empirical risk*

$$\widehat{\mathcal{R}}_n(\widehat{m}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\widehat{m}(\boldsymbol{x}_i), y_i\right),$$

Again, in the regression context, with a quadratic loss function, the empirical risk is the Mean Squared Error (MSE), defined as

$$\widehat{\mathcal{R}}_n(\widehat{m}) = \text{MSE}_n = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \widehat{m}(\boldsymbol{x}_i)\right)^2.$$

$\widehat{m}$, defined as the empirical risk minimizer, over a training sample and a collection of models, is also called M-estimator in Huber (1964).

In the context of a classifier, where $y \in \{0, 1\}$ as well as $\widehat{y}$, a natural loss is the so-called "0/1 loss",

$$\ell_{0/1}(y, \widehat{y}) = \mathbf{1}(y \neq \widehat{y}) = \begin{cases} 1 \text{ if } y \neq \widehat{y} \\ 0 \text{ if } y = \widehat{y}. \end{cases}$$

In the context of a classifier, the loss is a function on $\mathcal{Y} \times \mathcal{Y}$, i.e. $\{0, 1\} \times \{0, 1\}$, taking values in $\mathbb{R}_+$. But in many cases, we want to compute a "loss" between y and an estimation of $\mathbb{P}[Y = 1]$, instead of a predicted class $\widehat{y} \in \{0, 1\}$, therefore, it will be a function defined on $\{0, 1\} \times [0, 1]$. That will correspond to a "scoring rule" (see Definition 4.2.1). The empirical risk associated with the $\ell_{0/1}$ loss is the proportion of misclassified individuals, also named "classifier error rate". But it is possible to get more information: given a sample of size n, it is possible to compute the "confusion matrix", which is simply the contingency table of the pairs $(y_i, \widehat{y}_i)$, as in Table 3.1.

prediction \ actual value	0	1	total
0	true negative TN = n_{00}	false negative FN = n_{01}	$n_{0\bullet}$
1	false positive FP = n_{10}	true positive TP = n_{11}	$n_{1\bullet}$
total	$n_{\bullet 0}$	$n_{\bullet 1}$	n

prediction \ actual value	0	1	total
0	true negative TN = n_{00}	false negative FN = n_{01}	$\text{NPV} = \frac{\text{TN}}{\text{FN}+\text{TN}}$
1	false positive FP = n_{10}	true positive TP = n_{11}	$\text{PPV} = \frac{\text{TP}}{\text{FP}+\text{TP}}$
	$\text{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}}$	$\text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}}$	
	$\text{TNR} = \frac{\text{TN}}{\text{FP}+\text{TN}}$	$\text{FNR} = \frac{\text{FN}}{\text{TP}+\text{FN}}$	

Table 3.1: General representation on the "confusion matrix" on the left, with counts of $y = 0$ (column on the left, $n_{\bullet 0}$) and $y = 1$ (column on the right, $n_{\bullet 1}$), counts of $\widehat{y} = 0$, "negative" outcomes (row on top, $n_{0\bullet}$) and $\widehat{y} = 1$, "positive" outcomes (row below, $n_{1\bullet}$). On the right, expressions of the standard metrics (false positive rate, true positive rate, false negative rate, true negative rate, positive predictive value and negative predictive value).

Given a threshold t, one will get the confusion matrix, and various quantities can be computed. To illustrate, consider a simple logistic regression model, on $\boldsymbol{x}$ (and not s), a get predictions on $n = 40$ observations from `toydata2` (as in Table 8.1).

		actual value 0	actual value 1	total
prediction	0	true negative 12	false negative 3	15
prediction	1	false positive 8	true positive 17	25
	total	20	20	

		actual value 0	actual value 1	total
prediction	0	true negative 18	false negative 5	23
prediction	1	false positive 2	true positive 15	17
	total	20	20	

Table 3.2: Confusion matrices with threshold 30% and 50% for $n = 40$ observations from the `toydata2` dataset, and a logistic regression for m.

From Table 3.2, we can compute various quantities (as explained in Table 3.1). Sensitivity (true positive rate) is the probability of a positive test result, conditioned on the individual truly being positive. Thus, here we have

$$\text{TPR}(30\%) = \frac{17}{3+17} = 0.85 \text{ and } \text{TPR}(50\%) = \frac{15}{5+15} = 0.75,$$

while the miss rate (false negative rate)

$$\text{FNR}(30\%) = \frac{3}{3+17} = 0.15 \text{ and } \text{FNR}(50\%) = \frac{5}{5+15} = 0.25.$$

Specificity (true negative rate) is the probability of a negative test result, conditioned on the individual truly being negative,

$$\text{TNR}(30\%) = \frac{12}{8+12} = 0.6 \text{ and } \text{TNR}(50\%) = \frac{18}{2+18} = 0.9,$$

while the fall-out (false positive rate) is

$$\text{FPR}(30\%) = \frac{8}{8+12} = 0.4 \text{ and } \text{FPR}(50\%) = \frac{2}{2+18} = 0.1$$

The negative predictive value (NPV)

$$\text{NPV}(30\%) = \frac{12}{12+3} = 0.8 \text{ and } \text{NPV}(50\%) = \frac{18}{18+5} = 0.7826$$

while the precision (positive predictive value) is

$$\text{PPV}(30\%) = \frac{17}{17+8} = 0.68 \text{ and } \text{PPV}(50\%) = \frac{15}{15+2} = 0.8824.$$

Accuracy is the proportion of good predictions

$$\text{ACC}(30\%) = \frac{12+17}{12+8+3+17} = 0.725 \text{ and } \text{ACC}(50\%) = \frac{18+15}{18+2+5+15} = 0.825,$$

while "balanced accuracy" (see Langford and Schapire (2005)) is the average of TPR and TNR,

$$\text{BACC}(30\%) = \frac{0.85+0.6}{2} = 0.725 \text{ and } \text{BACC}(50\%) = \frac{0.75+0.9}{2} = 0.8250.$$

Finally, Cohen's kappa (from Cohen (1960), that is based on the accuracy assuming that y and $\widehat{y}$ are independent – as in the chi-square test),

$$\kappa(30\%) = \frac{\frac{29}{40} - \frac{20}{40}}{1 - \frac{20}{40}} = 0.45 \text{ and } \kappa(50\%) = \frac{\frac{33}{40} - \frac{20}{40}}{1 - \frac{20}{40}} = 0.65,$$

while Matthews correlation coefficient (see Definition 8.3.8) will be

$$\text{MCC}(30\%) = 0.464758 \text{ and } \text{MCC}(30\%) = 0.6574383.$$

One issue here is that the sample used to compute the empirical risk is the same as the one used to fit the model, also-called "in-sample risk"

$$\widehat{\mathcal{R}}_n^{\text{is}}(m) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(m(\boldsymbol{x}_i), y_i\right).$$

Thus, if we consider

$$\widehat{m}_n = \underset{m\in\mathcal{M}}{\text{argmin}}\left\{\widehat{\mathcal{R}}_n^{\text{is}}(m)\right\}$$

on a set $\mathcal{M}$ of admissible models, we will have a tendency to capture a lot of noise and to over-adjust the data: this is called "over-fitting." For example, on Figure 3.4, we have two fitted models $\widehat{m}$ such that the in-sample risk is null

$$\widehat{\mathcal{R}}_n^{\text{is}}(\widehat{m}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\widehat{m}(\boldsymbol{x}_i), y_i\right) = 0.$$

To avoid this problem, randomly dividing the initial database into a training database and a validation database. The training database, with $n_T < n$ observations, will be used to estimate the parameters of the model

$$\widehat{m}_{n_T} = \underset{m\in\mathcal{M}}{\text{argmin}}\left\{\widehat{\mathcal{R}}_{n_T}^{\text{is}}(m)\right\}$$

Then, the validation dataset, with $n_V = n - n_T$ observations, will be used to select the model, using the "out-of-sample risk"

$$\widehat{\mathcal{R}}_{n_V}^{\text{os}}(\widehat{m}_{n_T}) = \frac{1}{n_V}\sum_{i=1}^{n_V} \ell\left(\widehat{m}_{n_T}(\boldsymbol{x}_i), y_i\right).$$

Quite generally, given a loss function $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_+$, and a collection $\mathcal{D}_n$ of n independent observations drawn from $(\boldsymbol{X}, Y)$ (corresponding to the dataset) the risk is

$$\mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{\mathcal{D}_n}\left(\mathbb{E}_{Y|\boldsymbol{X}}\left[\ell(Y, \widehat{m}(\boldsymbol{X})|\mathcal{D}_n)\right]\right)\right],$$

that cannot be calculated without knowing the true distribution of $(Y, \boldsymbol{X})$.

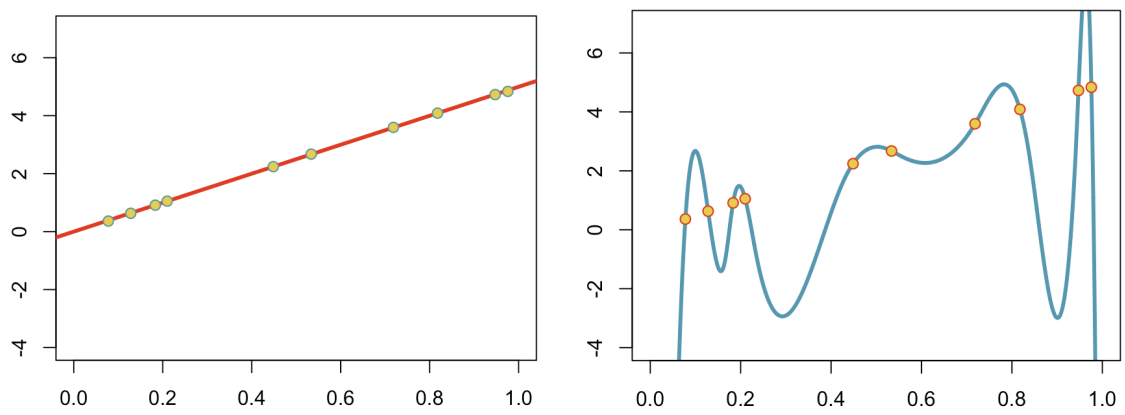


Figure 3.4: Two fitted models from a (fake) dataset $(x_1, y_1), \cdots, (x_{10}, y_{10})$, with a linear model on the left, and a polynomial model on the right, such that for both in-sample risk is null, $\widehat{\mathcal{R}}_n^{\text{is}}(\widehat{m}) = 0$.

If ℓ is the quadratic loss, $\ell_2(y, \widehat{y}) = (y - \widehat{y})^2$,

$$\begin{aligned}\mathcal{R}_2(\widehat{m}) &= \mathbb{E}_{\mathcal{D}_n}\left(\mathbb{E}_{Y|\boldsymbol{X}}\left[\ell(Y, \widehat{m}(\boldsymbol{X})|\mathcal{D}_n)\right]\right) \\ &= \left(\mathbb{E}_{Y|\boldsymbol{X}}(Y) - \mathbb{E}_{\mathcal{D}_n}\left[\widehat{m}(\boldsymbol{X})|D_n)\right]\right)^2 \\ &+ \mathbb{E}_{Y|\boldsymbol{X}}\left[\left(Y - \mathbb{E}_{Y|\boldsymbol{X}}(Y)\right)^2\right] \\ &+ \mathbb{E}_{\mathcal{D}_n}\left[\left(\widehat{m}(\boldsymbol{X})|\mathcal{D}_n) - \mathbb{E}_{\mathcal{D}_n}\left[\widehat{m}(\boldsymbol{X})|\mathcal{D}_n)\right]\right)^2\right].\end{aligned}$$

We recognize the square of the bias (bias^2), the stochastic error, and the variance of the estimator.

Here, so far, all observations in the training dataset have the same importance. But it is possible to include weights, for each observation, in the optimization procedure. A classical example could be the weighted least squares,

$$\sum_{i=1}^{n} \omega_i \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2,$$

for some positive (or null) weights $(\omega_1, \cdots, \omega_n) \in \mathbb{R}_+^n$. The weighted least square estimator is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{y}, \text{ where } \boldsymbol{\Omega} = \text{diag}(\boldsymbol{\omega}).$$

More generally, it is possible to consider a weighted version of the empirical risk

$$\widehat{\mathcal{R}}_{\boldsymbol{\omega}}(\widehat{m}) = \sum_{i=1}^{n} \omega_i \ell\left(\widehat{m}(\boldsymbol{x}_i), y_i\right).$$

We have considered here losses, that could be seen as a distance between a prediction $\widehat{y}$ and an actual observation y. But Huttegger (2017) claims that those losses – what was denoted $\ell(y, \widehat{y})$ – are maybe not a correct measure of "*epistemic accuracy*." The first extension is based on the fact that, instead of a point

estimate $\widehat{y}$, a confidence interval or a predictive distribution. Observing $y = 100$ when we predict $\widehat{y} = 50$ with a standard deviation on the prediction of 50 is not the same as observing $y = 100$ when we predict $\widehat{y} = 120$ with a standard deviation on the prediction of 10. Formally, it means that we want to quantify a distance between a distribution and a single point. That will be discussed in section 4.2.1, when we will introduce "scoring rules". Another important tool will be a "distance" between two distributions.

Definition 3.3.5 (Hellinger distance) *Hellinger (1909) For two discrete distributions p and q, Hellinger distance is*

$$d_{\mathrm{H}}(p,q)^2 = \frac{1}{2}\sum_i \left(\sqrt{p(i)} - \sqrt{q(i)}\right) = 1 - \sum_i \sqrt{p(i)q(i)} \in [0,1],$$

and for absolutely continuous distributions, if p and q are densities,

$$d_{\mathrm{H}}(p,q)^2 = \int_{\mathbb{R}} p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x \text{ or } \int_{\mathbb{R}^d} p(\boldsymbol{x})\log\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}.$$

For example, for two Gaussian distributions with means μ and variances σ^2,

$$d_{\mathrm{H}}^2(p_1,p_2) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\exp\left(-\frac{1}{4}\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}\right),$$

(that can be extended in higher dimension, as in Pardo (2018)), while for two exponential distributions with means μ_1 and μ_2,

$$d_{\mathrm{H}}^2(p_1,p_2) = 1 - \frac{2\sqrt{\mu_1\mu_2}}{\mu_1+\mu_2}.$$

A few years after, Saks (1937) introduced the concept of "total variation" (between measures) in the context of signed measures on a measurable space, and it can be used to define a total variation distance between probability measures (see Dudley (2010)). Quite generally, given two discrete distributions p and q, the total variation is the largest possible difference between the probabilities that the two probability distributions can assign to the same event:

Definition 3.3.6 (Total Variation) *Jordan (1881); Rudin (1966) For two discrete distributions p and q, the total variation distance between p and q is*

$$d_{\mathrm{TV}}(p,q) = \sup_{\mathcal{A}\subset\mathbb{R}} \left\{|p(\mathcal{A}) - q(\mathcal{A})|\right\}.$$

It should be stressed here that in the context of discrimination, Zafar et al. (2019) or Zhang and Bareinboim (2018) suggest to remove the symmetry property, to take into account that there is a favored and a disfavored group, and therefore to consider

$$D_{\mathrm{TV}}(p\|q) = \sup_{\mathcal{A}} \left\{p(\mathcal{A}) - q(\mathcal{A})\right\}.$$

Removing the standard property of symmetry (that we have on distances) yields the concept of "divergence", that is still a non-negative function, positive (in the sense that it is null if and only if "$p = q$"), and the triangle inequality is not satisfied (even if it could satisfy some sort of Pythagorean theorem, if an "inner product" can be derived). As Amari (1982) explains, it is mainly because divergences are generalizations of "squared distances", not "linear distances".

Definition 3.3.7 (Kullback–Leibler) *Kullback and Leibler (1951) For two discrete distributions p and q, Kullback–Leibler divergence of p, with respect to q is*

$$D_{\mathrm{KL}}(p\|q) = \sum_i p(i) \log \frac{p(i)}{q(i)},$$

and for absolutely continuous distributions,

$$D_{\mathrm{KL}}(p\|q) = \int_{\mathbb{R}} f px) \log \frac{p(x)}{q(x)} \,\mathrm{d}x \, or \int_{\mathbb{R}^d} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \,\mathrm{d}\boldsymbol{x},$$

in higher dimension.

This corresponds to the relative entropy from q to p. Interestingly Kullback (2004) mentioned that he preferred the term "discrimination information."

Notice that the ratio $\log(p/q)$ is sometimes called "*weight-of-evidence*," following Good (1950) and Ayer (1972), see also Wod (1985) or Weed (2005) for some surveys. Again, this is not a distance (even if it satisfies the nice property $p = q$ if and only if $D_{\mathrm{KL}}(p\|q) = 0$), so we will use the term "divergence" (and notation D instead of d).

Definition 3.3.8 (Mutual Information) *Shannon and Weaver (1949) For a pair of two discrete variables x and y with joint distributions p_{xy} (and marginal ones p_x and p_y), the mutual information is*

$$IM(x,y) = D_{\mathrm{KL}}(p_{xy}\|p_{xy}^{\perp}) = \sum_{i,j} p_{xy}(i,j) \log \frac{p_{xy}(i,j)}{p_{xy}^{\perp}(i,j)} = \sum_{i,j} p_{xy}(i,j) \log \frac{p_{xy}(i,j)}{p_x(i)p_y(j)},$$

where $p_{xy}^{\perp}$ is the independent version of p_{xy} ($p_{xy}^{\perp}(i,j) = p_x(i)p_y(j)$.

Observe that, for two Gaussian distributions,

$$D_{\mathrm{KL}}(p_1\|p_2) = \frac{1}{2}\left[\log\frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_1-\mu_2)^2}{\sigma_2^2} - 1\right],$$

and in higher dimension (say k),

$$D_{\mathrm{KL}}(p_1\|p_2) = \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \mathrm{tr}\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\} + (\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^{\top}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) - k\right].$$

As proved in Tsybakov (2009), it is possible to find relationships between those measures, such as

$$d_{\mathrm{TV}}(p,q) \leq \sqrt{1-\exp[D_{\mathrm{KL}}(p\|q)]} \text{ or } d_{\mathrm{H}}(p,q)^2 \leq d_{\mathrm{TV}}(p,q) \leq \sqrt{2} d_{\mathrm{H}}(p,q).$$

It is possible to derive a symmetric divergence measure by averaging with the so-called "dual divergence", also named "PSI index" (defined, as well as most functions in Siddiqi (2012), in the `scorecard` R package),

Definition 3.3.9 (Population Stability Index (PSI)) *Siddiqi (2012) PSI is a measure of population stability between two population samples,*

$$PSI = D_{\mathrm{KL}}(p_1 \parallel p_2) + D_{\mathrm{KL}}(p_2 \parallel p_1).$$

An alternative is to consider the following approach, with "Jensen-Shannon divergence",

Definition 3.3.10 (Jensen-Shannon) *Lin (1991) The Jensen-Shannon distance is a symmetric distance induced by Kullback-Liebler divergence,*

$$D_{\mathrm{JS}}(p_1, p_2) = \frac{1}{2} D_{\mathrm{KL}}(p_1 \| q) + \frac{1}{2} D_{\mathrm{KL}}(p_1 \| q),$$

where $q = \dfrac{1}{2}(p_1 + p_2)$.

Another popular distance is Wasserstein distance[5], also called Mallows' distance, from Mallows (1972),

Definition 3.3.11 (Wasserstein) *Wasserstein (1969). Consider two measures on* p *and* q *on* $\mathbb{R}^d$*, with a norm* $\|\cdot\|$ *(on* $\mathbb{R}^d$*). Then define*

$$W_k(p, q) = \left(\inf_{\pi \in \Pi(p,q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\boldsymbol{x} - \boldsymbol{y}\|^k \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}) \right)^{1/k}$$

where $\Pi(p, q)$ *is the set of all couplings of* p *and* q.

Without specifying, the Wasserstein distance will be W_2, and d is the Euclidean distance. W_1 is also called "*earth mover's distance*," see e.g. Levina and Bickel (2001). As mentioned in Villani (2009), the total variation distance arises quite naturally as the optimal transportation cost, when the cost function is, $\ell_{0/1}$, or $\mathbf{1}(x \neq y)$, since

$$d_{\mathrm{TV}}(p, q) = \inf_{\pi \in \Pi(p,q)} \left\{ \mathbb{P}[X \neq Y],\ (X, Y) \sim \pi \right\} = \inf_{\pi \in \Pi(p,q)} \left\{ \mathbb{E}[\ell_{0/1}(X, Y)],\ (X, Y) \sim \pi \right\}.$$

With Wasserstein-distance, we consider

$$\inf_{\pi \in \Pi(p,q)} \left\{ \mathbb{E}[\ell(X, Y)],\ (X, Y) \sim \pi \right\} \text{ or } \inf_{\pi \in \Pi(p,q)} \left\{ \int \ell(x, y) \pi(\mathrm{d}x, \mathrm{d}y) \right\}.$$

The connection with "transport" is obtained as follows: given $\mathcal{T} : \mathbb{R}^k \to \mathbb{R}^k$, define the "*push-forward*" measure,

$$\mathbb{P}_1(\mathcal{A}) = \mathcal{T}_{\#}\mathbb{P}_0(\mathcal{A}) = \mathbb{P}_0\big(\mathcal{T}^{-1}(\mathcal{A})\big),\ \forall \mathcal{A} \subset \mathbb{R}^k.$$

An optimal transport $\mathcal{T}^\star$ (in Brenier's sense, from Brenier (1991), see Villani (2009) or Galichon (2016)) from $\mathbb{P}_0$ towards $\mathbb{P}_1$ will be solution of

$$\mathcal{T}^\star \in \underset{\mathcal{T}:\mathcal{T}_{\#}\mathbb{P}_0 = \mathbb{P}_1}{\operatorname{arginf}} \left\{ \int_{\mathbb{R}^k} \ell(\boldsymbol{x}, \mathcal{T}(\boldsymbol{x})) \mathrm{d}\mathbb{P}_0(\boldsymbol{x}) \right\}.$$

In dimension 1 (distributions on $\mathbb{R}$), let F_0 and F_1 denote the cumulative distribution function, and F_0^{-1} and F_1^{-1} denote quantiles. Then

$$W_k(p_0, p_1) = \left(\int_0^1 \left| F_0^{-1}(u) - F_1^{-1}(u) \right|^k \mathrm{d}u \right)^{1/k},$$

[5]The original name, Васерштейн, is also written "Vaserstein" but since the distance is usually denoted "W", we will write "Wasserstein".

and one can prove that the optimal transport $\mathcal{T}^\star$ is a monotone transformation. More precisely,

$$\mathcal{T}^\star : x_0 \mapsto x_1 = F_1^{-1} \circ F_0(x_0).$$

For empirical measures, in dimension 1, the distance is a simple function of the order statistics:

$$W_k(\boldsymbol{x}, \boldsymbol{y}) = \left(\frac{1}{n} \sum_{i=1}^{n} |x_{(i)} - y_{(i)}|^k \right)^{1/k}.$$

Observe that, for two Gaussian distributions, and the Euclidean distance,

$$W_2(p_0, p_1)^2 = (\mu_1 - \mu_0)^2 + \big(\sigma_1 - \sigma_0\big)^2,$$

and in higher dimension,

$$W_2(p_0, p_1)^2 = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_2^2 + \operatorname{tr}\big(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1 - 2\big(\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_1^{1/2}\big)^{1/2}\big).$$

If variances are equal, we can write simply

$$\begin{cases} W_2(p_0, p_1)^2 = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_2^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \\ D_{\mathrm{KL}}(p_0\|p_1) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0). \end{cases}$$

And in that Gaussian case, there is an explicit expression for the optimal transport, which is simply an affine map (see Villani (2003) for more details). In the univariate case, $x_1 = \mathcal{T}_{\mathcal{N}}^\star(x_0) = \mu_1 + \dfrac{\sigma_1}{\sigma_0}(x_0 - \mu_0)$, while in the multivariate case, an analogous expression can be derived:

$$\boldsymbol{x}_1 = \mathcal{T}_{\mathcal{N}}^\star(\boldsymbol{x}_0) = \boldsymbol{\mu}_1 + \boldsymbol{A}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0),$$

where $\boldsymbol{A}$ is a symmetric positive matrix that satisfies $\boldsymbol{A}\boldsymbol{\Sigma}_0\boldsymbol{A} = \boldsymbol{\Sigma}_1$, which has a unique solution given by $\boldsymbol{A} = \boldsymbol{\Sigma}_0^{-1/2}\big(\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{1/2}\big)^{1/2}\boldsymbol{\Sigma}_0^{-1/2}$, where $\boldsymbol{M}^{1/2}$ is the square root of the square (symmetric) positive matrix $\boldsymbol{M}$ based on the Schur decomposition ($\boldsymbol{M}^{1/2}$ is a positive symmetric matrix), as described in Higham (2008).

In the non-Gaussian case, one can prove (see Alvarez-Esteban et al. (2018)) that

$$W_2(p_0, p_1)^2 = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_2^2 + W_2(\bar{p}_0, \bar{p}_1)^2,$$

where $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are the means of p_0 and p_1, and $\bar{p}_0$ and $\bar{p}_1$ are the corresponding centered probabilities. And if measure are not Gaussian, but have variances $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$, Gelbrich (1990) proved that

$$W_2(p_0, p_1)^2 \geq \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_2^2 + \operatorname{tr}\big(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1 - 2\big(\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_1^{1/2}\big)^{1/2}\big).$$

If variances are equal, we can write simply

$$\begin{cases} W_2(p_1, p_2)^2 = \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_2^2 = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \\ D_{\mathrm{KL}}(p_1\|p_2) = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \end{cases}$$

To conclude this part, Banerjee et al. (2005) suggested loss functions named "*Bregman distance functions*."

Definition 3.3.12 (Bregman distance functions) *Banerjee et al. (2005). Given a strictly convex differentiable function* $\phi : \mathbb{R} \to \mathbb{R}$*,*

$$B_\phi(y_1, y_2) = \phi(y_1) - \phi(y_2) - (y_1 - y_2)\phi'(y_2),$$

or if $\phi : \mathbb{R}^d \to \mathbb{R}$*,*

$$B_\phi(\boldsymbol{y}_1, \boldsymbol{y}_2) = \phi(\boldsymbol{y}_1) - \phi(\boldsymbol{y}_2) - (\boldsymbol{y}_1 - \boldsymbol{y}_2)^\top \nabla\phi(\boldsymbol{y}_2).$$

Note that B_ϕ is symmetric, positive, and $B_\phi(y_1, y_2) = 0$ if and only if $y_1 = y_2$. For example, if $\phi(t) = t^2$, $B_\phi(y_1, y_2) = \ell_2(y_1, y_2) = (y_1 - y_2)^2$. Huttegger (2017) pointed out that those functions have a "*nice epistemological motivation*." Consider some very general distance function ψ, such that, for any random variable Z,

$$\mathbb{E}\big[\psi(Y, \mathbb{E}[Y])\big] \leq \mathbb{E}\big[\psi(Y, Z)\big],$$

meaning that $\mathbb{E}[Y]$ is the "best estimate" of Y (according to this distance ψ). If $Y = \mathbf{1}_A$, it means that

$$\mathbb{E}\big[\psi(\mathbf{1}_A, \mathbb{P}[A])\big] \leq \mathbb{E}\big[\psi(\mathbf{1}_A, Z)\big],$$

meaning that $\mathbb{P}[A]$ is the "best" degree of belief of $\mathbf{1}_A$. If we suppose that ψ is continuously differentiable in its first argument, and $\psi(0, 0) = 0$, then ψ is a Bregman distance function. And one can write, if $\phi : \mathbb{R}^d \to \mathbb{R}$,

$$B_\phi(\boldsymbol{y}_1, \boldsymbol{y}_2) = (\boldsymbol{y}_1 - \boldsymbol{y}_2)^\top \nabla^2\phi(\boldsymbol{y}_t)(\boldsymbol{y}_1 - \boldsymbol{y}_2),$$

where ∇^2 denote the Hessian matrix, and where $\boldsymbol{y}_t = t\boldsymbol{y}_1 + (1-t)\boldsymbol{y}_2$, for some $t \in [0, 1]$. We recognize some sort of local Mahalanobis distance, induced by $\nabla^2\phi(\boldsymbol{y}_t)$.

3.3.2 Generalized Linear Models

Generalized Linear Models is a vast class of probabilistic models, that contains the logistic and the probit regression (for binary y) and the Poisson regression (when y corresponds to counts). For more than thirty years, Generalized Linear Models have been the most popular predictive technique for actuaries (see Haberman and Renshaw (1996), Denuit and Charpentier (2005), Denuit et al. (2007), Ohlsson and Johansson (2010) or Frees (2006); Frees et al. (2014a), among many others). The starting point is that the density of y (or the probability function if y is a discrete variable) should be in the exponential family:

Definition 3.3.13 (Exponential family) *McCullagh and Nelder (1989) Y is in the exponential family if its density (with respect to some appropriate measure) is*

$$f_{\theta,\varphi}(y) = \exp\left(\frac{y\theta - b(\theta)}{\varphi} + c(y, \varphi)\right),$$

where θ *is the canonical parameter,* φ *is a nuisance parameter, and* $b : \mathbb{R} \to \mathbb{R}$ *is some* $\mathbb{R} \to \mathbb{R}$ *function.*

The binomial, Poisson, Gaussian, Gamma, Inverse Gaussian distributions belong to this family (see McCullagh and Nelder (1989) for more examples). Consider some dataset $(y_i, \boldsymbol{x}_i)$ such that y_i is suppose to be a realization of some random variable Y_i with distribution $f_{\theta_i,\varphi}$, in the exponential family. More specifically, in this GLM framework, different quantities are used, namely the canonical parameter is θ_i, the prediction for y_i is

$$\mu_i = \mathbb{E}(Y_i) = b'(\theta_i),$$

the score associated with y_i is

$$\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

and the link function is g such that

$$\eta_i = g(\mu_i) = g(b'(\theta_i)).$$

For the "canonical link function", $g^{-1} = b'$ and then $\theta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \eta_i$ and $\mu_i = \mathbb{E}(Y_i) = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$. Inference is performed by finding the maximum of the log-likelihood, that is

$$\log \mathcal{L} = \frac{1}{\varphi} \sum_{i=1}^{n} \left(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - b(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \right) + \underbrace{\sum_{i=1}^{n} c(y_i, \varphi)}_{\text{independent of } \boldsymbol{\beta}},$$

and if $\widehat{\boldsymbol{\beta}}$ denotes the optimal parameter, the prediction is $\widehat{y}_i = \widehat{m}(\boldsymbol{x}_i) = g^{-1}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})$.

On Figure 3.5, we have an explanatory diagram of a Generalized Linear Model (GLM), starting from some predictor variables $\boldsymbol{x} = (x_1, \cdots, x_k)$ (on the left) and a target variable y (on the right). The score, $\eta = \boldsymbol{x}^\top \boldsymbol{\beta}$ is created from the predictors $\boldsymbol{x}$, and then the prediction is obtained by a nonlinear transformation, $m(\boldsymbol{x}) = g^{-1}(\boldsymbol{x}^\top \boldsymbol{\beta})$.

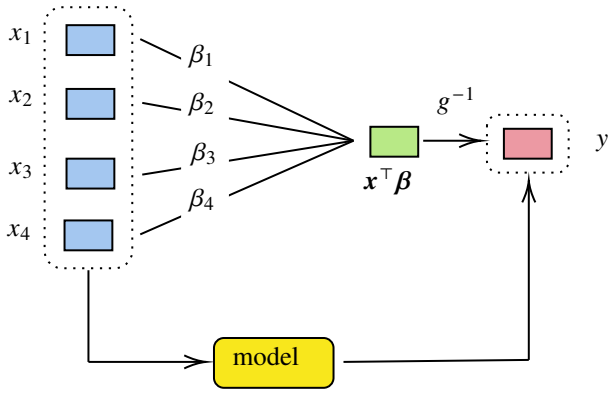


Figure 3.5: Explanatory diagram of a Generalized Linear Model (GLM), starting from some predictor variables $\boldsymbol{x} = (x_1, \cdots, x_k)$ and a target variable y.

With the canonical link function the first order condition is simply (with a standard matrix formulation)

$$\nabla \log \mathcal{L} = \boldsymbol{X}^T (\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{X}^T (\boldsymbol{y} - g^{-1}(\boldsymbol{X}\widehat{\boldsymbol{\beta}})) = \boldsymbol{0}.$$

This is the numerical equation solved numerically when calling `glm` in R (using Fisher's iterative algorithm, which is equivalent of a gradient descent, with Newton-Raphson iterative technique, where the explicit expression of the Hessian is used). In a sense, the probabilistic construction is simply a way to interpret the derivation. For example, a Poisson regression can be performed on positive observations y_i (not necessarily integers), which makes sense if we focus only on solving the first order condition (as done by the computer), not if we care about the interpretation. And actually, we can see this approach as a "machine learning" one. For convenience, let us write quite generally $\log \mathcal{L}_i(\widehat{y}_i)$ the contribution of the i-th observation to the log-likelihood.

As mentioned previously, with machine learning approach, the in-sample empirical risk is

$$\widehat{\mathcal{R}}_n(\widehat{m}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\widehat{m}(\boldsymbol{x}_i), y_i\right),$$

and it is possible to make connections between the maximum likelihood optimization and the minimization of the in-sample empirical risk, introducing the scaled deviance of the exponential model,

$$D^\star = \sum_{i=1}^{n} d^\star\left(\widehat{y}_i, y_i\right), \text{ where } d^\star\left(\widehat{y}_i, y_i\right) = 2\big(\log \mathcal{L}_i(y_i) - \log \mathcal{L}_i(\widehat{y}_i)\big).$$

Here, the first term $log\mathcal{L}_i(y_i)$ corresponds to the log-likelihood of a "perfect fit" (since $\widehat{y}_i$ is supposed to be equal to y_i), also called "saturated model". The unscaled deviance, $d = \varphi d^\star$ is used as a loss function. For the Gaussian model, $\ell\left(y_i, \widehat{y}_i\right) = \left(\widehat{y}_i, y_i\right)^2$, and the deviance corresponds to the ℓ_2 loss. For the Poisson distribution (with a log-link), the loss would be

$$\ell(y_i, \widehat{y}_i) = \begin{cases} 2\left(y_i \log y_i - y_i \log \widehat{y}_i - y_i + \widehat{y}_i\right) & y_i > 0 \\ 2\widehat{y}_i & y_i = 0, \end{cases}$$

while for the logistic regression,

$$\ell(y_i, \widehat{y}_i) = \begin{cases} 2\left[y_i \log\left(\dfrac{y_i}{\hat{y}_i}\right) + (1-y_i)\log\left(\dfrac{1-y_i}{1-\hat{y}_i}\right)\right] & y_i \in (0,1) \\ -2\log\left(1-\widehat{y}_i\right) & y_i = 0 \\ -2\log\left(\widehat{y}_i\right) & y_i = 1. \end{cases}$$

In insurance (and loss modeling), a classical model is the one introduced in Tweedie (1984), the "Tweedie model", that corresponds to the compound Poisson-gamma distribution. Thus, it is a distribution with a density on $(0, \infty)$, satisfying

$$\log[f(y)] = \frac{1}{\phi}\left(y\frac{\mu^{1-a}}{1-a} - \frac{\mu^{2-a}}{2-a}\right) + \text{constant}, \text{ for } y \in \mathbb{R}_+, \text{ where } a \in (1,2),$$

and with probability mass as 0 equal to $\exp\left(-\dfrac{\mu^{2-a}}{\phi(2-a)}\right)$. Here we use a parametrization not based on θ, but μ, so that it is easier to derive the "Tweedie loss" is, when $a \in (1,2)$,

$$\ell_a(y, \widehat{y}) = \frac{\widehat{y}^{2-a}}{2-a} - y\frac{\widehat{y}^{1-a}}{1-a}.$$

Among other complex models considered in actuarial literature, to model more precisely claims frequency, Cragg (1971) introduced the so called "hurdle model", while Lambert (1992) introduced "zero-inflated" count models (see Hilbe (2014) for a general survey on models for counts). In the first case, for the hurdle Poisson model (Welsh et al. (1996) referred to it as the "conditional Poisson model"),

$$\mathbb{P}(Y_i = y_i) = \begin{cases} \pi & \text{if } y = 0 \\ (1-\pi)\dfrac{\lambda^{y_i}}{[e^\lambda - 1]y_i!} & \text{if } y = 1, 2, \cdots \end{cases}$$

while for the zero-inflated Poisson,

$$\mathbb{P}(Y_i = y_i) = \begin{cases} \pi + (1-\pi)e^{-\lambda} & \text{if } y = 0 \\ (1-\pi)\dfrac{\lambda^{y_i}e^{-\lambda}}{y_i!} & \text{if } y = 1, 2, \cdots \end{cases}$$

A zero-inflated model can only increase the probability to have $y = 0$, but this is not a restriction in hurdle models. From those distributions, the natural idea is to consider a logistic regression for the binomial component, so that $\text{logit}(\pi_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_b$ and a Poisson regression for counts, $\exp(\lambda_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_p$. For the hurdle model, because there are two parameters, the log-likelihood can be separated in two terms (parameters are therefore estimated independently), so one can derive the associated loss functions. For example, the R package `countreg` contains loss functions that can be used in the package `mboost` for boosting.

A strong assumption of linear models is the linear assumption. Actually, the term "linear" is ambiguous, because $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle$ (using the standard geometric notations for inner products on vector spaces) is linear both in $\boldsymbol{x}$ and $\boldsymbol{\beta}$. If x is the age of a policyholder, we can consider a linear model in x, but also a linear model in $\log(x)$, $\sqrt{x}$, $(x-20)_+$ or any non-linear transformation. A natural extension will be based on more general functional forms, that will be linear in the parameters, but with functions of covariates. Thus,

$$\eta_i = g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

will become

$$\eta_i = g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \beta_0 + s_1(x_1) + \cdots + s_k(x_k),$$

where each function s_j is some unspecified function, introduced to make the model more flexible. Note that it is still additive here, so those models are named GAM, for "Generalized Additive Models". Of course, it is possible to have a mix of linear and non-linear functions. In the Gaussian case, where g is the identity function, instead of seeking a model $\widehat{m}(x_1, \cdots, x_k) = \beta_0 + s_1(x_1) + \cdots + s_k(x_k)$ such that

$$\widehat{m} \in \text{argmin}\left\{ \sum_{i=1}^{n} \left[y_i - (\beta_0 + s_1(x_{1i}) + \cdots + s_k(x_{ki})) \right]^2 \right\},$$

on a very general set of functions $s_1, \cdots, s_k$, it could be natural to ask functions to be sufficiently smooth, "un-smooth" meaning that the average value of $(s_j'')^2$ is too large, so we could add a penalty in the previous problem and try to solve

$$\widehat{m} \in \text{argmin}\left\{ \left(\sum_{i=1}^{n} \left[y_i - (\beta_0 + s_1(x_{1i}) + \cdots + s_k(x_{ki})) \right]^2 \right) + \sum_{j=1}^{k} \lambda_j \int s_j''(t_j)^2 \mathrm{d}t_j \right\},$$

that can be solve using simple numerical techniques, as discussed in Hastie and Tibshirani (1987). As mentioned in Verrall (1996) or Lee and Antonio (2015), in the context of actuarial applications, such nonlinear procedure is very useful, not only because they help to prevent misspecifications (and therefore possible incorrect predictions), but also because they provide information about each covariate and the outcome y. As on Figure 3.6 each variable x_j is converted into another one through a function s_j that can be expressed in a specific functional basis (in the `gam` function of the `mgcv` R package, `s` or `s` can denote either a thin plate spline, or a cubic spline).

On Figure 3.7, with a binary outcome y, a plain logistic regression is used on the left, with $\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ (and linear isodensity curves), and $\eta_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i})$ on the right. The true

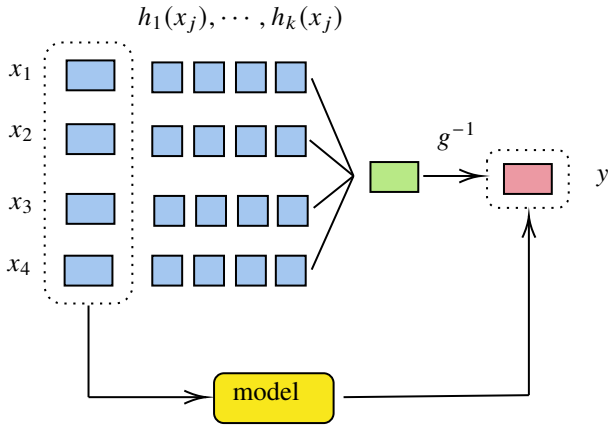


Figure 3.6: Explanatory diagram of a Generalized Additive Model (GAM), starting from the same predictor variables $\boldsymbol{x} = (x_1, \cdots, x_k)$ (on the left) and with the same target variable y (on the right). Each continuous variables x_j are converted into a function $h(x_j)$, expressed in some basis function (such as splines), $h_1(x_j), \cdots, h_k(x_j)$.
.

level curves can be visualized on Figure 1.4. One could also consider adding "interaction" terms. Following Friedman and Popescu (2008), given two variables x_j and $x_{j'}$, a model $\boldsymbol{x} \mapsto m(\boldsymbol{x})$ contains interactions between x_j and $x_{j'}$ if

$$\mathbb{E}\left[\frac{\partial^2 m(\boldsymbol{x})}{\partial x_j \partial x_{j'}}\bigg|_{\boldsymbol{x}=\boldsymbol{X}} \right]^2 > 0.$$

Classically, actuaries have considered a simple product, $(x_1, x_2) \mapsto x_1 x_2$ between the two variables to capture joint effect.

Another natural extension is the class of "linear mixed model" (LMM), with

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are the fixed effects design matrix, and fixed effects (fixed but unknown) respectively, while $\boldsymbol{Z}$ and $\boldsymbol{\gamma}$ are the random effects design matrix and random effects, respectively. The later is used to capture some remaining heterogeneity, that was not captured by the $\boldsymbol{x}$ variables. Here, $\boldsymbol{\varepsilon}$ contains the residuals components, that is supposed to be independent of $\boldsymbol{\gamma}$. And naturally, one can define a "generalized linear mixed model" (GLMM), as studied in McCulloch and Searle (2004),

$$g(\mathbb{E}[Y|\boldsymbol{\gamma}]) = \boldsymbol{x}^\top\boldsymbol{\beta} + z^\top\boldsymbol{\gamma},$$

as in Jiang and Nguyen (2007) and Antonio and Beirlant (2007). It is possible to make connections between credibility and generalized linear mixed model, as in Klinker (2010). The R package `glmm` can be used here.

3.3.3 Penalized Generalized Linear Models

In the context of linear models, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the least square estimate (OLS) of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ can be computed only if $\boldsymbol{X}$ is a full-rank matrix ($\boldsymbol{X}^\top\boldsymbol{X}$ has to be inverted). But when some variables are

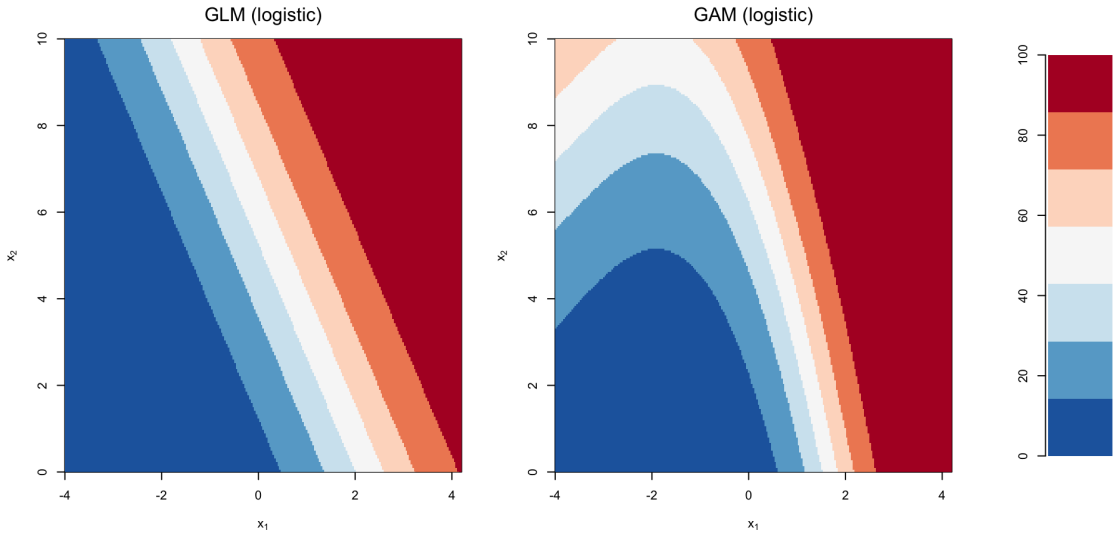


Figure 3.7: Evolution of $(x_1, x_2) \mapsto \widehat{m}(x_1, x_2, \texttt{A})$, on the `toydata2`, with a plain logistic regression on the left (GLM), and a Generalized Additive Model (GAM) on the right. The area in the lower left part corresponds to low probabilities for $\mathbb{P}[Y = 1|X_1 = x_1, X_2 = x_2]$, while the area in the uper right part corresponds to high probabilities. True values of $(x_1, x_2) \mapsto \mu(x_1, x_2, \texttt{A}) = \mathbb{E}[Y|x_1, x_2, \texttt{A}]$, are on the left of Figure 1.6. The scale can be visualized on the right (in %).

highly correlated, there can be numerical issues. Hoerl and Kennard (1970) suggested to use Tikhonov regularization, consisting in adding a (small) positive value on the diagonal of $\boldsymbol{X}^\top\boldsymbol{X}$, so that the matrix can be inverted. Thus, one could consider, for some $\lambda > 0$

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\mathbb{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y},$$

which can be also seen as the solution of the penalized objective

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \text{argmin}\left\{\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^k \beta_j^2\right\},$$

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \text{argmin}\Big\{\underbrace{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_{\ell_2}^2}_{\text{=empirical risk}} + \underbrace{\lambda\|\boldsymbol{\beta}\|_{\ell_2}^2}_{\text{=penalty}}\Big\}.$$

Here $\lambda \geq 0$ is a tuning parameter. It can be related to the Lagrangian in some constrained optimization problem,

$$\min\left\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_{\ell_2}^2\right\} \text{ subject to } \|\boldsymbol{\beta}\|_{\ell_2}^2 \leq k.$$

For the interpretation variables should have the same scales, so classically, variables are standardized, to have unit variance. In an OLS context, we want to solve

Definition 3.3.14 (Ridge Estimator (OLS)) *Hoerl and Kennard (1970)*

$$\widehat{\boldsymbol{\beta}}_{\lambda}^{ridge} = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{argmin}\left\{\frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{k}\beta_j^2\right\}.$$

or more generally (when maximizing the log-likelihood)

Definition 3.3.15 (Ridge Estimator (GLM))

$$\widehat{\boldsymbol{\beta}}_{\lambda}^{ridge} = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{argmin}\left\{-\sum_{i=1}^{n}\log f(y_i|\mu_i = g^{-1}(\boldsymbol{x}_i^\top\boldsymbol{\beta})) + \lambda\sum_{j=1}^{k}\beta_j^2\right\}.$$

See van Wieringen (2015) for much more results on Ridge regression. The "*least absolute shrinkage and selection operator*" regression, also called "LASSO", was introduced in Santosa and Symes (1986), and popularized by Tibshirani (1996) that extended Breiman (1995). Heuristically, the best subset selection problem can be expressed as

$$\min\left\{\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|_{\ell_2}^2\right\} \text{ subject to } \|\boldsymbol{\beta}\|_{\ell_0} \leq \kappa,$$

where $\|\boldsymbol{\beta}\|_{\ell_0}$ denotes a so-called "ℓ_0-norm", which is defined as $\|\boldsymbol{\beta}\|_{\ell_0} = k$ if exactly k components of $\boldsymbol{\beta}$ are nonzero. More generally, consider

$$\min\left\{\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|_{\ell_2}^2\right\} \text{ subject to } \|\boldsymbol{\beta}\|_{\ell_p} \leq \kappa,$$

with $\|\boldsymbol{\beta}\|_p = \left(\sum_{j=1}^{p}|\beta_j|^p\right)^{1/p}$. On the one hand, if $p \leq 1$, the optimisation problem can be seen as a variable section technique, since the optimal parameter has some null components (this corresponds to the statistical concept of "sparsity", see Hastie et al. (2015)). On the other hand, if $p \geq 1$, it is a convex constraint (strictly convex if $p > 1$), which simplifies computations. Thus, $p = 1$ is an interesting case. When the objective is the sum of the squares of residuals, we want to solve

Definition 3.3.16 (LASSO Estimator (OLS)) *Tibshirani (1996)*

$$\widehat{\boldsymbol{\beta}}_{\lambda}^{lasso} = \underset{}{argmin}\left\{\frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{k}|\beta_j|\right\}$$

.

or more generally (when maximizing the log-likelihood)

Definition 3.3.17 (LASSO Estimator (GLM))

$$\widehat{\boldsymbol{\beta}}_{\lambda}^{lasso} = \underset{}{argmin}\left\{-\sum_{i=1}^{n}\log f(y_i|\mu_i = g^{-1}(\boldsymbol{x}_i^\top\boldsymbol{\beta})) + \lambda\sum_{j=1}^{k}|\beta_j|\right\}$$

.

And it actually possible to consider the "elastic net method" that (linearly) combines the ℓ_1 and ℓ_2 penalties of the LASSO and ridge methods. Starting from the LASSO penalty, Zou and Hastie (2005) suggested to add a quadratic penalty term that makes the loss function strictly convex, and it therefore it has a unique minimum. In the OLS framework, consider

$$\widehat{\boldsymbol{\beta}}^{\text{elastic}}_{\lambda_1,\lambda_2} = \text{argmin}\left\{\frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + \lambda_1\sum_{j=1}^{k}|\beta_j| + \lambda_2\sum_{j=1}^{k}\beta_j^2\right\}.$$

In R, the package `glmnet` can be use to estimate those models, as on Figure 3.8, on the `germandata` training dataset[6]. We can visualize the shrinkage, and variable selection: if we want to consider only two variables the indicator associated with no checking account ($\mathbf{1}$(`no checking account`)) and the duration of the credit (`duration`) are supposed to be the "best" two. Note that in those algorithms, variables y and $\boldsymbol{x}$ are usually centered, to remove the intercept β_0 and also scaled. If we assume further that variables $\boldsymbol{x}$ are orthogonal with unit ℓ_2 norm, $\boldsymbol{X}^\top\boldsymbol{X} = \mathbb{I}$, then

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}}_{\lambda} = \frac{1}{1+\lambda}\widehat{\boldsymbol{\beta}}^{\text{ols}} \propto \widehat{\boldsymbol{\beta}}^{\text{ols}},$$

which is a "shrinkage estimator" of the least square estimator (since the proportionality coefficient is smaller than 1). Considering $\lambda > 0$ will induce a bias, $\mathbb{E}[\widehat{\boldsymbol{\beta}}^{\text{ols}}] \neq \mathbb{E}[\widehat{\boldsymbol{\beta}}^{\text{ridge}}_{\lambda}]$, but at the same time it could (hopefully) "reduce" the variance, in the sense that $\text{Var}[\widehat{\boldsymbol{\beta}}^{\text{ols}}] - \text{Var}[\widehat{\boldsymbol{\beta}}^{\text{ridge}}_{\lambda}]$ is a positive matrix. Theobald (1974) and Farebrother (1976) proved that such property holds for some $\lambda > 0$.

The variance reduction of those estimators comes with a price: those estimators of $\boldsymbol{\beta}$ deliberately biased, as well as predictions since $\widehat{y}_i = \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_\lambda$ (whatever the penalty considered) and therefore, those models are not well calibrated (in the sense discussed in Section 4.3.3). Nevertheless, as discussed in Steyerberg et al. (2001) for instance (in a classification context), those techniques might actually improve the calibration of predictions, especially when the number of covariates is large. And as proved by Fan and Li (2001), the LASSO estimate satisfies a nice consistency property, in the sense that the probability of estimating 0's for zero-valued parameters tends to one, when $n \to \infty$. The selects the correct variables and provides unbiased estimates of selected parameters, satisfying an "*oracle property*."

It is also possible to make a connection between credibility and penalize regression, as shown in Miller (2015a) and Fry (2015). Recall that Bühlmann credibility is the solution of a Bayesian model whose prior depends on the likelihood hypothesis for the target, as discussed in Jewell (1974), Klugman (1991) or Bühlmann and Gisler (2005). And penalized regression are the solution of a Bayesian model with either a normal (for Ridge) or a Laplace (for LASSO) prior.

3.3.4 Neural Nets

We will refer to Denuit et al. (2019b) for more details, and applications in actuarial science. Neural networks are first an architecture, that can be seen as an extension of the one we have seen with GLMs, and GAMs. On Figure 3.9, we have a neural network with two "hidden layers," between the predictor variables $\boldsymbol{x} = (x_1, \cdots, x_k)$, and the output. The first layer here consists in three "neurons" (or latent variables), and the second layer consists in two neurons.

To get a more visual understanding, one can consider the use of principal component analysis (PCA) to reduce dimension in a GLM, as on Figure 3.10. The single layer consists here in the collection of the k

[6]This is a simple example, with covariates that are both continuous and categorical. See Friedman et al. (2001) or Hastie et al. (2015) for more details.

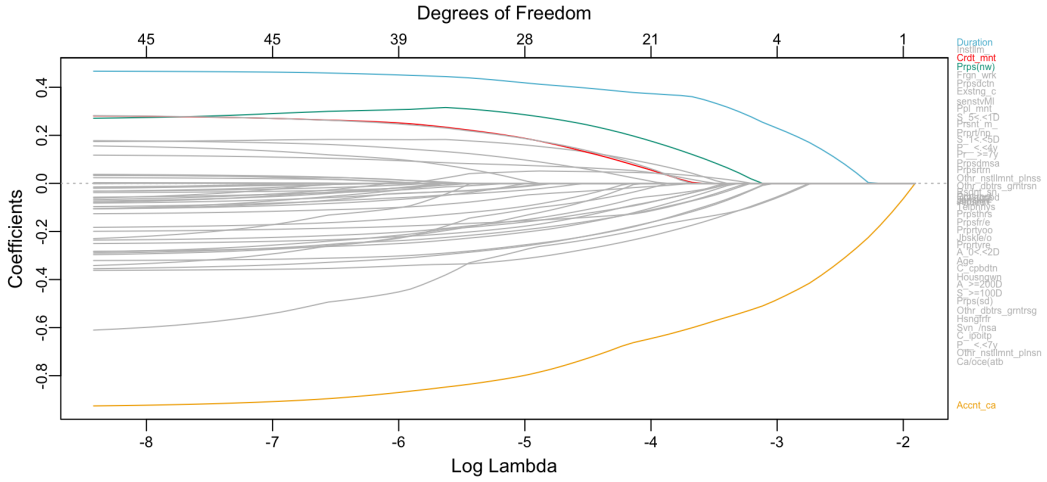


Figure 3.8: Evolution of $\lambda \mapsto \widehat{\boldsymbol{\beta}}_{\lambda}^{\text{lasso}}$, on the `germancredit` dataset on a logistic regression, with continuous variables `duration` and `Credit_amount`, as well as indicators $\mathbf{1}$(`no checking account`), or $\mathbf{1}$(`car (new)`).

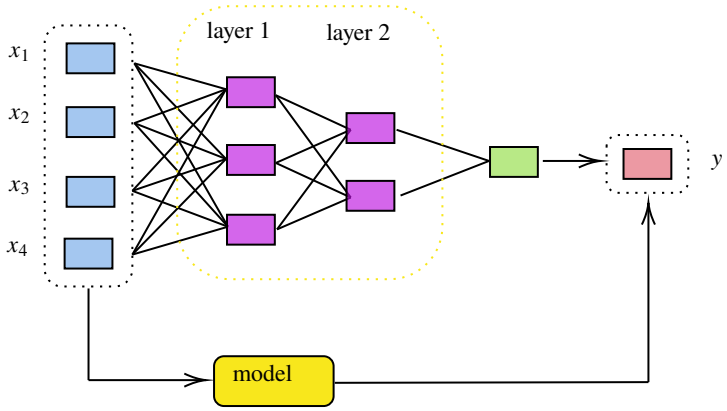


Figure 3.9: Explanatory diagram of a Neural Network, starting from the same predictor variables $\boldsymbol{x} = (x_1, \cdots, x_k)$ (actually, a normalized version of those variables, as explained in Friedman et al. (2001)) and with the same target variable y. The intermediate layers (of neurons) can be considered as the constitution of intermediate features which are then aggregated

.

principal components, obtained using simple algebra, so that each component z_j is a linear combination of the predictors $\boldsymbol{x}$. In this architecture, we consider a single layer, with k neurons, and only two will be used afterwards. Here, we conserve the idea of using linear combination of the variables.

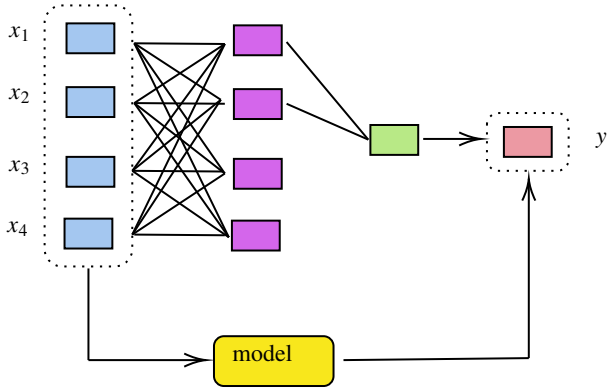


Figure 3.10: Explanatory diagram showing the use of principal component analysis (PCA, see section 3.4) to reduce dimension in a GLM, seen as a neural network architecture, starting from the same predictor variables $\boldsymbol{x} = (x_1, \cdots, x_k)$ and with the same target variable y. The intermediate layer are the k principal components. Then, the GLM is considers not on the k predictors, but on the first two principal components.

Once the architecture is fixed, we will not use feature engineering to get interpretative neurons in the intermediate layer, we will care only about accuracy, and the construction of intermediate variables is optimized.

As a starting point, consider some binary explanatory variables, $\boldsymbol{x} \in \{0,1\}^k$, McCulloch and Pitts (1943) suggested a simple model, with threshold b

$$y_i = h\left(\sum_{j=1}^{k} x_{j,i}\right) \text{ where } h(x) = \mathbf{1}(x \geq b),$$

or (equivalently)

$$y_i = h\left(\omega + \sum_{j=1}^{k} x_{j,i}\right) \text{ where } h(x) = \mathbf{1}(x \geq 0)$$

with weight $\omega = -b$. The trick of adding 1 as an input was very important (as the intercept in the linear regression) and can lead to simple interpretation. For instance, if $\omega = -1$ we recognize the "`or`" logical operator ($y_i = 1$ if $\exists j$ such that $x_{j,i} = 1$), while if $\omega = -k$, we recognize the "`and`" logical operator ($y_i = 1$ if $\forall j, x_{j,i} = 1$). Unfortunately, it is not possible to get the "`xor`" logical operator ($y_i = 1$ if $x_{1,i} \neq x_{2,i}$) with this architecture. Rosenblatt (1961) considered the extension where $\boldsymbol{x}$'s are real-valued (instead of binary), with "weight" $\boldsymbol{\omega} \in \mathbb{R}^k$ (the word is between quotation marks because here, weights can be negative)

$$y_i = h\left(\sum_{j=1}^{k} \omega_j x_{j,i}\right) \text{ where } h(x) = \mathbf{1}(x \geq b),$$

or (equivalently, with $\omega_0 = -b$)

$$y_i = h\left(\omega_0 + \sum_{j=1}^{k} \omega_j x_{j,i}\right) \text{ where } h(x) = \mathbf{1}(x \geq 0)$$

with "weights" $\boldsymbol{\omega} \in \mathbb{R}^{k+1}$. Even if there is no probabilistic foundations here, we recognize an expression close to $\widehat{y} = g^{-1}(\boldsymbol{x}^\top \boldsymbol{\beta})$.

Minsky and Papert (1969) proved that those perceptrons were linear separators, unfortunately not very powerful. The step function h was quickly replaced by the sigmoid function $h(x) = (1 + e^{-x})^{-1}$, that corresponds to the logistic function, already used by statisticians since Wilson and Worcester (1943) and Berkson (1944). This function h is called the "activation function". Scientists working in signal theory used both $y \in \{0, 1\}$ (off and on) and $y \in \{-1, +1\}$ (negative and positive). In that case, one can consider the hyperbolic tangent as an activation function

$$h(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})},$$

or the inverse tangent function. So here, for a classification problem,

$$y_i = h\left(\omega_0 + \sum_{j=1}^{k} \omega_j x_{j,i}\right) = m(\boldsymbol{x}_i).$$

But so far, there is nothing more, compared with econometric model (actually we have less since there is no probabilistic foundations, so no confidence intervals for instance). The interesting idea was to replicate those models, those neural models, into a network. Instead of mapping $\boldsymbol{x}$ and y, the idea is to map $\boldsymbol{x}$ and some latent variables z_j, and then to map $\boldsymbol{z}$ and y, using the same structure. And possible we can use a deeper network, not with one single layer, but much more. Consider Figure 3.11, which is a simplified version of Figure 3.9.

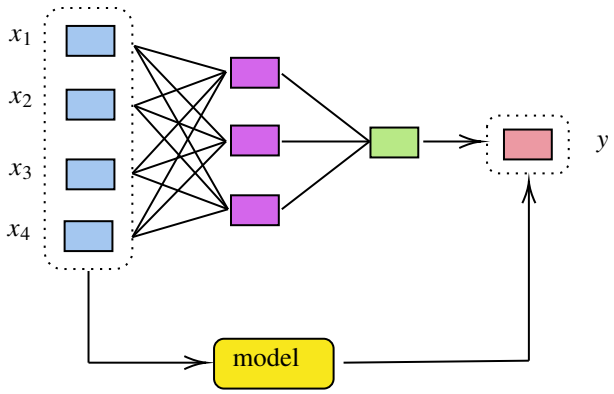


Figure 3.11: Neural network, starting from the same predictor variables $\boldsymbol{x} = (x_1, \cdots, x_k)$ and with the same target variable y, and a single layer, with variables $\boldsymbol{x} = (z_1, \cdots, z_J)$.

Here we consider a single layer, and $\boldsymbol{z} = (z_1, z_2, z_3)$,

$$\begin{cases} z_j = h_j\left(\omega_{j,0} + \boldsymbol{x}^\top \boldsymbol{\omega}_j\right), \ (\omega_{j,0}, \boldsymbol{\omega}_j) \in \mathbb{R}^{k+1} \\ y = h\left(\omega_0 + \boldsymbol{z}^\top \boldsymbol{\omega}\right), \ (\omega_0, \boldsymbol{\omega}) \in \mathbb{R}^{k+1}, \end{cases}$$

so that, when plugging-in,

$$y = m_{\boldsymbol{\omega}}(\boldsymbol{x}) = h\left(\omega_0 + \boldsymbol{z}^\top \boldsymbol{\omega}\right) = h\left(\omega_0 + \sum_{j=1}^{J} \omega_j h_j\left(\omega_{j,0} + \boldsymbol{x}^\top \boldsymbol{\omega}_j\right)\right).$$

Our model m_ω is now based on $(k+1)(J+1)$ parameters. Given a model m_ω, we can compute the quadratic loss function,

$$\sum_{i=1}^{n}(y_i - m_\omega(\boldsymbol{x}_i))^2,$$

the cross-entropy

$$\sum_{i=1}^{n}(y_i \log m_\omega(\boldsymbol{x}_i) + [1-y_i]\log[1-m_\omega(\boldsymbol{x}_i)]),$$

or any empirical risk, based on loss ℓ,

$$\sum_{i=1}^{n}\ell\left(y_i, m_\omega(\boldsymbol{x}_i)\right).$$

A natural idea is then to get optimal weights $\widetilde{\boldsymbol{w}}^\star = (\boldsymbol{w}, \boldsymbol{w}_1, \cdots, \boldsymbol{w}_J)$, by solving

$$\widetilde{\omega}^\star = \underset{\omega}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}\ell\left(y_i, m_\omega(\boldsymbol{x}_i)\right)\right\}.$$

This will be called a “back propagation" problem. Of course, theoretically, this can be performed whatever the architecture of the network, with multiple layers.

While classical regression models (such as GLM) are based on convex optimization problems, (deep) Neural Networks are usually not convex problems. And sometimes, the dimension of the dataset can be really big, so we cannot pass all the data to the computer at once to compute $m(\boldsymbol{x})$, we need to divide the data into smaller sizes and give it to our computer one by one (called “batches").

In the previous figures, we considered the case where layers were somehow related to some dimension reduction, with a number of hidden nodes $\boldsymbol{z}$ smaller than the number of predictors $\boldsymbol{x}$, but it is actually possible to consider more nodes in the hidden layers than the number of predictors. The point is to capture nonlinearity, and nonmonotonicity.

Without going too much further (see Denuit et al. (2019b) or Wüthrich and Merz (2022) for more more details), let us stress here that neural networks are intensively used because there have strong theoretical foundations and some “universal approximation theorems”, even in very general frameworks. Those results were obtained at the end of the 80’s, with Cybenko (1989), with an arbitrary number of artificial neurons, or Hornik et al. (1989); Hornik (1991), with multilayer feed-forward networks and only one single hidden layer. Later on, Leshno et al. (1993) proved that the universal approximation theorem was equivalent to having a non-polynomial activation function (see Haykin (1998) for more details on theoretical properties).

Unfortunately, a major drawback, especially in actuarial science, is the lack of probabilistic foundations of those models. This concern has raised some literature in the 80s, starting with Rumelhart et al. (1985) and Rumelhart et al. (1986), or more recently Hertz et al. (1991) and Buntine and Weigend (1991), where back-propagation is formalized in a Bayesian context, taken up by MacKay (1992) and Neal (1992) more than 30 years ago (or more recently Neal (2012) Theodoridis (2015), Gal and Ghahramani (2016) and Goulet et al. (2021)). Observe that similar issues were discussed in other machine learning techniques, such as “support vector machine," where the distance to the separation line is used as a score which can then be interpreted as a probability, as “Platt scaling" in Platt et al. (1999) or “isotonic regression" in Zadrozny and Elkan (2001, 2002) (see also Niculescu-Mizil and Caruana (2005a) where “good probabilities" are defined), as discussed in Section 4.3.3.

3.3.5 Trees and Forests

Again, let us briefly explain the general idea (see Denuit et al. (2020) for more details). Decision trees appeared in the statistical litterature in the 70's and the 80's, with Messenger and Mandell (1972) with THAID (Theta Automatic Interaction Detection) then, Breiman and Stone (1977) and Breiman et al. (1984) with CART (Classification And Regression Trees), as well as Quinlan (1986, 1987, 1993) with ID3 (Iterative Dichotomiser 3) that became later on C4.5 ("*a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date*" as wrote in Witten et al. (2016)). One should probably mention here the fact that idea of "decision trees" was mentioned earlier in psychology, as discussed in Winterfeldt and Edwards (1986), but without any details about the algorithmic construction (see also Lima (2014) for old representations of "decision trees"). Indeed, as already note in Laurent and Rivest (1976) constructing optimal binary decision trees is long and time consuming, and using some backward procedure will fasten the process. Starting with the entire training set, we select the most predictive variable, and we split the population in two, using an appropriate threshold and this predictive variable. And then we iterate within each sub-population. Heuristically, each sub-population should be as homogeneous as possible, with respect to y. In the methods considered previously, we consider all explanatory variables together, using linear algebra techniques to solve optimization problems more efficiently, but here, variables are used sequentially (or to be more specific, binary step functions, since x_j becomes $\mathbf{1}(x_j \leq t)$ for some optimally selected threshold). After some iterations, we end up with some subgroups, or sub-regions in $\mathcal{X}$ called either "leaves" or terminal nodes, while intermediary splits are called internal nodes. In the visual representation, segments connecting nodes are called "branches", and to continue with the arboricultural and forestry metaphors, we will evoke the pruning when we will trim the trees to avoid possible overfit.

As mentioned, after several iterations, we will split the population and the space $\mathcal{X}$ into a partition of J regions, $R_1, \cdots, R_J$, such that $R_i \cap R_j = \emptyset$ when $i \neq j$ and $R_1 \cup R_2 \cup \cdots \cup R_J = \mathcal{X}$. Then, in each region, prediction is performed simply by considering the average of y for observations in that specific region, in a regression context, if $|R_j|$ denotes the number of observations in region R_j,

$$\widehat{y}_{R_j} = \frac{1}{|R_j|} \sum_{i:\boldsymbol{x}_i \in R_j} y_i$$

or using a majority rule for a classifier. Classically, regions are (hyper) rectangles in $\mathcal{X} \subset \mathbb{R}^k$ (or orthants), in order to simplify the construction of the tree, and to have a simple and graphical interpretation of the model. In a regression context, the classical strategy is to minimize the mean squared error, which is the in-sample empirical risk for the ℓ_2-norm,

$$\text{MSE} = \sum_{j=1}^{J} \text{MSE}_j \text{ where } \text{MSE}_j = \sum_{i:\boldsymbol{x}_i \in R_j} \left(y_i - \widehat{y}_{R_j} \right)^2.$$

where $\widehat{y}_{R_j}$ is the prediction in region R_j. For a classification problem, observe that

$$\text{MSE}_j = \sum_{i:\boldsymbol{x}_i \in R_j} \left(y_i - \widehat{y}_{R_j} \right)^2 = n_{0,j}(0 - \widehat{y}_{R_j})^2 + n_{1,j}(1 - \widehat{y}_{R_j})^2,$$

so that, if $n_{0,j}$ and $n_{1,j}$ denote the number of observations such that $y = 0$ and $y = 1$, respectively, in the region R_j,

$$\max\{\text{MSE}_j\} = n_{0,j} \left(\frac{n_{0,j}}{n_{0,j} + n_{1,j}} \right)^2 + n_{1,j} \left(\frac{n_{0,j}}{n_{1,j} + n_{1,j}} \right)^2 = \frac{n_{0,j} n_{1,j}}{n_{0,j} + n_{1,j}},$$

and therefore

$$\text{MSE} = \sum_{j=1}^{J} \frac{n_{0,j} n_{1,j}}{n_{0,j} + n_{1,j}} = \sum_{j=1}^{J} n_j \cdot p_{1,j}(1 - p_{1,j}),$$

which corresponds to "Gini impurity index".

Formally, an impurity index is some function $\psi : [0, 1] \to \mathbb{R}_+$ positive, symmetric ($\psi(u) = \psi(1 - u)$), minimal in 0 and 1 (consider some normalised version $\psi(0) = \psi(1) = 0$). Classical indices are Gini index, $\psi(u) = u(1 - u)$, the missclassification function $\psi(u) = 1 - \max\{u, 1 - u\}$ and the (cross) entropy, $\psi(u) = -u \log u - (1 - u) \log(1 - u)$.

Instead of considering all possible partitions of $\mathcal{X}$ that are rectangles, some "top-down greedy approach" is usually considered, with some recursive binary splitting. The algorithm is top-down since we start with all observations in the same class (usually called the "root" of the tree) and then we split into regions that are smaller and smaller. And it is greedy since optimization is performed without looking back at the past. Formally, at the first stage, select variable x_κ and some cut-off point t and consider half spaces

$$R_1(\kappa, t) = \{\boldsymbol{x} = (x_1, \cdots, x_k) | x_\kappa < t\} \subset \mathcal{X}$$
$$R_2(\kappa, t) = \{\boldsymbol{x} = (x_1, \ldots, x_k) | x_\kappa \geq t\} \subset \mathcal{X}$$

We seek the regions that minimize the mean squared error,

$$\text{MSE}_\kappa(t) = \sum_{i:\boldsymbol{x}_i \in R_1(\kappa,c)} \left(y_i - \widehat{y}_{R_1(\kappa,t)}\right)^2 + \sum_{i:\boldsymbol{x}_i \in R_2(\kappa,c)} \left(y_i - \widehat{y}_{R_2(\kappa,t)}\right)^2.$$

or more generally any impurity function for a classification problem

$$n_1 \cdot \psi\left(\widehat{y}_{R_1(\kappa,t)}\right) + n_2 \cdot \psi\left(\widehat{y}_{R_2(\kappa,t)}\right).$$

Then find the best variable and the best cut-off, solution of

$$\min_{\kappa=1,\cdots,k} \left\{ \inf_{t \in \mathcal{X}_\kappa} \left\{\text{MSE}_\kappa(t)\right\}\right\}$$

At stage $j + 1$, repeat the previous procedure on all regions created at stage j: within region $\mathcal{X}_j$, identify the variable x_κ and cut-off point t that will minimize the empirical risk and yield to the split of $\mathcal{X}_j$ into two half spaces

$$\{\boldsymbol{x} \in \mathcal{X}_r | x_\kappa < t\} \text{ and } \{\boldsymbol{x} \in \mathcal{X}_r | x_\kappa \geq t\}.$$

And iterate. Ultimately, each leaf will contain a single observation, which will correspond to a null empirical risk, on the training dataset, but will hardly generate. To avoid this overfit, it will be necessary either to introduce a stopping criteria, and to prune the complete tree. In practice, we will stop when the leaves have reached a minimum number of observations, set beforehand. An alternative will be to stop if the variation (relative or absolute) of the objective function does not decrease enough. Formally, to decide whether a leaf $\{N\}$ should be divided into $\{N_L, N_R\}$, compute the impurity variation

$$\Delta\mathcal{I}(N_L, N_R) = \mathcal{I}(N) - \mathcal{I}(N_L, N_R) = \psi(\widehat{y}_N) - \left(\frac{n_L}{n}\psi(\widehat{y}_{N_L}) + \frac{n_R}{n}\psi(\widehat{y}_{N_L})\right),$$

We decide to split if $\Delta\mathcal{I}(N_L, N_R)/\mathcal{I}(N)$ exceeds some complexity parameter, usually denoted c_p (the default value in `rpart` in R is 1%).

On Figure 3.12 we can visualize the classification tree obtained on the `toydata2` dataset. Recall that in the entire training population, $\overline{y} = 40\%$. Then the tree grows as follows

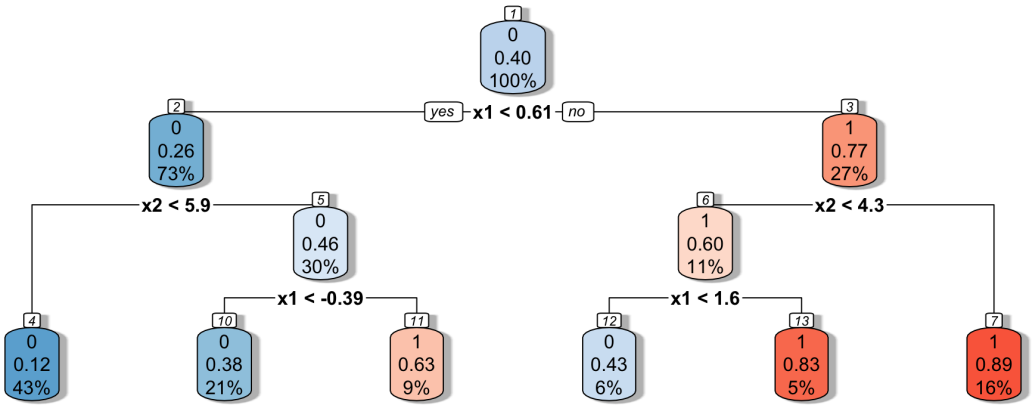


Figure 3.12: Classification tree on the `toydata2` dataset, with default pruning parameters of `rpart`. On top, the entire population (100%, $\overline{y} = 40\%$) and at the bottom six leaves, on the left, 43% of the population and $\overline{y} = 12\%$ and on the right, 16% of the population and $\overline{y} = 89\%$. The six regions (in space $\mathcal{X}$), associated to the six leaves, can be visualized on the right of Figure 3.15.

- if $x_1 < 0.61$ (first node, 73% of the population $\overline{y} = 26\%$)
  - if $x_2 < 5.9$ (first node, first branch, 43%, $\overline{y} = 12\%$, final leaf)
  - if $x_2 \geq 5.9$ (second node, first branch, 30%, $\overline{y} = 30\%$)
    - if $x_1 < -0.39$ (first node, second branch, 21%, $\overline{y} = 38\%$, final leaf)
    - if $x_1 \geq -0.39$ (second node, second branch, 9%, $\overline{y} = 63\%$, final leaf)
- if $x_1 \geq 0.61$ (second node, 27% of the population $\overline{y} = 77\%$)
  - if $x_2 < 4.3$ (first node, third branch, 11%, $\overline{y} = 60\%$)
    - if $x_1 < 1.6$ (first node, fourth branch, 6%, $\overline{y} = 43\%$)
    - if $x_2 \geq 1.6$ (second node, fourth branch, 5%, $\overline{y} = 83\%$, final leaf)
  - if $x_2 \geq 4.3$ (second node, third branch, 16%, $\overline{y} = 89\%$, final leaf)

Observe that only variables `x1` and `x2` are used here. It is then possible to visualize the prediction for a given pair (x_1, x_2) (the value of x_3 and s will have no influence). As on Figure 3.12. It is possible to visualize two specific predictions, as on Table 3.3 (we will discuss further interpretations for those two specific individuals in section 4.1).

On Figure 3.13 we can visualize the classification tree obtained on the `germancredit` dataset. Recall that in the entire training population ($n = 700$), $\overline{y} = 30.1\%$. Then the tree grows as follows

- if `Account_status` ≥ 200 (first node, % of the population $\overline{y} = 13.2\%$, final leaf)
- if `Account_status` < 200 (first node, 73% of the population $\overline{y} = 44.1\%$)
  - if `Duration` < 22.5 (first node, second branch, 43%, $\overline{y} = 12\%$, final leaf)

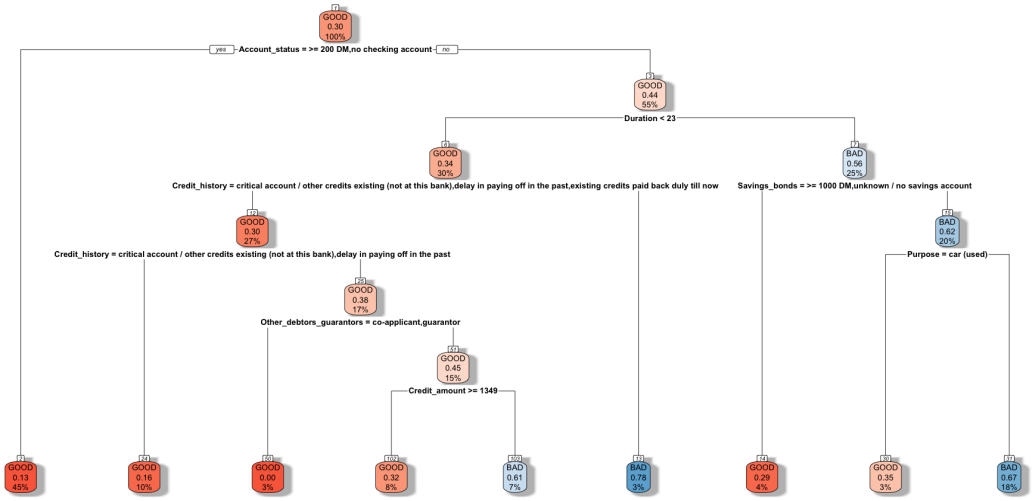


Figure 3.13: Classification tree on the `germancredit` dataset, with default pruning parameters of `rpart`. On top, the entire population (100%, $\overline{y} = 30\%$) and at the bottom nine leaves, on the left, 45% of the population and $\overline{y} = 13\%$ and on the right, 3% of the population and $\overline{y} = 35\%$.

- ...

For the pruning procedure, create a very large and deep tree, and then, cut some branches. Formally, given a large tree $\mathcal{T}_0$ identify a subtree $\mathcal{T} \subset \mathcal{T}_0$ that minimizes

$$\sum_{m=1}^{|\mathcal{T}|} \sum_{i:\mathbf{X}_i \in R_m} \left(Y_i - \widehat{Y}_{R_m}\right)^2 + \alpha|\mathcal{T}|,$$

where α is some complexity parameter, and $|\mathcal{T}|$ is the number of leaves in the subtree $\mathcal{T}$. Observe that it is similar to penalized methods described previously, used to get a tradeoff between bias and variance, between accuracy and parsimony.

Those classification and regression trees are easy to compute, to interpret. Unfortunately, those trees are rather unstable (even if the prediction is much more robust). A natural idea, introduced by Breiman (1996a) consists in growing multiple tree to get a collection of trees, or a “forest" , to improve classification by combining classifiers obtained from randomly generated training sets (using a “bootstrap" procedure – resampling, with replacement), and then to aggregate. This will correspond to “bagging", which is an ensemble approach.

3.3.6 Ensemble Approaches

We have seen so far how to estimate various models, and and classically, we use some metrics to select the best one. But rather than choosing the best among different models, it could be more efficient to combine them. Among those “ensemble methods," there will bagging, random forests, or boosting (see Sollich and

Krogh (1995), Opitz and Maclin (1999) or Zhou (2012)). Those techniques can be related to "Bayesian model averaging, that linearly combines submodels of the same family, with the posterior probabilities of each model, as described in Raftery et al. (1997) or Wasserman (2000), and "stacking", that involves training a model to combine the predictions of several other learning algorithms, as described in Wolpert (1992) or Breiman (1996c).

It should be stressed here that a "weak learner" is defined as a classifier that is only slightly correlated with the true classification (so to speak, it can label examples slightly better than random guessing). In contrast, a "strong learner" is a classifier that is arbitrarily well-correlated with the true classification. Long story short, we will discuss in this section the idea that combining "weak learner" could yield better results than seeking for a "strong learner".

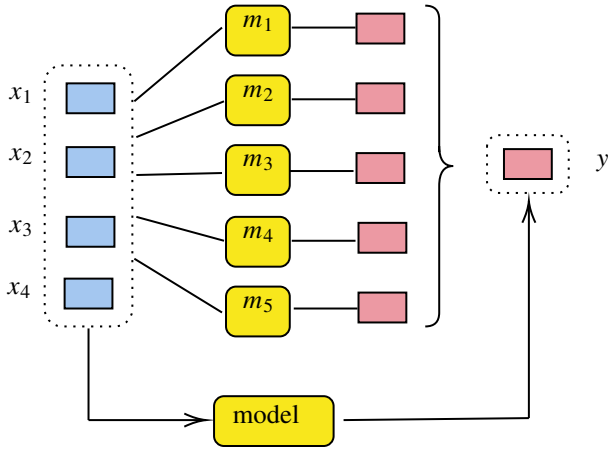


Figure 3.14: Explanatory diagram of parallel training, or "bagging" (bootstrap and aggregation), starting from the same predictor variables $\boldsymbol{x} = (x_1, \cdots, x_k)$ and with the same target variable y. Different models $m_j(\boldsymbol{x})$ are fitted, and the outcome is usually the average of the models.

A first approach is the one described on Figure 3.14: consider a collection of predictions, $\{\widehat{\boldsymbol{y}}^{(1)}, \cdots, \widehat{\boldsymbol{y}}^{(k)}\}$, obtained using k models (possibly from different families, GLM, trees, neural networks, etc), and consider a linear combination of those models, and solve a problem like

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^k} \left\{ \sum_{i=1}^{n} \ell\big(y_i, \boldsymbol{\alpha}^\top \widehat{\boldsymbol{y}}_i\big) \right\}.$$

In a classification problem, a popular aggregation function could be the "majority rule". The predicted class is the one most models predicted. Unfortunately this optimization problem can be rather unstable, since predictions obtained using the different models are usually highly (positively) correlated with one another. Therefore, it is rather natural to add a penalization in the previous problem, to use only a subset of models.

Historically, de Condorcet (1785) already suggested similar techniques to make decisions by taking into account a plurality of votes. Consider a jury, with k judges, and suppose that p is the probability to make a mistake (with a binary decision). If we suppose that mistakes are made independently, the probability that the majority makes a mistake is

$$\sum_{j\geq[(k+1)/2]} \binom{k}{j} p^j (1-p)^{k-j}.$$

For example, with $k = 11$ judges, if each judge has 30% chances to make the decision, the majority is wrong with only 8% chance. This probability will decrease with k, unless there is strong correlation. So in an ideal situation, ensemble techniques work better if predictions are independent from each other.

Galton (1907) suggested that technique while trying to *guess the weight of a ox* in a county fair in Cornwall, England, as recalled in Wallis (2014) and Surowiecki (2004). $k = 787$ participants provided guesses $\widehat{y}_1, \cdots, \widehat{y}_k$. The ox weighed 1,198 pounds, and the average of the estimates was 1,197 pounds. Francis Galton compared two strategies, either picking a single prediction $\widehat{y}_j$ or considering the average $\overline{y}$. If t if the truth, we can write

$$\mathbb{E}\big[(\widehat{y}_j - t)^2\big] = (\overline{y} - t)^2 + \frac{1}{k}\sum_{i=1}^{k}(\widehat{y}_i - \overline{y})^2,$$

so clearly, using the average prediction is better than seeking the “best one". From a statistical perspective, hopefully, ensemble generalises better than a single chosen model. From a computational perspective, averaging should be faster than solving an optimisation problem (seeking the “best" model). And interestingly, it will take us outside classical model classes.

In ensemble learning, we want models to be reasonably accurate, and as independent as possible. A popular example is “bagging" (for *b*ootstrap *agg*regat*ing*), introduced to increase stability of predictions and accuracy. It can also reduce variance, and overfit. The algorithm would be

1. generate k training datasets using boostrap (resampling), since subdividing the database into k smaller (independent) dataset will lead to small training dataset,
2. each dataset is used as a training sample to fit model $\widehat{m}^{(j)}$, with $j = 1, \cdots, k$, such as (deep) trees, with small bias and large variance (variance will decrease with aggregation, as shown below),
3. for a new observation $\boldsymbol{x}$, the aggregated prediction will be

$$\widehat{m}_{\text{bagging}}(\boldsymbol{x}) = \frac{1}{k}\sum_{j=1}^{k}\widehat{m}^{(j)}(\boldsymbol{x}).$$

Observe that here, we use the same weights for all models. The simple version of “random forest" is based on the aggregation of trees. On 3.15 we can visualize the predictions on the `toydata2` dataset, as a function of x_1 and x_2 (and $x_3 = 0$), with a classification tree on the left, and the aggregation of $k = 500$ trees on the right.

As explained in Friedman et al. (2001), the variance of aggregated models[7] is

$$\begin{aligned}\text{Var}\big(\widehat{m}_{\text{bagging}}(\boldsymbol{x})\big) &= \text{Var}\left(\frac{1}{k}\sum_{j=1}^{k}\widehat{m}^{(j)}(\boldsymbol{x})\right) = \frac{1}{k^2}\text{Var}\left(\sum_{j=1}^{k}\widehat{m}^{(j)}(\boldsymbol{x})\right)\\ &= \frac{1}{k^2}\sum_{j_1=1}^{k}\sum_{j_2=1}^{k}\text{Cov}\big[\widehat{m}^{(j_1)}(\boldsymbol{x}), \widehat{m}^{(j_2)}(\boldsymbol{x})\big]\\ &\leq \frac{1}{k^2}k^2\text{Var}\big[\widehat{m}^{(j)}(\boldsymbol{x}))\big] = \text{Var}(\widehat{m}^{(j)}(\boldsymbol{x}),\end{aligned}$$

[7]Variances and covariances are associated with random variables $U_{j,\boldsymbol{x}} = \widehat{m}^{(j)}(\boldsymbol{x})$'s for a given $\boldsymbol{x}$, based on the uncertainty of the training sample.

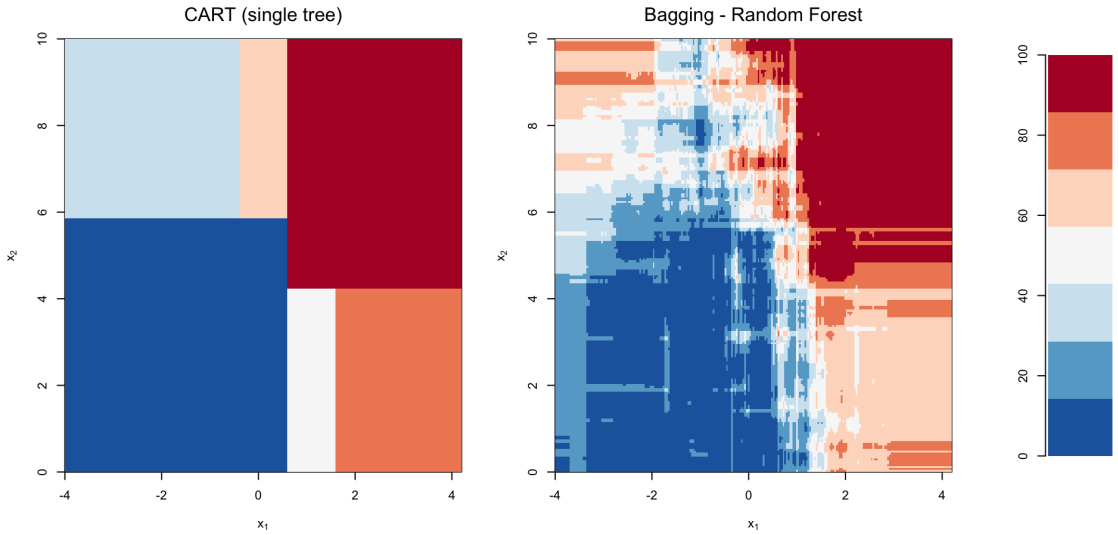


Figure 3.15: Evolution of $(x_1, x_2) \mapsto \widehat{m}(x_1, x_2, \texttt{A})$, on the `toydata2`, with a single classification tree on the left (the output being the probability to have $y = 1$), and a random forest on the right.

since, when $j_1 \neq j_2$, $\text{Corr}\left[\widehat{m}^{(j_1)}(\boldsymbol{x}), \widehat{m}^{(j_2)}(\boldsymbol{x})\right] \leq 1$. And actually, even if $\text{Var}(\widehat{m}^{(j)}(\boldsymbol{x})) = \sigma^2(\boldsymbol{x})$ and $\text{Corr}[\widehat{m}^{(j_1)}(\mathbf{x}), \widehat{m}^{(j_2)}(\boldsymbol{x})] = r(\boldsymbol{x})$,

$$\text{Var}\left(\widehat{m}_{\text{bagging}}(\boldsymbol{x})\right) = r(\boldsymbol{x})\sigma^2(\boldsymbol{x}) + \frac{1 - r(\boldsymbol{x})}{k}\sigma^2(\boldsymbol{x}).$$

The variance will be lower, when "more different" models are aggregated. And more generally, as explained in Denuit et al. (2020), Section 4.3.3,

$$\mathbb{E}\left[\ell\left(Y, \widehat{m}_{\text{bagging}}(\boldsymbol{X})\right)\right] \leq \mathbb{E}\left[\ell\left(Y, \widehat{m}^{(j)}(\boldsymbol{X})\right)\right].$$

Another type of ensemble model is related to sequential learning, with "boosting," described on Figure 3.16, introduced in Schapire (1990) and Breiman (1996b). Boosting is based on the question posed by Kearns and Valiant (1989) "*can a set of weak learners create a single strong learner?*." For a regression, the algorithm for boosting would be

1. initialization : k (number of trees), γ (learning rate), $m_0(\boldsymbol{x}) = \overline{y}$
2. at stage $t \geq 1$,
   - 2.1 compute "residuals" $r_{i,t} \leftarrow y_i - m_{t-1}(\boldsymbol{x}_i)$;
   - 2.2 fit a model $r_{i,t} \sim h(\boldsymbol{x}_i)$ for some weak learner (not too deep tree) h;
   - 2.3 update $m_t(\cdot) = m_{t-1}(\cdot) + \gamma h(\cdot)$;
   - 2.4 loop ($t \leftarrow t + 1$ and return to 2.1)

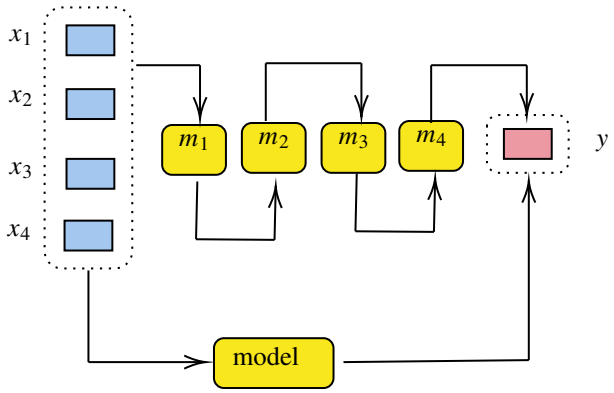


Figure 3.16: Explanatory diagram of sequential learning ("boosting") starting from the same predictor variables $\boldsymbol{x} = (x_1, \cdots, x_k)$ and with the same target variable y. Using "weak" learners h_t learning from the residuals of the previous models, and to improve sequentially the model (h_t is fitted to explain residuals $y - m_t(\boldsymbol{x})$, and the update is $m_{t+1} = m_t + h_{t+1}$).

or more generally,

1. initialization : k (number of trees), γ (learning rate), and $m_0(\boldsymbol{x}) = \underset{\varphi}{\text{argmin}} \sum_{i=1}^{n} \ell(y_i, \varphi)$;
2. at stage $t \geq 1$,
   - 2.1 compute "residuals" $r_{i,t} \leftarrow \left.\frac{\partial \ell(y_i, \widehat{y})}{\partial \widehat{y}}\right|_{\widehat{y}=m_{t-1}(\boldsymbol{x}_i)}$;
   - 2.2 fit a model $r_{i,t} \sim h_t(\boldsymbol{x}_i)$ for some weak learner $h \in \mathcal{H}$, $h_t = \underset{h \in \mathcal{H}}{\text{argmin}} \sum_{i=1}^{n} \ell(r_{i,t}, h(\boldsymbol{x}_i))$;
   - 2.3 update $m_t(\cdot) = m_{t-1}(\cdot) + \gamma h_t(\cdot)$;
   - 2.4 loop ($t \leftarrow t + 1$ and return to 2.1)

In the context of classification, Freund and Schapire (1997) introduced the "`adaboost`" algorithm (for "adaptive boosting"), based on updating weights (see also Schapire (2013) for additional heuristics). Consider some binary classification problem (where y takes values 0 or 1) with a training dataset $\mathcal{D} = (\{y_i; \boldsymbol{x}_i\}, i = 1, \ldots, n)$. Following Algorithm 10.1 in Friedman et al. (2001),

1. set weights $\omega_{i,1} = 1/n$, for $i = 1, \ldots, n$, $\gamma > 0$, $m_0(\boldsymbol{x}) = 0$, and define $\boldsymbol{\omega}_1$;
2. at stage $t \geq 1$,
   - 2.1 generate (by resampling, using weights $\boldsymbol{\omega}_t$) a training dataset $\mathcal{D}_t = (\mathcal{D}, \boldsymbol{\omega}_t)$;
   - 2.2 learn a model $y_i \sim h_t(\boldsymbol{x}_i)$ on $\mathcal{D}_t$, $h_t = \underset{h \in \mathcal{H}}{\text{argmin}} \sum_{i=1}^{n} \omega_{i,t} \ell(y_i, h(\boldsymbol{x}_i))$;

2.3 compute the "weighted error rate"

$$\bar{\epsilon}_t \leftarrow \frac{\sum_{i=1}^{n} \omega_{i,t} \cdot \mathbf{1}\left(h_t(\boldsymbol{x}_i) \neq y_i\right)}{\sum_{i=1}^{n} \omega_{i,t}}$$

on the training dataset $\mathcal{D}_t$, and compute $\alpha_t = \log((1-\bar{\epsilon}_t)/\bar{\epsilon}_t)$;

2.4 update the weights

$$\omega_{i,t+1} \leftarrow \omega_{i,t} \cdot \exp\left[\alpha_t \mathbf{1}\left(h_t(\boldsymbol{x}_i) \neq y_i\right)\right];$$

and update $m_{t+1}(\cdot) = m_t(\cdot) + \gamma\alpha_t h_t(\cdot)$;

2.5 loop ($t \leftarrow t+1$ and return to 2.1).

Friedman et al. (2000)and Niculescu-Mizil and Caruana (2005b) proved strong connections between additive models (GAM) and boosting. Since we consider a classifier, a natural loss is $\ell_{0/1}(y,\hat{y}) = \mathbf{1}(y \neq \hat{y})$. But for computations, a classical strategy consists in "convexifying" the loss function, by considering $\overline{\ell}(y,\hat{y}) = \exp(-y\hat{y})$, when $y \in \{-1,+1\}$.

On Figure 3.17, two boosting models are trained on the `toydata2` dataset, with different learning rates γ (respectively 2% and 10% on the left and on the right). With a higher learning rate (on the right), convergence is obtained faster. When the number of iterations t is too large, the model is "overfitting" , as we can see on the right, on the validation dataset (with function `gbm` in R, k-fold cross-validation is used, instead of the training/validation approach described previously, see Friedman et al. (2001) for more discussions). On the y-axis, the Bernoulli deviance is used. On Figure 3.18, we focus on predictions of two specific individuals `Andrew` and `Barbara` (as described in Table 3.3), with $t \mapsto m_t(\boldsymbol{x})$. The "optimal" predictions are respectively obtained with 450 and 100 trees, since those were values that minimize the overall error, on the validation samples (on Figure 3.18).

3.3.7 Application on the `toydata2` Dataset

On the `toydata2` dataset, we can visualize the prediction on two individuals, `Andrew` and `Barbara`, in Table 3.3.

	x_1	x_2	x_3	s	$\mu(\boldsymbol{x},s)$	$\widehat{m}_{\text{glm}}(\boldsymbol{x})$	$\widehat{m}_{\text{gam}}(\boldsymbol{x})$	$\widehat{m}_{\text{cart}}(\boldsymbol{x})$	$\widehat{m}_{\text{rf}}(\boldsymbol{x})$	$\widehat{m}_{\text{gbm}}(\boldsymbol{x})$
`Andrew`	-1	8	-2	A	0.366	0.379	0.372	0.384	0.310	0.354
`Barbara`	1	4	2	B	0.587	0.615	0.583	0.434	0.622	0.595

Table 3.3: Predictions for two individuals `Andrew` and `Barabara`, with models trained on `toydata2`, with the true value μ (used to generate data) then a plain logistic (glm), an additive logistic one (gam), a classification tree (cart), a random forest (rf) and the a boosting models (gbm, as on Figure 3.18).

The four models trained are fair by unawareness, in the sense that the sensitive attribute was not used. Our two individuals are characterized by $\boldsymbol{x}_{\text{A}}$ and $\boldsymbol{x}_{\text{B}}$. Consider some linear interpolation between those two individuals (`Andrew` on the left, `Barbara` on the right), $\boldsymbol{x}_t = t\boldsymbol{x}_{\text{B}} + (1-t)\boldsymbol{x}_{\text{A}}$, with $t \in [0,1]$. On Figure 3.19, we can visualize $t \mapsto \widehat{m}(\boldsymbol{x}_t)$.

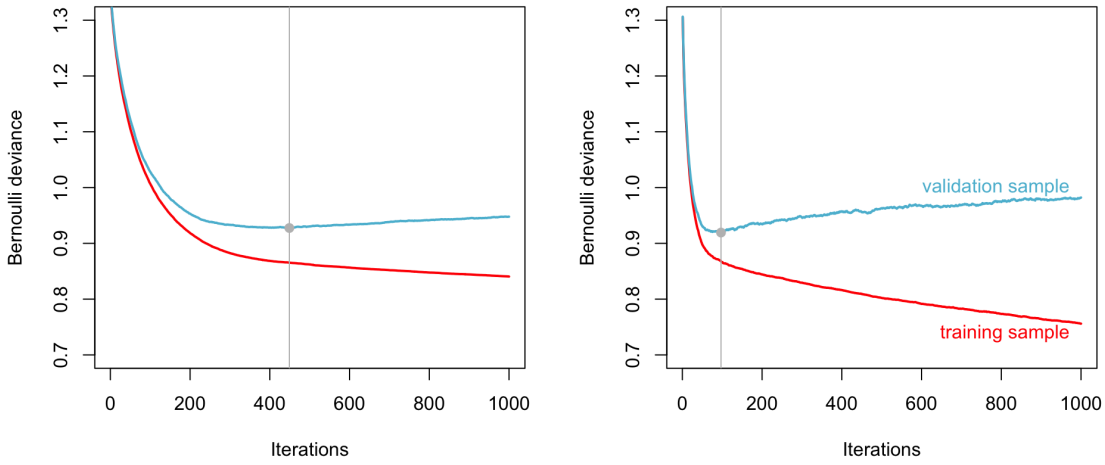


Figure 3.17: Evolution of Bernoulli deviance of m_t with a boosting learning (`adaboost`) on the `toydata2` dataset, as a function of t, on the validation sample (on top) and on the training sample (below), with two different learning rates (on the left and on the right). Vertical lines correspond to the "optimal" number of (sequential) trees.

3.3.8 Application on the `GermanCredit` Dataset

The classification tree on the `GermanCredit` training dataset can be visualized on Figure 3.13, with 9 "groups" (corresponding to terminal leaves). Confusion matrices, associated with various models (logistic regression and random forest), trained on the `GermanCredit` dataset, can be visualized on Table 3.4, with threshold $t = 30\%$ and 50%.

On Table 3.5 different metrics are used on a plain logistic regression and a random forest approach, with the accuracy (with a confidence interval) as well as specificity and sensitivity, computed with thresholds 70%, 50% and 30%. On the left, models with all features (including sensitive ones) are considered Then, from the left to the right, we consider models (plain logistic and random forest) without the gender, without the age, and without both.

ROC curves associated with the plain logistic regression and the random forest model (on the validation dataset) can be visualized on Figure 3.20, with models including all features, and then when possibly sensitive attributes ares not used.

More precisely, it is possible to compare models (here plain logistic and the random forest) not with global metrics, but by plotting $\widehat{m}_{\text{rf}}(\boldsymbol{x}_i, s_i)$ against $\widehat{m}_{\text{glm}}(\boldsymbol{x}_i, s_i)$, on the left of Figure 3.21. On the right, we can compare ranks among predictions. The (outlier) in the right lower corner corresponds to some individual i such that $\widehat{m}_{\text{glm}}(\boldsymbol{x}_i, s_i)$ is almost the 80% lowest (and would be seen as a "large risk", in the worst 20% tail with a plain logistic model) while $\widehat{m}_{\text{rf}}(\boldsymbol{x}_i, s_i)$ is almost the 20% lowest (and would we seen as a "small risk", in the top 20% tail with a random forest model).

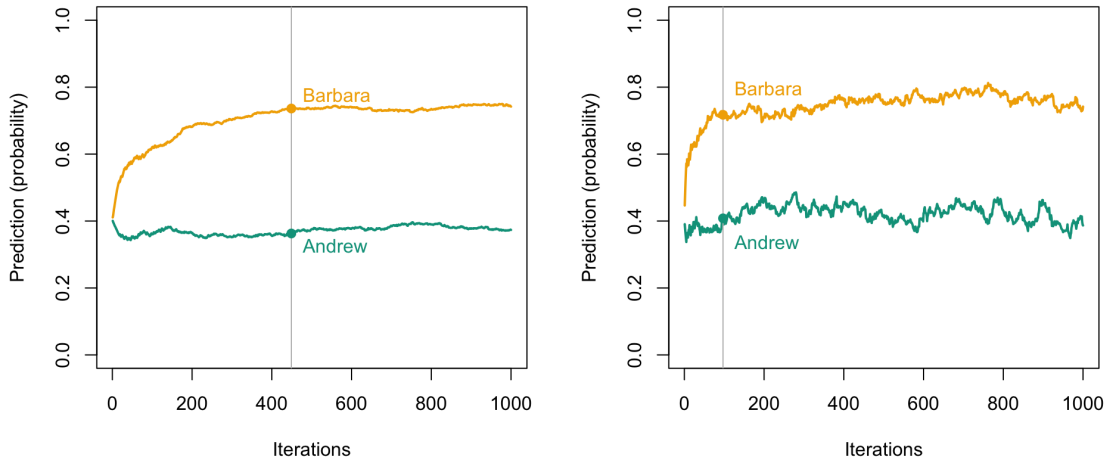


Figure 3.18: Evolution of $m_t(\boldsymbol{x})$ with a boosting learning (`adaboost`) on the `toydata2` dataset, as a function of t (for the models m_t from Figure 3.17), with x_{B} on top and x_{A} below. The difference between graphs on the left and on the right is the learning rate.

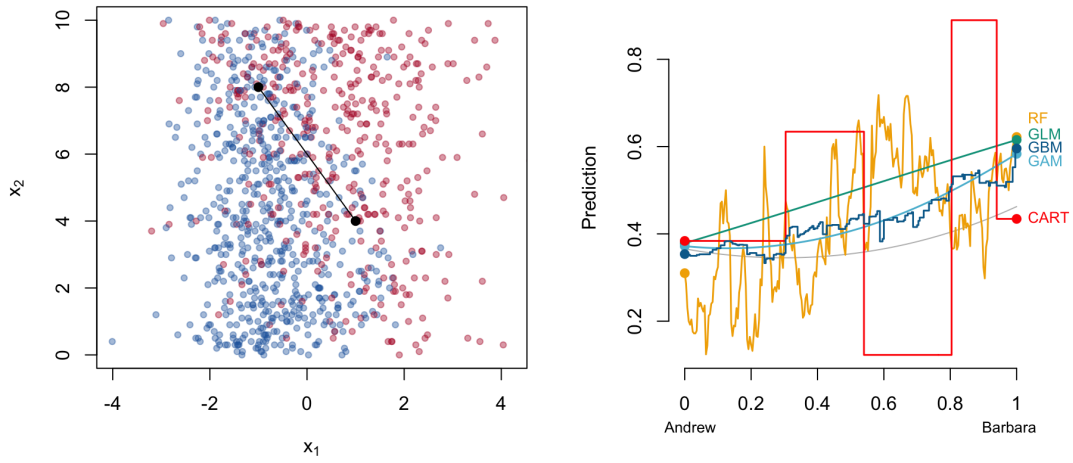


Figure 3.19: On the right, interpolation $t \mapsto \widehat{m}(\boldsymbol{x}_t)$ where $\boldsymbol{x}_t = t\boldsymbol{x}_{\text{B}} + (1-t)\boldsymbol{x}_{\text{A}}$, for $t \in [0,1]$ for five models on the `toydata2` dataset, corresponding to some individual in-between `Andrew` and `Barbara`. $\boldsymbol{x}_t$ correspond to fictitious individuals on the segments $[\boldsymbol{x}_{\text{A}}, \boldsymbol{x}_{\text{B}}]$ on the left, connecting Andrew and Barbara.

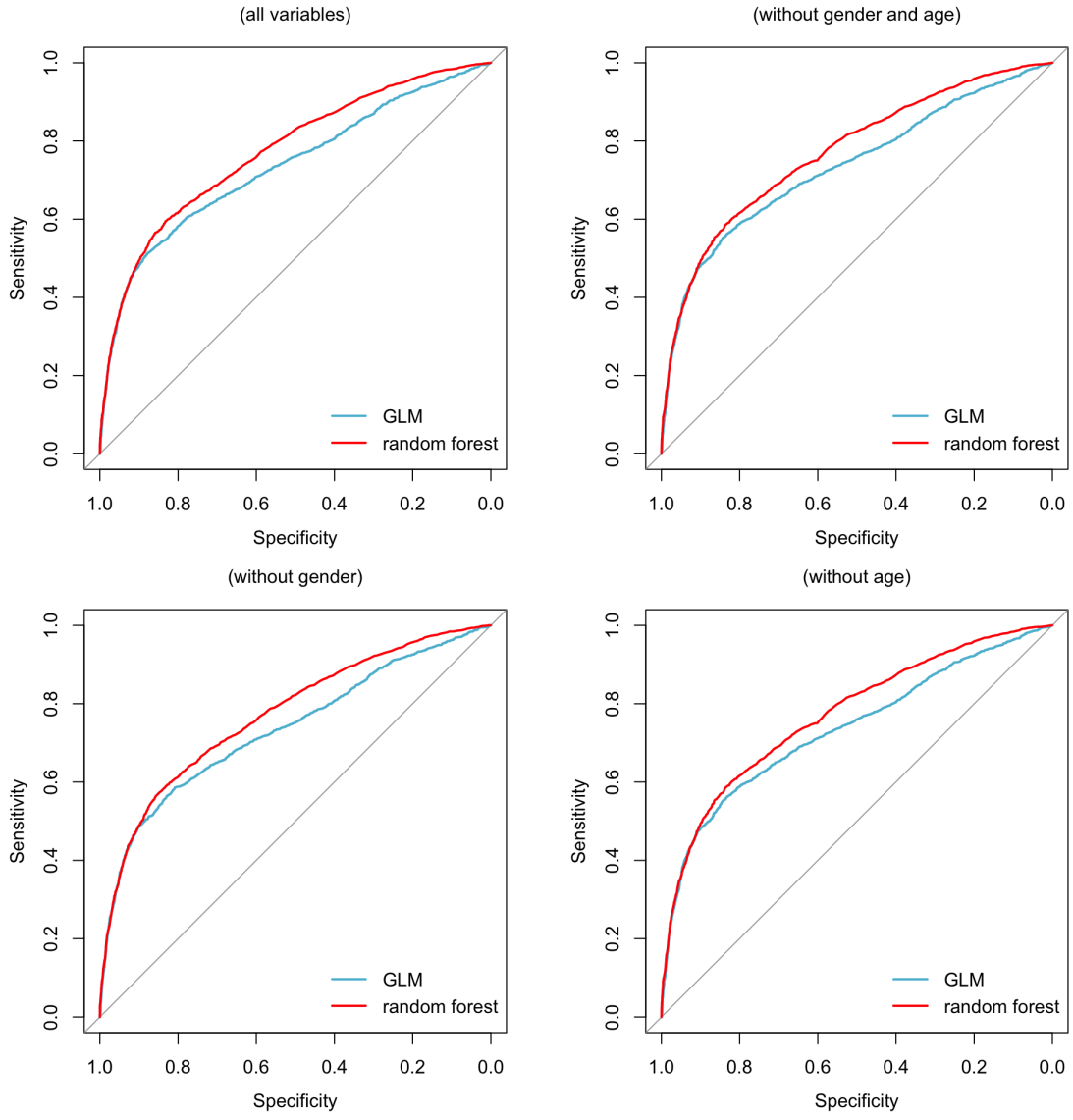


Figure 3.20: ROC curves, for different models, plain logistic regression and random forest on `GermanData`, with all variables $(\boldsymbol{x}, \boldsymbol{s})$ on the top left, and on $\boldsymbol{x}$ on the top right, without gender and age. Below, models are on $(\boldsymbol{x}, s)$ with a single sensitive attribute.

		actual value 0	actual value 1	total
prediction	0	true negative 160	false negative 29	189
prediction	1	false positive 51	true positive 60	111
total		211	89	

		actual value 0	actual value 1	total
prediction	0	true negative 143	false negative 28	171
prediction	1	false positive 68	true positive 61	129
total		211	89	

		actual value 0	actual value 1	total
prediction	0	true negative 185	false negative 46	231
prediction	1	false positive 26	true positive 43	69
total		211	89	

		actual value 0	actual value 1	total
prediction	0	true negative 197	false negative 57	254
prediction	1	false positive 14	true positive 32	146
total		211	89	

Table 3.4: Confusion matrices with threshold 30% (on top) and 50% (below) on 300 observations from the `germandata` dataset, with a logistic regression on the left, and a random forest on the right (without the sensitive attribute, the gender).

3.4 Unsupervised Learning

Supervised learning corresponds to the case we have a target y, we want to model and predict (as explained in the previous section). In the case of unsupervised learning, there is no target, and we only have a collection of variables $\boldsymbol{x}$. The two general problem we want to solve are dimension reduction (where we want the use a smaller number of features) and cluster construction (where we try to regroup individuals together to create groups).

For cluster analysis, see Hartigan (1975), Jain and Dubes (1988) or Gan and Valdez (2020) for theoretical foundations. Campbell (1986) applied cluster analysis to identify groups of car models with similar technical attributes for the purpose of estimating risk premium for individual car models, while Yao (2016) explored territory clustering for ratemaking in motor insurance.

Reducing dimension simply means that, instead of our initial vectors of data, $\boldsymbol{x}_i$, we want to consider a lower dimensional vector $\boldsymbol{x}_i'$. Instead of matrix $\boldsymbol{X}$, we consider $\boldsymbol{X}'$ of lower rank. In the case of a linear transformation, corresponding to principal component analysis (see Jolliffe (2002) or Hastie et al. (2015)) or linear auto-encoder (see Sakurada and Yairi (2014) or Wang et al. (2016)), the error is

$$\|\boldsymbol{X}-\boldsymbol{X}'\|^2=\|\boldsymbol{X}-\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{X}\|^2=\sum_{i=1}^{n}\left(\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{x}_i-\boldsymbol{x}_i\right)^\top\left(\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{x}_i-\boldsymbol{x}_i\right)$$

		all variables		– gender		– age		– gender and age	
		GLM	RF	GLM	RF	GLM	RF	GLM	RF
	AUC	0.793	0.776	0.790	0.783	0.794	0.790	0.794	0.790
$\tau = 70\%$	Accuracy	0.730	0.723	0.737	0.727	0.740	0.737	0.740	0.737
	Accuracy$^-$	0.676	0.669	0.683	0.672	0.686	0.683	0.686	0.683
	Accuracy$^+$	0.779	0.773	0.786	0.776	0.789	0.786	0.789	0.786
	Specificity	0.654	1.000	0.692	1.000	0.704	0.917	0.704	0.917
	Sensitivity	0.737	0.718	0.741	0.720	0.744	0.729	0.744	0.729
$\tau = 50\%$	Accuracy	0.757	0.773	0.760	0.767	0.753	0.787	0.753	0.787
	Accuracy$^-$	0.704	0.722	0.708	0.715	0.701	0.736	0.701	0.736
	Accuracy$^+$	0.804	0.819	0.807	0.813	0.801	0.832	0.801	0.832
	Specificity	0.618	0.756	0.623	0.721	0.609	0.778	0.609	0.778
	Sensitivity	0.797	0.776	0.801	0.774	0.797	0.788	0.797	0.788
$\tau = 30\%$	Accuracy	0.723	0.677	0.733	0.680	0.723	0.690	0.723	0.690
	Accuracy$^-$	0.669	0.621	0.679	0.624	0.669	0.634	0.669	0.634
	Accuracy$^-$	0.773	0.729	0.783	0.732	0.773	0.742	0.773	0.742
	Specificity	0.526	0.469	0.541	0.472	0.526	0.485	0.526	0.485
	Sensitivity	0.844	0.835	0.847	0.832	0.844	0.847	0.844	0.847

Table 3.5: Various statistics for two classifier, a logistic regression and a random forest on the `germandcredit` dataset (accuracy – with a confidence interval (lower and upper bounds) – specificity and sensitivity are computed with a 70%, 50% and 30% threshold), where all variables are considered, on the left. Then, from the left to the right, without the gender, without the age, and without both.

.

More generally, the error function of a nonlinear autoencoder is

$$\|\boldsymbol{X} - \boldsymbol{X}'\|^2 = \|\boldsymbol{X} - \psi \circ \varphi(\boldsymbol{X})\|^2 = \sum_{i=1}^{n} \big(\psi \circ \varphi(\boldsymbol{x}_i) - \boldsymbol{x}_i\big)^\top \big(\psi \circ \varphi(\boldsymbol{x}_i) - \boldsymbol{x}_i\big).$$

On Figure 3.22, we can visualize the general architecture of nonlinear autoencoder on top, and linear autoencoder below (corresponding to PCA, principal component analysis).

The error function of a linear auto-encoder is

$$\sum_{i=1}^{n} \text{trace}\left[\big(\boldsymbol{P}^\top\boldsymbol{P} - \mathbb{I}\big)\boldsymbol{x}_i\boldsymbol{x}_i^\top\big(\boldsymbol{P}^\top\boldsymbol{P} - \mathbb{I}\big)\right] = \text{trace}\left[\big(\boldsymbol{P}^\top\boldsymbol{P} - \mathbb{I}\big)\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\big(\boldsymbol{P}^\top\boldsymbol{P} - \mathbb{I}\big)\right],$$

the middle term is a covariance matrix, thus it is $\boldsymbol{V}\boldsymbol{\Delta}\boldsymbol{V}^\top$, and we recognize

$$\|\big(\boldsymbol{P}^\top\boldsymbol{P} - \mathbb{I}\big)\boldsymbol{V}\boldsymbol{\Delta}^{1/2}\|_F^2,$$

where $\|\cdot\|_F$ denotes the Froebenius norm, corresponding to the elementwise ℓ_2 norm of a matrix,

$$\|\boldsymbol{M}\|_F^2 = \sum_{i,j} M_{i,j}^2 = \text{trace}\big(\boldsymbol{M}\boldsymbol{M}^\top\big) \text{ where } \boldsymbol{M} = \big[M_{i,j}\big].$$

Let $\|\cdot\|_F$ denote the Froebenius / elementwise ℓ_2 norm of a matrix,

$$\|\boldsymbol{M}\|_F^2 = \sum_{i,j} M_{i,j}^2 \text{ where } \boldsymbol{M} = \big[M_{i,j}\big].$$

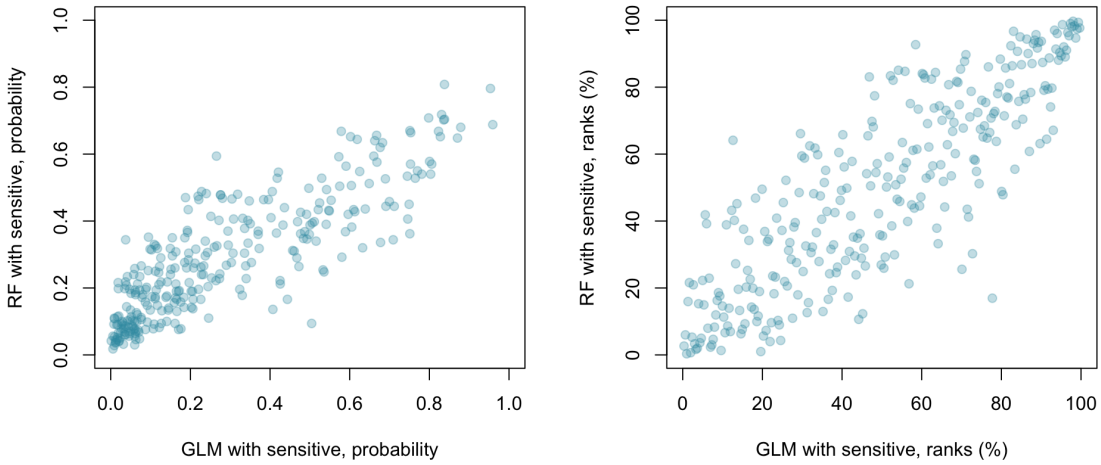


Figure 3.21: Scatterplot of $(\widehat{m}_{\text{glm}}(\boldsymbol{x}_i), \widehat{m}_{\text{rf}}(\boldsymbol{x}_i))$ on the left, on the `germancredit` dataset, and the associated ranks on the right (with a linear transformation to have "ranks" on 0-100, correponding to the empirical copula).

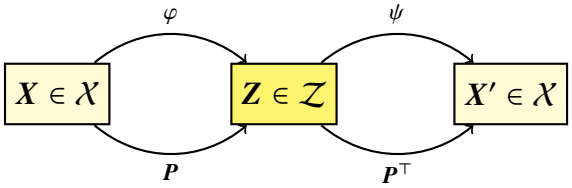


Figure 3.22: Nonlinear autoencoder on top, and linear autoencoder below (PCA, principal component analysis). $\boldsymbol{X}$ is the original dataset, $\boldsymbol{Z}$ the embedded version in a smaller latent space (principal factors) and $\boldsymbol{X}'$ is the reconstruction.

Consider the following problem,

$$\min_{\boldsymbol{Y}} \left\{ \|\boldsymbol{X} - \boldsymbol{Y}\|_F^2 \right\} \text{ subject to } \text{rank}(\boldsymbol{Y}) = k \quad (\leq \text{rank}(\boldsymbol{X})).$$

If $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{V}^\top$ then $\boldsymbol{Y} = \boldsymbol{U}_k\boldsymbol{\Delta}_k\boldsymbol{V}_k^\top$ where we keep the first k columns of $\boldsymbol{U}$, $\boldsymbol{V}$ and $\boldsymbol{\Delta}$.

One can rewrite

$$\min_{\boldsymbol{P} \in \Pi} \left\{ \|\boldsymbol{X} - \boldsymbol{P}\boldsymbol{X}\|_F^2 \right\} \text{ subject to } \text{rank}(\boldsymbol{P}) = k,$$

where Π is the set of projection matrices.

If $\boldsymbol{S} = \boldsymbol{X}^\top\boldsymbol{X}$, we can write (equivalently)

$$\max_{\boldsymbol{P} \in \Pi} \left\{ \text{trace}(\boldsymbol{S}\boldsymbol{P}) \right\} \text{ subject to } \text{rank}(\boldsymbol{P}) = k$$

or $\max_{\boldsymbol{P}\in\mathcal{P}}\left\{\text{trace}(\boldsymbol{SP})\right\}$ where $\mathcal{P}=\left\{\boldsymbol{P}\in\Pi:\text{eigenvalues}(\boldsymbol{P})\in\{0,1\}\text{ and trace}(\boldsymbol{P})=k\right\}$.

As explained in Samadi et al. (2018), if $\|\boldsymbol{X}-\boldsymbol{X}'\|_F^2$ can be seen as the reconstruction error of $\boldsymbol{X}$ with respect to $\boldsymbol{X}'$, the natural way to define a rank k-reconstruction loss, consider

$$\ell_k(\boldsymbol{X},\boldsymbol{X}')=\|\boldsymbol{X}-\boldsymbol{X}'\|_F^2-\|\boldsymbol{X}-\boldsymbol{X}^\star\|_F^2 \text{ where } \boldsymbol{X}^\star=\underset{\boldsymbol{Y}}{\text{argmin}}\left\{\|\boldsymbol{X}-\boldsymbol{Y}\|_F^2\right\} \text{ subject to } \text{rank}(\boldsymbol{Y})=k.$$

Chapter 4

Models: Interpretability, Accuracy and Calibration

In this chapter, we will present important concepts when dealing with predictive models. We will start with a discussion about the interpretability and explainability of models and algorithms, presenting different tools that could help understanding “why" the predicted outcome of the model is the one we got. Then, we will discuss accuracy, which is usually the ultimate target of most machine learning technique. But as we will see, the most important concept is the “well-calibration" of the model, which means that we want to have, locally, a balanced portfolio, and that the probability predicted by the model is, indeed, related to the true risk.

In a popular book on the philosophy of science, Nancy Cartwright, examines the relationship between theoretical models, experiments, explanations and various concepts linked to causal relations. And she tells the following anecdote, “*My newly planted lemon tree is sick, the leaves yellow and dropping off. I finally explain this by saying that water has accumulated in the base of the planter: the water is the cause of the disease. I drill a hole in the base of the oak barrel where the lemon tree lives, and foul water flows out. That was the cause. Before I had drilled the hole, I could still give the explanation and to give that explanation was to present the supposed cause, the water. There must be such water for the explanation to be correct. An explanation of an effect by a cause has an existential component, not just an optional extra ingredient*," Cartwright (1983).

4.1 Interpretability and Explainability

Interpretability is about transparency, about understanding exactly why and how the model is generating predictions, and therefore, it is important to observe the inner mechanics of the algorithm considered. This leads to interpreting the model’s parameters and features used to determine the given output. Explainability is about explaining the behavior of the model in human terms.

Definition 4.1.1 (Ceteris paribus) *Marshall (1890)* Ceteris paribus *(or more precisely* ceteris paribus sic stantibus*) is a Latin phrase, meaning “all other things being equal" or “other things held constant".*

The *ceteris paribus* approach is commonly used to consider the effects of a cause, in isolation, by assuming that any other relevant conditions are absent. On Figure 4.1, the output of a model, $\widehat{y}$ can be influenced by x_1 and x_2, and in the *ceteris paribus* analysis of the influence of x_1 on $\widehat{y}$, we isolate the effect of x_1 on $\widehat{y}$. In the *mutatis mutandis* approach, if x_1 and x_2 are correlated, we add to the "direct effect" (from x_1 to $\widehat{y}$) a possible "indirect effect" (through x_2.)

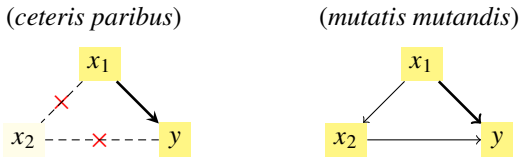


Figure 4.1: On the left, the *ceteris paribus* approach (only the direct relationship from x_1 to y is considered, and x_2 is supposed to remain unchanged) and the *mutatis mutandis* approach (a change in x_1 will have a direct impact on y, and there could be an additional effect via x_2).

Definition 4.1.2 (Mutatis mutandis) Mutatis mutandis *is a Latin phrase meaning "with things changed that should be changed" or "once the necessary changes have been made".*

In order to illustrate, let $(X_1, X_2, \varepsilon)^\top$ denote some Gaussian random vector, where the first two components are correlated, and ε is some unpredictable random noise, independent of the pair $(X_1, X_2)^\top$

$$\begin{pmatrix} X_1 \\ X_2 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & r\sigma_1\sigma_2 & 0 \\ r\sigma_1\sigma_2 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix} \right)$$

Suppose that $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ (as in a standard linear model), then for some $\boldsymbol{x}^* = (x_1^*, x_2^*)$,

$$\mathbb{E}_{Y|\boldsymbol{X}}[Y|\boldsymbol{x}^*] = \mathbb{E}_{\boldsymbol{X}}[Y|x_1^*, x_2^*] = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^*,$$

while $\mathbb{E}_Y[Y] = \beta_0 + \beta_1\mu_1 + \beta_2\mu_2$. Then, on the one hand, if we compute the standard conditional expected value of X_2, conditional on X_1, we have

$$\mathbb{E}_{X_2|X_1}[X_2|x_1^*] = \mu_2 + \frac{r\sigma_2}{\sigma_1}(x_1^* - \mu_1),$$

and therefore

$$\mathbb{E}_{Y|X_1}[Y|x_1^*] = \beta_0 + \beta_1 x_1^* + \beta_2 \left(\mu_2 + \frac{r\sigma_2}{\sigma_1}(x_1^* - \mu_1) \right) \quad : \textit{mutatis mutandis}.$$

On the other hand, in the *ceteris paribus* approach, "isolating" the effect of x_1 to other possible causes means that we pretend that X_1 and X_2 are now independent. Therefore, formally, instead of (X_1, X_2), we consider $(X_1^\perp, X_2^\perp)$ a "copy" with independent components and the same marginal distributions[1], then $\mathbb{E}_{Y|X_2^\perp|X_1^\perp}[Y|x_1^*] = \mu_2$, and

$$\mathbb{E}_{Y|X_1^\perp}[Y|x_1^*] = \beta_0 + \beta_1 x_1^* + \beta_2\mu_2 \quad : \textit{ceteris paribus}$$

[1] in the sense that $X_2^\perp = X_2$, almost surely, and $X_1^\perp \overset{\mathcal{L}}{=} X_1$, and $X_1^\perp \perp\!\!\!\perp X_2$.

Therefore, we have clearly the direct effect (*ceteris paribus*), and the indirect effect,

$$\underbrace{\mathbb{E}_{Y|X_1}[Y|x_1^*]}_{\textit{mutatis mutandis}} = \underbrace{\mathbb{E}_{Y|X_1^\perp}[Y|x_1^*]}_{\textit{ceteris paribus}} + \beta_2 \frac{r\sigma_2}{\sigma_1}(x_1^* - \mu_1).$$

As expected, if variables x_1 and x_2 are independent, $r = 0$, and the *mutatis mutandis* and the *ceteris paribus* approaches are identical. Later on, when presenting various techniques in this chapter, we might use notation $\mathbb{E}_{X_1}$ and $\mathbb{E}_{X_1^\perp}$, instead of $\mathbb{E}_{Y|X_1}$ or $\mathbb{E}_{Y|X_1^\perp}$, respectively, to avoid too heavy notations.

And more generally, from a statistical perspective, if we consider a non-linear model $\mathbb{E}_{Y|\boldsymbol{X}}[Y|\boldsymbol{x}^*] = \mathbb{E}_{\boldsymbol{X}}[Y|x_1^*, x_2^*] = m(x_1^*, x_2^*)$, a natural estimate *ceteris paribus* of the effect of x_1 on the prediction is

$$\mathbb{E}_{Y|X_1^\perp}[m(X_1^\perp, X_2^\perp)|x_1^*] \text{ is } \frac{1}{n}\sum_{i=1}^{n} m(x_1^*, x_{i,2})$$

while to estimate *mutatis mutandis*, we need a local version, to take into account a possible (local) correlation between x_1 and x_2, i.e.

$$\mathbb{E}_{Y|X_1}[m(X_1, X_2)|x_1^*] \approx \frac{1}{\|\mathcal{V}_\epsilon(x_1^*)\|}\sum_{i\in\mathcal{V}_\epsilon(x_1^*)} m(x_1^*, x_{i,2}),$$

where $\mathcal{V}_\epsilon(x_1^*) = \left\{i : |x_{i,1} - x_1^*| \leq \epsilon\right\}$ is a neighborhood of x_1^*. It should be stressed that notations "$\mathbb{E}_{Y|X_1}[m(X_1, X_2)|x_1^*]$" and "$\mathbb{E}_{Y|X_1^\perp}[m(X_1^\perp, X_2^\perp)|x_1^*]$" have yet not measure theory foundations, but they will be useful to highlight that in some cases, metrics and mathematical objects "pretend" that explanatory variables are independent.

4.1.1 Variable importance

When introducing random forests, Breiman (2001) suggested a simple technique to rank the importance of variables, in a natural way. This technique has been improved, in Helton and Davis (2002), Azen and Budescu (2003), Rifkin and Klautau (2004) and Saltelli et al. (2008), in the context of classification and regression trees, and random forests. The general definition, for other models, could be the following,

Definition 4.1.3 (VI_j or "permutation VI_j") *Fisher et al. (2019). Given a loss function ℓ and a model m, the importance of the j-th variable is*

$$VI_j = \mathbb{E}\big[\ell(Y, m(\boldsymbol{X}_{-j}, X_j))\big] - \mathbb{E}\big[\ell(Y, m(\boldsymbol{X}_{-j}, X_j^\perp))\big]$$

and the empirical version is

$$\widehat{VI}_j = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, m(\boldsymbol{x}_{i,-j}, x_{i,j})) - \ell(y_i, m(\boldsymbol{x}_{i,-j}, \tilde{x}_{i,j}))$$

for some permutation $\tilde{\boldsymbol{x}}_j$ or $\boldsymbol{x}_j$.

On the `todydata2` dataset, with three explanatory variables (x_1, x_2 and x_3) and a sensisitive attribute (s), $\widehat{\text{VI}}_j$ can be computed using function `variable_importance` in the `DALEX` package (see Biecek and Burzykowski (2021) for more details). By default, the loss considered is the one associated with 1 − AUC for

classification (`loss_one_minus_auc`, as here), but cross entropy can be used for multilabel classification, while RMSE is the default loss for regression. On Figures 4.2 and 4.3 we can visualize variable importance for the four models (including some confidence band), respectively for model without and with the sensitive attribute s. This measure can be quantified a some "drop-out loss of AUC" as a measure of variable importance. One could also use `FeatureImp` in the `iml` R package, based on Molnar (2023). Observe on those Figures that all models are comparable, here.

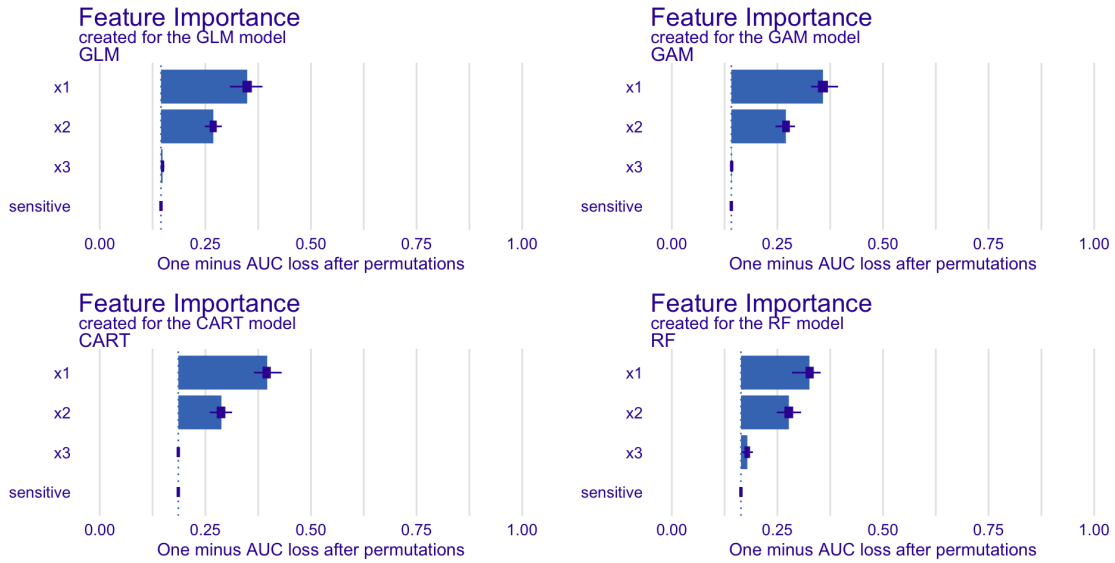


Figure 4.2: Variable importance for different models trained on `toydata2`, without the sensitive attribute s, with four variables, x_1, x_2, x_3 and s.

		full	x_1	x_2	x_3	s	base
loss (1-AUC)							
glm	$m(\boldsymbol{x})$	0.145	0.347	0.268	0.150	0.147	0.502
gam	$m(\boldsymbol{x})$	0.141	0.356	0.269	0.143	0.143	0.499
cart	$m(\boldsymbol{x})$	0.186	0.394	0.287	0.187	0.188	0.501
rf	$m(\boldsymbol{x})$	0.164	0.324	0.276	0.180	0.165	0.502
loss root mean square							
glm	$m(\boldsymbol{x})$	0.385	0.502	0.458	0.388	0.386	0.573
gam	$m(\boldsymbol{x})$	0.381	0.516	0.454	0.383	0.383	0.579
cart	$m(\boldsymbol{x})$	0.399	0.525	0.458	0.400	0.401	0.569
rf	$m(\boldsymbol{x})$	0.397	0.496	0.466	0.406	0.398	0.586

		full	x_1	x_2	x_3	s	b
loss (1-AUC)							
$m(\boldsymbol{x}, s)$		0.140	0.311	0.262	0.141	0.159	0
$m(\boldsymbol{x}, s)$		0.138	0.324	0.262	0.139	0.153	0
$m(\boldsymbol{x}, s)$		0.186	0.395	0.286	0.187	0.188	0
$m(\boldsymbol{x}, s)$		0.164	0.297	0.275	0.171	0.191	0
loss root mean square							
$m(\boldsymbol{x}, s)$		0.382	0.484	0.454	0.382	0.394	0
$m(\boldsymbol{x}, s)$		0.379	0.498	0.451	0.380	0.389	0
$m(\boldsymbol{x}, s)$		0.399	0.525	0.457	0.400	0.401	0
$m(\boldsymbol{x}, s)$		0.398	0.475	0.463	0.401	0.415	0

Table 4.1: Variable importance for the four models on `toydata2`, with and without the sensitive attribute s, respectively on the right, and on the left.

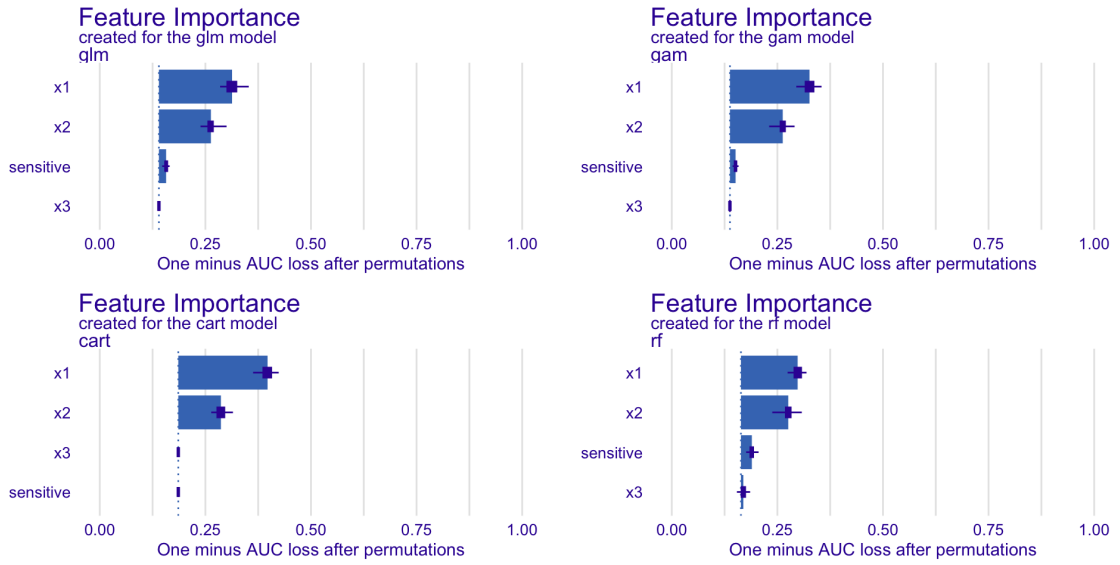


Figure 4.3: Variable importance for different models trained on `toydata2`, with the sensitive attribute s, with four variables, x_1, x_2, x_3 and s.

4.1.2 Ceteris Paribus Profiles

Instead of a global measure, some local metrics can be considered. Goldstein et al. (2015) defined the "individual conditional expectation" directly derived from *ceteris paribus* functions, and coined "*ceteris-paribus profile*" in Biecek and Burzykowski (2021)

Definition 4.1.4 (Ceteris Paribus profile $z \mapsto m_{\boldsymbol{x}^*,j}(z)$**)** *Goldstein et al. (2015). Given* $\boldsymbol{x}^* \in \mathcal{X}$*, define on* $\mathcal{X}_j$

$$z \mapsto m_{\boldsymbol{x}^*,j}(z) = m(\boldsymbol{x}^*_{-j}, z) = m(x_1^*, \cdots, x_{j-1}^*, z, x_{j+1}^*, \cdots, x_p^*).$$

Define then the difference when component j takes generic value z and x_j^*,

$$\delta m_{\boldsymbol{x}^*,j}(z) = m_{\boldsymbol{x}^*,j}(z) - m_{\boldsymbol{x}^*,j}(x_j^*).$$

Definition 4.1.5 $(dm_j^{\mathsf{cp}}(\boldsymbol{x}^*))$ *The mean absolute deviation associated with the* j*-th variable, at* $\boldsymbol{x}^*$*, is* $dm_j(\boldsymbol{x}^*)$*,*

$$dm_j(\boldsymbol{x}^*) = \mathbb{E}\big[|\delta m_{\boldsymbol{x}^*,j}(X_j)|\big] = \mathbb{E}\big[|m(\boldsymbol{x}^*_{-j}, X_j) - m(\boldsymbol{x}^*_{-j}, x_j^*)|\big]$$

Definition 4.1.6 $(\widehat{dm}_j^{\mathsf{cp}}(\boldsymbol{x}^*))$ *The empirical mean absolute deviation associated with the* j*-th variable, at* $\boldsymbol{x}^*$*, is*

$$\widehat{dm}_j(\boldsymbol{x}^*) = \frac{1}{n}\sum_{i=1}^{n} |m(\boldsymbol{x}^*_{-j}, x_{i,j}) - m(\boldsymbol{x}^*_{-j}, x_j^*)|.$$

On Figures 4.4 and 4.5, we can visualize "*ceteris-paribus profiles*" on our four models, on `toyxdata2`, with $j = 1$ (variable x_1) with the plain logistic regression, the GAM, the classification tree, and the random

forest, $z \mapsto m_{\boldsymbol{x}^\star,1}(z)$. On Figure 4.4, it is $z \mapsto m_{\boldsymbol{x}^\star,1}(z)$ associated with `Andrew` (when $(\boldsymbol{x}^\star, s^\star) = (-1, 8, -2, \texttt{A})$) and on Figure 4.5, it is $z \mapsto m_{\boldsymbol{x}^\star,1}(z)$ associated with `Barbara` (when $(\boldsymbol{x}^\star, s^\star) = (1, 4, 2, \texttt{B})$). Bullet points indicate the values $m_{\boldsymbol{x}^\star,1}(x_1^\star)$ for `Andrew` on Figure 4.4, and `Barbara` on Figure 4.5. On top left, function is monotonic, with a "logistic" shape. On the right, we see that a GLM will probably miss a non linear effect, with a (caped) J shape.

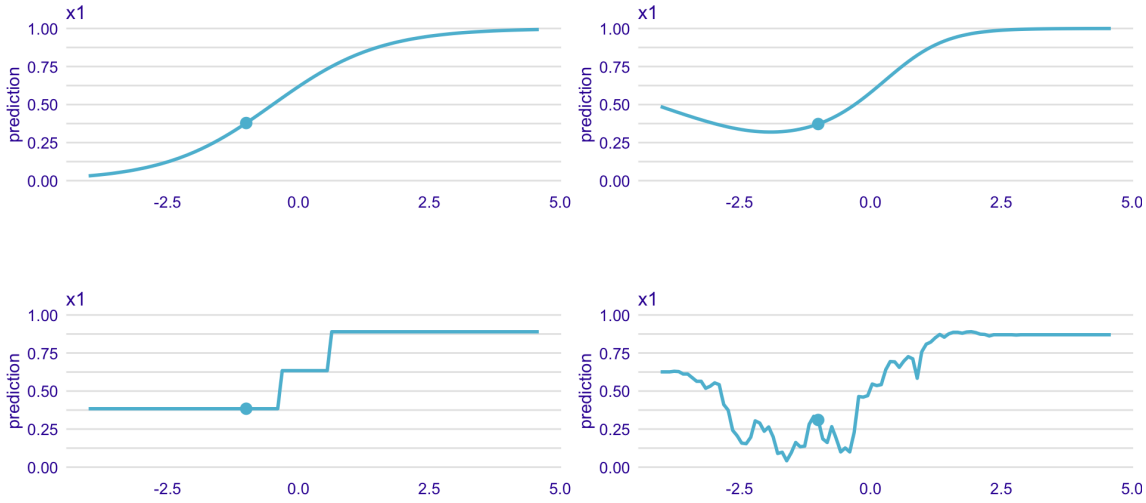


Figure 4.4: "*ceteris-paribus profiles*" for `Andrew` for different models trained on `toydata2` (see Table 3.3 for numerical values, for variable x_1, here $\boldsymbol{z}^\star = (\boldsymbol{x}^\star, s^\star) = (-1, 8, -2, \texttt{A})$).

4.1.3 Breakdowns

For a standard linear model, observe that

$$\widehat{m}(\boldsymbol{x}^*) = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}^* = \widehat{\beta}_0 + \sum_{j=1}^{k} \widehat{\beta}_j x_j^* = \overline{y} + \sum_{j=1}^{k} \underbrace{\widehat{\beta}_j \left(x_j^* - \overline{x}_j\right)}_{=v_j(\boldsymbol{x}^*)}$$

where $v_j(\boldsymbol{x}^*)$ is interpreted as the contribution of the j-th variable on the prediction for individual with characteristics $\boldsymbol{x}^*$. More generally, Robnik-Šikonja and Kononenko (1997, 2003, 2008) defined the (additive) contribution of the j-th variable on the prediction for individual with characteristics $\boldsymbol{x}^*$

$$v_j(\boldsymbol{x}^*) = m(x_1^*, \cdots, , x_{j-1}^*, x_j^*, x_{j+1}^*, \cdots, x_k^*) - \mathbb{E}_{X_j^\perp}[m(x_1^*, \cdots, x_{j-1}^*, X_j, x_{j+1}^*, \cdots, x_k^*)],$$

so that

$$m(\boldsymbol{x}^*) = \mathbb{E}\big[m(\boldsymbol{X})\big] + \sum_{j=1}^{k} v_j(\boldsymbol{x}^*),$$

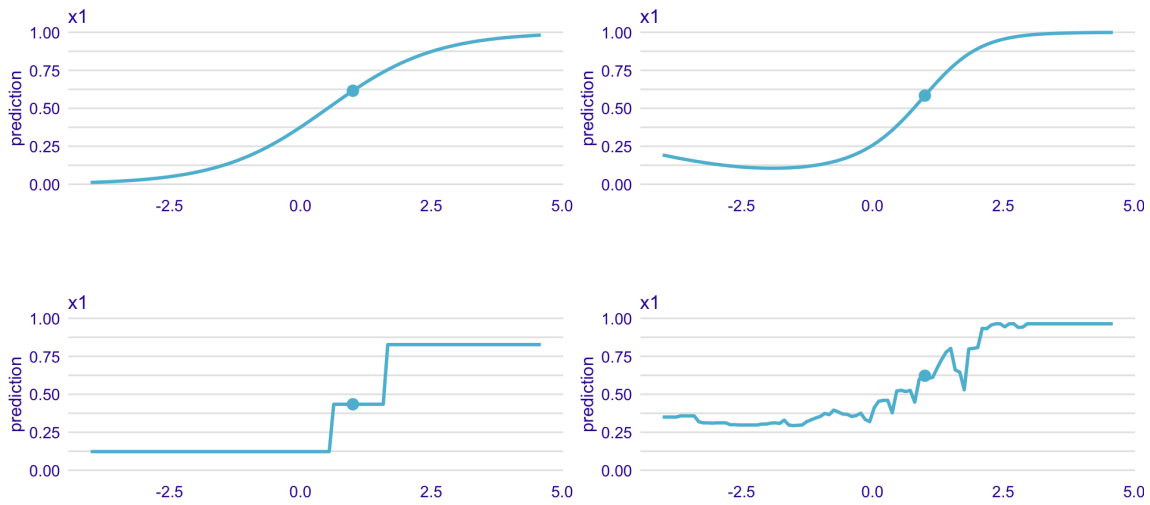


Figure 4.5: "*ceteris-paribus profiles*" for `Barbara` for different models trained on `toydata2` (see Table 3.3 for numerical values, for variable x_1, here $\boldsymbol{z}^\star = (\boldsymbol{x}^\star, s^\star) = (1, 4, 2, \mathtt{B})$).

and for the linear model $v_j(\boldsymbol{x}^*) = \beta_j\left(x_j^* - \mathbb{E}_{X_j^\perp|\boldsymbol{X}_{-j}}[X_j^\perp|\boldsymbol{X}_{-j} = \boldsymbol{x}_{-j}^*]\right)$, and $\widehat{v}_j(\boldsymbol{x}^*) = \widehat{\beta}_j\left(x_j^* - \overline{x}_j\right)$.

More generally, $v_j(\boldsymbol{x}^*) = m(\boldsymbol{x}^*) - \mathbb{E}_{X_j^\perp|\boldsymbol{X}_{-j}}[m(\boldsymbol{x}_{-j}^*, X_j))]$, where we can write $m(\boldsymbol{x}^*)$ as $\mathbb{E}[m(\boldsymbol{x}^*)]$, i.e.

$$v_j(\boldsymbol{x}^*) = \begin{cases} \mathbb{E}\left[m(\boldsymbol{X})\middle|x_1^*, \cdots, x_k^*\right] - \mathbb{E}_{X_j^\perp|\boldsymbol{X}_{-j}}\left[m(\boldsymbol{X})\middle|x_1^*, \cdots, x_{j-1}^*, x_{j+1}^*, \cdots, x_k^*\right] \\ \mathbb{E}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}^*\right] - \mathbb{E}_{X_j^\perp|\boldsymbol{X}_{-j}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}_{-j}^*\right]. \end{cases}$$

Definition 4.1.7 ($\gamma_j^{\mathsf{bd}}(\boldsymbol{x}^*)$) *Biecek and Burzykowski (2021). The breakdown contribution of the j-th variable, at $\boldsymbol{x}^*$, is*

$$\gamma_j^{bd}(\boldsymbol{x}^*) = v_j(\boldsymbol{x}^*) = \mathbb{E}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}^*\right] - \mathbb{E}_{X_j^\perp|\boldsymbol{X}_{-j}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}_{-j}^*\right].$$

"*In other words, the contribution of the j-th variable is the difference between the expected value of the model's prediction conditional on setting the values of the first j variables equal to their values in $\boldsymbol{x}^*$ and the expected value conditional on setting the values of the first $j1$ variables equal to their values in $\boldsymbol{x}^*$*," as said Biecek and Burzykowski (2021).

We can rewrite the contribution of the j-th variable, at $\boldsymbol{x}^*$,

$$v_j(\boldsymbol{x}^*) = \begin{cases} \mathbb{E}\left[m(\boldsymbol{X})\middle|x_1^*, \cdots, x_k^*\right] - \mathbb{E}_{X_j^\perp|\boldsymbol{X}_{-j}}\left[m(\boldsymbol{X})\middle|x_1^*, \cdots, x_{j-1}^*, x_{j+1}^*, \cdots, x_k^*\right] \\ \mathbb{E}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}^*\right] - \mathbb{E}_{X_j^\perp|\boldsymbol{X}_{-j}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}_{-j}^*\right]. \end{cases}$$

Definition 4.1.8 ($\Delta_{j|S}(\boldsymbol{x}^*)$) *The contribution of the j-th variable, at $\boldsymbol{x}^*$, conditional on a subset of variables, $S \subset \{1, \cdots, k\}\backslash\{j\}$, is*

$$\Delta_{j|S}(\boldsymbol{x}^*) = \mathbb{E}_{\boldsymbol{X}_{S\cup\{j\}}^\perp}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}_{S\cup\{j\}}^*\right] - \mathbb{E}_{\boldsymbol{X}_S^\perp}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}_S^*\right]$$

so that $v_j(\boldsymbol{x}^*) = \Delta_{j|\{1,2,\cdots,k\}\backslash\{j\}} = \Delta_{j|-j}$.

On the `toydata2` dataset, we can compute contributions of x_1, x_2 and x_3 for two individuals, `Andrew` and `Barbara`, as on Figures 4.6 and 4.7, respectively, using `type = "break_down"` in the `predict_parts` function of the `DALEX` R package. For `Andrew` (on Figure 4.6), the starting point is the average value on the entire population (close to 40%). The large value of x_2 (here 8) yield about +0.18 on the prediction, while the negative value of x_1 (here -1) yield about from -0.19 to -0.14 on the prediction. Here s has no impact, since we consider models trained without the sensitive attribute.

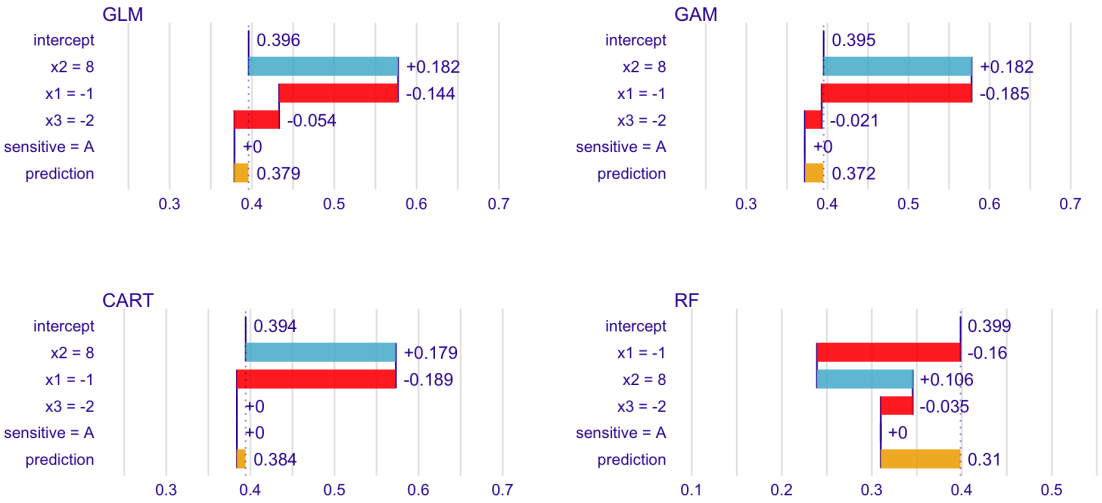


Figure 4.6: Breakdown decomposition $\widehat{\gamma}_j^{\text{bd}}(z_{\texttt{A}}^*)$ for `Andrew` for different models trained on `toydata2` (see Table 3.3 for numerical values, here $z_{\texttt{A}}^\star = (\boldsymbol{x}_{\texttt{A}}^\star, s^\star) = (-1, 8, -2, \texttt{A})$).

4.1.4 Shapley Value and Shapley Contributions

In order to get a robust way to define contributions, in the context of predictive modeling, Lipovetsky and Conklin (2001) suggested to use Shapley value in statistics, to decompose the R^2 of a linear regression into additive contributions of each single covariate. Then Štrumbelj and Kononenko (2010, 2014) suggested to use Shapley values to decompose predictions into feature contribution, and more recently, Lundberg and Lee (2017) provided a unified version.

Recall that the "Shapley value", as defined in Shapley (1953), is based on coalitional game, with k players, and a "value function" (also named "characteristic function") $\mathcal{V}$ that can be defined on any coalition of players, $S \subset \{1, 2, \cdots, k\}$. Given a coalition $S \subset \{1, 2, \cdots, k\}$ of players, then $\mathcal{V}(S)$ corresponds to the "worth of coalition S", that should reflect payoffs the members of S would obtain from this cooperation. In the context of games, assuming that all players collaborate, the Shapley value is one way (among many others) to distribute the total gains among all players. In game theory literature (starting with Shapley and Shubik (1969) but then emphizised by Moulin (1992) and Moulin (2004)), it can be referred as a "fair" mechanism, in the sense that it is the only distribution with certain desirable properties. The Shapley value describes contribution to the payout, weighted and summed over all possible feature value combinations, as

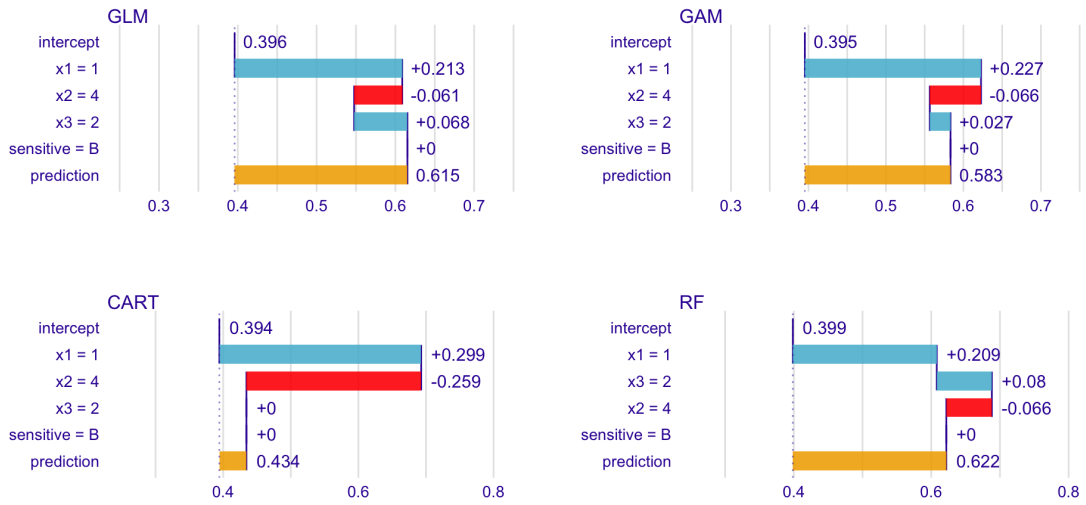


Figure 4.7: Breakdown decomposition $\widehat{\gamma}_j^{\text{bd}}(z_{\text{B}}^*)$ for `Barbara` for different models trained on `toydata2` (see Table 3.3 for numerical values, here $z^\star = (\boldsymbol{x}^\star, s^\star) = (1, 4, 2, \text{B})$).

follows

$$\phi_j(\mathcal{V}) = \frac{1}{k} \sum_{S \subseteq \{1,\ldots,k\} \setminus \{j\}} \frac{|S|!\,(k - |S| - 1)!}{k!} \left(\mathcal{V}\left(S \cup \{j\}\right) - \mathcal{V}(S)\right),$$

As explained in Ichiishi (2014), if we suppose that coalitions are being formed one player at a time, at step j, it should be fair for player j to be given $\mathcal{V}\left(S \cup \{j\}\right) - \mathcal{V}(S)$ as a fair compensation for joining the coalition. And then for each actor, to take the average of this contribution over all possible different permutations in which the coalition can be formed. Which is exactly the expression above,

$$\phi_j(\mathcal{V}) = \frac{1}{\text{number of players}} \sum_{\text{coalitions including } j} \frac{\text{marginal contribution of } j \text{ to coalition}}{\text{number of coalitions excluding } j}.$$

The goal, in Shapley (1953), was to find contributions $\phi_j(\mathcal{V})$, for some value function $\mathcal{V}$, that satisfies a series of desirable properties, namely

- "efficiency": $\sum_{j=1}^{k} \phi_j(\mathcal{V}) = \mathcal{V}(\{1, \ldots, k\})$,
- "symmetry": if $\mathcal{V}\left(S \cup \{j\}\right) = \mathcal{V}\left(S \cup \{j'\}\right)$ $\forall S$, then $\phi_j = \phi_{j'}$,
- "dummy" (or "null player"): if $\mathcal{V}\left(S \cup \{j\}\right) = \mathcal{V}\left(S\right)$ $\forall S$, then $\phi_j = 0$,
- "additivity": if $\mathcal{V}^{(1)}$ and $\mathcal{V}^{(2)}$ have decomposition $\boldsymbol{\phi}(\mathcal{V}^{(1)})$ and $\boldsymbol{\phi}(\mathcal{V}^{(2)})$, then $\mathcal{V}^{(1)} + \mathcal{V}^{(2)}$ has decomposition $\boldsymbol{\phi}(\mathcal{V}^{(1)} + \mathcal{V}^{(2)}) = \boldsymbol{\phi}(\mathcal{V}^{(1)}) + \boldsymbol{\phi}(\mathcal{V}^{(2)})$. "Linearity" will be obtained if we add $\boldsymbol{\phi}(\lambda \cdot \mathcal{V}) = \lambda \cdot \boldsymbol{\phi}(\mathcal{V})$.

In the context of predictive models, S denotes some subset of features used in the model ($S \subset \{1, 2, \cdots, k\}$), $\boldsymbol{x}$ is some vector of features. Here, it could be natural to suppose that $\mathcal{V}_{\boldsymbol{x}}$ denotes the prediction for feature values in set S that are marginalized, over features that are not included in set S. Štrumbelj and Kononenko (2014) suggested Monte Carlo sampling to compute contributions, while Lundberg and Lee (2017)introduced a Kernel Shapley value.

Here, we will use $\mathcal{V}_{\boldsymbol{x}^*}(S) = \mathbb{E}_{\boldsymbol{X}_{\bar{S}}}\big[m(\boldsymbol{X})\big|\boldsymbol{x}_S^*\big]$, as value function, for any set S of variables, so that $\Delta_{j|S}(\boldsymbol{x}^*) = \mathcal{V}_{\boldsymbol{x}^*}(S \cup \{j\}) - \mathcal{V}_{\boldsymbol{x}^*}(S)$, from Definition 4.1.8,

Definition 4.1.9 (Shapley contributions $\gamma_j^{\textsf{shap}}(\boldsymbol{x}^*)$) *The Shapley contribution of the j-th variable, at $\boldsymbol{x}^*$, is*

$$\gamma_j^{shap}(\boldsymbol{x}^*) = \frac{1}{k}\sum_{S\subseteq\{1,\ldots,k\}\setminus\{j\}}\binom{k-1}{|S|}^{-1}\Delta_{j|S}(\boldsymbol{x}^*) = \phi_j(\mathcal{V}_{\boldsymbol{x}^*}).$$

Interestingly, for a linear regression with k uncorrelated features, and mean centered,

$$m(\boldsymbol{x}^*) = \underbrace{\beta_0}_{=\mathbb{E}\left[m(\boldsymbol{X})\right]} + \underbrace{\beta_1 x_1^*}_{\gamma_1^{\textsf{shap}}(\boldsymbol{x}^*)} + \underbrace{\beta_2 x_2^*}_{\gamma_2^{\textsf{shap}}(\boldsymbol{x}^*)} + \cdots + \underbrace{\beta_k x_k^*}_{\gamma_k^{\textsf{shap}}(\boldsymbol{x}^*)},$$

as discussed in Aas et al. (2021).

More generally, these contributions satisfy the following properties

- "local accuracy": $\sum_{j=1}^{k}\gamma_j^{\textsf{shap}}(\boldsymbol{x}^*) = m(\boldsymbol{x}^*) - \mathbb{E}\big[m(\boldsymbol{X})\big]$
- "symmetry": if j and k are interchangeable, $\gamma_j^{\textsf{shap}}(\boldsymbol{x}^*) = \gamma_k^{\textsf{shap}}(\boldsymbol{x}^*)$
- "dummy": if X_j does not contribute in the model, $\gamma_j^{\textsf{shap}}(\boldsymbol{x}^*) = 0$

Here, the interpretation of the additivity principle is not easy to derive (and to legitimate as a "desirable property", in the context of models). Observe that if there are two variables, $k = 2$, $\gamma_1^{\textsf{shap}}(\boldsymbol{x}^*) = \Delta_{1|2}(\boldsymbol{x}^*) = \gamma_1^{\textsf{bd}}(\boldsymbol{x}^*)$. And if $p \gg 2$, computations can be heavy. Štrumbelj and Kononenko (2014) suggested an approach based on simulations.

Given $\boldsymbol{x}^*$ and some individual $\boldsymbol{x}_i$, define

$$\tilde{x}_{i,j'} = \begin{cases} x_{j'}^* \text{ with probability } 1/2 \\ x_{i,j'} \text{ with probability } 1/2 \end{cases} \quad \text{and} \quad \begin{cases} \boldsymbol{x}_i^{*+} = (\tilde{x}_{i,1}, \cdots, x_j^*, \cdots, \tilde{x}_{i,k}) \\ \boldsymbol{x}_i^{*-} = (\tilde{x}_{i,1}, \cdots, x_{i,j}, \cdots, \tilde{x}_{i,k}) \end{cases}$$

Observe that $\gamma_j^{\textsf{shap}}(\boldsymbol{x}^*) \approx m(\boldsymbol{x}_i^{*+}) - m(\boldsymbol{x}_i^{*-})$, and therefore

$$\widehat{\gamma}_j^{\textsf{shap}}(\boldsymbol{x}^*) = \frac{1}{s}\sum_{i\in\{1,\cdots,n\}} m(\boldsymbol{x}_i^{*+}) - m(\boldsymbol{x}_i^{*-})$$

(we pick at each step individual i in the training dataset, s times).

In the context of our `toydata2` dataset, it is possible to compute Shapley values for two individuals (`Andrew` and `Barbara`), as on Figures 4.8 and 4.9, respectively, obtained using option `type = "shap"` in

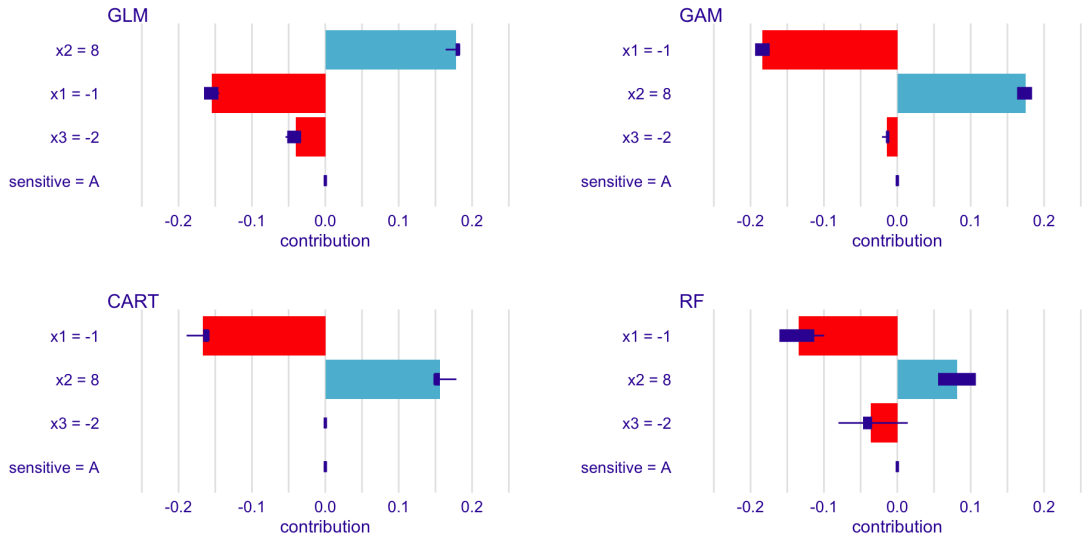


Figure 4.8: Shapley contributions $\widehat{\gamma}_j^{\text{shap}}(z_{\text{A}}^\star)$ for `Andrew` for different models trained on `toydata2` (see Table 3.3 for numerical values, here $z^\star = (\boldsymbol{x}^\star, s^\star) = (-1, 8, -2, \text{A})$).

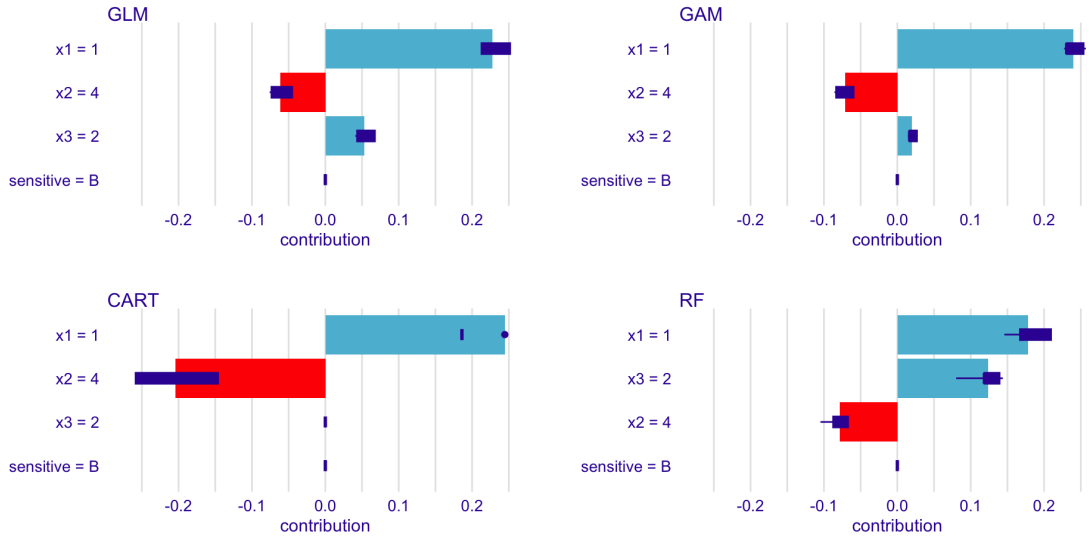


Figure 4.9: Shapley contributions $\widehat{\gamma}_j^{\text{shap}}(z_{\text{B}}^\star)$ for `Barbara` for different models trained on `toydata2` (see Table 3.3 for numerical values, here $z^\star = (\boldsymbol{x}^\star, s^\star) = (1, 4, 2, \text{B})$).

function `predict_parts` of package `DALEX`, as in Biecek and Burzykowski (2021). Observe that, at least,

signs of contributions are consistent among models: x_1^* has a negative contribution while x_2^* has a positive one, for `Andrew`, while it is the oppositive for `Barbara`.

Štrumbelj and Kononenko (2014) and Lundberg and Lee (2017) suggested to use that decomposition to get a global contribution of each variable, instead of a local version

Definition 4.1.10 (Shapley contribution $\overline{\gamma}_j^{\text{shap}}$) *The contribution of the j-th variable is*

$$\overline{\gamma}_j^{shap} = \frac{1}{n}\sum_{i=1}^{n} \gamma_j^{shap}(\boldsymbol{x}_i).$$

One interesting feature about Shapley value is that the contribution can be extended, from a single player j to any coalition, for example two players $\{i, j\}$. This will yield the concept of "Shapley interaction",

Definition 4.1.11 (Shapley interaction $\gamma_{i,j}^{\text{shap}}(\boldsymbol{x}^*)$) *The interaction contribution between the i-th and the j-th variable, at $\boldsymbol{x}^*$, is*

$$\gamma_{i,j}(\boldsymbol{x}^*) = \sum_{S\subseteq\{1,\ldots,k\}\setminus\{i,j\}} \frac{|S|!\,(k-|S|-2)!}{2\,k!}\Delta_{i,j|S}(\boldsymbol{x}^*)$$

where

$$\begin{aligned}\Delta_{i,j|S}(\boldsymbol{x}^*) &= \mathbb{E}_{\boldsymbol{X}_{S\cup\{i,j\}}^{\perp}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}_{S\cup\{i,j\}}^*\right] - \mathbb{E}_{\boldsymbol{X}_{S\cup\{j\}}^{\perp}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}_{S\cup\{j\}}^*\right] \\ &- \mathbb{E}_{\boldsymbol{X}_{S\cup\{i\}}^{\perp}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}_{S\cup\{i\}}^*\right] + \mathbb{E}_{\boldsymbol{X}_{S}^{\perp}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}_{S}^*\right].\end{aligned}$$

4.1.5 Partial Dependence

The "partial dependence plot", formally defined and coined in Friedman (2001), is simply the average of "ceteris paribus profiles",

Definition 4.1.12 (PDP $p_j(x_j^*)$ and $\widehat{p}_j(x_j^*)$) *The Partial Dependence Plot associated with the j-th variable is the function $\mathcal{X}_j \to \mathbb{R}$ defined as*

$$p_j(x_j^*) = \mathbb{E}_{X_j^{\perp}}\left[m(\boldsymbol{X})|x_j^*\right],$$

and the empirical version is

$$\widehat{p}_j(x_j^*) = \frac{1}{n}\sum_{i=1}^{n} m(x_j^*, \boldsymbol{x}_{i,-j}) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{m_{\boldsymbol{x}_i,j}(x_j^*)}_{\textit{ceteris paribus}}.$$

See Greenwell (2017) for the implementation in R, with the `pdp` package. One can also use `type = "partial"` in the `predict_parts` function of the `DALEX` package, as in Biecek and Burzykowski (2021). On Figure 4.10 we can visualize $\widehat{p}_1$ (associated with variable x_1) in dataset `toydata2`, the average of $m(x_j^*, \boldsymbol{x}_{i,-j})$ when $i = 1, \cdots, n$, including all $m(x_j^*, \boldsymbol{x}_{i,-j})$'s on Figure 4.11.

Interestingly, instead of the sum over the n predictions, subsums can be considered, with respect to some criteria. For example, on Figures 4.12, 4.13 and 4.14, sums over $s_i =$ `A` or $s_i =$ `B` are considered,

$$\widehat{p}_j^{\texttt{A}}(x_j^*) = \frac{1}{n_{\texttt{A}}}\sum_{i:s_i==\texttt{A}} m(x_j^*, \boldsymbol{x}_{i,-j}) \text{ and } \widehat{p}_j^{\texttt{B}}(x_j^*) = \frac{1}{n_{\texttt{B}}}\sum_{i:s_i==\texttt{B}} m(x_j^*, \boldsymbol{x}_{i,-j}).$$

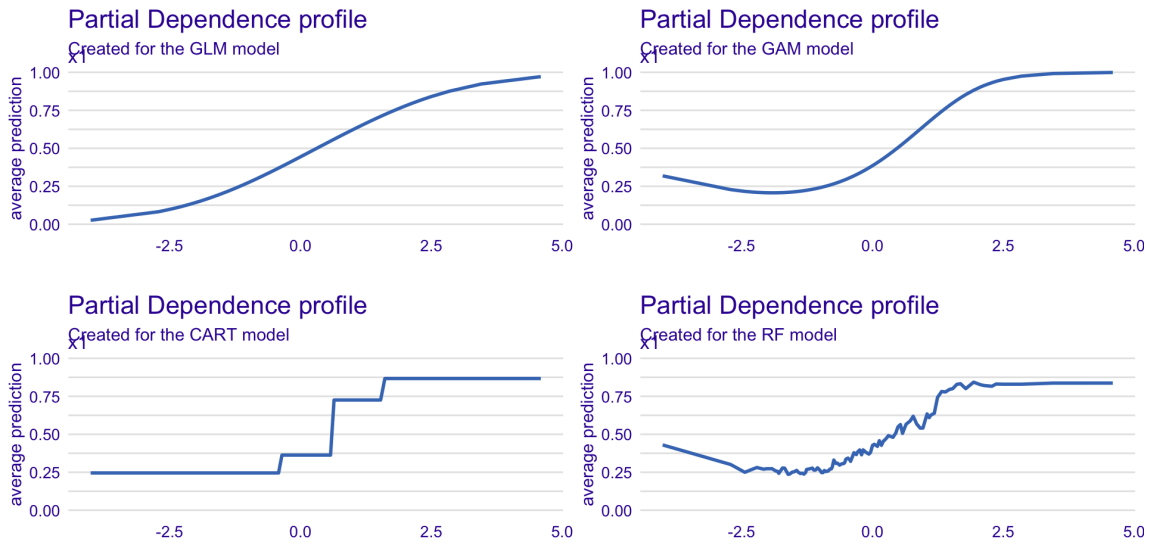


Figure 4.10: Partial dependence profile $\widehat{p}_1$ associated with variable x_1, for four different models trained on `toydata2`.

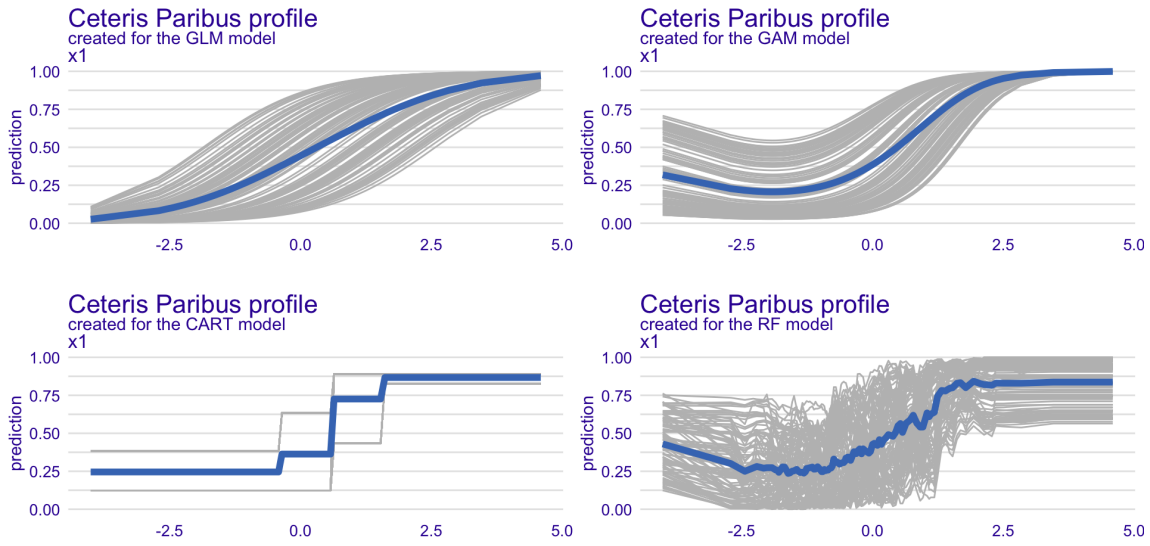


Figure 4.11: Partial dependence profile $\widehat{p}_1$ associated with variable x_1, seen as the average of ceteris paribus profiles $m(x_j^*, \boldsymbol{x}_{i,-j})$'s (in gray) for different models trained on `toydata2`.

On the `toydata2` data, the three variables j (namely x_1^*, x_2^* and x_3^*) are used, on Figures 4.12, 4.13 and 4.14, respectively. If x_3^* has a very flat impact (on Figure 4.14), and no influence on the outcome, one should

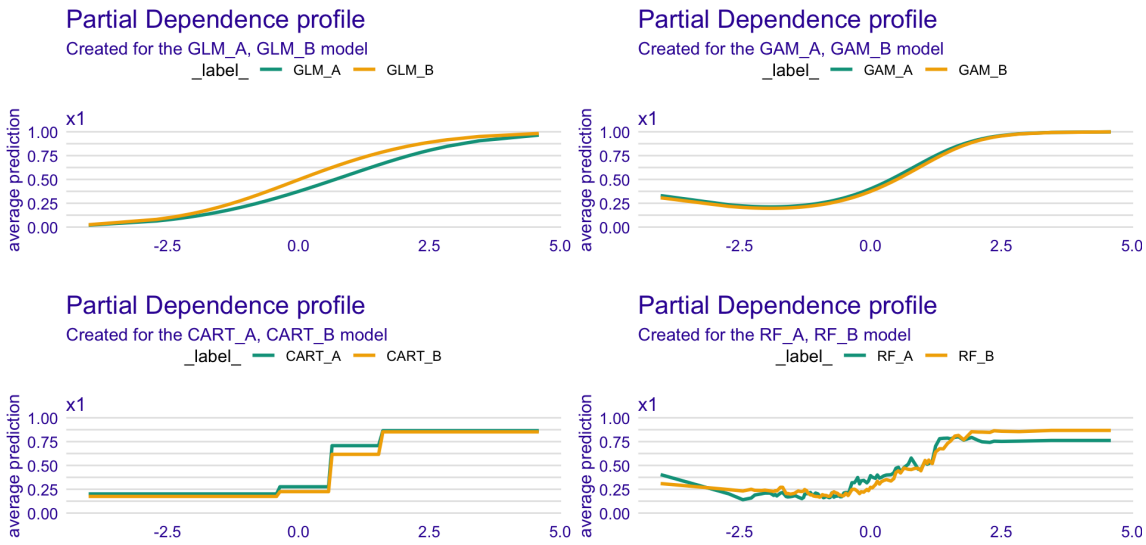


Figure 4.12: Partial dependence profiles $\widehat{p}_1^{\mathtt{A}}$ and $\widehat{p}_1^{\mathtt{B}}$, for x_1, when the sensitive attribute s is either A or B, as the average of subgroups (s_i being either A or B) for different models trained on `toydata2`.

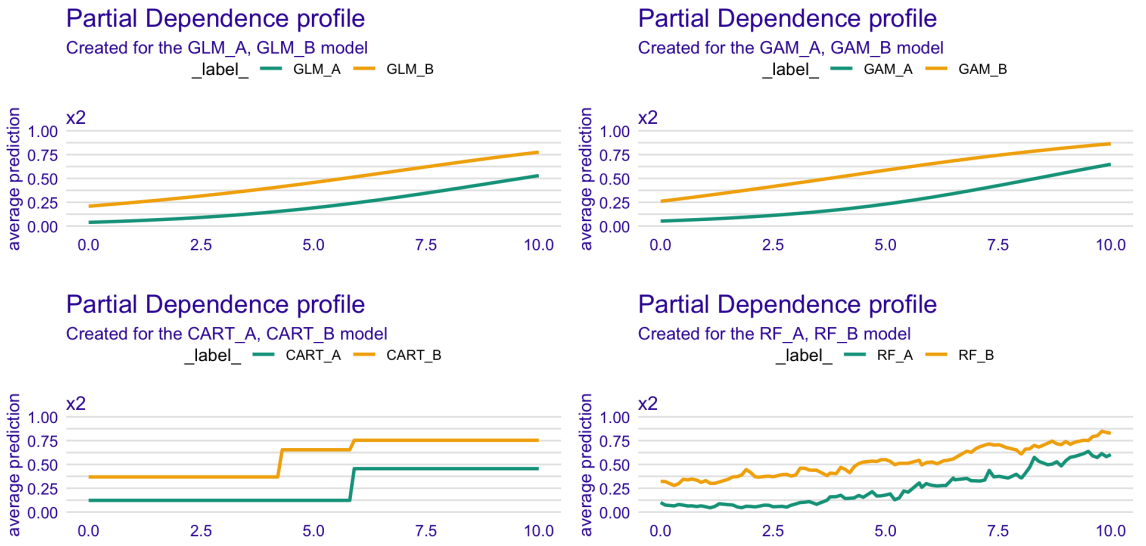


Figure 4.13: Partial dependence profiles $\widehat{p}_2^{\mathtt{A}}$ and $\widehat{p}_2^{\mathtt{B}}$, for x_2, when the sensitive attribute s is either A or B, as the average of subgroups (s_i being either A or B) for different models trained on `toydata2`.

observe that $\widehat{p}_j^{\mathtt{A}}(x_3^*)$ and $\widehat{p}_j^{\mathtt{B}}(x_3^*)$ are significantly different.

But instead of those *ceteris paribuss* dependence plots, it could be interesting to consider some local

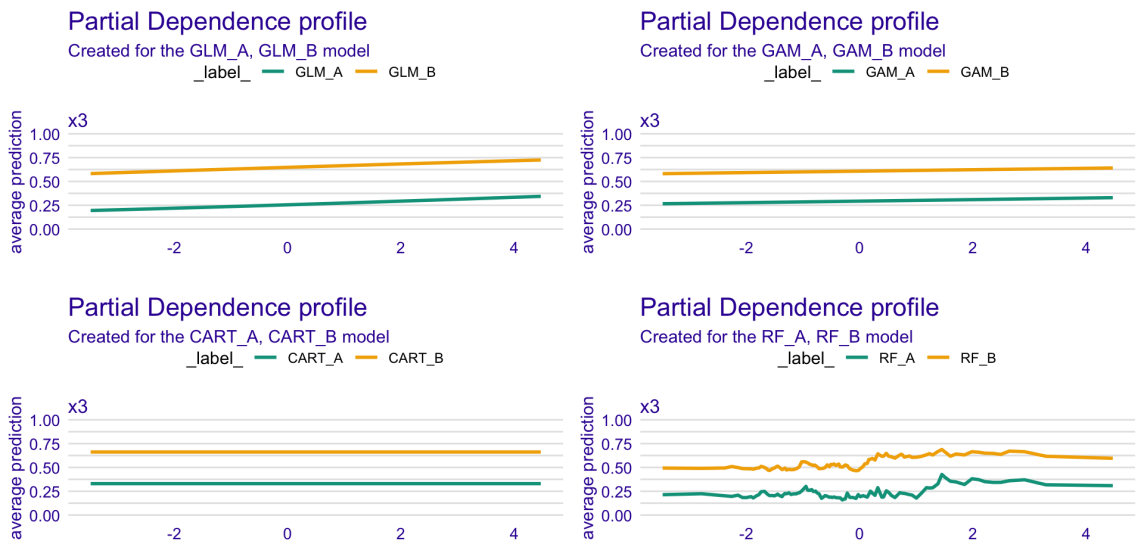


Figure 4.14: Partial dependence profiles $\widehat{p}_3^{\mathtt{A}}$ and $\widehat{p}_3^{\mathtt{B}}$, for x_3, when the sensitive attribute s is either A or B, as the average of subgroups (s_i being either A or B) for different models trained on `toydata2`.

versions, or *mutatis mutandis* dependence plots. Apley and Zhu (2020) introduced the "local dependence plot" and the "accumulated local plot," defined as follows

Definition 4.1.13 (Local Dependence Plot $\ell_j(x_j^*)$ and $\widehat{\ell}_j(x_j^*)$) *The local dependence plot is defined as*

$$\ell_j(x_j^*) = \mathbb{E}_{X_j}\left[m(\boldsymbol{X})|x_j^*\right]$$

$$\widehat{\ell}_j(x_j^*) = \frac{1}{card(V(x_j^*))}\sum_{i\in V(x_j^*)} m(x_j^*, \boldsymbol{x}_{i,-j}) \text{ where } V(x_j^*) = \left\{i : d(x_{i,j}, x_j^*) \leq \epsilon\right\},$$

$$\text{or } \widetilde{\ell}_j(x_j^*) = \frac{1}{\sum_i \omega_i(x_j^*)}\sum_{i=1}^{n} \omega_i(x_j^*) m(x_j^*, \boldsymbol{x}_{i,-j}) \text{ where } \omega_i(x_j^*) = K_h(x_j^* - x_{i,j}),$$

for a smooth version, for some kernel K_h.

Apley and Zhu (2020) suggessted instead to use

Definition 4.1.14 (Accumulated Local $a_j(x_j^*)$)

$$a_j(x_j^*) = \int_{-\infty}^{x_j^*} \mathbb{E}_{X_j}\left[\frac{\partial m(x_j, \boldsymbol{X}_{-j})}{\partial x_j}\middle| x_j\right] \mathrm{d}x_j.$$

The following estimate was considered

Definition 4.1.15 (Accumulated Local function $\widehat{a}_j(x_j^*)$)

$$\widehat{a}_j(x_j^*) = \alpha + \sum_{u=1}^{k_j^*} \frac{1}{n_u} \sum_{u:x_{i,j}\in(a_{u-1},a_u]} \left[m(a_k, \boldsymbol{x}_{i,-j}) - m(a_{k-1}, \boldsymbol{x}_{i,-j})\right]$$

(where α is some normalization constant, since $\mathbb{E}[\widehat{a}_j(X_j)] = 0$).

Observe on Figure 4.15, the three dependence profiles for x_1, for the random forest model, with respectively the "partial dependence plot" on the left, the "local dependence plot" in the middle, and the "accumulated local plot" on the right, on the `toydata2` dataset, with options `type = "accumulated"` in function `predict_parts`, as in Biecek and Burzykowski (2021). One could also use the `FeatureEffect` function in the `iml` R package, based on Molnar (2023), respectively with `method = "pdp"`, `"ale"` and `"ice"`,

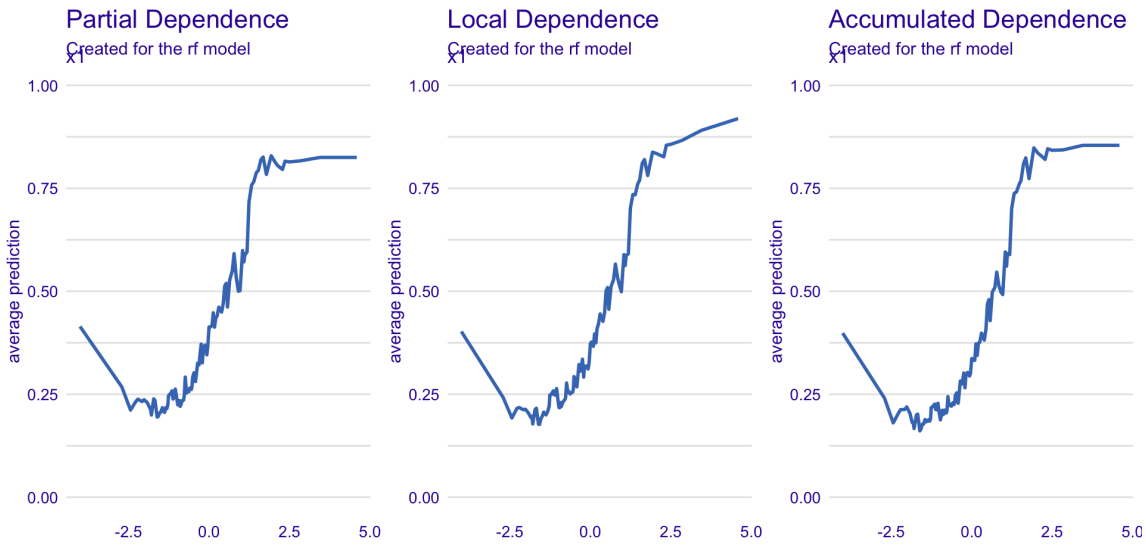


Figure 4.15: Partial dependence plot $\widehat{p}_1$ on the left, local dependence plot $\widehat{\ell}_1$ in the middle, and accumulated local function $\widehat{a}_1$ on the right, for x_1, for the random forest model m, trained on `toydata2`.

4.1.6 Application on the `GermanCredit` Dataset

In order to illustrate, consider four models estimated on the `GermanCredit` dataset: a plain GLM (estimated using `glm`), a gradient boosting model GBM (estimated using `adaboost` algorithm, in `gbm`, with the default tuning parameters), a classification tree, CART (estimated using `rpart` and the default tuning parameters) and a random forest, RF (estimated using `randomForest`, based on 500 trees). On the `germancredit` dataset, consider two individuals with a loan, `Barbara` and `Andrew`, in Table 4.2.

On Figures 4.16 and 4.17 we can visualize breakdown decomposition $\widehat{\gamma}_j^{\text{bd}}(z_{\text{A}}^*)$ for `Andrew` and $\widehat{\gamma}_j^{\text{bd}}(z_{\text{B}}^*)$ for `Barbara`, respectively, on four models trained on `germancredit`.

	gender s	age	housing	property	Credit amount	account status	duration	employment since	job	purpose
`Barbara`	Female	63	for free	no property	13,756	none	60	≥ 7 years	highly qualified	new car
`Andrew`	Male	28	rent	car	4,113	[0, 200]	24	< 1 year	skilled employee	old car

	GLM				boosting (GBM)				CART				
	with s		without s		with s		without s		with s		without s		
	Prob.	(Rank)	Prob.	Rank	Prob.	(Rank)	Prob.	(Rank)	Prob.	(Rank)	Prob.	(Rank)	Prob
`Barbara`	38.4%	(69%)	40.0%	(69%)	38.4%	(76%)	40.6%	(73%)	66.7%	(77%)	66.7%	(77%)	46.4%
`Andrew`	46.7%	(74%)	44.2%	(72%)	46.7%	(60%)	27.7%	(58%)	15.4%	(54%)	15.4%	(54%)	29.0%

Table 4.2: Information about `Barbara` and `Andrew`, two policyholders (names are fictional) from the `germancredit` dataset.

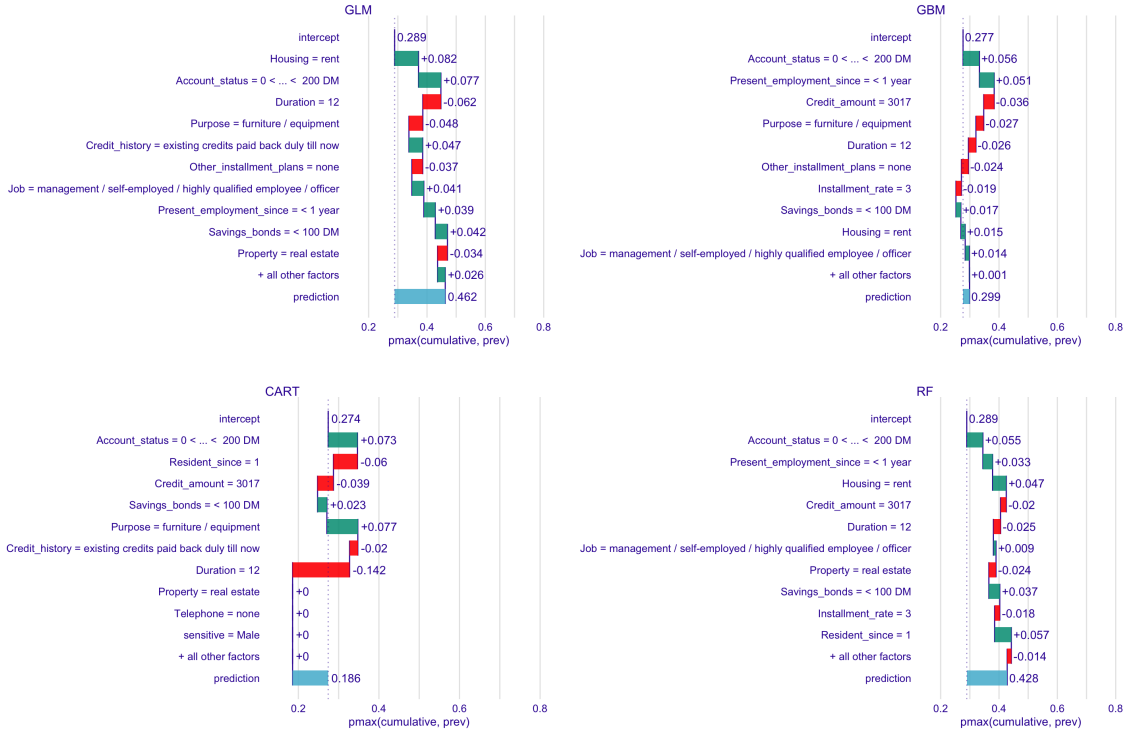


Figure 4.16: Breakdown decomposition $\widehat{\gamma}_j^{\text{bd}}(z_{\text{A}}^*)$ for `Andrew`, `germancredit`, on four models.

On Figures 4.18 and 4.19, we can visualize Shapley contributions $\gamma_j^{\text{shap}}(z_{\text{A}}^*)$ for `Andrew` and $\gamma_j^{\text{shap}}(z_{\text{B}}^*)$ for `Barbara`, respectively.

On Figure 4.20, partial dependence plots $\widehat{p}^{\text{A}}_{\texttt{duration}}$ and $\widehat{p}^{\text{B}}_{\texttt{duration}}$ for the `duration` variable, for male (s = A) and female (s = B), respectivelty, on the `germancredit` dataset, with the plain logistic and `adaboost` (GBM) on top, a classification tree and a random forest, below. All functions are increasing, indicating that

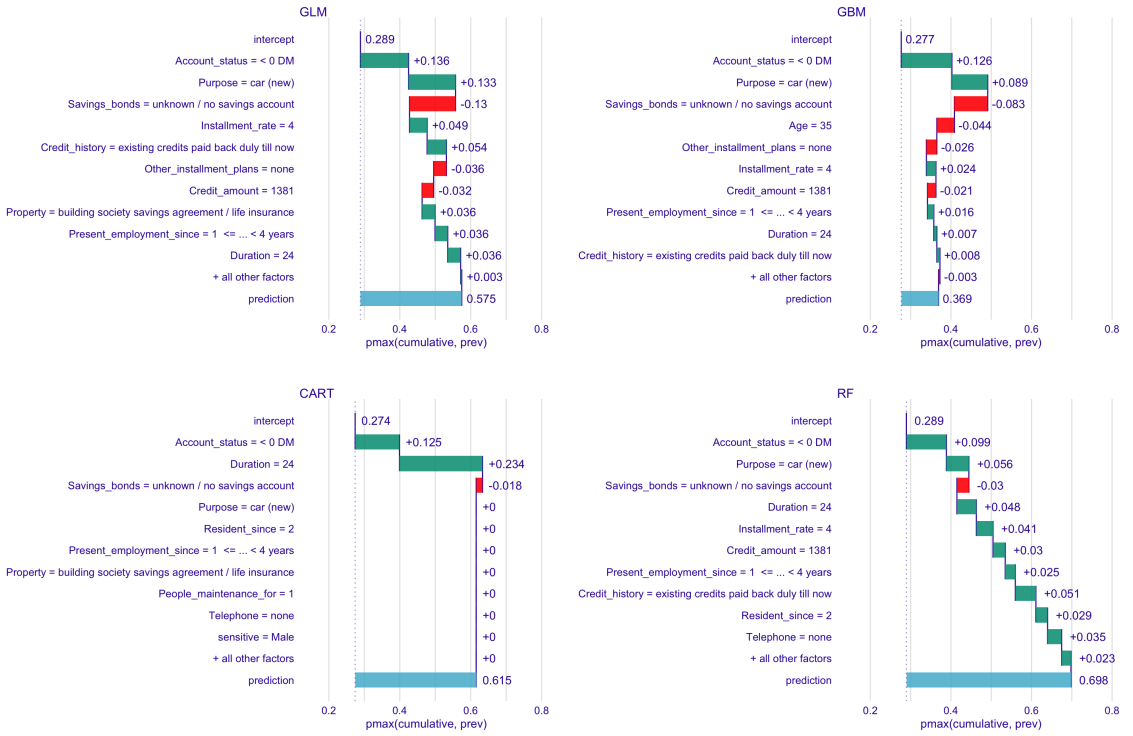


Figure 4.17: Breakdown decomposition $\widehat{\gamma}_j^{\text{bd}}(z_{\text{B}}^*)$ for `Barbara`, `germancredit`, on four models.

the probability of having a default on a loan should increase with the duration.

4.1.7 Application on the `frenchmotor` Dataset

On the `frenchmotor` dataset, we model the occurrence of a claim (variable `claim`), with four models: a plain GLM (estimated using `glm`), a gradient boosting model (estimated using `gam` from the `mgcv` R package, with splines – function `s()` – on the age of the driving licence, `LicAge`, the driver's age `DrivAge`, the bonus-malus coefficient `BonusMalus` and the `RiskVar` variable), a classification tree (estimated using `rpart` and the default tuning parameters) and a random forest (estimated using `randomForest`, based on 500 trees). The sensitive attribute s is here the gender. Two individuals are considered in Table 4.3, on top (named respectively `Barbara` and `Andrew`). Predictions of the probability to claim a loss, for those two individuals are given below. For `Barbara`, the classification tree (including the sensitive attribute s, here the gender) predicts a probability to claim a loss of 6.6% (seeing `Barbara` as a median driver, since 6.6% corresponds to the 49% quantile).

On Figures 4.21 and 4.22, we can visualize breakdown decomposition $\widehat{\gamma}_j^{\text{bd}}(z_{\text{A}}^*)$ for `Andrew` and $\widehat{\gamma}_j^{\text{bd}}(z_{\text{B}}^*)$ for `Barbara`, respectively, on four models trained on `frenchmotor`.

On Figures 4.23 and 4.24, we can visualize Shapley contributions (including confidence intervals) $\gamma_j^{\text{shap}}(z_{\text{A}}^*)$ for `Andrew` and $\gamma_j^{\text{shap}}(z_{\text{B}}^*)$ for `Barbara`, respectively, for models estimated using the `frenchmotor` dataset. One could also use `Shapley` in the `iml` R package (see Molnar et al. (2018))

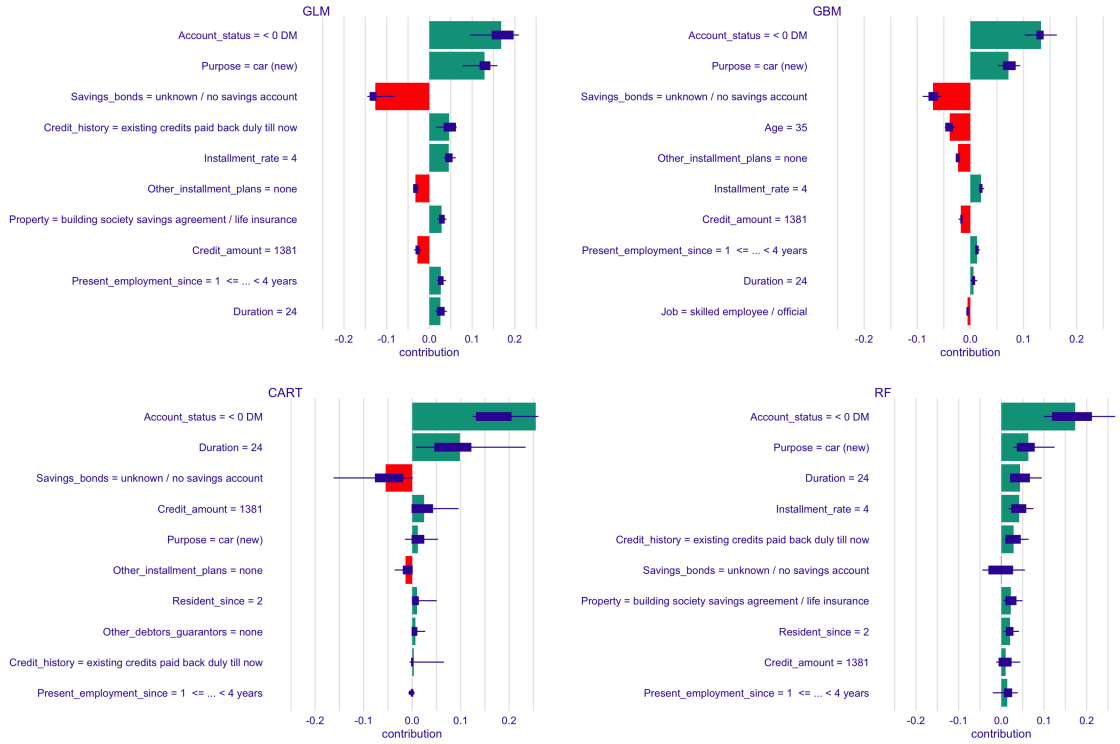


Figure 4.18: Shapley contributions for `Andrew`, on the `germancredit` dataset.

	gender *s*	age (years)	marital status	social category	licence (months)	car age	car use	car class	car gas	bonus malus	risk score	claim *y*
`Barbara`	Female	26	Alone	CSP5	67	10+	Private+office	M1	regular	76	7	0
`Andrew`	Male	36	Couple	CSP1	206	0	Professional	M2	diesel	78	19	0

	GLM				GAM				CART				
	with *s*		without *s*		with *s*		without *s*		with *s*		without *s*		with
	Prob.	(Rank)	Prob.	Rank	Prob.	(Rank)	Prob.	(Rank)	Prob.	(Rank)	Prob.	(Rank)	Prob.
`Barbara`	3.4%	(11%)	3.3%	(10%)	3.8%	(12%)	3.8%	(12%)	6.6%	(49%)	6.6%	(43%)	2.4%
`Andrew`	9.8%	(64%)	9.9%	(65%)	7.8%	(46%)	7.8%	(46%)	17.1%	(93%)	17.1%	(93%)	80.2%

Table 4.3: Information about `Barbara` and `Andrew`, two policyholders (names are fictional) from the `frenchmotor` dataset. CSP1 corresponds 'farmers and breeders' (employees of their farm) while CSP5 corresponds to 'employees'. The risk variable (`RiskVar`) is an internal score going from 1 (low) to 20 (high). A bonus of 50 is the best (lowest) level, and 100 is usually the entering level for new drivers. Both have no garage. For the predictions, a 'rank' of 11% means that the policyholder is perceived as almost in the top 10% of all drivers in the validation database.

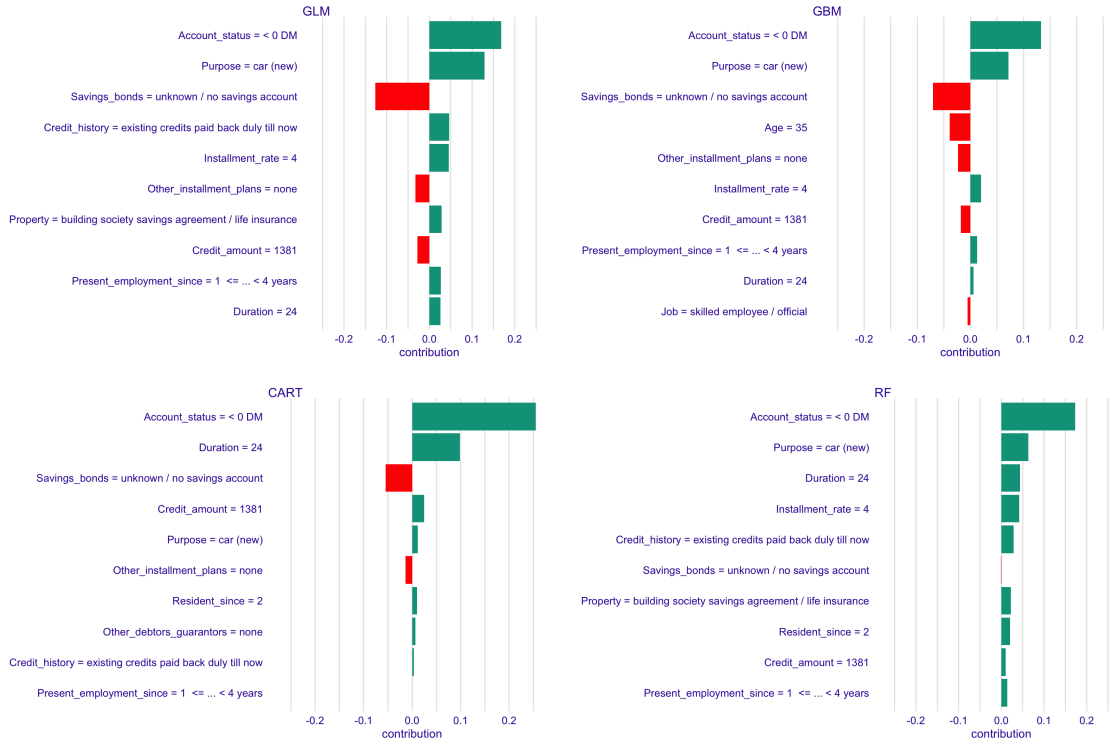


Figure 4.19: Shapley contributions for `Barbara`, on the `germancredit` dataset.

Figure 4.25 is the partial dependence plot for a specific (continuous) variable, the licence age (`licence_age`), with the average of ceteris paribus profiles, for males and females, respectively. Observe that, even if s is not included in the models, partial dependence plots are different in the two groups: for GLM and GAM, predicted probabilities for males are higher than for females, while for the random forest, predicted probabilities for females are higher than for males.

Note that there are more or more connections between interpretation and causal models, as discussed in Feder et al. (2021), Geiger et al. (2022) or Wu et al. (2022). We will get back on those approaches in chapters 7 (on experimental and observational data) and 9 (on individual fairness, and counterfactual).

4.1.8 Counterfactual Explanation

To conclude this section, let us briefly mention here a concept that will be discussed further in the context of fairness and discrimination, related to the idea of "counterfactuals" (as named in Lewis (1973)). The word "counterfactual" can be either an adjective describing something "*thinking about what did not happen but could have happened, or relating to this kind of thinking*," or a noun defined as "*something such as piece of writing or an argument that considers what would have been the result if events had happened in a different way to how they actually happened*." Therefore, a counterfactual is a statement of the form "*had `A` occurred then `B` would have occurred*." For example Wachter et al. (2017) suggested to use such counterfactual statements as a psychology-grounded approach for explaining black-box decision rule. It is

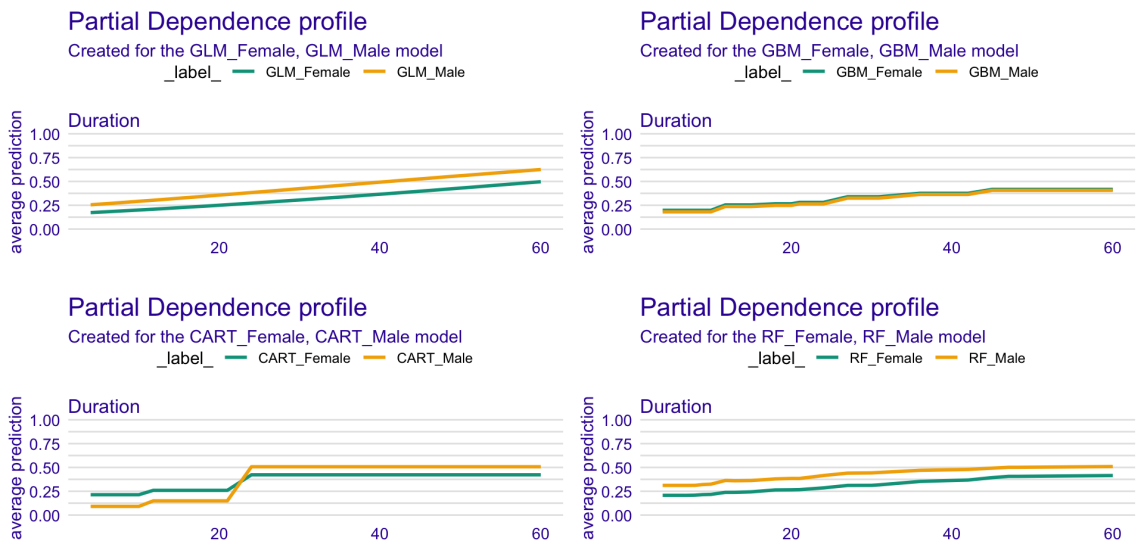


Figure 4.20: Partial dependence plots $\widehat{p}^{\texttt{A}}_{\texttt{duration}}$ and $\widehat{p}^{\texttt{B}}_{\texttt{duration}}$ for the `duration` variable, for male (s = A) and female (s = B), on the `germancredit` dataset, with the plain logistic and `adaboost` (GBM) on top, a classification tree and a random forrest, below.

clearly related to the idea of deriving explanations.

To conclude, here, we provided some tools that could be used because, in real life, some insurers believe that accuracy (as discussed in the next section) is the ultimate goal when creating a predictive model, even if the price to pay is a very dark box. As Rudin (2019) wrote, "*some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society (...) Explanations are often not reliable, and can be misleading (...) If we instead use models that are inherently interpretable, they provide their own explanations, which are faithful to what the model actually computes*."

4.2 Accuracy of Actuarial Models

For classification problems, calibration measures how well your model's scores can be interpreted as probabilities, according to Cameron (2004) or Degroot (2004). Accuracy measures how often your model produces correct answers, as defined Schilling (2006).

4.2.1 Accuracy and Scoring Rules

Accurate, from Latin *accuratus* (past participle of *accurare*), means "*done with care*." With a statistical perspective, accuracy is how far off a prediction ($\widehat{y}$) is from its true value (y). Thus, a model is accurate if the errors ($\widehat{\varepsilon} = y - \widehat{y}$) are small. In the least-square approach, accuracy can be measured simply by looking

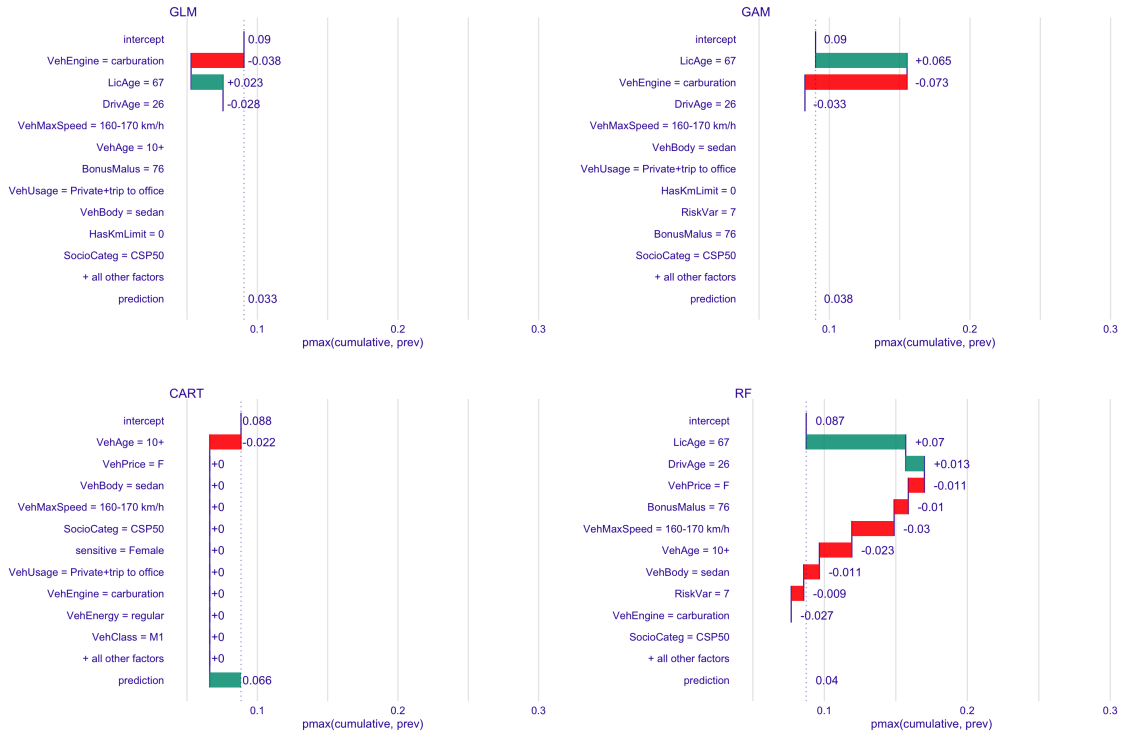


Figure 4.21: Breakdown decomposition $\widehat{\gamma}_j^{\text{bd}}(z_{\text{A}}^*)$ for `Andrew` for different models trained on `frenchmotor`.

at the loss, or mean squared error, that is the empirical risk

$$\widehat{\mathcal{R}}_n = \frac{1}{n}\sum_{i=1}^{n} \ell_2(y_i, \widehat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varepsilon}_i^2.$$

One could also consider the mean absolute error

$$\frac{1}{n}\sum_{i=1}^{n} \ell_1(y_i, \widehat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} |\widehat{\varepsilon}_i|,$$

or the symmetric mean absolute percentage error, as introduced by Armstrong (1985)

$$\sum_{i=1}^{n} \frac{|\widehat{\varepsilon}_i|}{|y_i| + |\widehat{y}_i|}.$$

But those metrics a based on a loss function defined on $\mathcal{Y} \times \mathcal{Y}$, that measures a distance between the observation y and the prediction $\widehat{y}$, seen as a pointwise prediction. But in many applications, the prediction can be a distribution. So instead of a loss defined on $\mathcal{Y} \times \mathcal{Y}$, one could consider a scoring function defined on $\mathcal{P}_y \times \mathcal{Y}$, where $\mathcal{P}_y$ denotes a set of distributions on $\mathcal{Y}$. As defined in Gneiting and Raftery (2007)

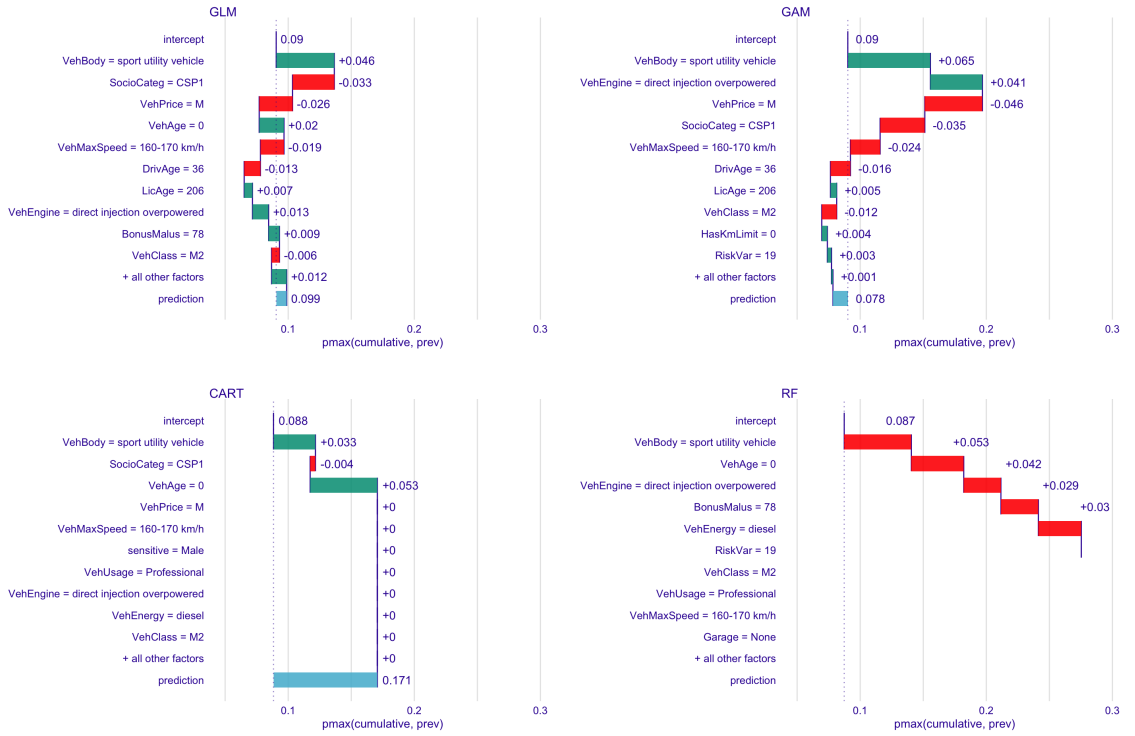


Figure 4.22: Breakdown decomposition $\widehat{\gamma}_j^{\mathsf{bd}}(z_{\mathrm{B}}^*)$ for `Barbara` for different models trained on `frenchmotor`.

Definition 4.2.1 (Scoring Rules) *Gneiting and Raftery (2007). A scoring rule is a function* $s : \mathcal{Y} \times \mathcal{P}_y \to \mathbb{R}$ *that quantifies the error of reporting* $\mathbb{P}_y$ *when the outcome is* y*. The expected score when belief is* $\mathbb{Q}$ *is* $S(\mathbb{P}_y, \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\big[s(\mathbb{P}_y, Y)\big]$.

Definition 4.2.2 (Proper Scoring Rules) *Gneiting and Raftery (2007). A scoring rule is proper if* $S(\mathbb{P}_y, \mathbb{Q}) \geq S(\mathbb{Q}, \mathbb{Q})$ *for all* $\mathbb{P}_y, \mathbb{Q} \in \mathcal{P}_y$*, and strictly proper if equality holds only when* $\mathbb{P}_y = \mathbb{Q}$.

Let us start with the binary case, when $y \in \mathcal{Y} = \{0, 1\}$ and $\mathcal{P}_y$ is the set Bernoulli distributions, $\mathcal{B}(p)$, with $p \in [0, 1]$. The scoring rule is $s(p, y) \in \mathbb{R}$, and the expected score is $S(p, q) = \mathbb{E}(s(p, Y))$ with $Y \sim \mathcal{B}(q)$, for some $q \in [0, 1]$. For example Brier scoring rule is defined as follows : let $f_q(y) = q^y(1-q)^{1-y}$, and define

$$s(f_q, y) = -2f_q(y) + \sum_{y=0}^{1} f_q(y)^2$$

that can be written

$$s(q, y) = -q^y(1-q)^{1-y} + \frac{1}{2}\big(q^2 + (1-q)^2\big).$$

This is a proper scoring rule. And we can define $G(q) = S(q, q)$ that will satisfy then $G(q) \leq S(p, q)$ for all p, q. Interestingly, we can recover S from G. Because the expected value is a linear operator, $q \to S(p, q)$

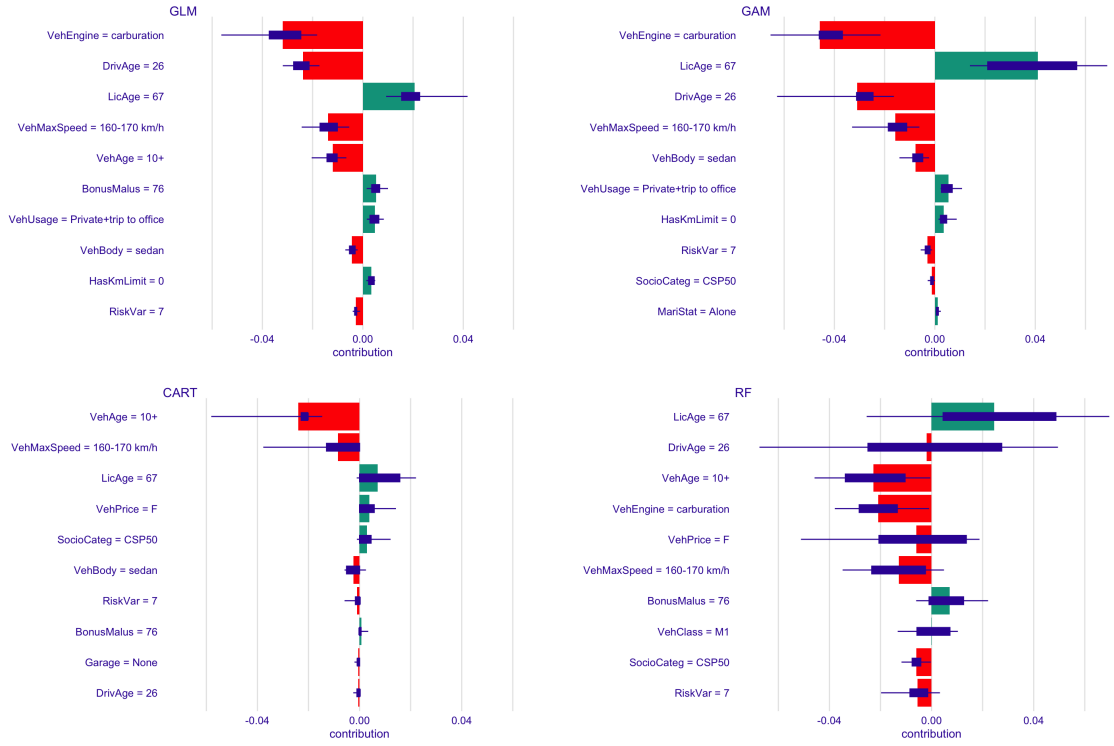


Figure 4.23: Shapley contributions $\widehat{\gamma}_j^{\text{shap}}(\boldsymbol{z}_{\text{A}}^\star)$ for `Andrew`, on the `frenchmotor` dataset.

is linear, as mentioned in Parry (2016). S is proper if $S(q,q) \leq S(p,q)$ for all p, q. And in that case, the divergence $d(p,q) = S(p,q) - S(q,q)$ cannot be negative. And for a proper scoring rule, $d(p,q) = 0$ if and only if $p = q$. Under some mild conditions on $\mathcal{P}$, S is a proper scoring rule if and only if there exists a concave function G such that

$$s(y,p) = G(p) + (y-p)\partial G(p)$$

and $G(q) = S(q,q)$. For example, if $G(p) = -p\log p - (1-p)\log(1-p)$ (corresponding to the entropy),

$$S(y,q) = -y\log q - (1-y)\log(1-q)$$

In that case, the deviance is

$$d(p,q) = q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$$

which is Kullback-Leibler divergence (see Definition 3.3.7). If $G(p) = p(1-p)$ (corresponding to Gini index) yields $S(y,q) = (y-q)^2$, which is Brier score. In that case, the deviance is the mean squared difference between probabilities, $d(p,q) = (p-q)^2$.

On can use the R package `scoringRules` to compute those quantities, see Jordan et al. (2019). Thus, scoring rules provide summary measures of predictive performances, assigning a numerical score to probabilistic forecasts. An interesting property is the "sharpness" of the predictive distributions, or its concentration. The

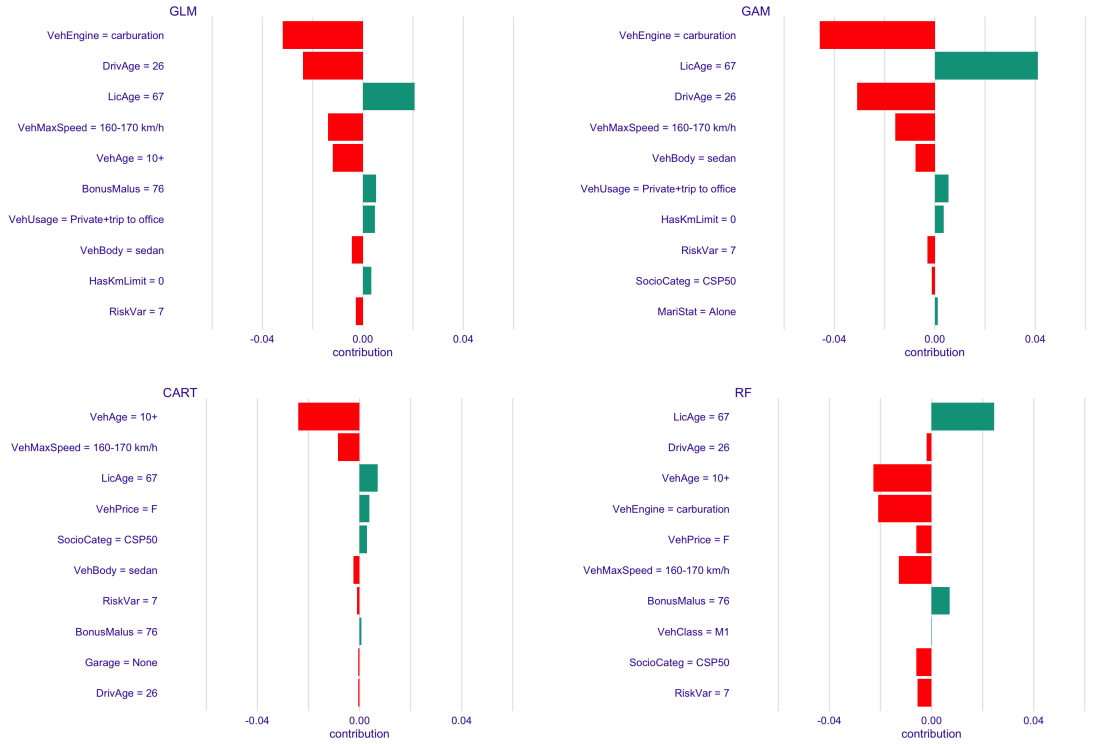


Figure 4.24: Shapley contributions $\widehat{\gamma}_j^{\text{shap}}(z_{\text{B}}^\star)$ for `Barbara`, on the `frenchmotor` dataset.

less variable the predictions, the more concentrated the predictive distribution is. See Candille and Talagrand (2005) and Gneiting et al. (2007) for more details. Note that Duan et al. (2020) suggested to use a gradient of the scoring rule with respect to the distribution of the prediction, in a boosting procedure (instead of the standard gradient of the loss, as in Section 3.3.6).

To conclude this section on the accuracy, let us recall that in a classification problem, “accuracy" has a precise quantitative definition (and not only an abstract concept): it is the fraction of predictions our model got right. Therefore, it is a property of m_t (for a given threshold τ) and not of the model m. Note that instead of $m_t = \mathbf{1}(m > t)$ it is also possible to define the “decision boundary function" $d = m - t$ so that m_t is 1 if d is positive, 0 otherwise.

Definition 4.2.3 (Accuracy (classifier)) *The accuracy of a classifier m_t is the number of correct prediction over the total number of prediction*

$$accuracy(m_t) = \frac{TP + TN}{TP + TN + FP + FN}.$$

In some heavily unbalanced data (for instance to detect fraudulent transactions), with $\mathbb{P}[Y = 1] = 1\%$, observe that model m that always predict 0 will have (on average) a 99% accuracy. Instead of looking a classifier m_t, for a given t, we can consider the overall scoring function m (for all t in $[0, 1]$).

$$(\mathbb{P}[m(\boldsymbol{X}) > t|Y = 0], \mathbb{P}[m(\boldsymbol{X}) > t|Y = 1])_{t\in[0,1]},$$

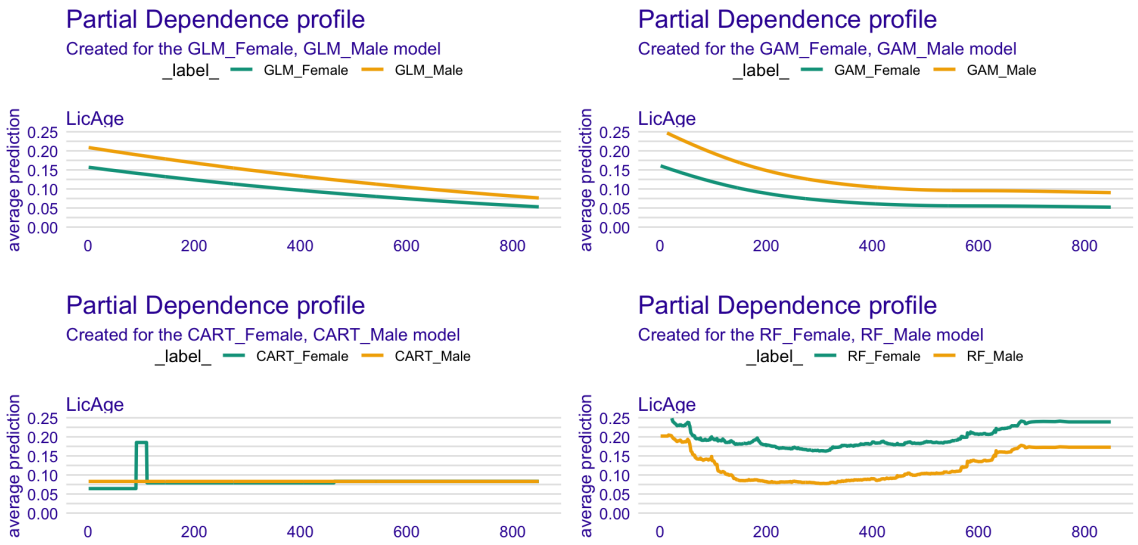


Figure 4.25: Partial dependence plots $\widehat{p}^{\texttt{A}}_{\texttt{licence\_age}}$ and $\widehat{p}^{\texttt{B}}_{\texttt{licence\_age}}$ for the `licence_age` variable, for male ($s = \texttt{A}$) and female ($s = \texttt{B}$), on the `frenchmotor` dataset.

since $\widehat{y} = m_t(\boldsymbol{x}) = \mathbf{1}_{m(\boldsymbol{x})>t}$ for threshold t,

$$(\mathbb{P}[\widehat{Y} = 1|Y = 0], \mathbb{P}[\widehat{Y} = 1|Y = 1]) = (\text{FPR}, \text{TPR}).$$

The ROC curve is the curve obtained by representing the true positive rates according to the false positive rates, by changing the threshold. This can be related to the "discriminant curve" in the context of credit scores, in Gourieroux and Jasiak (2007).

Definition 4.2.4 (ROC curve) *The ROC curve is the parametric curve*

$$\big\{\mathbb{P}[m(\boldsymbol{X}) > t|Y = 0], \mathbb{P}[m(\boldsymbol{X}) > t|Y = 1]\big\}, \text{ for } t \in [0, 1],$$

when the score $m(\boldsymbol{X})$ and Y evolve in the same direction (a high score indicates a high risk).

$$C(t) = TPR \circ FPR^{-1}(t),$$

where

$$\begin{cases} FRP(t) = \mathbb{P}[m(\boldsymbol{X}) > t|Y = 0] = \mathbb{P}[m_0(\boldsymbol{X}) > t] \\ TPR(t) = \mathbb{P}[m(\boldsymbol{X}) > t|Y = 1] = \mathbb{P}[m_1(\boldsymbol{X}) > t]. \end{cases}$$

In other words, the ROC curve is obtained from the two survival functions of $m(\boldsymbol{X})$ FPR and TPR (respectively conditional on $Y = 0$ and $Y = 1$). The AUC, the area under the curve, is then written as follows,

Definition 4.2.5 (AUC, area under the ROC curve) *The area under the curve is defined as the area below the ROC curve,*

$$AUC = \int_0^1 C(t)\mathrm{d}t = \int_0^1 TPR \circ FPR^{-1}(t)\mathrm{d}t.$$

On Figure 4.26, we can visualize on the left a classification tree, when we try to predict the gender of a driver using telematic information (from the `telematic` dataset), and on the right, ROC curves associated to three models, a smooth logistic regression (GAM), `adaboost` (boosting, GBM) and a random forest (RF), trained on 824 observations, and ROC curves are based on the 353 observations of the validation dataset.

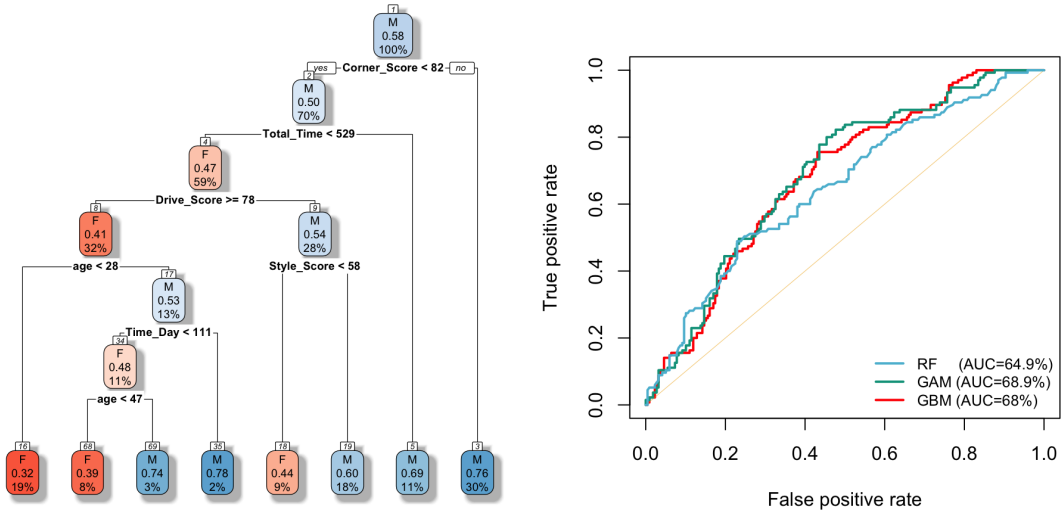


Figure 4.26: On the right, ROC curves $\widehat{C}_n(t)$, for various models on the `telematic` (validation) dataset, where we try to predict the gender of the driver using telematic data (and the age). AUC using GAM is close to 69%. A classification tree is also plotted on the left.

The AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. Indeed, assume for simplicity that the score (actually m_0 and m_1) has a derivative, so that the true positive rate and the false positive rate are given by

$$\text{TPR}(t) = \int_t^1 m_1'(x)\,\mathrm{d}x \text{ and } \text{FPR}(t) = \int_t^1 m_0'(x)\,\mathrm{d}x,$$

then

$$\text{AUC} = \int_0^1 \text{TPR}\big(\text{FPR}^{-1}(t)\big)\,\mathrm{d}t = \int_\infty^{-\infty} \text{TPR}(u)\text{FPR}'(u)\,\mathrm{d}u,$$

with a simple change of variable, and therefore

$$\text{AUC} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}(v>u)m_1'(v)m_0'(u)\,\mathrm{d}v\,\mathrm{d}u = \mathbb{P}\big(M_1 > M_0\big),$$

where M_1 is the score for a positive instance and M_0 is the score for a negative instance. Therefore, as discussed in Calders and Jaroszewicz (2007), the AUC is very close to the Mann–Whitney U test, used to test the null hypothesis that, for randomly selected values Z_0 and Z_1 from two populations, the probability of

Z_0 being greater than Z_1 is equal to the probability of Z_1 being greater than Z_0, that is written, empirically

$$\frac{1}{n_0 n_1} \sum_{i:z_i=0} \sum_{j:z_j=1} \mathbf{1}(z_j > z_i).$$

	training data				validation data			
	GLM	CART	GAM	RF	GLM	CART	GAM	RF
$\widehat{m}(\boldsymbol{x}, s)$	85.3	82.7	86.1	100.0	86.0	81.4	86.3	83.6
$\widehat{m}(\boldsymbol{x})$	85.0	82.7	85.9	100.0	85.5	81.4	85.9	83.6

Table 4.4: AUC for various models on the `toydata2`.

Therefore, the ROC curve and the AUC have more to do with the rankings of individuals than values of predictions: if h is some strictly increasing function, m and $h \circ m$ have the same ROC curves. Thus, accuracy for classifiers has to do with the ordering of predictions, not their actual value (that will be related to calibration, and discussed in section 4.3.3). Austin and Steyerberg (2012) used the term "concordance statistic" for the AUC. Note that the AUC is also related to Gini index.

On the `germancredit` dataset, the variable of interest y is the default, taking values in $\{0, 1\}$, and the protected attribute p is the gender (binary, with male and female). We consider four models, either on both $\boldsymbol{x}$ and p, or only on $\boldsymbol{x}$ (without the sensitive attribute, corresponding to fairness through unawareness, as defined in chapter 8) : (1) a logistic regression, or GLM (2) a classification tree, (3) a boosted model and (4) a bagging model, corresponding to a random forest. The AUC for those models is given in Table 4.5 and ROC curves are in Figure 4.27

	GLM	tree	boosting	bagging
$\widehat{m}(\boldsymbol{x}, s)$	79.339%	72.922%	79.488%	77.914%
$\widehat{m}(\boldsymbol{x})$	78.992%	72.922%	79.035%	78.287%

Table 4.5: AUC for various models on the `germancredit` (validation subsample) dataset, predicting the default. On top, models including the sensitive variable (the gender), and below, models without the sensitive variable, corresponding to fairness through unawareness.

If y is a categorical variable in more than 2 classes, different scoring rules can be used, as suggested by Kull and Flach (2015).

4.3 Calibration of Predictive Models

Somehow, calibration is related to ideas mentioned previously, since we simply want that "probabilities" given by a model make sense. If that property was somewhat natural in GLMs, it is usually not the case with machine learning ones. According to Wang et al. (2021), "*deep neural networks tend to be overconfident and poorly calibrated after training*", and more recently, Guo et al. (2017) "*have shown that modern neural networks are poorly calibrated and over-confident despite having better performance*" (as wrote in Müller et al. (2019)).

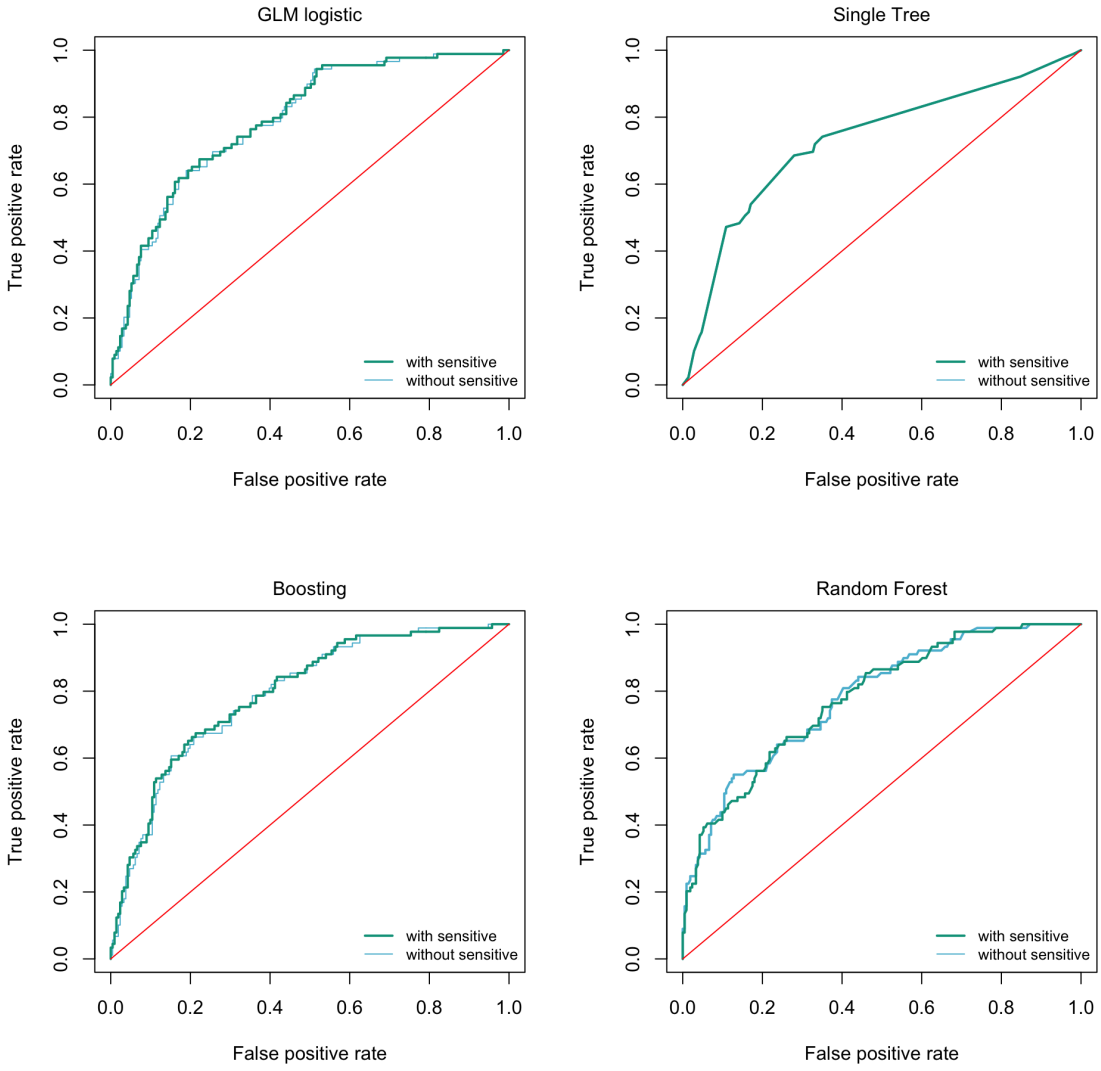


Figure 4.27: ROC curves $\widehat{C}_n(t)$, for various models on the `germancredit` (validation) dataset, predicting a default. Large lines correspond to models including the sensitive variable (the gender), and thin lines, models without the sensitive variable, corresponding to "fairness through unawareness".

4.3.1 From accuracy to calibration

According to Heckert et al. (2002), "*accuracy is a qualitative term referring to whether there is agreement between a measurement made on an object and its true (target or reference) value. Bias is a quantitative term describing the difference between the average of measurements made on the same object and its true*

value. ". And "calibration" could be related to this "bias". Silver (2012) raised that issue in the context of probabilities and forecasting, "*out of all the times you said there was a 40 percent chance of rain, how often did rain actually occur? If, over the long run, it really did rain about 40 percent of the time, that means your forecasts were well calibrated.*" The same idea can be found in Scikit Learn (2017), "*well calibrated classifiers are probabilistic classifiers for which the output can be directly interpreted as a confidence level. For instance, a well calibrated (binary) classifier should classify the samples such that among the samples to which it gave a*[predicted probability] *value close to 0.8, approximately 80% actually belong to the positive class.*"

As defined in Kuhn et al. (2013), when seeking for well-calibration, "*we desire that the estimated class probabilities are reflective of the true underlying probability of the sample* ". This was initially written in the context of posterior distribution (in a Bayesian setting), as in Dawid (1982): "*suppose that a forecaster sequentially assigns probabilities to events. He is well calibrated if, for example, of those events to which he assigns a probability 30 percent, the long-run proportion that actually occurs turns out to be 30 percent.* " Van Calster et al. (2019) gave a nice state of the art.

4.3.2 Lorenz and Concentration Curves

Before defining properly "calibration", let us mention Lorenz curve, as well as "concentration curves", in the context of actuarial models. Frees et al. (2011, 2014b) suggested to use Lorenz curve and Gini index to provide accuracy measures of for a classifier, and a regression. In economics, the Lorenz curve is a graphical representation of the distribution of income or of wealth, and it was popularized to represent inequalities in a wealth distribution. It simply shows the proportion of overall income, or wealth, owned by the bottom u% of the people. The first diagonal (obtained if $L(u)$) is called "*line of perfect equality*," in the sense that the bottom u% of population always gets u% of the income. Inequalities arise when the bottom u% of population gets v% of the income, with $v < u$ (because population is sorted, based on incomes, the Lorenz curve cannot rise above the line of perfect equality). Formally, we have the following definition

Definition 4.3.1 (Lorenz curve) *Lorenz (1905). If Y is a positive random variable, with quantile function $Q = F^{-1}$, where F is the cumulative distribution function,*

$$L(u) = \frac{\mathbb{E}\big[Y \cdot \mathbf{1}(Y \leq Q(u))\big]}{\mathbb{E}\big[Y\big]} = \frac{\int_0^u Q(t)\mathrm{d}t}{\int_0^1 Q(t)\mathrm{d}t},$$

and the empirical version is

$$\widehat{L}_n(u) = \frac{\sum_{i=1}^{[nu]} y_{(i)}}{\sum_{i=1}^{n} y_{(i)}},$$

for a sample $\{y_1, \cdots, y_n\}$, where $y_{(i)}$ denote the order statistics, in the sense that $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n-1)} \leq y_{(n)}$.

On Figure 4.28, we can visualize some Lorenz curves on various models fitted on the `toydata2` (validation) dataset. Those functions can be obtained using the `Lc` function in the `ineq` R package.

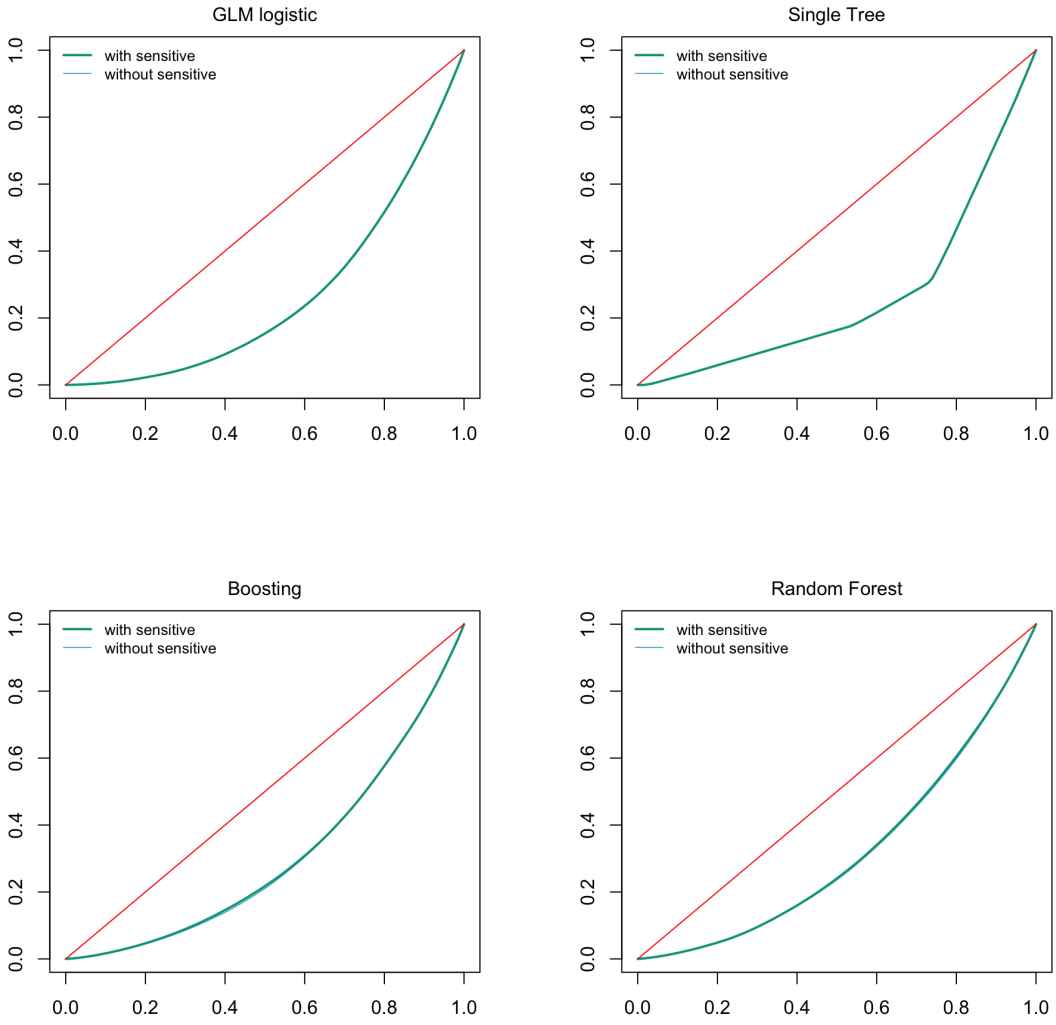


Figure 4.28: Lorenz curves $\widehat{L}_n(u)$ for various models on the `germancredit` (validation) dataset. Large **lines** correspond to models including the sensitive variable (the gender), and thin lines, models without the sensitive variable, corresponding to "fairness through unawareness".

In order to measure the concentration of wealth Lorenz (1905) suggested the following function $[0,1] \to [0,1]$, for positive variable y_i.

The expression of Lorenz curve, $u \mapsto L(u)$, reminds of the expected shortfall, where the denominator is the quantile, and not the expected value. For example, if Y has a lognormal distribution $LN(\mu, \sigma^2)$, $L(u) = \Phi\big(\Phi^{-1}(u) - \sigma\big)$, while if Y has a Pareto distribution with tail index α (i.e. $\mathbb{P}[Y > y] = (x/x_0)^{-\alpha}$),

$L(u) = 1 - (1-u)^{(\alpha-1)/\alpha}$ (see Cowell (2011)). It is rather common to summarize the Lorenz curve into a single parameter, Gini index, introduced in Gini (1912), corresponding a linear transformation of the area under the Lorenz curve. More precisely, $G = 1 - 2\text{AUC}$, so that a small AUC correspond to a distribution with small concentration $G \sim 1$, and AUC $= 1/2$ correspond to identical values for y_i, and $G = 0$. Thus,

$$G = 1 - 2\int_0^1 L(u)\mathrm{d}u = \frac{2\mathrm{Cov}\big[Y, F(Y)\big]}{\mathbb{E}\big[Y\big]} = \frac{\mathbb{E}\big[|Y - Y'|\big]}{2\mathbb{E}\big[Y\big]}$$

where Y' is an independent version of Y, and the empirical version is

$$\widehat{G}_n = \frac{1}{2\overline{y}}\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n |y_i - y_j|.$$

The numerator in the computation of Gini index is the mean absolute difference, also named "Gini mean difference" in Yitzhaki and Schechtman (2013),

$$\gamma = \mathbb{E}\big[|Y - Y'|\big] \text{ where } Y, Y' \sim F \text{ are independent copies.}$$

If this framework can be used in the case where $y \in \mathbb{R}_+$, it is also possible to consider the case where $y \in \{0, 1\}$. Hence, if F corresponds to a Bernoulli variable $\mathcal{B}(p)$,

$$\gamma = 2p(1-p) = 1 - p^2 - (1-p)^2,$$

that corresponds to Brier score as in Gneiting and Raftery (2007). Consider now more generally some classification problem, with a training sample $(\boldsymbol{x}_i, y_i)$, Murphy (1973), Murphy and Winkler (1987), Dawid (2004) and more recently Pohle (2020), introduced a so-called "Murphy decomposition." For a squared loss,

$$\mathbb{E}\big[(X-Y)^2\big] = \underbrace{\mathrm{Var}[Y]}_{\text{UNC}} - \underbrace{\mathrm{Var}\big[\mathbb{E}[Y|X]\big]}_{\text{RES}} + \underbrace{\mathbb{E}\big[\big(X - \mathbb{E}[Y|X]\big)^2\big]}_{\text{CAL}}$$

where the first term, UNC, is the unconditional entropy uncertainty, which represents the "uncertainty" in the variable of interest and does not depend on the predictions (also called "pure randomness"); the second component, RES, is called "resolution", and corresponds to the part of the uncertainty in Y that can be explained by the prediction, so it should reduces the expected score by that amount (compared to the unconditional forecast); the last part, CAL corresponds to "miscalibration," or "reliability."

As explained in Murphy (1996), in original decomposition was derived as a partition of Brier score, which can be interpreted as the MSE for probability forecasts. For Brier score in a binary setting, as discussed in Bröcker (2009), with a calibration function $\pi(p) = \mathbb{P}[Y = 1|p]$, where p is the forecast probability, and y the observed value, if $\overline{\pi} = \mathbb{P}[Y = 1]$, Brier score is decomposed as

$$\underbrace{\overline{\pi}(1-\overline{\pi})}_{\text{UNC}} - \underbrace{\mathbb{E}\big[\pi(p) - \overline{\pi}\big]^2}_{\text{RES}} + \underbrace{\mathbb{E}\big[p - \pi(p)\big]^2}_{\text{CAL}}.$$

A second way of assessing models is to study the Kullback-Leibler divergence (see Definition 3.3.7). For a classification problem, if $\widehat{m}$ is the fitted model while μ is the true model, Kullback-Leibler divergence (also named "discrimination information," as in Kullback (2004)) can be defined as follows

$$D_{\mathrm{KL}}(\mu(\boldsymbol{x}_i)\|\widehat{m}(\boldsymbol{x}_i)) = \mu(\boldsymbol{x}_i)\log\left(\frac{\mu(\boldsymbol{x}_i)}{\widehat{m}(\boldsymbol{x}_i)}\right) + \big(1-\mu(\boldsymbol{x}_i)\big)\log\left(\frac{1-\mu(\boldsymbol{x}_i)}{1-\widehat{m}(\boldsymbol{x}_i)}\right).$$

Given a collection of fitted models, $\widehat{m} \in \mathcal{M}$, we select the one that minimize the divergence, i.e.

$$\min_{\widehat{m}\in\mathcal{M}} \left\{ \frac{-1}{n} \sum_{i=1}^{n} \left[\mu(\boldsymbol{x}_i) \log\left(\widehat{m}(\boldsymbol{x}_i)\right) + \left(1-\mu(\boldsymbol{x}_i)\right) \log\left(1-\widehat{m}(\boldsymbol{x}_i)\right) \right] \right\}.$$

Since the true model is m is unknown, this problem cannot be solved, and a "natural" idea is to replace $m(\boldsymbol{x}_i)$ by the true observed values, that is solve

$$\min_{\widehat{m}} \left\{ \frac{-1}{n} \sum_{i=1}^{n} \left[y_i \log\left(\widehat{m}(\boldsymbol{x}_i)\right) + \left(1-y_i\right) \right] \log\left(1-\widehat{m}(\boldsymbol{x}_i)\right) \right\}.$$

which means that we select the model that minimize Bernoulli deviance loss, in the training sample.

Logistic regression satisfies the "balance property," that could be seen as some global "unbiased" estimator property, as discussed in Section 3.3.2

$$\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{m}(\boldsymbol{x}_i) \right) = 0$$

on the training sample.

It is possible to use Lorenz curve on scores $\{\widehat{m}(\boldsymbol{x}_1), \cdots, \widehat{m}(\boldsymbol{x}_n)\}$. Let $\widehat{F}_m$ denote the empirical cumulative distribution function,

$$u \mapsto \widehat{F}_m(u) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\widehat{m}(\boldsymbol{x}_i) \leq u),$$

while Lorenz curve is

$$u \mapsto \widehat{L}(u) = \frac{\sum_{i=1}^{[nu]} p_{(i)}}{\sum_{i=1}^{n} p_{(i)}}$$

Where $m_i = \widehat{m}(\boldsymbol{x}_i)$ and $m_{(i)}$ is the order statistics. Observe that some mirrored Lorenz curve can also be used, with $\widehat{M} : u \mapsto 1 - \widehat{L}(1-u)$,

$$u \mapsto \widehat{M}(u) = \frac{\sum_{i=[n(1-u)]}^{n} m_{(i)}}{\sum_{i=1}^{n} m_{(i)}} = \frac{\sum_{i=1}^{n} m_i \mathbf{1}(m_i > m_{i_u^\star})}{\sum_{i=1}^{n} m_i}$$

with $i_u^\star = [n(1-u)]$ so that the curve should be in the upper corner (as for the ROC curve), as suggested in Tasche (2008). If m_i is identical for all individual, $u \mapsto \widehat{M}(u)$ is on the first diagonal. With a "perfect model" (or "saturated model"), when $m_i = y_i$, we have a a piece-wise linear function, from $(0,0)$ to $(\overline{y},1)$ and from $(\overline{y},1)$ to $(1,1)$. In the context of motor insurance, it does not mean that we have perfectly estimated the probability, but it means that our model was able to predict without any mistake *who* would have claimed a loss.

Ling and Li (1998) introduced a "gain curve," also called "(cumulative) lift curve," defined as

$$u \mapsto \widehat{\Gamma}(u) = \frac{\sum_{i=1}^{n} y_i \mathbf{1}(m_i > m_{i_u^\star})}{\sum_{i=1}^{n} y_i}$$

where observed outcomes y_i are aggregated in the ordering of their prediction $m_i = \widehat{m}(\boldsymbol{x}_i)$.

Definition 4.3.2 (Concentration curve) *Gourieroux and Jasiak (2007) and Frees et al. (2011, 2014b). If Y is a positive random variable, observed jointly with $\boldsymbol{X}$, and if $m(\boldsymbol{X})$ and $\mu(\boldsymbol{X})$ denote a predictive model, and the regression curve, the concentration curve of μ with respect to m is*

$$\Gamma(u) = \frac{\mathbb{E}\big[\mu(\boldsymbol{X}) \cdot \mathbf{1}(m(\boldsymbol{X}) \leq Q_m(u)\big]}{\mathbb{E}\big[\mu(\boldsymbol{X})\big]} = \frac{\mathbb{E}\big[Y \cdot \mathbf{1}(m(\boldsymbol{X}) \leq Q_m(u))\big]}{\mathbb{E}\big[Y\big]},$$

where Q_m is the quantile function of $m(\boldsymbol{X})$, i.e.

$$Q_m(u) = \inf\big\{y : \mathbb{P}[m(\boldsymbol{X}) \leq t] > u\big\},$$

and the empirical version is

$$\widehat{\Gamma}_n(u) = \frac{\sum_{i=1}^{[nu]} y_{(i):m}}{\sum_{i=1}^{n} y_{(i):m}},$$

for a sample $\{y_1, \cdots, y_n\}$, where $y_{(i):m}$ are ordered with respect to m, in the sense that $m(\boldsymbol{x}_{(1):m}) \leq m(\boldsymbol{x}_{(2):m}) \leq \cdots \leq m(\boldsymbol{x}_{(n-1):m}) \leq m(\boldsymbol{x}_{(n):m})$.

This function could be seen as the extension of the Lorenz curve, in the sense that $L(u)$ was the proportion of wealth owned by the lower u% of the population, $\Gamma(u)$ represent the proportion of the total true expected loss, corresponding to the u% of the policyholder with smallest u% premiums (pure premium, computed using model m). Such a function provides more information than some simple "accuracy" plot (as the ROC curve), and it can be related to $\mathbb{E}[Y|m(\boldsymbol{X}) \leq t]$, or more interesting, $\mathbb{E}[Y|m(\boldsymbol{X}) = t]$, that corresponds to a "calibration" curve.

4.3.3 Calibration, Global and Local Biases

According to Kuhn et al. (2013), Section 11.1, "*we desire that the estimated class probabilities are reflective of the true underlying probability of the sample. That is, the predicted class probability (or probability-like value) needs to be well-calibrated. To be well-calibrated, the probabilities must effectively reflect the true likelihood of the event of interest*." That is the definition we will consider here

Definition 4.3.3 (Well-calibrated (1)) *Van Calster et al. (2019), Krüger and Ziegel (2021). The forecast X of Y is a well-calibrated forecast of Y if $\mathbb{E}(Y|X) = X$ almost surely, or $\mathbb{E}[Y|X = x] = x, \ \forall x$.*

Definition 4.3.4 (Well-calibrated (2)) *Zadrozny and Elkan (2002); Cohen and Goldszmidt (2004). The prediction $m(\boldsymbol{X})$ of Y is a well-calibrated prediction if $\mathbb{E}[Y|m(\boldsymbol{X}) = \widehat{y}] = \widehat{y}, \ \forall \widehat{y}$.*

Definition 4.3.5 (Calibration plot) *The calibration plot associated with model m is the function $\widehat{y} \mapsto \mathbb{E}(Y|m(\boldsymbol{X}) = \widehat{y})$. The empirical version is some local regression on $\{y_i, m(\boldsymbol{x}_i)\}$.*

Such plot can be obtained either using binning techniques, as in Hosmer and Lemesbow (1980). Given a partition $\{I_1, \cdots, I_k, \cdots\}$ of $\mathcal{Y}$, we would have a well calibrated model if, for any k,

$$\frac{\sum_{i=1}^{n} m(\boldsymbol{x}_i)\mathbf{1}(m(\boldsymbol{x}_i) \in I_k)}{\sum_{i=1}^{n} \mathbf{1}(m(\boldsymbol{x}_i) \in I_k)} \approx \frac{\sum_{i=1}^{n} y_i\mathbf{1}(m(\boldsymbol{x}_i) \in I_k)}{\sum_{i=1}^{n} \mathbf{1}(m(\boldsymbol{x}_i) \in I_k)}$$

One can also consider k-nearest neighbors approach, or a local regression (using the `locfit` R package), as in Denuit et al. (2021). On Figures 4.29, 4.30 and 4.31, we can visualize some empirical calibration plots on the `toydata2` (Figure 4.29), for the GLM and the random forest, and on the four models on the `germancredit` dataset (Figures 4.30 and 4.31). Function `calibPlot` in package `predtools` can also be used.

We introduced in Section 2.5.2 the idea of a "balanced" model, with Definition 2.5.2, which corresponds to a property of "globally unbiased",

Definition 4.3.6 (Globally unbiased model m**)** *Denuit et al. (2021) Model m is globally unbiased if* $\mathbb{E}[Y] = \mathbb{E}[m(\boldsymbol{X})]$.

But it is possible to consider a local version,

Definition 4.3.7 (Locally unbiased model m**)** *Denuit et al. (2021) Model m is locally unbiased at* $\widehat{y}$ *if* $\mathbb{E}[Y|m(\boldsymbol{X}) = \widehat{y}] = \widehat{y}$.

It means that the model is balanced "locally" on the group of individuals such that $\boldsymbol{x}$'s satisfy $m(\boldsymbol{X}) = \widehat{y}$ (and therefore no cross-financing between groups).

On the `germancredit` dataset, the variable of interest y is the default, taking values in $\{0, 1\}$, and the protected attribute p is the gender (binary, with male and female). We consider four models, either on both $\boldsymbol{x}$ and p, or only on $\boldsymbol{x}$ (without the sensitive attribute, corresponding to fairness through unawareness, as defined in chapter 8): (1) a logistic regression, or GLM (2) a classification tree, (3) a boosted model and (4) a bagging model, corresponding to a random forest. On Figure 4.30, we can visualize the calibration graphs for the four models, when $\boldsymbol{x}$ and p are used, and on Figure 4.31, when $\boldsymbol{x}$ only is used.

In the context of GLMs, as mentioned in Vidoni (2003), some theoretical results can be derived. Consider some sample $(y_i, \boldsymbol{x}_i)$ such that Y_i is supposed to have distribution f, in the exponential family, as discussed in section 3.3.2, i.e.

$$f(y_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\varphi} + c(y_i, \varphi)\right).$$

In the GLM framework, different quantities are used, namely the natural parameter for y_i (θ_i), the prediction for y_i ($\widehat{y}_i = \mu_i = \mathbb{E}(Y_i) = b'($
the score associated with y_i ($\eta_i = \boldsymbol{x}_i^\top\boldsymbol{\beta}$) and link function : g such that $\eta_i = g(\mu_i) = g(b'(\theta_i))$. The first order conditions can be written, using the standard chain rule technique

$$\frac{\partial \log \mathcal{L}_i}{\partial \beta_j} = \frac{\partial \log \mathcal{L}_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = 0$$

where the four terms are

$$\frac{\partial \log \mathcal{L}_i}{\partial \beta_j} = \frac{y_i - \mu_i}{\varphi} \cdot \frac{1}{V(\mu_i)} \cdot x_{i,j} \cdot \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1} = (g'(\mu_i))^{-1}.$$

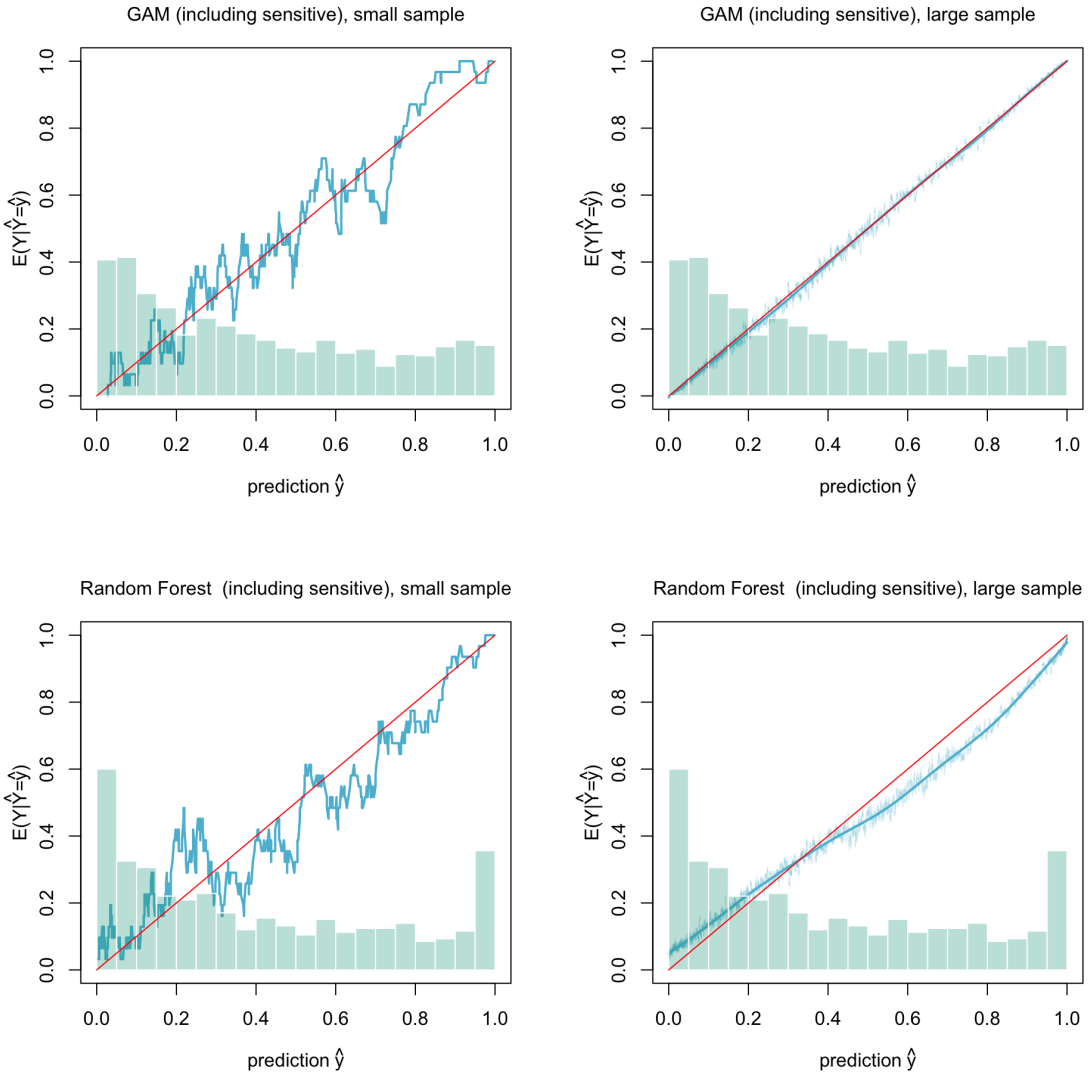


Figure 4.29: The blue line is the empirical calibration plot $\widehat{y} \mapsto \mathbb{E}(Y|m(\boldsymbol{X}) = \widehat{y})$ on two models estimated on the `toydata2` dataset (based on k-nearest neighbors), with the initial validation dataset on the left ($n = 1,000$) and some simulated larger sample ($n = 100,000$) on the right, with the GAM on top, and random forest model below. Bars in the back correspond to some histogram of $m(\boldsymbol{X})$.

In that case, with canonical link $g_\star = b'^{-1}$, i.e. $\eta_i = \theta_i$, the first order condition is (with notation $\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{\mu}}$),

$$\nabla \log \mathcal{L} = \boldsymbol{X}^\top(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{0},$$

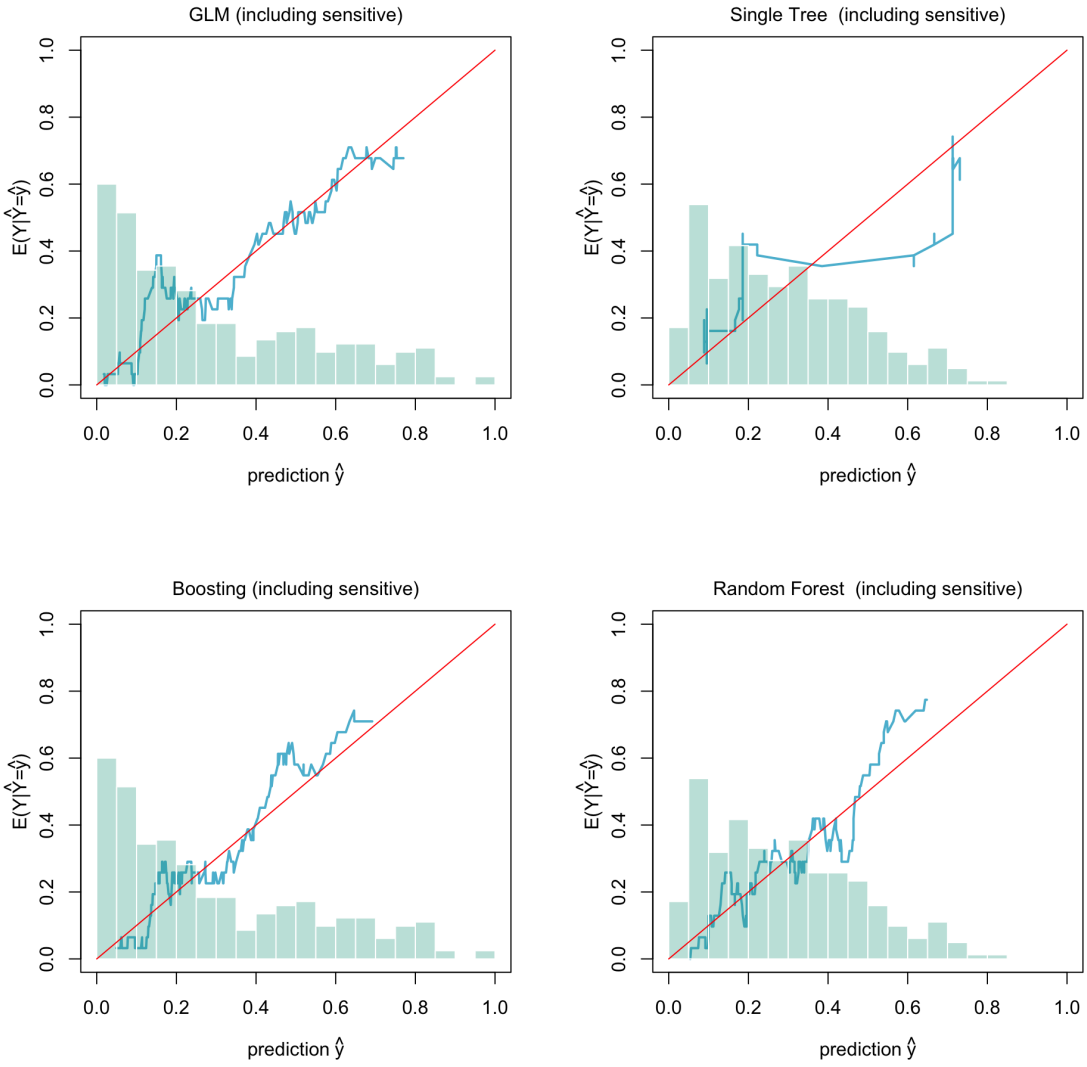


Figure 4.30: The blue line is the empirical calibration plot $\widehat{y} \mapsto \mathbb{E}[Y|\widehat{Y} = \widehat{y}]$, on the `germancredit` dataset (based on k-nearest neighbors), with a GLM, a classification tree (on top), a boosting and a bagging model (below), when $\widehat{m}(\boldsymbol{x}, s)$, including the sensitive attribute (here the gender, s).

so, if there is an intercept, $\mathbf{1}^\top(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = 0$, i.e. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i$, which is the empirical version (training dataset) of $\mathbb{E}[Y] = \mathbb{E}[\widehat{Y}]$. If a non-canonical link function is used (which is the case for for the Tweedie model or the

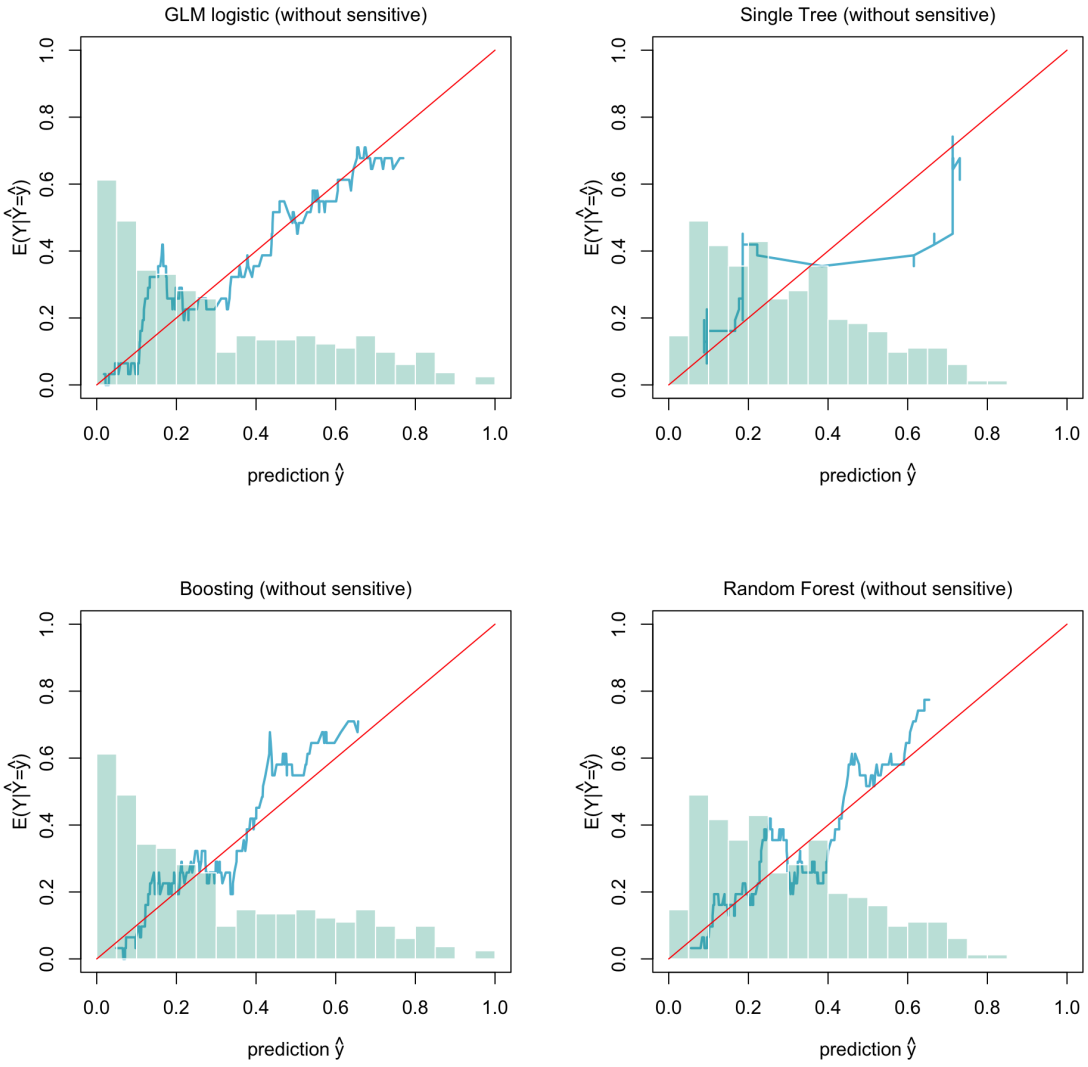


Figure 4.31: Calibration plot $\widehat{y} \mapsto \mathbb{E}[Y|\widehat{Y} = \widehat{y}]$, on the `germancredit` dataset (based on k-nearest neighbors), with a GLM, a classification tree (on top), a boosting and a bagging model (below), when $\widehat{m}(\boldsymbol{x})$, excluding the sensitive attribute (the gender s).

gamma model with a logarithm link function), the first order condition is

$$\nabla \log \mathcal{L} = \boldsymbol{X}^T \boldsymbol{\Omega}(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{0},$$

where $\boldsymbol{\Omega}$ is a diagonal matrix[2], so we no longer have $\mathbb{E}[Y] = \mathbb{E}[\widehat{Y}]$ (unless we consider a change of measure).

Propoition 4.3.1 *In the GLM framework with the canonical link function, $\widehat{m}(\boldsymbol{x}) = g_\star^{-1}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})$ is balanced, or globally unbiased, but possibly locally biased.*

This property can be visualized in Table 4.6.

	training data					validation data				
	$\overline{y}$	GLM	CART	GAM	RF	$\overline{y}$	GLM	CART	GAM	RF
$\widehat{m}(\boldsymbol{x}, s)$	8.73	8.73	8.73	8.73	8.27	8.55	9.05	9.03	8.84	8.70
$\widehat{m}(\boldsymbol{x})$	8.73	8.73	8.73	8.73	8.29	8.55	9.05	9.03	8.84	8.73

Table 4.6: Balance property on the `frenchmotor`, where y is the occurence of a claim within the year, with $\frac{1}{n}\sum_{i=1}^{n} \widehat{m}(\boldsymbol{x}_i, s_i)$ on top, and $\frac{1}{n}\sum_{i=1}^{n} \widehat{m}(\boldsymbol{x}_i)$ below (in %). Recall that on the training dataset $\frac{1}{n}\sum_{i=1}^{n} y_i = 8.73\%$.

If a model is not well-calibrated, several techniques have been considered in the literature, such as a logistic correction, Platt scaling, and the "isotonic regression", as discussed in Niculescu-Mizil and Caruana (2005b), in the context of boosting. Friedman et al. (2000) showed that `adaboost` builds an additive logistic regression model, and that the optimal m satisfies

$$m^\star(x) = \frac{1}{2}\log\frac{\mathbb{P}[Y=1|\boldsymbol{X}=\boldsymbol{x}]}{\mathbb{P}[Y=0|\boldsymbol{X}=\boldsymbol{x}]} = \frac{1}{2}\log\frac{\mu(\boldsymbol{x})}{1-\mu(\boldsymbol{x})},$$

which suggests applying a "logistic correction" in order to get back the conditional probability. Platt et al. (1999) suggested the use of a sigmoid function

$$m'(\boldsymbol{x}) = \frac{1}{1+\exp[am(\boldsymbol{x})+b]},$$

where a and b are obtained using maximum likelihood techniques. Finally, The isotonic (or monotonic) regression, introduced in Robertson et al. (1988) and Zadrozny and Elkan (2001), that consider a nonparametric increasing transformation from $m(\boldsymbol{x}_i)$'s to y_i's (that can be performed either with `Iso` package, or `rfUtilities` R package, and `probability.calibration` function).

On Figure 4.32, we can visualize smooth calibration curves (estimated using using a local regression, with the `locfit` package) on three models estimated from the from the `toydata1` dataset. On top, crude calibration curves, fitted on $\{\widehat{y}_i, y_i\}$'s, and below, calibration curves after some "isotonic correction" (obtained with the `probability.calibration` function from the `rfUtilities` R package).

[2] $\boldsymbol{\Omega} = \boldsymbol{W}\boldsymbol{\Delta}$, where $\boldsymbol{W} = \text{diag}\big((V(\mu_i)g'(\mu_i)^2)^{-1}\big)$ and $\boldsymbol{\Delta} = \text{diag}\big(g'(\mu_i)\big)$, so that we recognize Fisher information - corresponding to the Hessian matrix (up to a negative sign) – $\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}$.

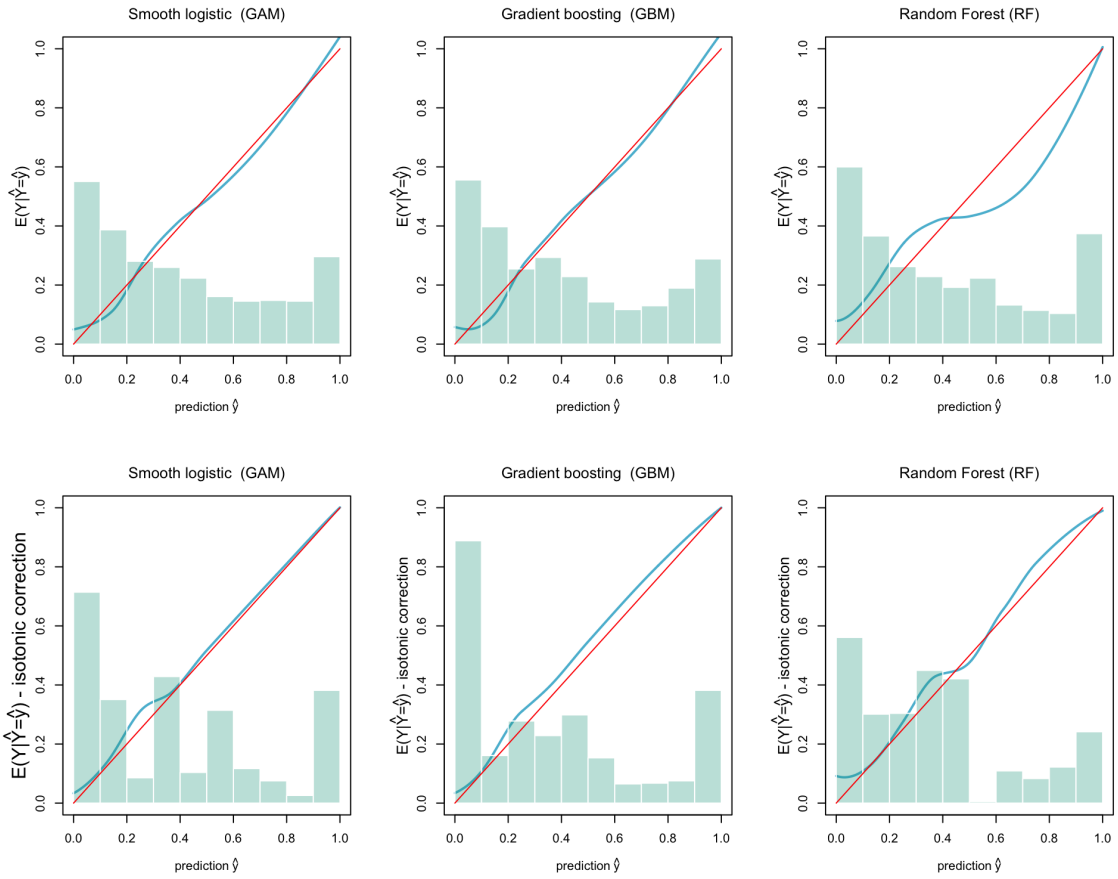


Figure 4.32: Calibration plots $\widehat{y} \mapsto \mathbb{E}(Y|m(\boldsymbol{X}) = \widehat{y})$ (based on smooth local regression) on three models estimated on the `toydata1` dataset, on top (GLM, GBM and RF, from the left to the right), and the calibration after the "isotonic correction" below.

Part II

Data

"*Insurance companies are in the business of discrimination. Insurers attempt to segregate insureds into separate risk pools based on the differences in their risk profiles, first, so that different premiums can be charged to the different groups based on their differing risks and, second, to incentivize risk reduction by insureds. This is why we let insurers discriminate. There are limits, however, to the types of discrimination that are permissible for insurers. But what exactly are those limits and how are they justified?,*" Avraham et al. (2014)

– Tu la troubles, reprit cette bête cruelle,
Et je sais que de moi tu médis l'an passé.
– Comment l'aurais-je fait si je n'étais pas né ?
Reprit l'Agneau, je tette encor ma mère.
– Si ce n'est toi, c'est donc ton frère.
– Je n'en ai point.
– C'est donc quelqu'un des tiens
de La Fontaine (1668), *Le Loup et l'Agneau*.

– You roil it, said the wolf; and, more, I know
You cursed and slander'd me a year ago.
– O no! how could I such a thing have done!
A lamb that has not seen a year,
A suckling of its mother dear?
– Your brother then.
– But brother I have none.
– Well, well, what's all the same,
Twas some one of your name
de La Fontaine (1668), *The Wolf and the Lamb*[3].

[3](Αισωπος) Aesop's original fables did not use this motivation as *casus belli*, even if the translation by Jacobs (1894) is "*if it was not you, it was your father, and that is all one*".

Chapter 5

What Data?

Actuaries now collect all kinds of information about policyholders, which can be used to refine a premium calculation, but also to carry out prevention operations. We will return here to the choice of relevant variables in pricing, with emphasis on actuarial, operational, legal and ethical motivations. In particular, we will discuss the idea of capturing information on the behaviour of an insured person, and the difficult reconciliation with the strong constraints of privacy, but also of fairness.

In Part I, we used the symbolic notation $\boldsymbol{x} = (x_1, \cdots, x_k)$ to formalize "predictive variables" or "pricing variables", corresponding to variables that could be used to compute the premium. In this chapter, we will briefly discuss which variables could be used. As explained in Flanagan (1985), although a wide range of diverse criteria can be utilized for premium differentiation, insurers have traditionally relied upon the personal characteristics of an insured in motor insurance, including factors such as age, gender, and marital status. These personal characteristics are used as grounds for assessing risk because they are convenient: they are accessible, cheap, verifiable, stable and most of them are "*reliably correlated with many aspects of behaviour important in insurance*." For example, motor vehicle accident rates are statistically higher for very young drivers and for elderly drivers. This is because young drivers are more likely to have less driving experience and have a higher propensity to engage in high-risk driving conduct. And on the other hand, elderly drivers, are likely to suffer from sensory or cognitive impairments, which negatively impact their driving abilities, as explained in Kelly and Nielson (2006) or Brown et al. (2007). In this Chapter, we discuss data, personal data, sensitive data, behavioral data, data related to past historical data, protected data, etc. But before starting, it is important to keep in mind that "*all data is man-made*," as Christensen et al. (2016) wrote it. "*Somebody, at some point, decided what data to collect, how to organize it, how to present it, and how to infer meaning from it—and it embeds all kinds of false rigor into the process. Data has the same agenda as the person who created it, wittingly or unwittingly. For all the time that senior leaders spend analyzing data, they should be making equal investments to determine what data should be created in the first place. (...) Data has an annoying way of conforming itself to support whatever point of view we want it to support*."

5.1 Data (a Brief Introduction)

"*All data is credit data*," said Merrill (2012) in a conference. And if credit institutions collect a lot of "information," so do insurance companies, to assess and prevent risk, target ideal customers, accurately price policies, provide quotes, conduct investigations, follow trends, create new products, etc. Such information is now called "data" (from the Latin *datum*, *data* being the plural, past participle of *dare* "to give", used in the XVII-th century to designate a fact given as the basis for calculation, in mathematical problems[1])

Definition 5.1.1 (Data) *Wikipedia (2023). In common usage, data is a collection of discrete or continuous values that convey information, describing the quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted.*

A few years ago, the term "statistics" was also popular, as introduced in Achenwall (1749). It was based on Latin *statisticum (collegium)*," meaning "(lecture course on) state affairs", and Italian *statista* "one skilled in statecraft," and the German term "statistik" designated the analysis of data about the state, signifying the "science of state" (corresponding to "political arithmetic" in English). So in a sense, "statistics" was used to designate official data, collected for instance by the Census, with a strict protocol, as explained in Bouk (2022), while the term "data" would correspond to any kind of information that could be scraped, stored and used.

The "big data" hype has given us the opportunity to talk about its large volume, value, but also its variety (and all sorts of words beginning with the letter "v"). While for actuaries, data has often been "tabular data", corresponding to matrix numbers as seen in Part I, in the last few years the variety of data types has become more apparent. There will naturally be text, starting with a name, an address (which can be converted into spatial coordinates), but also possibly drug names in prescriptions, telephone conversations with an underwriter or claims manager, or in the case of companies, contracts with digitised clauses, etc. We can have images, such as a picture of the car after a fender-bender or of the roof of a house after a fire, medical images (X-rays, MRI), a satellite image of a field for a crop insurance contract, or of a village after a flood, etc. Finally, there will also be information associated with connected objects, data obtained from devices in a car fleet, from a water leak detector or from chimney monitoring and control devices. However, statistical summaries of "scores" are often based on this raw data (frequently not available to the insurer) such as the number of kilometres driven in a given week by a car insurance policyholder, or an acceleration score. This data, which is much more extensive than tabular variables with predefined fields (admittedly, sometimes with definition issues, as Desrosières (2016) points out), can provide sensitive information that can be exploited by a "black box" algorithm, possibly without the knowledge of the actuary.

In non-commercial insurance, the policyholder is an individual, a person (even in property insurance), and some part of the information collected will be considered as "*personal data*." In Europe, "*personal data*" is any information relating to a natural person who is identified or can be identified, directly or indirectly. The definition of personal data is specified in Article 4 of the GDPR (General Data Protection Regulation). This information can be an identifier (a name, an identification number, location data, for example) or one or more specific elements specific to the physical, physiological, genetic, mental, economic, cultural or social identity of the person. Among the (non-exhaustive) list given by the French CNIL[2], there may be the surname, first name, telephone number, license plate, social security number, postal address, e-mail, a voice recording, a photograph, etc. Such information is relevant, and important, in the insurance business.

[1] Euclid's treatise on plane geometry was named δεδομενα, translated as "data".

[2] CNIL is the *Commission Nationale de l'Informatique et des Libertés*, an independent French administrative regulatory body whose mission is to ensure that data privacy law is applied to the collection, storage, and use of personal data, in France.

"Sensitive data" is a subset of personal data, that include religious beliefs, sexual orientation, union involvement, ethnicity, medical status, criminal convictions and offences, biometric data, genetic information or sexual activities. According to the GDPR, in 2016[3], "*processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs or trade union membership, as well as processing of genetic data, bio-metric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning the sex life or sexual orientation of a natural person are prohibited.*" Such information will be considered "sensitive". In Europe, the 2018[4] "Convention 108" (or "*convention for the protection of individuals with regard to automatic processing of personal data*") further clarifies the contours.

5.2 Personal and Sensitive Data

5.2.1 Personal and Non-Personal Data

Classically, we can distinguish between different types of insurance contracts. Life insurance is a contract (between a policyholder and an insurance company) where the insurer promises to pay a designated beneficiary a sum of money upon the death of a person (or possibly some critical illness, including disability insurance). Health insurance is a type of insurance that covers the whole (or a part of the risk) of a person incurring medical expenses. Property insurance provides protection against most risks to property, such as fire, theft and some weather damage. Casualty insurance is the term which broadly encompasses insurance not directly concerned with life insurance, health insurance, or property insurance. There is clearly a difference between the first cases (life insurance or health insurance) where the object of the contract is a person, and the last cases (property and casualty, or general insurance) where the object of the contract is a material good (such as a house, or a car – in the latter case, we include damage caused by the use of the latter, including to third parties, including people). In the first two cases, actuaries consider the information provided by the person in question to be important, if not essential, for valuing the contract. In the other two cases, information about the car is important, but information about the driver is almost as important. For household insurance, personal data can be used to identify previous insurance coverage, to assess fraudulent activities (false claim), etc.

Obviously, a lot of personal information can be used to assess the probability to die, to become disabled, or to face medical expenses, including biological factors (attained age, and health information), genetic factors (sex, individual genotype, that will be discussed in Section 6.4 in the next chapter), lifestyle (smoking, drinking habits, etc.), social factors (marital status, social class, occupation, etc.), geographical factors (region, postal code).

5.2.2 Sensitive and Protected Data

From Avraham et al. (2013), in the U.S., we can get Table 5.1 provides a general perspective about variables that prohibited across all States, with the "*highest average level of strictness*," for different types of insurance. If there is a strong consensus about religion and "race", it is more complicated for other criteria.

Table 5.2 presents some variables traditionally used in car insurance, in the United States[5] and in Canada[6].

[3] See https://gdpr-info.eu/.

[4] See https://www.coe.int/en/web/data-protection/convention108-and-protocol.

[5] CA: California, HI: Hawaii, GA: Georgia, NC: North Carolina, NY: New York, MA: Massachusetts, PA: Pennsylvania, FL: Florida, TX: Texas.

[6] AL: Alberta, ON: Ontario, NB: New Brunswick, NL: Newfoundland and Labrador, QC: Québec.

	auto	P&C	disability	health	life
race (or origin)	✖	✖	✖	✖	✖
religion	✖	✖	✖	✖	✖
gender	○	●	○	○	●
sexual orientation	●	●	○	○	○
age	○*	●	○	○	●
credit score	○	○	●	●	●
zip code	○	○	●	●	●
genetics	●	●	○	✖	○+

Table 5.1: (U.S.) State Insurance Antidiscrimination Laws . A factor is either considered “permitted" or there is no specific regulation (●) usually because factor is not relevant to the risk. “prohibited" (✖) or there could be variation across states (○). * means limited regulation; + specifically permitted because of adverse selection.
(Source: Avraham et al. (2013).)

Unlike most European countries, that have a civil law culture (where the main source of law is found in legal codes), the Canadian provinces and the states of the United States of America have a common law system (where rules are primarily made by the courts as individual decisions are made). Québec uses a mixed law. Most states and provinces have documents listing the Prohibited Rating Variables, such as Alberta’s Automobile Insurance Rate Board (2022). This heterogeneity makes it possible to highlight the cultural character of what is considered as “discriminatory".

	CA	HI	GA	NC	NY	MA	PA	FL	TX	AL	ON	NB	NL	QC
Gender	✖	✖	●	✖	●	✖	✖	●	●	●	●	✖	✖	●
Age	✖	✖	●	✖*	●	✖	●	●	●	●*	●	✖	✖	●
Driving experience	●	✖	●	●	●	●	●	●	●	●	●	●	●	●
Credit history	✖	✖	●	●	●	✖	●*	●	●	✖*	✖	●*	✖	●
Education	✖	✖	✖	✖	✖	✖	●	●	●	●	●	●	●	●
Profession	✖	✖	✖	●	✖	✖	●	●	●	●	●	●	●	●
Employment	✖	✖	✖	●	✖	✖	●	●	●	●	●	●	●	●
Family	●	✖	●	●	●	✖	●	●	●	●	●	●	●	●
Housing	✖	✖	●	●	●	✖	●	●	●	✖	✖	●	●	●
Address/ZIP code	●	●	●	●	●	●	●	●	●	✖	✖	●	●	●

Table 5.2: A factor is considered “permitted" (●) when there are no laws or regulatory policies in the state or province that prohibit insurers from using that factor. Otherwise, it will be “prohibited" (✖). In North Carolina, age is only allowed when giving a discount to drivers 55 years of age and older. In Pensylvania, credit score can be used for new business and to reduce rates at renewal, but not to increase rates at renewal. In Alberta, credit score and driver’s license seniority cannot be used for mandatory coverage (but can be used on optional coverage). In Labrador, age cannot be used before 55, and beyond that, it must be a discount (as in North Carolina).
(Source: in the United States, The Zebra (2022) and in Canada, Insurance Bureau of Canada (2021).)

In Québec (Canada), as stated in section 20.1 of the Charter of Human Rights and Freedoms, “*a distinction based on age, sex or marital status is permitted when it is based on a factor that allows a risk*

Box 5.1 **Sensitive data**, European Commission (1995)

Sensitive data is data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs or trade union membership, data concerning health or sex life. The prohibition on the processing of sensitive data does not apply if:
(a) the data subject has given his explicit consent to the processing of those data, except where the laws of the Member State provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject's giving his consent; or
(b) processing is necessary for the purposes of carrying out the obligations and specific rights of the controller in the field of employment law in so far as it is authorized by national law providing for adequate safeguards; or
(c) processing is necessary to protect the vital interests of the data subject or of another person where the data subject is physically or legally incapable of giving his consent; or
(d) processing is carried out in the course of its legitimate activities with appropriate guarantees by a foundation, association or any other non-profit-seeking body with a political, philosophical, religious or trade-union aim and on condition that the processing relates solely to the members of the body or to persons who have regular contact with it in connection with its purposes and that the data are not disclosed to a third party without the consent of the data subjects; or
(e) the processing relates to data which are manifestly made public by the data subject or is necessary for the establishment, exercise or defence of legal claims.

to be determined. For example, an insurance company may ask you questions about your age and sex to determine your premium." In California, as noted by Butler and Butler (1989), many of the rating variables cannot be used by insurers if they are not causally related to the risk of accidents and their cost (we will get back on this "causal" issue in Chapter 7).

In Europe, Article 5-2 of Directive 2004 / 113 based on the Charter of Fundamental Rights provided that "*Member States may decide (...) to allow proportional differences in premiums and benefits for individuals where the use of sex is a determining factor in the assessment of risk, on the basis of relevant and accurate actuarial and statistical data*," as recalled by Laulom (2012).

In Box 5.1, we have a definition of "sensitive data", according to the European directive on the protection of personal data, in Europe.

A variable that has been intensively discussed is the "gender", from the Council Directive 2004 / 113 / EC (see Section 6.2). In France, Article L. 111-7 of the Insurance Code states that "*the Minister in charge of the economy may authorize, by decree, differences in premiums and benefits based on the taking into account of sex and proportionate to the risks when relevant and precise actuarial and statistical data establish that sex is a determining factor in the evaluation of the insurance risk*." Here, "*determining factor*" echoes the "causal" effect requiered in California. In Box 5.2, we have a description of legal aspects regarding discrimination in the context of insurance, in France by Rodolphe Bigot, that are explicitly described (age, family status and sexual orientation – page 150 – pregnancy, maternity and "social insecurity" – page 151 – and finally sex– page 151).

As Debet (2007) said, "*in order to fight against discrimination, it is still necessary to be able to identify it; in order to identify it, it seems natural to proceed with the statistical observation of differences, of diversity*." And as we have seen in Box 5.1, in European regulation "*sensitive data is data revealing racial or ethnic origin, political opinions, etc*." meaning that there could be discrimination if sensitive inferences can be made about individuals, as discussed in Wachter and Mittelstadt (2019).

Box 5.2 **Discrimination & Insurance in France**, by Rodolphe Bigot

To fight against discrimination, there are general prohibitions, formulated in the Penal Code and the Civil Code, and special prohibitions in insurance law, inserted in the Insurance Code. Both are subject to variable geometry adjustments linked to the specificities of the insurance business. In civil matters, a general prohibition applicable to insurance is set forth in Article 16-13 of the Civil Code, which provides that "*no one may be discriminated against on the basis of his or her genetic characteristics*." In criminal matters, the refusal or subordination of the provision of a service or good based on one of the criteria - discriminatory - listed in article 225-1 or 225-1-1 of the French Criminal Code constitutes reprehensible discrimination (C. pén., art. 225-1 et seq.). Since 2008, all direct or indirect discrimination based on such criteria has been covered (L. 27 May 2008; modified by L. 18 Nov. 2016 de modernisation de la justice du xxıe siècle), with a general derogation in the presence of differences in treatment "*justified by a legitimate aim*" and if "*the means of achieving this aim are necessary and appropriate*." A reversal of the burden of proof is provided for in actions brought before the civil courts (L. 27 May 2008, art. 4), unlike the rules of evidence applicable before the criminal courts, where it is up to the plaintiff to prove the discrimination.
First, a distinction based on **age** in particular (C. pén., art. 225-1 and 225-2) makes up a criminally sanctioned discrimination if it tends to apply a difference in treatment for access to insurance coverage or for the termination of insurance benefits: it is therefore prohibited in the case of risks of loss of employment for the purpose of securing a bank loan, for example. However, an exemption is provided for the pricing of life insurance contracts (death insurance and life annuities), by applying mortality tables (C. assur., art. A 132-18). The Defender of Rights admitted this discrimination for a 74-year-old health insurance applicant whose enrollment was refused because the policy limited it to 70 years of age (opinion no. MLD / 2012-150, November 16, 2012: "*when objectively justified by actuarial and statistical elements, age limits in access to a personal insurance contract do not constitute discrimination*" (The random nature of the insurance contract and the principles of risk selection and pooling may justify taking age into account in personal insurance. Article 2 of the law of May 27, 2008 provides for two other exceptions which should be confirmed by the proposed directive on equal treatment between persons (COM 2008 / 426 of July 2, 2008): 1) differences in treatment justified by a legitimate aim, if the means of achieving this aim are necessary and appropriate; 2) when the differences in treatment are provided for and authorized by the laws and regulations in force.
Second, discrimination based on **family status** or **sexual orientation** is also prohibited (C. pén., art. 225-1 and 225-2). In the case of a homosexual couple, this would include an employer's refusal to pay a death benefit to an employee's partner in a civil union or to the employee in a civil union in the event of his partner's death (...)

5.2.3 Sensitive Inferences

Sometimes, data is "derived" (e.g. country of residency derived from the subject's postcode) or "inferred" data (e.g. credit score, outcome of a health assessment, results of a personalization or recommendation process) and not "provided" by the data subject actively or passively, but rather created by a data controller or third party from data provided by the data subject and, in some cases, other background data, as explained in Wachter and Mittelstadt (2019). According to Abrams (2014), inferences can be considered personal data, since they derive from personal data. Another precaution that should be kept in mind relates to the distinction between what "*reveals*" and what "*is likely to reveal*," as Debet (2007) states (see also Van Deemter (2010), "*in praise of vagueness* "). Some information is self-reported, and some is inferred data. For instance it could be possible to ask for "sex at birth" to collect a `sex` variable, but in most cases a variable is based on civility (where "Mrs" or "Mr" are proposed), so the information is more a `gender` variable. But it can be more complex, since some models can be use to infer some information. One can imagine that "being pregnant" could be a sensitive information in many situations. This information exists in some health organizations databases (or health care reimbursements). But as shown by Duhigg (2019) (in *how companies learn your secrets*), there are organizations that try to infer this information, from purchases. This is the famous story of the man, in the Minneapolis area, who was surprised that coupons for various products for young mothers were addressed to his daughter. In this story, the inference by the model had been correct. We can imagine,

(...) Thirdly, discrimination based on a **woman's pregnancy and maternity** is prohibited. In insurance, women cannot be treated less favourably in terms of premiums and benefits as a result of expenses related to pregnancy and maternity (C. assur., art. L. 111-7, I, paragraph 2), because, in principle, both sexes are affected by maternity. In other words, insurance contracts, whatever the risk covered (health, provident fund, etc.), must no longer include differentiated treatment of these factors. Clauses instituting a longer waiting period for the coverage of hospitalization costs when the latter is consecutive to a pregnancy are therefore no longer legal. However, if there is relevant actuarial and statistical data, i.e. objective considerations, "*contractual provisions that are more favorable to women may remain*." Failure to comply with these provisions exposes the employer to the penalties provided for in the Criminal Code for acts of discrimination. (Lamy Assurances, 2021, n° 3806).
Fourth, the refusal to provide a good or service because of a person's **place of residence** makes up discrimination in the penal sense (C. pén., art. 225-1). While it is still possible to adjust the amount of the premium according to the place of insured's residence, the insurer is prohibited from refusing an applicant for insurance because of this factor. This does not apply to subscriptions or memberships made by residents of States a French insurer is not authorized to contract with (Lamy Assurances, 2021, No. 3808).
Fifth, refusal to insure on the basis of the **insured's social insecurity** is prohibited, which falls within the scope of the offence of discrimination for refusing to supply a good or service on the basis of "*particular vulnerability resulting from the economic situation, apparent or known to its author*" (C. pén., art. 225-1, amended by L. 24 June 2016 aimed at combating discrimination on the grounds of social insecurity). The insurer cannot reject a membership or subscription on this basis, but it can nevertheless take this factor into account to modulate the amount of the premium. Note that since June 26, 2016, indirect discrimination can result from a provision, criterion or practice that is neutral on the surface, but likely to result in a particular disadvantage for individuals. Moreover, there is a complex system of protection for vulnerable persons in insurance: "*the movement for vulnerable persons is a dance that alternates between special protection and undifferentiated protection*" (Noguéro (2010), p. 633). (...)

(...) Discrimination on the basis of **sex** is subject to a renewed regime dedicated to new contracts concluded as of December 21, 2012 (except for retirement, health and accident contracts subscribed by an employer and to which the employee is a compulsory member; C. séc. soc. L. 911-1) and aimed at uniformly applying the unisex rule - prohibiting any direct or indirect discrimination based on sex - to insurance contracts within the European Union (Parléani (2012), p. 563). Article A. 111-6 of the Insurance Code incorporated the European Commission's guidelines (Arr. 18 Dec. 2012, NOR: EFIT1238658A, relating to equality between men and women in insurance, JO 20 Dec., mod. by Arr. 3 Feb. 2014, NOR: EFIT1400411A, JO 11 Feb.).
The calculation of premiums and benefits falls within the scope of the unisex rule (in insurance operations classified, by reference to Article R. 321-1, in the branches: 1 "*Accidents (including occupational accidents and diseases)*" (art. a. 111-2), 2 "*Sickness*" (art. A. 111-3), 3 "*Land vehicle bodies (other than railways)*" (art. A. 111-4), 10 "*Civil liability for motor vehicles*" (Art. A. 111-4), 20 "*Life and death*" (Art. A. 111-5), 22 "*Insurance linked to investment funds*" (Art. A. 111-5), 23 "*Tontine operations*" (Art. A. 111-5), 26 "*Any operation of a collective nature defined in Section I of Chapter I of Title IV of Book IV*" (Art. A. 111-5)). By way of derogation, the criterion of sex may be used "*as a factor in the assessment of risks in general to collect, store and use information on sex or related to sex for internal provisioning and pricing, marketing and advertising, and pricing of reinsurance*." "*The use of sex as indirect discrimination is also allowed for the pricing of certain real risks, such as, for example, a differentiation of premiums based on the size of a car's engine, even though the most powerful cars are in fact bought more by men*" (Lamy Assurances, 2021, n° 3803).
To bring French law into compliance with European rules, Article L. 111-7 of the Insurance Code was rewritten with the law of 26 July 2013. A paragraph II bis was added: "*The derogation provided in the last paragraph of I is applicable to contracts and memberships in group insurance contracts concluded or made no later than December 20, 2012 and to such contracts and memberships tacitly renewed after that date. The derogation does not apply to contracts and memberships mentioned in the first paragraph of this IIa which have been substantially modified after December 20, 2012, requiring the parties' agreement, other than a modification that at least one of the parties cannot refuse*." In terms of collective supplementary employee benefits, no discrimination on the basis of gender can be made. However, insurers still have the possibility to offer options in policies or insurance products according to gender in order to cover conditions that exclusively or mainly concern men or women. Differentiated coverage is therefore possible for breast cancer, uterine cancer or prostate cancer.

for marketing reasons, that some insurers might be interested in knowing such information.

Recently, Lovejoy (2021) recalls that in June 2020, LinkedIn had a massive breach (exposing the data of 700 million users), with a database of records including phone numbers, physical addresses, geolocation data and... “inferred salaries". Again, knowing the salary of policyholders could be seen as interesting for some insurers (any micro-economic model used to study insurance demand is based on the “wealth" of the policyholder, as in Section 2.6.1). Much more information can be inferred from telematics data, albeit with varying degrees of uncertainty. As mentioned by Bigot et al. (2019), by observing that a person parks almost every Friday morning near a mosque, one could say that there is a high probability that he or she is Muslim (based on surveys of Muslim practices). But it is possible that this inference is completely wrong, and that this person actually goes to the gym, across the street from the mosque, and moreover, is a regular attendee. Facebook may be able to infer protected attributes such as sexual orientation, race (Speicher et al. (2018)), as well as political views (Tufekci (2018)) and impending suicide attempts (Constine (2017)), while third parties have used Facebook data to decide eligibility for loans (Taylor and Sadowski (2015)) and infer political positions on abortion (Coutts (2016)). Susceptibility to depression can also be inferred from Facebook and Twitter usage data. Microsoft can also predict Parkinson’s (Allerhand et al. (2018)) and Alzheimer’s (White et al. (2018)) disease from search engine interactions. Other recent invasive applications include predicting pregnancy by Target, assessing user satisfaction from mouse tracking (Chen et al. (2017)). Even if such inferences are most of the time impossible to understand (since they are provided by opaque black-box models), or refute, they could impact deeply on our private lives, reputation and identity, as discussed in Wachter and Mittelstadt (2019) Therefore, inferences is very close to data available, and to “attacks", as coined in the literature about privacy.

5.2.4 Privacy

As explained by Kelly (2021), “*often the location data is used to determine what stores people visit. Things like sexual orientation are used to determine what demographics to target*," (in a marketing context). Each type of data can reveal something about our interests and preferences, our opinions, our hobbies and our social interactions. For example, a MIT study[7] demonstrated how email metadata can be used to map our lives, showing the changing dynamics of our professional and personal networks. This data can be used to infer personal information, including a person’s background, religion or beliefs, political views, sexual orientation and gender identity, social relationships or health. For example, it is possible to infer our specific health conditions simply by connecting the dots between a series of phone calls. For Mayer et al. (2016), the law currently treats call content and metadata separately and makes it easier for government agencies to obtain metadata, in part because it assumes that it should not be possible to infer specific sensitive details about individuals from metadata alone.Chakraborty et al. (2013) reminds us that current approaches to privacy protection, typically defined in multi-user contexts, rely on anonymization to prevent such sensitive behaviour from being traced back to the user - a strategy that does not apply if the user’s identity is already known. In time, a tracking system may be accurate enough to place a person in the vicinity of a bank, bar, mosque, clinic or other privacy-sensitive location. In 2015, as told in Miracle (2016), Noah Deneau wondered if it would be possible to identify devout Muslim drivers in New York City by looking at anonymised data and inactive drivers during the five times of the day they are supposed to pray. He quickly searched for drivers who were not very active during the 30-45 minute Muslim prayer period and was able to find four examples of drivers who might fit this pattern. This brings to mind Gambs et al. (2010), who conducted an investigation on a dataset containing mobility data of taxi drivers in the San Francisco Bay Area. By finding places where the taxi’s GPS sensor was turned off for a long period of time (e.g. two hours), they were able

[7]Project https://immersion.media.mit.edu/.

to infer the interests of the drivers. For 20 of the 90 analysed users, they were able to locate a plausible home in a small neighbourhood. They even confirmed these results for 10 users by using a satellite view of the area: It showed the presence of a yellow taxi parked in front of the driver's supposed home. Dalenius (1977) introduced an interesting concept of privacy. Nothing about an individual should be learned from a dataset if it cannot be learned without having access to the dataset. We will return to this idea when we define the fairness criteria, and when we require that the protected variable s cannot be predicted from the data, and from the predictions.

5.2.5 Right to be Forgotten

The "*right to be forgotten*" is the right to have private information about a person be removed from various directories, as discussed in Rosen (2011), Mantelero (2013) or Jones (2016). It is also named "*right to oblivion*" in de Andrade (2012) and "*right to erasure*" in Ausloos (2020).

As explicitly mentioned in Mbungo (2014), "*individuals with a criminal record constitute a minority that cannot escape discrimination*." Criminal records can have a wide range of collateral consequences which often extend well beyond any sentence imposed by the courts, and in many countries, the mark of a criminal record follows a person long after he or she has served a prison sentence or paid a fine, Pager (2003, 2008). As explained in Henley (2014), a criminal record thus becomes a legitimate reason to refuse cover, or to massively increase of insurance premiums. According to Bath and Edgar (2010), more than four in five ex-prisoners said that their previous convictions made it harder for them to get insurance and that, even when they were successful, they were charged far more. In Québec, it is not considered discriminatory to refuse to insure a person who has a criminal record, within the meaning of the Québec Charter of Human Rights and Freedoms, simply because criminal history is not included in the list of prohibited grounds for discrimination, as it is, for example, for ethnic origin or religion. On the other hand, if you are refused insurance because you are living with a spouse or parent who has a criminal record, this could constitute discrimination based on your civil status (and civil status is on the list of prohibited grounds of discrimination in the Québec Charter, including marriage, civil union and de facto union, but also filiation). There is nothing in Québec law that prohibits insurers from asking you about your criminal record and making their decision based on this factor. In fact, as mentioned in Sanche and Roberge (2023), the province's Civil Code even requires you to declare - in good faith - factors that could influence the insurer's assessment of the risk you represent to it. If you fail to disclose such information, it is possible that a court could rule in favour of your insurer if it eventually refuses to compensate you following a loss, or if it revokes your policy. In Québec, some companies will refuse to insure a person with a criminal record even if the offence is not related to the insurance applied for. Other companies will offer insurance, but the premium will be significantly higher (double or quadruple). This premium surcharge applies to all persons living in the same household as a person who has been prosecuted.

Medical historical data can also be problematic. "*The financial burden of cancer can extend decades after diagnosis*," wrote Dumas et al. (2017). Some countries in Europe (France, Belgium, Luxembourg, the Netherlands, Portugal or Romania) adopted national legislative initiatives to recognise a "*right to be forgotten*" for cancer survivors. On January 26 2016, a law (2016-41, commonly referred as "the right to be forgotten") was adopted in France. According to this law, survivors do not have to disclose their past history of cancer to the insurer after a fixed number of years post-treatment: 5 years for patients diagnosed under the age of 18 at time of diagnosis and 10 years for patients aged 18 or above (in 2022, it became 5 years for adults, too). Before 2016 and the adoption of this law, insurers could impose higher premiums or could refuse to insure survivors because of their past history of cancer, even when they had no health problem at the time they purchase insurance.

5.3 Internal and External Data

5.3.1 Internal Data

Collecting “internal data" usually starts with a a “form", from Latin *forma* (form, contour, figure, shape), that start to designate in the XIV-th century a legal agreement[8]. Forms are used by insurers in the underwriting process or to handle claims. “*Think of a form to be filled in, on paper or a screen, intended to gather information that can later be quantified*" wrote Bouk (2022). Almost paraphrasing Christensen et al. (2016), he adds that “*someone, somewhere, designed that form, deciding on the set of questions to be asked, or the spaces to be left blank. Maybe they also listed some possible answers or limited the acceptable responses (...) The final resulting form and all that is written upon it as well as all negotiations that shaped it, weather backstage of offscreen, so to speak – all of this is data too. The data behind the numbers. To find stories in the data, we must widen our lens to take in not only the numbers but also the processes that generated those numbers*."

Traditionally, with some simple perspective, insurance companies use two kinds of databases: an underwriting database (one line represents a policy, with information on the policyholder, the insured property, etc.) and a claims database (one line corresponds to a claim, with the policy number, and the last view of the associated expenses), as in Figure 5.1. These two bases are linked by the policy number. But it is possible to use other “keys" (as coined in database management systems), corresponding to a single variable (or set of them) that can uniquely identify rows. For internal data, classical keys are the policy number (to connect the underwriting and the claim database), a “client number" (to connect different policies that could hold the same perso), or a claim number (usually to connect the claim database to financial records). But it is also possible to use some “keys" to connect to other databases, such as the licence plate of the car, the model of a car (“2018 Honda Civic Touring Sedan"), the address of a house, etc.

5.3.2 Connecting Internal and External Data

In recent years, however, companies have increasingly relied on data obtained from a wide variety of external sources. This data is either about the insured property, with information about the car model, or about the house, obtained from the address, as in Figure 5.1. The address historically allowed to have (aggregated) information on the neighborhood, with numbers of violations, on past floods, on the distance to the nearest fire station, etc. We can also use satellite images (via Google Earth) or information from OpenStreetMap (we will mention those in Section 6.8). And insurers rely on data that is becoming more and more extensive, with sensors deployed everywhere, in the car, or cell phones, as Prince and Schwarcz (2019) recall. This “data boom" raises the question of whether an increasingly detailed insight into the lives of policyholders can lead to more accurate pricing of risks.

For online pricing, for example, new customers who do extensive research and read policy information may be offered lower prices than loyal customers, as noted by Minty (2016).

Hand (2020) introduced the concept of “*dark data*," noting that all data has a hidden side, which could potentially generate bias. We will formalize in the following sections these notions of bias, some of which can be visualized on Figure 5.2, inspired by Suresh and Guttag (2019). The “*historical bias*" is the one that exists in the world as it is. This is the bias evoked in Garg et al. (2018) in contextualization in textual

[8]Instead of the Latin *formula* that could designate a contract. Actually, “formula" will refer nowadays to “mathematical formulas" as seen in Chapters 3 and 4, or “magic formulas", the two being very close for many people (see e.g. the introduction of O’Neil (2016) explaining how mathematics “*was not only deeply entangled in the world’s problems but also fueling many of them. The housing crisis, the collapse of major financial institutions, the rise of unemployment - all had been aided and abetted by mathematicians wielding magic formulas*."

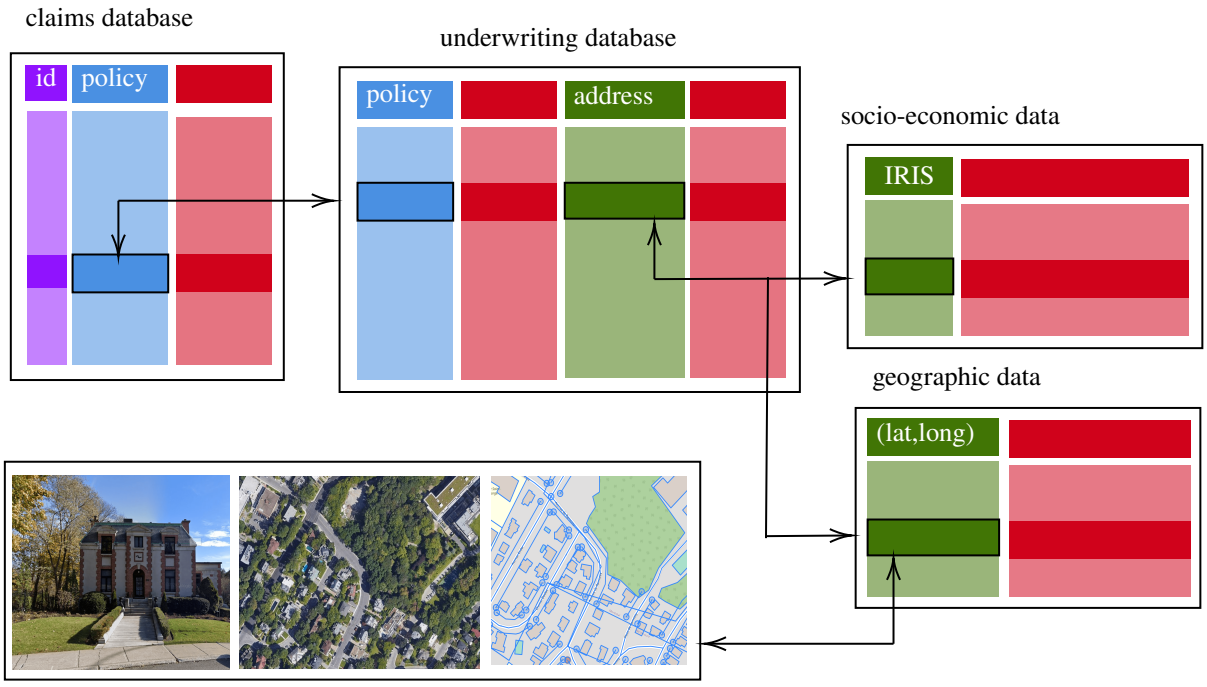
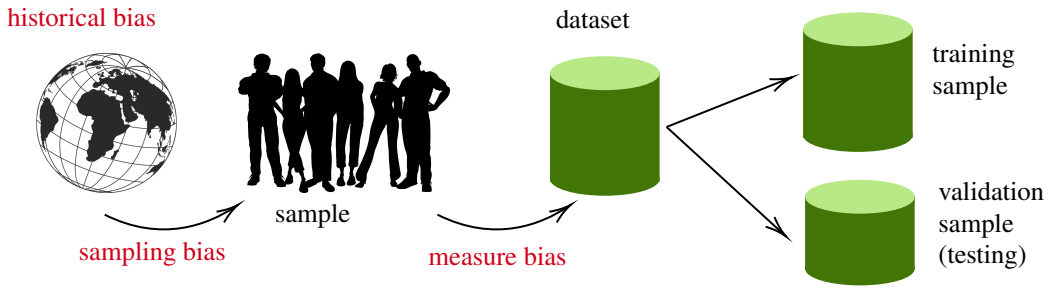
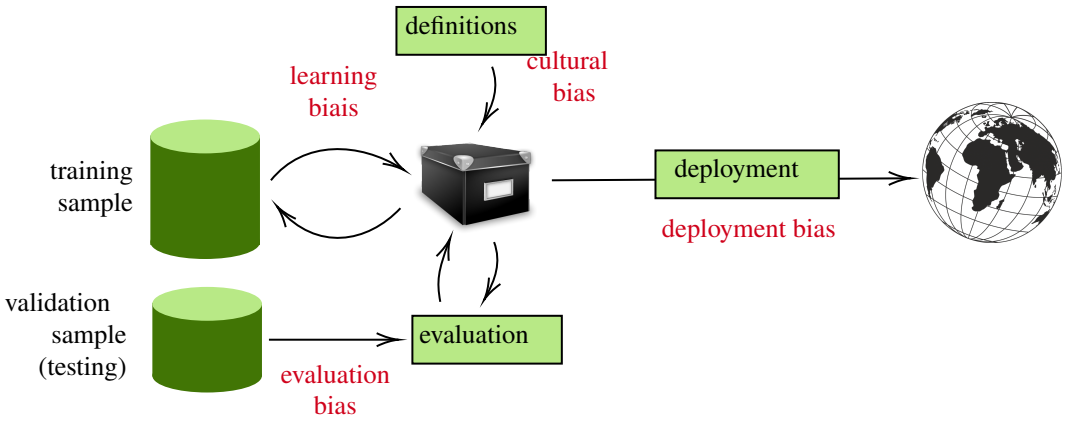


Figure 5.1: The databases of an insurer, with an underwriting database (in the center), with one line per insurance policy. This database will be linked to the claims database, which contains all the claims, and an associated policy number. Other data can be used to enhance the database, for example based on the insured's home address, with either socio-economic information (aggregated by neighborhood, wealth, number of crimes, etc.) but also other information, extracted from maps, satellite images (distance to the nearest fire hydrant, presence of a swimming pool, etc.) In car insurance, it is possible to find the value of a car from its characteristics, etc.

analysis ("*word embedding*") where the vectorization reflects the biases existing between men and women (the word "*nurse*" (non-gendered) is often associated with words associated with women, while "*doctor*" is often associated with words associated with men), or towards minorities. The "*sampling bias*" is the one mentioned in Ribeiro et al. (2016), where a classification algorithm is trained, to distinguish dogs from wolves, except that all the images of wolves are taken in the snow, and the algorithm just looks at the background of the image to assign a label. For "*measurement bias*," Dressel and Farid (2018) refers to reoffending in predictive justice, which is sometimes measured not as a new conviction, but as a second arrest. The "*cultural bias*" (called "*aggregation bias*" in Suresh and Guttag (2019)) refers to the following problem: a particular dataset might represent people or groups with different backgrounds, cultures or norms, and a given variable can mean something quite different for them. Examples include irony in textual analysis, or a cultural reference (which the algorithm cannot understand). Hooker et al. (2020) observe that compression amplifies existing algorithmic biases (where compression is similar to tree pruning, when attempting to simplify models). Another example is Bagdasaryan et al. (2019) who point out that data anonymisation techniques can be problematic. Differential privacy learning mechanisms such as gradient clipping and noise



(a) Data generation



(b) Modeling process

Figure 5.2: Bias in data generation, and in model building (loosely based on Suresh and Guttag (2019)).

addition have a disproportionate effect on under-represented and more complex subgroups. This phenomenon is called "*learning bias*." Evaluation bias takes place when the reference data used for a particular task is not representative. This can be seen in facial recognition algorithms, trained on a population that is very different from the real one. Buolamwini and Gebru (2018) note that darker-skinned women make up 7.4% of the Adience database (published in Eidinger et al. (2014), with 26,580 photos across 2,284 subjects with a binary gender label and one label from eight different age groups), and this lack of representativeness of certain populations can be an issue (e.g. for an algorithm to detect skin cancers). Finally, "*deployment bias*" refers to the gap between the problem a model is expected to solve, and the way it is actually used. This is what Collins (2018) or Stevenson (2018) show, describing the harmful consequences of risk assessment tools for actuarial sentencing, particularly in justifying an increase in incarceration on the basis of individual characteristics. In Box 5.3, Olivier l'Harridon discusses "*decision bias*."

Box 5.3 **Decision bias**, by Olivier l'Harridon[9]

The literature developed in behavioural economics and risk psychology has highlighted several "*biases*" that affect decision-making in times of uncertainty. Typically, a decision bias corresponds to an observed deviation from a norm of rational behaviour or judgment. Traditionally, economic theory provides, with the expected utility theory of von Neumann, Morgenstern and Savage and Bayes' theorem, a norm of rational behaviour in the face of uncertainty, whether the probabilities associated with events are known or not. Any deviation from this clearly defined norm is generally qualified as a decision bias, even if this deviation can be justified by the particular decision context, the decision-making speed, or the lack of probabilistic information.
Among decision biases, there are numerous choice heuristics, which are mental shortcuts corresponding to intuitive and rapid mental operations. It is important to keep in mind that these shortcuts can be justified by the need to obtain a satisfactory answer, having the advantage of being fast without necessarily being optimal. The most famous choice heuristic in the field of uncertainty is certainly the representativeness heuristic. It occurs when individuals misjudge the frequency of an event by an abusive generalization of a similar past event. The direct translation of the representativeness heuristic is the use of stereotypes to make predictions: events that have only been observed on a limited sample are generalized to the whole population, extreme events are trivialized, phenomena for which more details are known are considered more probable, data on recent events are favoured over initial data on the situation, etc. This last characteristic can be reinforced by the availability heuristic, where individuals that have a better memory of events are more easily available in their representations, i.e. they are significant for them, such as frequent events, more recent events, or even larger events. The last major heuristic identified by the literature is the anchoring heuristic, which refers to the tendency to set beliefs at a certain level and to maintain this anchor in subsequent evaluations of uncertainty, leading to a conservatism bias. The use of these heuristics is fundamentally linked to the decision-making process: by providing frequent feedback on past decisions, or by emphasising decisions, the impact of these heuristics on decision-making can be significantly reduced.
Alongside these heuristics, motivational biases are more difficult to consider, particularly in terms of possible interventions to reduce them. For example, behavioural economics has shown that individuals tend to establish a separate mental accounting for each small risk to which they are subjected, without considering their potential compensations. Moreover, even for small risks, individuals are much more sensitive to those that lead to losses than to those that lead to gains, demonstrating a strong aversion to losses. Individuals therefore tend to over-insure themselves for these small risks but also to choose low-deductible contracts even if they are less advantageous for them. Note that loss aversion also generates a form of risk conservatism by encouraging individuals to stay at their status quo, their point of reference, rather than risk losses, even for the sake of greater prospects of gains. In a different vein, risk decisions are also biased by the strong tendency of individuals to overestimate small probabilities and underestimate large probabilities, which are important sources of optimism and pessimism about subjective views of luck and uncertainty. (...)

5.3.3 External Data

Insurers have the feeling that they can use external data to get valuable information about their policyholders. And not only insurers, all major companies are fascinated by external data, especially the one collected by big tech companies. As Hill (2022) wrote it, "*Facebook defines who we are, Amazon defines what we want, and Google defines what we think*."

But it is not new. Scism and Maremont (2010a,b) reported that a U.S. insurer had teamed up with a consulting firm, to look at 60,000 recent insurance applicants (health related) and they found that a predictive model based partly on consumer-marketing data, was "*persuasive*" in its ability to replicate traditional underwriting techniques (based on costly blood and urine testing). So called "external information" was personal and family medical history, as well as information shared by the industry from previous insurance applications, and data provided by Equifax Inc., such as likely hobbies, television viewing habits, estimated income, etc. Among "good" risk assessment factors, "*foreign traveler*," "*healthy food choice*," "*outdoor enthusiast*," "*strong ties to community*," while "bad" risk factors were "*long commute*," "*high television consumption*," "*purchases tied to obesity*" (among many others).

(...) In addition to these heuristics, motivational biases are more difficult to account for, particularly in respect of how to reduce them. For instance, behavioural economics has demonstrated that individuals tend to set up a separate mental accounting for each small risk they face, without taking into account their potential offsets. Moreover, for even small risks, individuals are much more sensitive to those that lead to losses than those that lead to gains, thereby displaying a strong aversion to losses. Individuals therefore tend to over-insure themselves for these small risks but also to choose low-deductible policies even if they are less beneficial for them. It is worth noting that loss aversion also generates a kind of risk conservatism by encouraging individuals to remain in their status quo, their reference point, rather than risk losses, even for greater prospects of gains. On a different note, risk decisions are also biased by the strong tendency of individuals to overestimate small probabilities and underestimate large probabilities, both of which are important sources of optimism and pessimism regarding the subjective vision of luck and uncertainty.
Moreover, the absence of information, or the need to own it before making a decision, also results in a certain number of decision biases. For example, individuals may react strongly to ambiguity, typically when they are aware of the absence of information that might be available but is not given to them. Some individuals also have difficulty taking into account new information, positive or negative, that they learn about risk factors, and these difficulties can have particularly important consequences when they involve risk factors affecting health, such as heredity, smoking, or information on cardiovascular risks.
Finally, decision biases arising from fields related to uncertainty, typically involving the perception of the present and the future, can strongly interact with risky decision making. Therefore, the immediacy bias, which is a strong variation in impatience towards immediate consequences, coupled with loss aversion leads individuals to overestimate present costs despite future benefits outcome and to neglect prevention and predictability, whether in the monetary or health domain.

This is possible because of “data brokers" or “data aggregators", as discussed in Beckett (2014) and Spender et al. (2019). These companies collect data on a grand scale from various sources, independently, they clean the data, link it and use use machine learning techniques to extract information. As mentioned by Harcourt (2015a), one of the largest data brokers in the U.S. sells lists of “*Elderly Opportunity Seekers*" (with 3.3 million older people “*looking for ways to make money*"), “*Suffering Seniors*" (with 4.7 million people with cancer or Alzheimer’s disease), or “*Oldies but Goodies*" (with 500,000 gamblers over 55 years old). As explained in Schneier (2015), Lauer (2017) and Zuboff (2019), those profiling companies are replacing the control companies that took over from the discipline companies. Nakashima (2018) revealed that Google continued to track movements even when a user explicitly asked it not to, by locking location tracking in Android, for example. Because knowing the main movements and locations gives a lot of information about a person, often sensitive information. Beyond home and work, it is possible to infer sexual preferences, religion, etc.

This aggregation of data is problematic, as discussed in O’Neil (2016) or Lauer (2017), without even mentioning any ethical considerations. This could explain the interest of insurers to connected objects, also named “Internet-of-Things", as described in Iten et al. (2021). Not only those devices can help to collect new information, but overall, three main business opportunities have been identified: customer engagement, risk reduction and risk assessment. If we think of health insurance, contacts between the insurance company and a future policyholder start with sickness-related questions. Applications of smartphones could, in a way, appear as less intrusive, and Barbosa (2019) recalls that offering a free fitness tracker, increased “customer engagement". Those wearable devices offer an opportunity to encourage customers to have a more healthy lifestyle, by simple push notifications to nudge the customer to leave their sofa, and go for a walk, or by setting and tracking workout goals to motivate the policyholder to pursue a regular exercise regime, Spender et al. (2019). They can also be used for prevention, to detect a problem before it leads to very costly medical procedures. Those devices are also used in motor insurance, with different types of insurance, usage-based insurance (UBI), pay-as-you-drive (PAYD), pay-how-you-drive (PHYD) and manage-how-you-drive (MHYD). UBI is the simplest one, and the only difference with traditional insurance is that instead of estimating the mileage beforehand the exact milage is calculated at the end of the years, and adjustments are

made. PAYD takes not only the exact mileage, but also other non-behavioral risk factors into account, e.g. kind of road or time of day. Such data can come from a cellphone. PHYD adds behavioral risk factors, such as speeding and driving style, measured by an on-board device. The main difference is that a cellphone is usually related to a single person, while the on-board device is related to a car. Finally, MHYD is in principle the same as PHYD, except that it is based on dynamic interactions. The driver gets suggestions on how to drive more safely and therefore lower the premium. Such “gamification" introduces feedback biases, and makes the data harder to analyse.

5.4 Typology of Ratemaking Variables

In Section 3.2, we have mentioned the historical importance of “categorization" for insurers, and now, we will look more carefully at the variables used as ratemaking variables.

5.4.1 Ratemaking Variables in Motor Insurance

In France, as recalled in Delaporte (1962, 1965), in the aftermath of the Second World War, the price of motor liability insurance was defined using the “*méthode de la tarification à la moyenne*" (or “*average pricing method*"). The work of the actuaries consisted in defining large categories of insureds deemed to be homogeneous. The construction of these equivalence classes is based on the grouping of three criteria: (1) the zone where the vehicle is parked, (2) its fiscal power, (3) the insured’s profession and the use of the vehicle. In 1962, the rate structure of the Groupement Technique Accidents (GTA) of the Association Générale des Sociétés d’Assurances contre les Accidents (AGSAA), the professional organization of automobile insurance companies, in France, was base on six geographic zones, twelve vehicle groups, as well as three usages and six professional categories (namely, for the three usages, “business or commercial", "leisure", and "public transport of goods", and the six professional categories were (1) craftsmen, (2) ministerial officers (lawyers, bailiffs, notaries, etc.), (3) travelers, representatives and salesmen, (4) full-time employees of commercial, industrial and craft enterprises, ministerial offices and liberal professions, (5) clergymen, (6) civil servants, magistrates and members of the education system). According to Delaporte (1965), “*the grouping of observations into classes, while it masks all or part of the inter-individual differences, does not eliminate them*." At the turn of the 1950s and 1960s, a “risk-modelled premium" – “*prime modelée sur le risque*" – is introduced. The estimation of the risk should therefore take into consideration, on the one hand, the class in which the insured is classified, and on the other hand, the number of accidents observed during the previous years. On the basis of these data, coefficients are calculated which should be applied to the premium for each insured. “*This is the ‘risk-modelled premium’, which represents the probable value of an individual risk*," as discussed in Chapter 2. This model would therefore not only be more technically correct, but also socially correct. According to Delaporte (1965) “*if the risks of the insured are not all equal, it is normal to ask each insured a premium If the risks of the insured are not all equal, it is normal to ask each insured a premium proportional to the risk that he or she makes the mutual insurance company bear*." Within the same the same class, the bad risks, a minority, would significantly increase the premium of the good the premium of the good risks.

More recently, Jean Lemaire presented classification variables commonly used, in Lemaire (1985), see Box 5.4.

Banham (2015) and Karapiperis et al. (2015), cited in Kiviat (2021), mentioned that in car insurance, policymakers have investigated the use of credit scores, web browser history, groceries store purchases, zip codes, the number of past addresses, speeding tickets, education level, data from devices that track driving in real time, etc. In Box 5.4, we reproduce a list of variable given by David Becker, insurer in Florida, U.S.

Box 5.4 **Classification variables**, Lemaire (1985)

"the main task of the actuary who sets up a new tariff is to make it as fair as possible by partitioning the policies into homogeneous classes, with all policyholders belonging to the same class paying the same premium (...) Most developed countries use several classification variables to differentiate premiums among automobile third-party liability policyholders. Typical variables include age, sex, and occupation of the main driver, the town where he resides, and the type and use of his car. More exotic variables, such as the driver's marital status and smoking behavior, or even the color of his car, have been introduced in some countries. Such variables are often called prior *rating variables, as their values can be determined before the policyholder starts to drive. The main purpose for their use is to subdivide policyholders into homogeneous classes. If, for instance, females are proved to cause significantly fewer accidents than males, equity arguments suggest that they should be charged a lower premium. (...) Despite the use of many a priori variables, very heterogeneous driving behaviors are still observed in each tariff cell (..) Hence the idea came in the mid-1950s to allow for* posterior *premium adjustments, after having observed the claims history of each policyholder. Such practices, called experience rating, merit-rating, no-claim discount, or bonus-malus systems (BMS), penalize the insureds responsible for one or more accidents by an additional premium or malus, and reward claim-free policyholders, by awarding a discount or bonus.",*

5.4.2 Criteria for Variable Selection

In Chapter 3 when discussed variable selection in the context of machine learning and statistics, in the sense that a variable is legitimate in a pricing model if the association with the outcome y (claim occurrence or frequency, average cost or aggregated losses) is significant. In Chapter 4, we have seen that interpretability was important, so, including that variable "should make sense". If we want to go further, several factors should be taken into account. In an interview mentioned in Meyers and Van Hoyweghen (2018), a senior reinsurance manager says "*for example, we are not allowed to use genetic testing because you are born like that. You have no say in it. No matter what you eat, you have no influence at all on your genetics. So in the long run the regulator and consumer interest groups will probably shut down us using things that people have no control over. Now, the one thing people have clear control over, is their behaviour. And that is a very easy discussion to have with people to say: 'You know you are doing wrong. Yes I do. Why don't you change it? I don't feel like it.' Then you say: 'You understand, I cannot actually reward you for that or accept you as a client. There is not much you can argue because all you have to do is act in a positive way and that solves the problem.' It is something they have full control over.* " Therefore, if people can control how they behaved, they should take responsibility for the outcomes that their behaviour produces. This was actually the statement made by a Dutch telematics based motor insurer in a video[10], "*Other peoples' bad luck, recklessness and carelessness determine how much you pay for your insurance. Even if you never cause any damage, you end up paying for people who do. At Fairzekering, we don't think that is fair. How badly or, more importantly, how well you drive should make a difference*." It is a classical actor-observer bias, as defined in Jones and Nisbett (1971), corresponding the tendency to attribute the behavior of others to internal causes, while attributing our own behavior to external causes. We often think that it is the structure, or circumstances, that constrain our own choices, but at the same time, that it is the behavior of others that explain theirs. And the story is biased because if we don't want to pay for a bad neighbor's claims, we have to accept that he too, in return, doesn't want to pay for ours. Furthermore, our behaviors can certainly be influenced by our own decisions, but they can also most certainly be influenced by the decisions of others. Our behaviors are not isolated. They are the product of our own actions and the product of our interactions with others. And this applies when we are driving a car, as well as in a million and one other activities.

As explained in Finger (2006) and Cummins et al. (2013), the risk factors of a risk classification system

[10] See https://vimeo.com/123722811

Box 5.4 **Motor Insurance Pricing**, Backer (2017)

Examples of factors are:

- **Characteristics of the automobile** – Age, manufacturer, value, safety features (anti-lock brakes, anti-theft devices, adaptive cruise control, lane departure feature) all matter to insurance companies.
- **The coverage selected** – Liability limits, uninsured motorist, collision, comprehensive coverages vary significantly depending on your willingness to take on more risk.
- **Deductible you select** – Higher deductibles mean that the driver pays more to repair his vehicle before the insurance kicks in, thus reducing your premium.
- **Profile of the driver** – Age, gender, marital status, place of residence, driving record all help determine your ultimate premium.
- **Usage of the car** – Do you commute to work, use your vehicle for business or pleasure only?

The following are individual risk factors that will determine if you are charged more or less money by your auto insurance company:

1. **Credit Rating** – Study after study indicates that a person's credit rating can determine their propensity to have an accident. The best automobile insurance companies insure drivers with the best credit scores.
2. **Payment History** – Similar to your credit history, if you pay your automobile premiums on time, you may be able to reduce your premium. Age – Around 30% of all vehicle injuries in the U.S. are caused by drivers aged 15-24. Those in the 16-19 group are three times as likely as those over age 20 to be in a fatal car crash. Motor vehicle accidents are the leading cause of death for teenagers in the country. For this reason, drivers age 16-19 pay 50% more on average for automobile insurance than drivers aged 20-24. As drivers age, their rates typically decline until, by age 55, drivers can enjoy senior discounts.
3. **Driving Record** – Driving history has shown to be an accurate indicator of future claims. Drivers with a clean driving record enjoy discounts not granted to those with tickets or accidents.
4. **Gender** – Men, especially young ones, are much more likely to be involved in automobile accidents than women, regardless of their age. For this reason, men pay significantly more for their automobile insurance over the years than women.
5. **Nature of Employment** – Drivers who use their cars for business, like real estate salespeople, pay higher rates than drivers who work from home. While it is difficult to quantify, drivers who work in high stress jobs for long hours, like doctors, tend to have more accidents than those in low stress occupations.
6. **Vehicle type** – Cars that are more expensive to repair or which are most likely to be stolen carry higher rates than others. Some high performance automobiles are very difficult and expensive to insure. Cars with extra safety features may be subject to additional credits.
7. **Location of your home** – Drivers who live in high risk urban areas of the country and those who live in high-crime neighborhoods where their vehicles are more likely to be stolen or vandalized, pay higher rates than others.
8. **Marital Status** – Married people tend to be better drivers than singles and pay lower premiums. In many cases, the multi-car discount kicks in, thus reducing your premium.

have to meet many requirements, they have to be (i) statistically (actuarially) relevant, (ii) accepted by society, (iii) operationally suitable, and (iv) legally accepted.

5.4.3 An Actuarial Criterion

A classification variable is considered actuarially fair, in the sense of Cummins et al. (2013), if it is accurate, provides homogeneity among members, has statistical credibility, and is reliable over time. A classification variable will be said to be "*accurate*" if it allocates policyholders in such a way that they each pays a premium

proportional to its expected cost of claims. To prevent adverse selection, this is probably the most important criterion. Homogeneity requires that all policyholders in the same risk class have the same expected claim costs. A large number of insureds in each group is needed to make the group's loss history statistically credible, and to make pooling still meaningful. Too few members result in losses that vary widely from year to year and cause premiums to fluctuate in the same way. Finally, a reliable classification variable produces cost differences between different groups that remain relatively stable over time. Being perceived as a "*good risk*" in 2020 and then as a "*bad risk*" in 2021, without having felt that one's behaviour has changed in the meantime, will generally not be well perceived.

5.4.4 An Operational Criterion

Some actuarially fair risk classification variables cannot be applied in practice because they do not meet the operational standards of objectivity, low cost of implementation and handling difficulty. Data that are laborious and costly to collect and verify rarely make good classification variables. In connection with the objective of low administrative costs, data used for another purpose make good risk classification variables. The use of a variable that is reported or collected by other agencies reduces the likelihood of it being manipulated and, as Cummins et al. (2013) point out, reduces the cost of verification. A classification variable needs to offer minimal ambiguity between insureds, and the total categories described by the variable should be mutually exclusive and comprehensive.

5.4.5 A Criterion of Social Acceptability

A third consideration in the selection of risk classification variables is social acceptability. According to the classification of Cummins et al. (2013), the four main criteria are privacy, causality, controllability, and affordability/availability. Privacy affects the willingness of individuals to disclose certain information, which in turn affects the accuracy of a risk classification variable as well as the ease with which it can be collected and verified (more at the end of this section). Causality requires more than an intuitive relationship between the classification variable and expected losses. A good risk classification variable must encourage individuals to reduce the expected frequency and/or severity of their losses - the "controllability" criterion. The social criterion of affordability/availability requires that those who need to purchase insurance coverage be able to do so reasonably. Social acceptability seems to be even greater when the risk is linked to a criterion that is a matter of choice for the insured.

5.4.6 A legal Criterion

Finally, in practice, the use or prohibition of certain classification variables is most often imposed by law (or regulation). In Canada, provincial laws, which are generally more restrictive than in most states in the United States, generally require that classification variables not be unfairly discriminatory, i.e., the actuarial fairness test must be demonstrated. However, classification variables have sometimes been prohibited because there is only a correlation, not a causal relationship, between the classification variable and expected claims costs, or because they have been deemed socially unacceptable. The legal criteria, of course, vary by state and province, as noted previously.

5.5 Behaviors and Experience Rating

Algorithmic differentiation can also be used to differentiate behavioral controls, with usually a distinction between “hard" and “soft" behavioral controls. “Hard" behavioral controls are based on strict rules, and obligations, while “soft" behavioral controls are achieved through financial incentives, recommendations, “nudges", etc, as discussed in Yeung (2018a,b). Even if it is usually based on technological tools, the idea echoes “experience rating", as coined in Keffer (1929), that occurs when the premium charged for an insurance policy is explicitly linked to the previous claim record of the policy or the policyholder. Actually, claim history has been the most important rating factor in motor insurance, over the past 60 years. Lemaire et al. (2016) argued that annual mileage and claim history (such as a bonus-malus class) are the two main powerful rating variables.

Experience rating has to do with “merit-based" insurance pricing, “merit rating" as coined by Van Schaack (1926), studied in Wilcox (1937) or Rubinow (1936), and popularized by Bailey and Simon (1960). It seems like “merit" has always been seen as a morally valid predictor. But recently, Sandel (2020) criticized this “*tyranny of merit*," “*the meritocratic ideal places great weight on the notion of personal responsibility Holding people responsible for what they do is a good thing, up to a point. It respects their capacity to think and act for themselves, as moral agents and as citizens. But it is one thing to hold people responsible for acting morally; it is something else to assume that we are, each of us, wholly responsible for our lot in life*." “*What matters for a meritocracy is that everyone has an equal chance to climb the ladder of success; it has noting to say about how far apart the rungs of the ladder should be. The meritocratic ideal is not a remedy for inequality; it is a justification of inequality*." Hence, behavioural fairness corresponds to the fairness of merit. But as pointed out by Szalavitz (2017), the narrative that legitimizes the idea of merit is an often biased narrative, with a classical “actor-observer bias". Personalization is close to this idea, since it means looking at that individual based on her own risk profile, claims experience, and characteristics, rather than viewing him or her as part of a set of similar risks. Hyper-personalization takes personalization further.

5.6 Omitted Variable Bias and Simpson’s Paradox

Omitted variable bias occurs when a regression model is fitted without considering an important (predictor) variable. For example, Pradier (2011) noted that actuarial textbooks (such as Depoid (1967)) state that “*the pure premium of women in the North American market would be equal to that of men if it were conditioned on mileage*."

5.6.1 Omitted Variable in a Linear Model

The sub-identification corresponds to the case where the true model would be $y_i = \beta_0 + \boldsymbol{x}_1^\top \boldsymbol{\beta}_1 + \boldsymbol{x}_2^\top \boldsymbol{\beta}_2 + \varepsilon_i$, but the estimated model is $y_i = b_0 + \boldsymbol{x}_1^\top \boldsymbol{b}_1 + \eta_i$ (i.e., the variables $\boldsymbol{x}_2$ are not used in the regression). The least square estimate of $\boldsymbol{b}_1$, in the mispecified model, is (with the standard matrix writing in econometrics, like Davidson et al. (2004) or Charpentier et al. (2018))

$$\begin{aligned}
\widehat{\boldsymbol{b}}_1 &= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{y} \\
&= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top [\boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}] \\
&= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_1 \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{\beta}_2 + (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta}_1 + \underbrace{(\boldsymbol{X}_1' \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{\beta}_2}_{\boldsymbol{\beta}_{12}} + \underbrace{(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{\varepsilon}}_{\nu_i},
\end{aligned}$$

so that $\mathbb{E}[\widehat{\boldsymbol{b}}_1] = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{12}$, the bias (what we have noted $\boldsymbol{\beta}_{12}$) being null only in the case where $\boldsymbol{X}_1^\top \boldsymbol{X}_2 = \boldsymbol{0}$ (that is to say $\boldsymbol{X}_1 \perp \boldsymbol{X}_2$): we find here a consequence of the Frisch-Waugh theorem (from Frisch and Waugh (1933)). If we simplify a little, let us suppose that the real underlying model of the data

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

where x_1 and x_2 are explanatory variables, y is the target variable, and ε is random noise. The estimated model by removing x_2 gives

$$\widehat{y} = \widehat{b}_0 + \widehat{b}_1 x_1.$$

One can think of a missing significant variable x_2, or the case where x_2 is a protected variable. Estimates of the regression coefficients obtained by least squares are (usually) biased, in the sense that

$$\widehat{b}_1 = \frac{\widehat{\text{cov}}[x_1, y]}{\widehat{\text{Var}}[x_1]} = \frac{\widehat{\text{cov}}[x_1, \beta_0 + \beta_1 x_1 + \beta x_2 + \varepsilon]}{\widehat{\text{Var}}[x_1]},$$

or

$$\widehat{b}_1 = \beta_1 \cdot \underbrace{\frac{\widehat{\text{cov}}[x_1, x_1]}{\widehat{\text{Var}}[x_1]}}_{=1} + \beta_2 \cdot \frac{\widehat{\text{cov}}[x_1, x_2]}{\widehat{\text{Var}}[x_1]} + \underbrace{\frac{\widehat{\text{cov}}[x_1, \varepsilon]}{\widehat{\text{Var}}[x_1]}}_{=0} = \beta_1 + \beta_2 \cdot \frac{\widehat{\text{cov}}[x_1, x_2]}{\widehat{\text{Var}}[x_1]}.$$

When x_2 is omitted, $\widehat{b}_1$ is biased especially since x_1 and x_2 are correlated. Therefore, in most realistic cases, not only does the removal of the sensitive variable (if $x_2 = p$) not make the regression models fair, but on the contrary, such a strategy is likely to amplify the discrimination. For example, in labor economics, if immigrants tend to have lower levels of education, then the regression model would "punish" low education even more by offering even lower wages to those with low levels of education (who are mostly immigrants). Žliobaite and Custers (2016) suggests that a better strategy for sorting regression models would be to learn a model on complete data that includes the sensitive variable, then remove the component containing the sensitive variable and replace it with a constant that does not depend on the sensitive variable. A study on discrimination prevention for regression, Calders and Žliobaite (2013), is related to the topic, but with a different focus. Our goal is to analyze the role of the sensitive variable in suppressing discrimination, and to demonstrate the need to use it for discrimination prevention. Calders and Žliobaite (2013) does, of course, use the sensitive variable to formulate non-discrimination constraints, which are applied during model fitting. But a discussion of the role of the sensitive variable is not the focus of their study. Similar approaches have been discussed in economic modeling, as in Pope and Sydnor (2011), where the focus was on the sanitization of regression models, our focus in this paper is on the implications for data regulations.

Returning to our example, there are cases where $\widehat{b}_1 < 0$ (for example) while in the true model, $\beta_1 > 0$. This is called a (Simpson) paradox or spurious correlation (in ecological inference) in the sense that the direction of impact of a predictor variable is not clear.

5.6.2 School Admission and Affirmative Action

While examples of such paradoxes are numerous (Alipourfard et al. (2018) compiles a substantial list), Simpson's paradox has historically been on college admissions, in Bickel et al. (1975), as in Table 5.3.

With mathematical notations, Simpson's paradox can be written as

$$\begin{cases} \mathbb{P}[\mathcal{A}|\mathcal{B} \cap C] < \mathbb{P}[\mathcal{A}|\overline{\mathcal{B}} \cap C] \\ \mathbb{P}[\mathcal{A}|\mathcal{B} \cap \overline{C}] < \mathbb{P}[\mathcal{A}|\overline{\mathcal{B}} \cap \overline{C}] \\ \mathbb{P}[\mathcal{A}|\mathcal{B}] > \mathbb{P}[\mathcal{A}|\overline{\mathcal{B}}] \end{cases}$$

	Total	Men	Women	Proportions
Total	5233/12763 ~ 41%	3714/8442 ~ **44**%	1512/4321 ~ 35%	66%-34%
Top 6	1745/4526 ~ 39%	1198/2691 ~ **45**%	557/1835 ~ 30%	59%-41%
A	597/933 ~ 64%	512/825 ~ 62%	89/108 ~ **82**%	88%-12%
B	369/585 ~ 63%	353/560 ~ 63%	17/ 25 ~ **68**%	96%- 4%
C	321/918 ~ 35%	120/325 ~ **37**%	202/593 ~ 34%	35%-65%
D	269/792 ~ 34%	138/417 ~ 33%	131/375 ~ **35**%	53%-47%
E	146/584 ~ 25%	53/191 ~ **28**%	94/393 ~ 24%	33%-67%
F	43/714 ~ 6%	22/373 ~ 6%	24/341 ~ **7**%	52%-48%

Table 5.3: Admission statistics on the six largest programs graduating at Berkeley, with the number of admits/number of applications received the percentage of admissions. The bolded numbers indicate, by row, which men or women have the highest admission rate. The proportion column shows the male-female proportions in the application submissions. The total corresponds to the 12763 applications in 85 graduate programs; the six largest programs are detailed below, and the "top 6" line is the total of these six programs, i.e. 4526 applications (source: Bickel et al. (1975)).

(where the bar denotes the complement of the event) or analytically,

$$\frac{a_1}{c_1} < \frac{a_2}{c_2} \text{ and } \frac{b_1}{d_1} < \frac{b_2}{d_2} \text{ while } \frac{a_1 + b_1}{c_1 + d_1} > \frac{a_2 + b_2}{c_2 + d_2}$$

The conclusion of Bickel et al. (1975) emphasizes "*the bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seems quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects*." In other words, the source of the gender bias in admissions was a field problem: through no fault of the departments, women were "separated by their socialization" that occurred at an earlier stage in their lives.

5.6.3 Survival of the Sinking of the Titanic

Another illustration is found in the Titanic data, specifically when using information about crew members to passengers. Table 5.4 shows the same paradox, in the context of survival following the sinking.

	Total	Women	Men
third class	181/709 ~ **25.5**%	106/216 ~ 49.1%	75/493 ~ 15.2%
crew member	211/890 ~ 23.7%	20/ 23 ~ **86.9**%	191/867 ~ **22.0**%

Table 5.4: Survival statistics for Titanic passengers conditional on two factors, crew / passenger (x_1) and gender (x_2).

But let's look at the survival rates for men and women separately, as presented in Table 5.4. For men, among the crew, there were 885 men, of whom 192 survived, for a rate of 21.7%. Among the third class passengers, 462 were men, and 75 survived, for a rate of 16.2%. For women, among the crew, there were

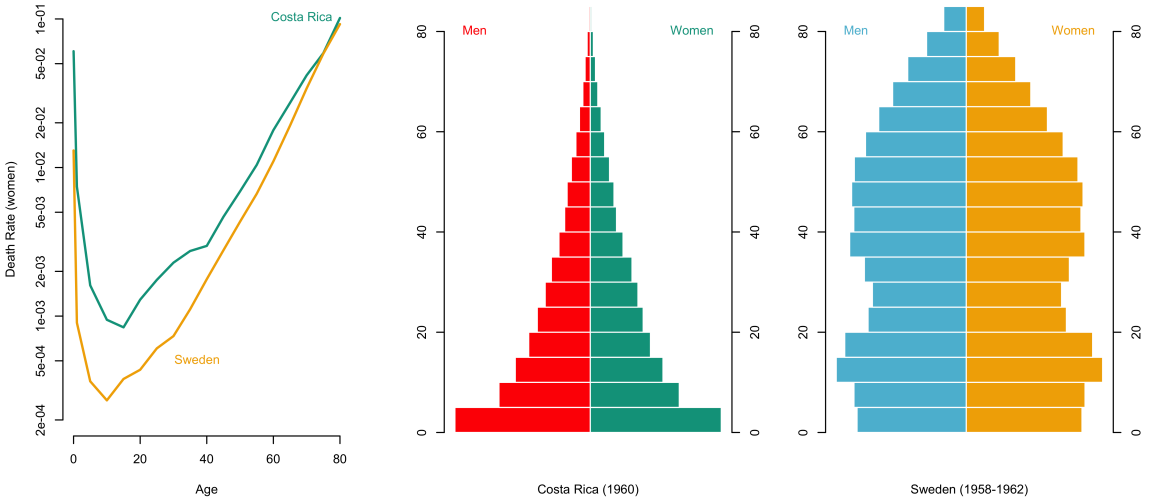


Figure 5.3: Annual mortality rate for women, Costa Rica and Sweden, and age pyramid for both countries. (Inspired by Cohen (1986), data source: Keyfitz et al. (1968))

23 women, and 20 of them survived, for a rate of 87.0%. And among the third class passengers, 165 were women, and 76 survived, for a rate of 46.1%. In other words, for males and females separately, the crew had a higher survival rate than the third-class passengers; but overall, the crew had a lower survival rate than the third-class passengers. As with the admissions, there is no miscalculation, or catch. There is simply a misinterpretation, because gender x_2 and status x_1 (passenger/crew member) are not independent, just as gender x_2 and survival y are not independent. Indeed, while women represent 22% of the total population, they represent more than 50% of the survivors... and 2.5% of the crew.

5.6.4 Simpson's Paradox in Insurance

In a demographic context, Cohen (1986) looked at mortality in Costa Rica and Sweden (Sweden was then known, and promoted, for its excellent life expectancy). Not surprisingly, he found that in 1960, the mortality rate for women in all age groups was higher in Costa Rica than in Sweden, as shown in Figure 5.3. Yet the overall mortality rate for women in Costa Rica was lower than in Sweden, with a mortality rate of 8.12 in Costa Rica, compared to 9.29 in Sweden. The explanation is related to Simpson's paradox, and it comes from the different structure of the populations. The population of Costa Rica is much younger on average than that of Sweden, and therefore the younger age groups (which have a low mortality rate) weigh more in the average for Costa Rica than for Sweden, leading to a fairly low overall mortality rate in Costa Rica, despite a fairly bad rate in each age group, as seen in the age pyramid, right of the Figure 5.3.

Davis (2004) studied the relationship between the number of pedestrian-vehicle accidents and the average speed of vehicles at various locations in a city. Specifically, the study assessed the value of imposing stricter speed limits on vehicles in cities, and unexpectedly, the model showed that lowering the speed limit from 30mph to 25mph would increase the number of accidents. The explanation is that an unfortunate aggregation of the data (which did not take into account that the number of accidents was much lower in residential areas,

for example) led to a paradoxical conclusion (the number of accidents is expected to decrease when the speed limit is reduced).

5.6.5 Ecological Fallacy

Problems similar to Simpson's paradox also occur in other forms. For example, the ecological paradox (analyzed by Freedman (1999), Gelman (2009), King et al. (2004)) describes a contradiction between a global correlation and a correlation within groups. A typical example was described by Robinson (1950). The correlation between the percentage of the foreign-born population and the percentage of literate people in the 48 states of the United States in 1930 was +53%. This means that states with a higher proportion of foreign-born people were also more likely to have higher literacy rates (more people who could read, at least in American English). Superficially, this value suggests that being foreign born means that people are more likely to be literate. But if we look at the state level, the picture is quite different. Within states, the average correlation is −11%. The negative value means that being foreign born means that people are less likely to be literate. If the within-state information had not been available, an erroneous conclusion could have been drawn about the relationship between country of birth and literacy.

5.7 Self-Selection, Feedback Bias and Goodhart's Law

Non-response is a form of self-selection where individuals refuse to be part of the learning sample. In fact, this situation is increasingly found in administrative files, as pointed out by Westreich (2012). Until very recently, our data was often stored automatically, without our knowledge, and without us having any say in the matter. The European Union's General Data Protection Regulation (GDPR) has changed that, as you probably realize from all the invitations to check boxes indicating that you understand and consent to the storage of your personal data when you browse websites. In application of this principle, many countries have passed laws giving those who request it the opportunity to have their data deleted. This concept of "opting-out" is restrictive, and can strongly bias the retained data. One can think of Dilley and Greenwood (2017) who noted that the number of abandoned emergency calls (to 999[11]) in the U.K. had more than doubled in 2016. How do we account for these unsuccessful calls if we want to seriously study feelings of insecurity?

This problem of self-selection can also resurface when we try to model admission to a university program (to echo the problem presented in the previous section). Here we consider three indicator variables $(y_{1:i}, y_{2:i}, y_{3:i})$ where (1) the first variable indicates whether a person has applied to a program (2) the second indicates admission to a program (3) and the third indicates enrolment in a program. Assume two more that the records are evaluated on the basis of two quantities $(x_{1:i}, x_{2:i})$, or $\boldsymbol{x}_i$. In the examples from the United States, the two main measures studied are the GPA score (*Grade Point Average*, often ranging from 1 to 4, respectively for a D and an A or an A+) and the SAT score (*Scholastic Aptitude Test*, ranging from 400 to 1600). For simplicity, let us assume that both scores are normalized on a scale of 0 to 100, as in Figure 5.4. Suppose that very simple rules are adopted for each of the classes (j): $y_{j:i} = \mathbf{1}[\boldsymbol{x}^\top \boldsymbol{\beta}_j > s_j]$. For example no student whose sum $x_1 + x_2$ does not exceed 60 is admitted ($y_{1:i} = \mathbf{1}[x_{1:i} + x_{2:i} > 60]$), any student whose sum of scores exceeds 120 is admitted ($y_{2:i} = \mathbf{1}[x_{1:i} + x_{2:i} > 120]$) and students who are too good do not come in the end ($y_{3:i} = \mathbf{1}[x_{1:i} + x_{2:i} \in (120, 160]]$). As can be seen in the figure 5.4, depending on the data used, the correlation between the variables x_1 and x_2 can strongly change: globally the variables are strongly correlated, positively, but on the sub-base of the students enrolled in the program (i.e. $\{i : y_{3:i} = 1\}$), the variables x_1 and x_2 are strongly correlated, but negatively this time

[11] Equivalent, in the U.K., of 911 in North America, 112 in many European countries, or 0118 999 881 999 119 725 3.

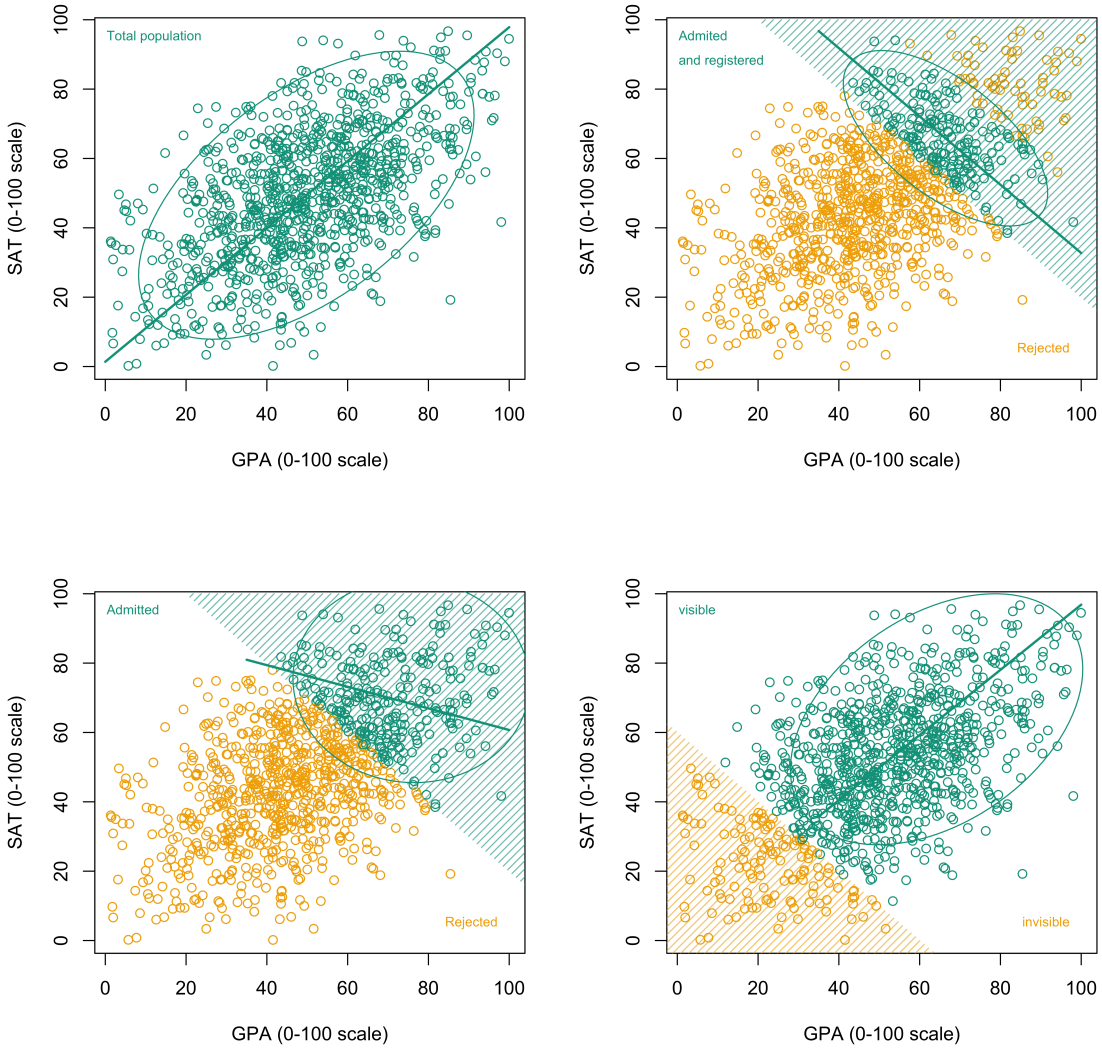


Figure 5.4: Relationship between the two variables, x_1 (GPA) and x_2 (SAT), as a function of the population studied: total population top left ($r \sim 0.55$), observable population top right, i.e. students who applied to the program ($r \sim 0.45$), population of students admitted to the program at the bottom left ($r \sim -0.05$), and population of students enrolled in the program on the bottom right ($r \sim -0.7$).
(source: author, dummy data)

This example on school admissions can be found in many cases, in insurance or banking. In credit risk, we try to estimate the probability that an underwriter will default. But we can wonder about the data available to estimate this probability (or to build a credit score). The first barrier is self-selection, as some people do

not apply for credit, for several reasons.

This self-selection is all the more disturbing since we often have no information on the people who do not apply for credit[12]. We can also think of speed cameras to identify areas where speeding would be frequent. Are all the speeds of the cars measured and recorded? or only the cars that have committed a speeding offence? Do the cameras operate permanently, or only at times when children are playing in the street? Does the device record only speed or also other suspicious behaviour? One can also think of a relatively popular database of road traffic accidents, where is recorded "*every road traffic accident known to the police is the subject of a Bulletin d'Analyse d'Accident Corporel (BAAC)*." We can imagine that some accidents do not appear in the database, and that some information is only partially reported, such as information related to blood alcohol content, as pointed out by Carnis and Lassarre (2019): does a missing information mean that the test was negative or that it was not done? It is crucial to know how the data was collected before you begin to analyze it. Missing information is also common in health insurance, as medical records are often relatively complex, with all kinds of procedure codes (that vary from one hospital to another, from one insurer to another). In France, the majority of drugs are pre-packaged (in boxes of 12, 24, 36), which makes it difficult to quantify the "real" consumption of drugs.

5.7.1 Goodhart's Law

When we want to measure a population's wealth, we can think of surveys or tax data. The former are expensive and complex to organize. The latter are biased, especially for very high incomes, which practice tax optimization (and that will change over time). In this area, the imagination is grand. For example, in the U.K., we can avoid inheritance tax by borrowing against a taxable asset (e.g., your home) and investing the loan in a non-taxable asset, such as a woodland; one can buy property through an offshore company, since non-U.K. companies and residents do not pay tax in the U.K. When loopholes in a tax system are discovered and people begin to make extensive use of it, but often this only leads to even more elaborate structures, which in turn have their own loopholes. This is Goodhard's Law.

As Marilyn Strathern has stated, Goodhart's Law says that "*once a measure becomes a goal, it ceases to be a good measure*." In U.S. Healthcare, Poku (2016) notes that starting in 2012, under the Affordable Care Act, Medicare began charging financial penalties to hospitals with "*higher than expected*" 30-day readmission rates. Consequently, the average 30-day hospital readmission rate for fee-for-service beneficiaries decreased. Is this due to improved efforts by hospitals to transition and coordinate care, or is it related to the increase in "observed" stays over the same period? Very often, setting a target based on a precise measure (here the 30-day readmission rate) makes this variable completely unusable to quantify the risk of getting sick again, but also has a direct impact on other variables (in this case the number of "observation" stays), making it difficult to monitor over time. On the Internet, algorithms are increasingly required to sort content, judge the defamatory or racist nature of tweets, see if a video is a deepfake, put a reliability score on a Facebook account, etc. Many people ask that algorithms be used to evaluate websites, to assess the quality of their content. A lot of people are asking for the reliability of content to be measured. Many people also demand that algorithms be made transparent, so that they know how these scores are created. Unfortunately, as noted by Dwoskin (2018) "*not knowing how* [Facebook is] *judging us is what makes us uncomfortable.But the irony is that they can't tell us how they are judging us - because if they do, the algorithms that they built will be gamed*," exactly as Goodhart's law implies.

As Desrosières (2016) wrote, "*quantitative indicators retroact on quantified actors*." In the spring of 2020, television news channels were giving, in continuous time, the number of people in intensive care, and

[12]To continue the analogy, in credit risk, we find the three previous levels, with (1) those who do not apply for credit, (2) those to whom the institution does not offer credit and (3) those who are not interested in the offer made.

the number of deaths in hospitals, measures that we will then find in the form of graphs, updated every week, or even every evening, on dedicated websites. In this period of crisis, at the height of hospital saturation, the National Health Service (NHS) in England asked each hospital to estimate its bed capacity in order to reallocate resources globally. Announcing that few beds were available was the best strategy to obtain more funding. This raises the question of how full the system really is, with each hospital having understood the rule and manipulating the measure as it sees fit. And just as troubling, while governments focused on hospitals (providing the official data used to make most indicators), nursing homes experienced disastrous death tolls, which took a long time to be quantified, as told by Giles (2020).

But aside from data errors, the idea of crime maps raises more subtle issues related to hidden data and feedback. Attention was drawn to these problems when the British insurance group Direct Line conducted a survey that found that "*10% of all British adults would definitely or probably consider not reporting a crime to the police because it would appear on an online crime map, which could have a negative impact on their ability to rent/sell or reduce the value of their property*." Instead of showing where incidents have occurred, the maps may show where people don't care to report them. This is quite different, and anyone making decisions based on this data could easily be misled. O'Neil (2016) also recalls the selection bias in early telematics data systems: "*in these early days, the auto insurers' tracking systems are opt-in. Only those willing to be tracked have to turn on their black boxes. They get rewarded with a discount of between 5% and 50% and the promise of more down the road. (And the rest of us subsidize those discounts with higher rates).*." More than twenty years ago, this problem had already been talked about, e.g. Morrison (1996), which recalls (quoting Stein (1994)), that certain insurance companies, in the United States, used the proof of domestic violence to discriminate against the victim by refusing her access to any form of insurance. The argument was the same as the one used centuries ago to prohibit life insurance: "*There is some fear that if the beneficiary is the batterer, we would be providing a financial incentive, if it's life insurance, for the proceeds to be paid for him to kill her*," reported Seelye (1994). Therefore, victims realized that if they sought medical or police protection, the resulting records could compromise their insurability. As a result, Morrison (1996) asserts that victims could stop seeking help, or reporting incidents of violence, in order to preserve their insurance coverage.

Another example is connected objects, finally providing insight into the behaviour of policyholders. Car insurers have known for a long time that the risk is very strongly linked to the behaviour of the motorist, but, as Lancaster and Ward (2002) pointed out, this hunch has never been used. "*As certainty replaces uncertainty*," in the words of Zuboff (2019), "*premiums that once reflected the necessary unknowns of everyday life can now rise and fall from millisecond to millisecond, informed by the precise knowledge of how fast you drive to work after an unexpectedly hectic early morning caring for a sick child or if you perform wheelies in the parking lot behind the supermarket*." Policyholders are rewarded when they improve their driving behaviour, "*relative to the broader policy holder pool*" as stated by Friedman and Canaan (2014). This approach, sometimes referred to as "gamification", may even encourage drivers to change their behaviour and risks. Jarvis et al. (2019) goes so far as to assert that "*insurers can eliminate uncertainty by shaping behaviour*."

5.7.2 Other Biases and "*Dark Data*"

In attempting to typify dark data biases, Hand (2020) listed dozens of other existing biases. Beyond the missing variables mentioned above, there is a particularly important selection bias. In 2017, during one of the debates at the NeurIPS conference on interpretability,[13], an example on pneumonia detection was mentioned: a deep neural network is trained to distinguish low-risk patients from high-risk patients, in order to determine

[13]Called "*The Great AI Debate: Interpetability is necessary for machine learning*," opposing Rich Caruana and Patrice Simard (for) to Kilian Weinberger and Yann LeCun (against) https://youtu.be/93Xv8vJ2acI.

who to treat first. The model was extremely accurate on the training data. Upon closer inspection, it turns out that the neural network found out that patients with a history of asthma were extremely low risk, and did not require immediate treatment. This may seem counter-intuitive, as pneumonia is a lung disease and patients with asthma tend to be more inclined to it (typically making them high-risk patients). Looking more in detail, asthma patients in the training data did have a low risk of pneumonia, as they tended to seek medical attention much earlier than non-asthma patients. In contrast, non-asthmatics tended to wait until the problem became more severe before seeking care.

Survival bias is another type of bias that is relatively well known and documented. The best known example is the one presented by Mangel and Samaniego (1984). During World War II, engineers and statisticians (British) were asked how to strengthen bombers that were under enemy fire. The statistician Abraham Wald began to collect data on cabin impacts. To everyone's surprise, he recommended armoring the aircraft areas that showed the least damage. Indeed, the aircraft used in the sample had a significant bias: only aircraft that returned from the theatre of operations were considered. If they were able to come back with holes in the wingtips, it meant that the parts were strong enough. And since no aircraft came back with holes in the propeller engines, those are the parts that needed to be reinforced. Another example is patients with advanced cancer. To determine which of two treatments is more effective in extending life spans, patients are randomly assigned to the two treatments and the average survival times in the two groups are compared. But inevitably, some patients survive for a long time - perhaps decades - and we don't want to wait decades to find out which treatment is better. So the study will probably end before all the patients have died. That means we won't know the survival times of patients who have lived past the study end date. Another concern is that over time, patients may die of causes other than cancer. And again, the data telling us how long they would have survived before dying of cancer is missing. Finally, some patients might drop out (for reasons unrelated to the study, or not). Once again, their survival times are missing data. Related to this example, we can return to another important example: why more people are dying from Alzheimer's disease than in the past? One answer may seem paradoxical: it is due to the progress of medical science. Thanks to medical advances, people who would have died young are now surviving long enough to be vulnerable to potentially long-lasting diseases, such as Alzheimer's disease. This raises all sorts of interesting questions, including the consequences of living longer.

Chapter 6

Some Examples of Discrimination

We will return here to the usual protected, or sensitive, variables that can lead to discrimination in insurance. We will mention direct discrimination, with race and ethnic origin, gender and sex, or age. We will also discuss genetic-related discrimination, and since several official protected attributes are not related to biology, but to social identity, we will get back to this concept. We will also discuss other inputs used by insurers, that could be related to sensitive attributes, with text, pictures and spatial information, and could be seen as some discrimination by proxy. We will also mention the use of credit scores and network data.

In this chapter, we will present and discuss some examples of discrimination. Again, and that was stressed previously, for example in Section 1.1.3, a pricing model can be seen as "discriminatory" (based on a criteria that we will present in Part III) without any intention. Using a sensitive attribute, or a variable correlated to the later, because it increases the accuracy of the model is making "rational discrimination" as discussed in Section 1.1.5, also named "*unintentional proxy discrimination*" in Prince and Schwarcz (2019). But could be illegal, and wrong. Assessing whether it is legal or not requires a jurist, and whether it is wrong or not a philosopher. Hellman (2011) addresses that issue, answering to the question "*when is discrimination wrong?*." In this chapter, we will try to present some features that are usually seen either as "sensitive" or as "correlated to a sensitive attribute", and discuss the issue in an insurance context.

6.1 Racial Discrimination

In a chapter dedicated to examples of discrimination, a natural and simple starting point could be "racism," corresponding to discrimination and prejudice towards people based on their race or ethnicity. As mentioned earlier (such as Table 5.1 for the U.S.), racism is commonly recognized as a wrongful form of discrimination.

According to Dennis (2004), "*racism is the idea that there is a direct correspondence between a group's values, behavior and attitudes, and its physical features ... Racism is also a relatively new idea: its birth can be traced to the European colonization of much of the world, the rise and development of European capitalism, and the development of the European and US slave trade.*" Nevertheless, we can probably also mention Aristotle (350-320 before our area), according to whom, Greeks (or Hellenes, Ελληνες) were free by nature, while "barbarians" (βαρβαρος, non-Greeks) were slaves by nature. But in a discussion on Aristotle

and racism, Puzzo (1964) claimed that "*racism rests on two basic assumptions: that a correlation exists between physical characteristics and moral qualities; that mankind is divisible into superior and inferior stocks. Racism, thus defined, is a modern conception, for prior to the* XVI*-th century there was virtually nothing in the life and thought of the West that can be described as racist. To prevent misunderstanding a clear distinction must be made between racism and ethnocentrism ... The Ancient Hebrews, in referring to all who were not Hebrews as Gentiles, were indulging in ethnocentrism, not in racism. ... So it was with the Hellenes who denominated all non-Hellenes*." We could also mention the "*blood purity laws*" ("*limpieza de sangre*") that were once commonplace in the Spanish empire (that differentiated converted Jews and Moors (*conversos* and *moriscos*) from majority Christians, as discussed in Kamen (2014)), and require each candidate to prove, with family tree in hand, the reliability of his or her identity through public disclosure of his or her origins.

6.1.1 A Sensitive Variable Difficult to Define

There are dozens of books, and dedicated entries in encyclopedias, discussing "race," and its scientific grounds, such as Memmi (2000), Dennis (2004), Ghani (2008) or Zack (2014). Historically, we should probably start in the XIX-th century, with Georges-Louis Leclerc de Buffon ("*the father of all thought in natural history in the second half of the* XVIII*-th century*," Mayr (1982)) and Carl Linnaeus ("*father of modern taxonomy*," Calisher (2007)), who defined "varietas" or "species" – the largest group of organisms in which any two individuals of the appropriate sexes or mating types can produce fertile offspring – and "subspecies" – rank below species, used for populations that live in different areas and vary in size, shape, or other physical characteristics. Following Keita et al. (2004), we can consider "race" as a synonym for "subspecies." At the same period, Johann Friedrich Blumenbach explored the biodiversity of humans, in Blumenbach (1775), mainly by comparing skull anatomy and skin color, an suggested five human "races," or more precisely "*generis humani varietates quinae principes, species vero unica*" (one species, and five principle varieties of humankind): the "*Caucasian*" (or white race, for Europeans, including Middle Easterners and South Asians in the same category), the "*Mongolian*" (or yellow race, including all East Asians), the "*Malayan*" (or brown race, including Southeast Asians and Pacific Islanders), the "*Ethiopian*" (or black race, including all sub-Saharan Africans), and the "*American*" (or red race, including all Native Americans), as discussed in Rupke and Lauer (2018). Johann Friedrich Blumenbach and Carl Linnaeus investigated the idea of "*human race*," from an empirical perspective, and at the same time, Immanuel Kant became one of the most notable Enlightenment thinkers to defend racism, from a philosophical and scientific perspective, as discussed in Eze (1997) or Mills (2017) (even if Kant (1795) ultimately rejected racial hierarchies and European colonialism).

A more simple version of "racism" is the discrimination based on skin color, also known as "colourism", or "shadeism", which is a form of prejudice and discrimination in which people who are perceived as belonging to a darker skinned race are treated differently based on their darker skin color. Somehow, this criteria could be seen as more objective, since it is based on a color. Telles (2014) as defined by the Fitzpatrick Skin Scale, used by dermatologists and researchers. Types 1 through 3 are widely considered lighter skin, and 4 through 6 as darker skin. On could also consider a larger palette, as on Figure 6.1.

This link between "racism" and "colourism" is an old one. Kant (1775), entitled "*Of the Different Human Races*" as translated in English (the original title was "*Rassender Menschen*" in German), was the preliminary work for his lectures on physical geography (collected in Rink (1805)). In Kant (1785), translated as "*Determination of the Concept of a Human Race*," his initial position, on the existence of human races, was confirmed. The first essay was published a few months before Johann Friedrich Blumenbach's "*de generis humani varietate nativa*, and proposed a classification system less complex than the one in Blumenbach (1775), based almost solely on color. Both used color as a way of differentiating the races, even if it was

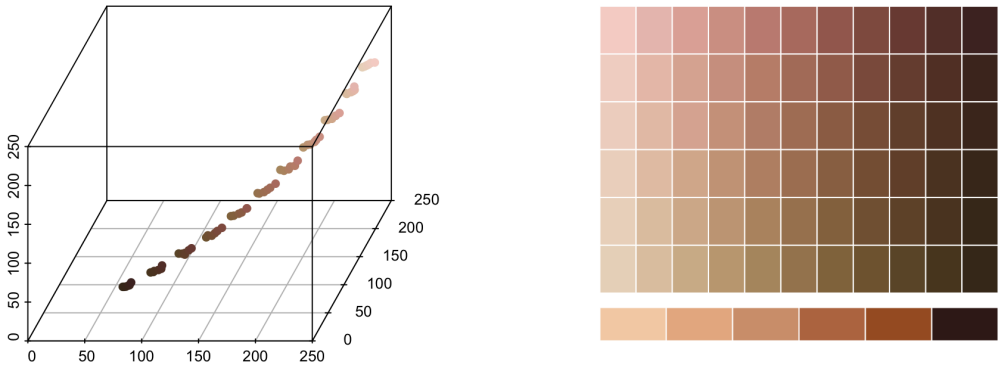


Figure 6.1: On the right, color chart showing various shades of skin color, by L'Oréal (2022), as well as Fitzpatrick Skin Scale at the bottom (six levels), and on the left, the RGB (red-green-blue) decomposition of those colors.
(source Telles (2014))

already quite criticized. For Immanuel Kant, there were four "*races*," Whites, Blacks, Hindustanic, and Kalmuck, and the explanation of such differences were the effects of air and sun. For example, he argued that by the solicitude of nature, human beings were equipped with seeds ("*keime*") and natural predispositions ("*anlagen*") that were developed, or held back, depending on climate.

The use of colors could be seen as simple and practical, with a more objective definition than the one underlying the concept of "race", but it is also very problematic. For example, Hunter (1775) classified as "*light brown*" Southern Europeans, Sicilians, Abyssinians, the Spanish, Turks, and Laplanders, and as "*brown*" Tartars, Persians, Africans on the Mediterranean, and the Chinese. In the xx-th century in the U.S., simple categories such as "*white*" and "*black*" was not enough. As mentioned by Marshall (1993), in New York City, populations who spoke Spanish was not usually referred to as belonging either to the "*white race*" or the "*black race*," but were designated as '*Spanish*," '*Cuban*," or '*Puerto Ricans*." Obviously, the popular racial typologies used in the U.S. were not based on any competent genetic studies. And it was the same almost everywhere. In Brazil for instance, descendants played a negligible role in establishing racial identity. Harris (1970) has shown that siblings could be assigned to different racial categories. He counted more than forty racial categories that where used in Brazil. Phenotypical attributes (such as skin color, hair form, and nose or mouth shape) were entering into the Brazilian racial classification, but the most important determinant of racial status was the socio-economic position. It was the same for the Eta in Japan, Smythe (1952) and Donoghue (1957). As Atkin (2012) wrote, an "*east African will be classified as 'black' under our ordinary concept but this person shares a skin colour with people from India and a nose shape with people from northern Europe. This makes our ordinary concept of race look to be in bad shape as an object for scientific study – it fails to divide the world up as it suggests it should.*"

From a scientific and biological perspective, races do not exist (there are no "*biobehavioral racial essences*," as Mallon (2006) wrote it), and that in the sociological sense, races are created by racism. Racism

Box 6.1 **Race and Ethnicity in the U.S.**, by Elisabeth Vallet[1]

Since the early days of the Republic, the U.S. Census has identified individuals according to broad racial and ethnic categories. In addition to the traditional federal categories (White/Caucasian, Black/African American, Asian American, American Indian/Alaska Native, Native Hawaiian/Pacific Islander, Multiracial), the decennial census adds the Hispanic/Latino category. It is also expected that eventually the Middle East and North Africa category will be included in the census.

This racial classification, which now serves as the basis for statistical analysis in the country, has its roots in slavery and a classification designed to establish a form of racial purity, when the "single drop of blood" and "blood quantum" rules (Villazor (2008)) were applied to determine racial group membership and helped shape the population groups involved and their own self-perceptions. Moreover, the complexity of this classification lies in the fact that it is not always based on constant elements. On the one hand, the categories, as stated in the forms, change over time. Therefore, the nature of the questionnaire and the way in which the questions are formulated, because they evolve, sometimes alter the overall picture of the population over time.

On the other hand, this classification is now based on the self-identification of individuals, which may also vary from one census to another. Therefore, some studies have shown significant changes in the self-identification of certain individuals, which must be taken into account when using these data. According to Liebler et al. (2017), for example, there was significant variation for the same individuals between the 2000 and 2010 censuses, particularly in categories that could report multiple ethnic or Hispanic backgrounds. For example, the study shows that the consistency of responses was most pronounced among non-Hispanic whites, blacks, and Asians. Identification instability was more pronounced among people who identified as Native American, Pacific Islanders, people with multiple backgrounds, and Hispanics. There are many reasons for this, but some of them have to do with the evolution of the integration of immigrant communities, or with the fact that miscegenation can lead to the prevalence of one identity over the other over time. Therefore, analyses must include the fact that the very idea of race/ethnicity as a social construct must be included as such in the analysis of statistical data.

here is a set of processes that create (or perpetuate) inequalities based on racialising some groups, with a '"privileged" group that will be favoured, and a "racialized" group that will be disadvantaged. Given the formal link between racism and perceived discrimination, it is natural to start with this protected variable (even if it is not clearly defined at the moment). Historically, in the United States, the notion of race has been central in discussions on discrimination (Anderson (2004) provides a historical review in the United States), see also Box Box 6.1, by Elisabeth Vallet.

6.1.2 Race and Risk

As recalled by Wolff (2006), in 1896, Frederick L. Hoffman, an actuary with Prudential Life Insurance, published a book demonstrating, with statistics, that the American black man was uninsurable (see Hoffman (1896)). Du Bois (1896) wryly noted that the death rate of blacks in the United States was only slightly higher (but comparable) to that of white citizens in Munich, Germany, at the same time. But above all, the main criticism is that it aggregated all kinds of data, preventing a more refined analysis of other causes of (possible) excess mortality (this is also the argument put forward by O'Neil (2016)). At that time, in the United States, several states were passing anti-discrimination laws, prohibiting the charging of different premiums on the basis of racial information. For example, as Wiggins (2013) points out, in the summer of 1884, the Massachusetts state legislature passed the Act to Prevent Discrimination by Life Insurance Companies Against People of Color. This law prevented life insurers operating in the state from making any distinction or discrimination between white persons and colored persons wholly or partially of African descent, as to the premiums or rates charged for policies upon the lives of such persons. The law also required

[1]Professor, Director of the Raoul Dandurand Center in Montréal, Canada

insurers to pay full benefits to African American policyholders. It was on the basis of these laws that the uninsurability argument was made: insuring blacks at the same rate as whites would be statistically unfair, argued Hoffman (1896), and not insuring blacks was the only way to comply with the law (see also Heen (2009)). As Bouk (2015) points out "*industrial insurers operated a high-volume business; so to simplify sales they charged the same nickel to everyone. The home office then calculated benefits according to actuarially defensible discrimination, by age initially and then by race. In November 1881, Metropolitan decided to mimic Prudential, allowing policies to be sold to African Americans once again, but with the understanding that black policyholders' survivors only received two-thirds of the standard benefit.*"

In the credit market, Bartlett et al. (2021), following up on Bartlett et al. (2018), shows that in the U.S., discrimination based on ethnicity has continued to exist in the U.S. mortgage market (for African Americans and Latinos), both traditional and algorithm-based lending. But algorithms have changed the nature of discrimination from one based on prejudice, or human dislike, to illegitimate applications of statistical discrimination. Moreover, algorithms discriminate not by denying loans, as traditional lenders do, but by setting higher prices or interest rates. In health care, Obermeyer et al. (2019) shows that there is discrimination, based on ethnicity or "racial" bias in a commercial software program widely used to assign patients requiring intensive medical care to a managed care program. White patients were more likely to be assigned to the care program than black patients in comparable health status. The assignment was made using a risk score generated by an algorithm. The calculation included data on total medical expenditures in a given year and fine-grained data on health service use in the previous year. The score therefore does not reflect expected health status, but predicts the cost of treatment. Bias, stereotyping, and clinical uncertainty on the part of health care providers can contribute to racial and ethnic disparities in health care, as noted by Nelson (2002). Finally, in auto insurance, Heller (2015) found that predominantly African American neighbourhood pay 70% more, on average, for auto insurance premiums than other neighbourhoods. In response, the Property Casualty Insurers Association of America responded[2] in November 2015 that "*insurance rates are color-blind and solely based on risk*." This position is still held by actuarial associations in the United States, for whom questions about discrimination are meaningless. Larson et al. (2017) obtained 30 million premium quotes, by zip code, for major insurance companies across the United States, and confirmed that a gap existed, albeit a smaller one. Also, in Illinois, insurance companies charged on average more than 10% more in auto liability premiums for "*majority minority*" zip codes (in the sense that the rate of minorities was the highest) than in majority white zip codes. Historically, as recalled by Squires (2003), many financial institutions have used such discrimination by refusing to serve predominantly African American geographic areas.

While such analyses have recently proliferated (Klein (2021) provides a relatively comprehensive literature review), this potential racial discrimination issue was analysed by Klein and Grace (2001), for instance, who offered the possibility of controlling covariates correlated with race, and showed that there was no statistical evidence of geographic redlining. This conclusion was consistent with the analysis of Harrington and Niehaus (1998), and was subsequently echoed by Dane (2006), Ong and Stoll (2007) or Lutton et al. (2020) among many others. It should be noted here that redlining is not only associated with an antisocial criterion, but very often with an economic criterion. A recent case of statistical discrimination is currently being investigated in Belgium, as mentioned by Orwat (2020). In this country, the energy supplier EDF Luminus refuses to supply electricity to people living in a certain postal code district. For the energy supplier, this postal code area represents a zone where many people with bad credit history live. Redlining, which relies on the proxy variable "*place of residence*," was named so because it encircles areas with red lines, as noted by Barocas and Selbst (2016).

If the term "ethnic statistics" is a sensitive subject in France, the censuses ask, traditionally (for more than a century), the nationality at birth, therefore distinguishing French by birth and French by adoption. And since

[2]Online on their website, https://www.pciaa.net/, see https://bit.ly/43ls6eb.

1992, the variable “parents’ country of birth" has been introduced in a growing number of public surveys. In French statistics, the word “ethnic" in the anthropological sense (sub-national or supra-national human groups whose existence is proved even though they do not have a state) has long had a place, particularly in surveys on migration between Africa and Europe. Nevertheless, in legal texts, in which it is sometimes a euphemistic substitute for *empire*. The 1978 Data Protection Act therefore uses the expression “*racial or ethnic origins*." In this sense, “*ethnic origin*" means any reference to a foreign origin, whether it is a question of nationality at birth, the parents’ nationality, or “reconstructions" based on the family name or physical appearance. Some general intelligence and judicial police files contain personal information on an individual’s physical characteristics, and in particular on their skin color, as recalled by Debet (2007). Some medical research files (e.g. in dermatology) may contain similar information. INSEE had initially refused to introduce a question on the parent’s birthplace in its 1990 family survey, which could have served as a sampling frame. It was not until the 1999 survey that it was introduced, as recalled by Tribalat (2016). Another concern that may arise is that the difference that may exist in insurance premiums between ethnic origins is not a reflection of different risks, but of different treatments. Therefore, Hoffman et al. (2016) shows that racial prejudice and false beliefs about biological differences between black and white people continue to shape the way we perceive and treat the former - they are associated with racial disparities in pain assessment and treatment recommendations.

6.2 Sex and Gender Discrimination

Aristotle (350-320 before our area) suggested that a woman actually is a man that had failed to develop to his full potential, Horowitz (1976) and **?**. Aristotle’s *Politics* (Πολιτικα) was a standard textbook in medieval and early modern universities On the other hand, Plato (380-350 before our area) encouraged common education of women and men for military, intellectual, and political leadership, in *Republic* (Πολιτεια) unfortunately, it was not widely read in the West (until its translation from Greek to Latin) during the early XVth century, as mentioned in Allen (1975).

According to online version the Encyclopædia Britannica, “sexism" is a “*prejudice or discrimination based on sex or gender, especially against women and girls*," that echoes the New Oxford American Dictionary, where “sexism" is a “*prejudice, stereotyping, or discrimination, typically against women, on the basis of sex.*" For Cudd and Jones (2005), “*"sexism refers to a historically and globally pervasive form of oppression against women.*"

Before discussing the “gender directive", in Box 6.2, Émilie Biland-Curinier explains the differences between “sex" and “gender".

6.2.1 Sex or Gender ?

In Box 6.3, we list several concepts related to sexual identity. As an anecdote, in 2018, a cisgender Canadian man changed the gender identification on his driver’s license from male to female for the sole purpose of benefiting from lower motor vehicle liability insurance premiums available to “female drivers", said Ashley (2018).

[3]Professor at Science-Po Paris

Box 6.2 **Sex and Gender**, by Émilie Biland-Curinier[3]

We traditionally differentiate sex, which is a biological characteristic (related to physical and physiological features, e.g. chromosomes, gene expression, hormone levels and function, etc.) and gender, which refers to the sexual identity of an individual. Sex and gender are often described in binary terms (girl/woman or boy/man). However, the diversity of sexual development and atypical formulas is important, whether they are of chromosomal, hormonal or environmental origin. In reality, the sex/gender debates are numerous, and are refracted in many fields of knowledge and public action.
In the social sciences, sex is today considered less as a "*biological reality*" than as a social and, above all, legal construction: sex is the one attributed to each individual on his or her birth certificate and then on all of his or her more or less official papers. Most people keep this native and legal sex all their life but some change it (a large part of "trans" people, whether they have undergone genital surgery or not). In the case of intersex/intersex people, the assignment of legal sex involves a good deal of arbitrariness (or more sociologically, medical discretion), since the biological attributes do not fit into any of the binary female/male categories. Most countries base this legal category of sex on this duality, but some have recently opened up other possibilities (e.g.: possible to fill 'X' in the Netherlands, and 'various' in Germany). In these countries, we speak of neutral sex/gender, in others of neutral third.
In statistics, the "sex" category is the one that is mostly used. This category is indeed declarative: it reflects most often the legal sex, but in the case where individuals are asked to fill it in (e.g.: census), there may be very few discrepancies. Interesting fact: Statistics Canada has recently adapted its categories to take into account gender identity (i.e. to make trans people visible).
As a result, gender relations in sociology and economics (and in particular inequalities between women and men) are mainly analyzed quantitatively on the basis of the sex variable. For a brief presentation of this issue in the French context Grobon and Mourlot (2014), otherwise Amossé and De Peretti (2011). Finally, the definition gender as a social science concept, one can quote philosopher Elsa Dorlin: "*The concept of gender has made it possible to historicize the identities, the roles and the symbolic attributes of the feminine and the masculine, defining them, not only as the product of a differentiated socialization of the individuals, specific to each society and variable in time, but also as the effect of an asymmetrical relation, of a power relation*."
In this sense, the gender relationship can be defined with Joan Scott as follows: "*Gender is a primary way of signifying power relationships. The categories of masculine and feminine, such as "men" and "women", therefore have meaning and existence only in their antagonistic relationship, and not as "identities" or as "essences" taken separately*" (Dorlin (2005)).

Box 6.3 **Gender Identity**, in Canada

The gender identity categories offered as potential responses represent the considerable diversity in how individuals and groups understand, experience and express gender identity.
Gender fluid refers to a person whose gender identity or expression changes or shifts along the gender spectrum.
Man refers to a person who internally identifies and/or publicly expresses as a man. This may include cisgender and transgender individuals. Cisgender means that one's gender identity matches one's sex assigned at birth.
Nonbinary refers to a person whose gender identity does not align with a binary understanding of gender such as man or woman.
Trans man refers to a person whose sex assigned at birth is female, and who identifies as a man.
Trans woman refers to a person whose sex assigned at birth is male, and who identifies as a woman.
Two-Spirit is a term used by some North American Indigenous people to indicate a person who embodies both female and male spirits or whose gender identity, sexual orientation or spiritual identity is not limited by the male/female dichotomy.
Woman refers to a person who internally identifies and/or publicly expresses as a woman. This may include cisgender and transgender individuals. Cisgender means that one's gender identity matches one's sex assigned at birth.

6.2.2 Sex, Risk and Insurance

Many books, published in the XVIII-th century, mention that men and women have very different behaviors when it comes to insurance. According to Fish (1868), "*upon no class of society do the blessings of life*

Figure 6.2: Probability of dying between 30 and 70 years old, on the left, for most countries, and life expectancy at birth, on the right. Women are on x-axis while men are on the y-axis. The size of the dots is related to the population size.
(data source Ortiz-Ospina and Beltekian (2018)).

insurance fall so sweetly as upon women. And yet" agents "*have more difficulty in winning them to their cause than their husbands*." Phelps (1895) asked explicitly "*do women like insurance?*" while Alexander (1924) collects fables and short stories, published by insurance companies, with the idea of scaring women by dramatizing the "*disastrous economic consequences of their stubbornness*," as Zelizer (2018) named it.

Women live longer than men across the world and scientists have by and large linked the sex differences in longevity with biological foundation to survival. A new study of wild mammals has found considerable differences in life span and aging in various mammalian species. Among humans, women's life span is almost 8% on average longer than men's life span. But among wild mammals, females in 60% of the studied species have, on average, 18.6% longer lifespans. Everywhere in the world women live longer than men – but this was not always the case (see also life tables in Chapter 2). On Figure 6.2, we can compare on the left, the probability (at birth) to die between 30 and 70 years also, in several countries, for women (x-axis) and men (y-axis). On the right, we compare life expectancy at birth.

6.2.3 The "Gender Directive"

The 2004 EU Goods and Services Directive, Council of the European Union (2004), aimed to reduce gender gaps in access to all goods and services, discussed for example by Thiery and Van Schoubroeck (2006). A special derogation in Article 5(2) allowing insurers to set gender-based prices for men and women. Indeed, "*Member States may decide (...) to allow proportionate differences in premiums and benefits for individuals where the use of sex is a determining factor in the assessment of risk, on the basis of relevant and accurate actuarial and statistical data.*" In other words, this clause allowed an exception for insurance companies, provided that they produced actuarial and statistical data that established that sex was an objective factor

Box 6.4 **Gender Inequality and Discrimination**, by Avner Bar-Hen[4]

We all have our own ideas about inequality and talk about it as if it were something simple. Let's take a look at the making of gender disparity indicators. Traditionally there are two main approaches to measuring gender inequality indices: (i) The first method measures gender equality using surveys, usually as a complement to other questions. In this case, knowledge of gender equality is not the main objective. (ii) The second approach is based on targets and statistics to quantify them. The first international indices only date back to 1995 and are heavily based on the UNDP Human Development Index (HDI). The HDI is a well-theorized and widely disseminated concept; this makes it possible to know fairly precisely what is being measured, but international comparisons are difficult. The complementary variables to the HDI that have been proposed are often interesting, but above all complicate the readability of the results.

Since the 2000s, several alternatives have been proposed. Among the most popular are the Gender Equity Index (GEI, introduced by Social Watch in 2004) and the Global Gender Gap Index (GGGI proposed by the World Economic Forum in 2006). These indices claim to be measures of gender equality, but their general concepts are not clearly formulated. For example, the GGGI ignores the underlying causes of gender inequality, such as health. Other proposals such as the Social Institutions and Gender Index (SIGI, proposed by the OECD in 2007) focus on social institutions that affect gender equality, but also on family codes or property rights. This index can be seen as a combined measure of women's disadvantages relative to men regarding certain basic rights and the fulfillment of distinctive rights for women. The lack of symmetry in the SIGI indicators and the scales of the indicators makes it ultimately unclear what is being measured.

The purpose of these different indices is to assess multiple aspects of gender disparities, not only in academic research on the causes and consequences of gender inequality, but also to inform public policy debates. A single index is never the solution to the problem caused by the large number of indicators: it is therefore necessary to compare different gender indices. However, the description of methodologies uses different terminologies, does not adequately describe the methodological choices and is often silent on potential sources of measurement error. Finally, index developers are rarely explicit about the overall concept they seek to measure. Another important point is that the use of gender data raises questions about the quality of the data. If they are of questionable quality, the legitimacy of their use for collective purposes is questionable. The methodological choices underlying the construction of indices often show that what the index measures is different from what it claims to determine!

The valuable contributions made by the proponents of gender indices must be recognized. They are a resource that has made it possible to describe gender disparities and promote women's rights in a more effective way than was possible before 1995. Having a comprehensive empirical index is better than not having one, even if current measures suffer from methodological weaknesses. However, producing multiple indices also creates a problem for users: they do not have the means to compare them with each other.

in the assessment of risk. The European Court of Justice canceled this legal exception in 2011, in a ruling discussed at length by Schmeiser et al. (2014) or Rebert and Van Hoyweghen (2015), for example. This regulation, which generated a lot of debate in Europe in 2007 and then in 2011, also raised many questions in the United States, several decades earlier, in the late 1970s, with Martin (1977), Hedges (1977) and Myers (1977). For example, in Los Angeles, Department of Water and Power vs. Manhart, the Supreme Court considered a pension system in which female employees made higher contributions than males for the same monthly benefit because of longer life expectancy. The majority ultimately determined that the plan violated Title VII of the Civil Rights Act of 1964 because it assumes that individuals conform to the broader trends associated with their gender. The court suggested such discrimination is troubling from a civil rights perspective because it does not treat individuals as such, as opposed to merely group members they belong to. These laws were driven, in part, by the fact that employment decisions are generally individual: A specific person is hired, fired, or demoted, based on his or her past or expected contribution to the employer's mission. In contrast, stereotyping of individuals based on group characteristics is generally more tolerated in fields such as insurance, where individualized decision making does not make sense.

In Box 6.4, Avner Bar-Hen discusses measurement of gender related inequalities.

[4]Professor at the Conservatoire National des Arts et Métiers in Paris

6.3 Age Discrimination

As Robbins (2015) said, "*if you are not already part of a group disadvantaged by prejudice, just wait a couple of decades—you will be*". As explained in Ayalon and Tesch-Römer (2018), "agism" is a recent concept, defined in a very neutral way in Butler (1969) as "*a prejudice by one age group against another age group.*" Butler (1969) argued that ageism represents discrimination by the middle-aged group against the younger and older groups in society, because the middle-aged group is responsible for the welfare of the younger and older age groups, which are seen as dependent. But according to Palmore (1978), ageing is seen as a loss of functioning and abilities, and therefore, it carries a negative connotation. Accordingly, terms such as "old" or "elderly" have negative connotations and thus should be avoided.

Age contrasts with race and sex in that our membership of age-based groups changes over time, as mentioned in Daniels (1998)). And unlike caterpillars becoming butterflies, human aging is usually considered as a continuous process. Therefore, the boundaries of age-defined categories (such as the "under 18" or the "over 65") will always be based on arbitrariness.

6.3.1 Young or Old ?

According to of the European Union (2018), "*on the grounds of age do not constitute discrimination* (...) *if age is a determining factor in the assessment of risk for the service in question and this assessment is based on actuarial principles and relevant and reliable statistical data*." In the United States, the idea that age can be a ground for discrimination was reflected in the 1967 Age Discrimination in Employment Act, which followed the 1964 Civil Rights Act, which focused primarily on ethical and racial issues, as shown in the example of Macnicol (2006). In the majority of cases, age discrimination is considered from an employment perspective, as Duncan and Loretto (2004) or Adams (2004). In terms of insurance, age is considered "*less discriminatory*" than gender, as we have seen, because as Macnicol (2006) observes, age is not a club in which one enters at birth, and it will change with time. We can note for some insurers refusing to discriminate according to age is an important factor, a form of "*raison d'être*" (in the sense given by the 2019 French Pact law). In France, for example, Mutuelle Générale is committed to strengthening solidarity between generations. This should be reflected in a rejection of discrimination, segmenting, according to this criterion.

But just as a distinction exists between biological sex and gender, some suggest distinguishing between biological age and perceived (or subjective) age, such as Stephan et al. (2015) or Kotter-Grühn et al. (2016). Uotinen et al. (2005) showed that this subjective age would be a better predictor of mortality than biological age. As Beider (1987) points out, it can be argued that if people do not have a fair chance based on their age, because not everyone ages equally, people die at different ages. Bidadanure (2017) reminds us that age discrimination is always perceived as less "preoccupying" than other kinds. The aging process, from birth to adulthood, is correlated with various developmental and cognitive processes that make it relevant to assign different responsibilities, consent capacities and autonomy to children, young adults or the elderly. But unlike sex and race, age is not a discrete and immutable characteristic. As the saying goes, age is not a club we are born in. We expect to go through the different stages of a life and old age is a club we know we will most likely be joining one day. Therefore, differential treatment on age does not necessarily generate inequalities among people over time while differential treatment on ethnicity and gender does: "*a society that relentlessly discriminates against people because of their age can still treat them equally throughout their lives. Everyone's turn [to be discriminated against] is coming*" said Gosseries (2014). Citing a decision of the Court of Appeal of 2008, Mercat-Bruns (2020) recalls that "*the legislator was careful to make a distinction between age and health, and these two grounds cannot be confused by considering that advanced*

age necessarily implies poor health"."

In automobile insurance, as recalled by Cooper (1990), Liisa (1994) and Clarke et al. (2010), the increase in the number of claims among the elderly can be explained by a loss of sensory and motor acuity, the use of medication (in particular psychotropic drugs), and a decrease in reflexes. But the elderly also tend to drive less, as pointed out by Fontaine (2003). Figure 6.3 shows the frequency of accidents and the number of fatalities, per million kilometres driven, for different age groups. The risk of personal injury (or death) in a car accident increases significantly from age 60 onwards and continues to increase rapidly with age, as shown by

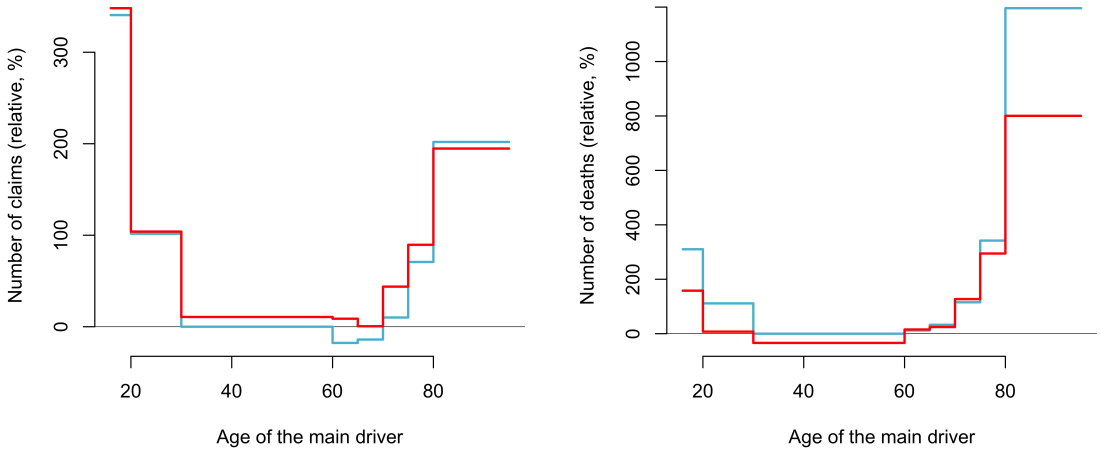


Figure 6.3: Number of crashes (left) and number of fatalities (right), per million miles driven, for both males and females (males in blue and females in red), by driver age. The reference (0) are men aged 30-60 years. The number of accidents is three times higher (+200%) for those over 85, and the number of deaths more than ten times higher (+900%).
(source: Li et al. (2003))

But in the majority of countries, the average risk of older people does not appear to be particularly important (as long as older people are seen as a homogeneous group). The Figure 6.4 shows the evolution of the (annual) frequency of claims, the average cost of claims, and the average premium for automobile insurance in Quebec, as an age function of insured (by age group), for collision or upset coverage.

To go further, Dulisse (1997), Meuleners et al. (2006) and Cheung and McCartt (2011) note that the share of responsibility in accidents also increases with age, in particular more accidents on the right-hand side, often reflecting failure to give way. Many countries have raised the question of regulation in relation to very advanced ages (over 80 years). In terms of disability, insurers are not allowed to discriminate on the basis of disability if the person is allowed to drive. But for degenerative diseases, a few laws explicitly prohibit driving, for example for someone with an established disease, such as Parkison's disease, Crizzle et al. (2012). The fact that older people are more responsible for accidents raises many moral questions, as putting oneself at risk as a driver is one thing, but potentially injuring or killing others is less acceptable.

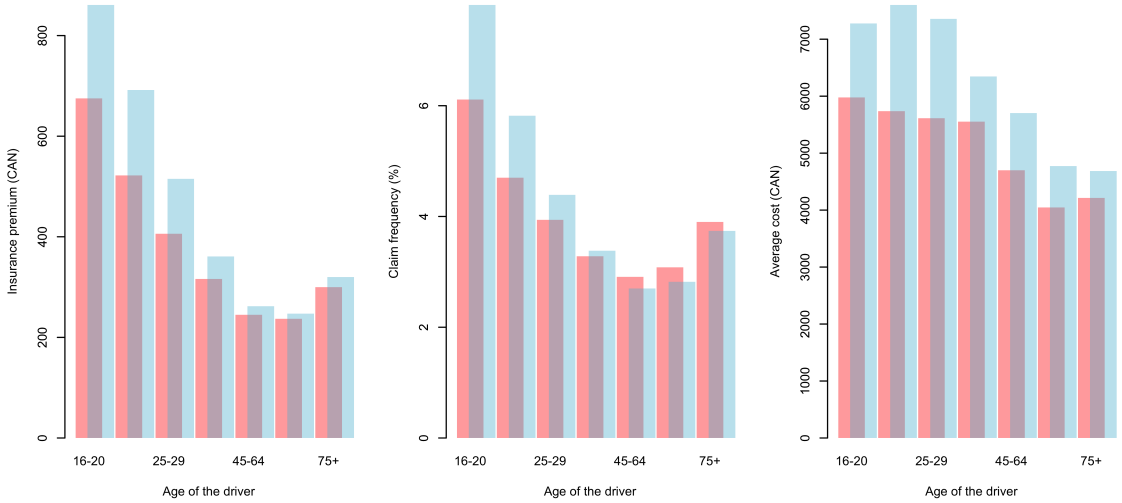


Figure 6.4: From left to right, average written premium (Canadian dollars), claims frequency, and average claims cost (Canadian dollars), by age group (x-axis) and gender (males in blue and females in red) in Quebec. (source, Groupement des Assureurs Automobiles (2021))

6.4 Genetics versus Social Identity

6.4.1 Genetic Related Discrimination

Genetic discrimination, or “genoism", occurs when people treat others (or are treated) differently because they have or are perceived to have a gene mutation(s) that causes or increases the risk of an inherited disorder. According to Ajunwa (2014, 2016), “*genetic discrimination should be defined as when an individual is subjected to negative treatment, not as a result of the individual's physical manifestation of disease or disability, but solely because of the individual's genetic composition*." This concept is related to “genetic determinism" (as defined in de Melo-Martín (2003) and Harden (2023)) or more recently “genetic essentialism" (as in Peters (2014)). Dar-Nimrod and Heine (2011) defined “genetic essentialism" as the belief that people of a same group share some set of genes that make them physically, cognitively, and behaviorally uniform, but different from others. Consequently, “genetic essentialists" believe that some traits are not influenced (or only little) by the social environment. But as explained in Jackson and Depew (2017), essentialism is genetically inaccurate since, not only it overestimates the amount of genetic differentiation between human races , but also, it underestimates the amount of genetic variation among same race individuals (Rosenberg (2011); Graves Jr (2015)).

An important issue here is that “*genetic discrimination should be defined as when an individual is subjected to negative treatment, not as a result of the individual's physical manifestation of disease or disability, but solely because of the individual's genetic composition*," Ajunwa (2014). And that is usually difficult to assess (as we can see, for instance, with obesity). But other definitions are more vague. For example, according to Erwin et al. (2010), “*the denial of rights, privileges, or opportunities or other adverse treatment based solely on genetic information, including family history or genetic test results*," could be seen

Box 6.5, **Genetic tests: individuals and insurers**, Bélisle-Pipon et al. (2019)

"*'Genetic tests' largely refers to germline (not somatic) DNA changes that are discovered in apparently healthy individuals, through direct-to-consumer genetic testing, consumer-facing but physician-mediated genetic testing over the internet or the rare but growing phenomenon of predictive genomics clinics in conventional medical environments. This testing may result in the identification of disease risk information through either the presence of monogenic risk variants, Torkamani et al. (2018), or extremes in polygenic risk scores, Vassy et al. (2017). DNA testing is not fully accurate at predicting disease, and this inaccuracy can arise in two ways. The techniques used (generally next-generation sequencing) can have analytic errors, producing incorrect 'calls' such that the variants identified are simply wrong. These types of errors are increasingly rare as the technology improves. However, simply carrying a pathogenic variant or a high polygenic risk score does not mean that someone will eventually get the disease, as genetic markers have variable 'penetrance', Cooper et al. (2013). Thus, genetic markers, even well-accepted pathogenic variants, may occur in individuals who will never develop the disease in question. While these points are true, the presence of both pathogenic variants for monogenic diseases and very high polygenic risk scores may increase the probability that an individual will develop the disease in question and thus can form the basis of the discrimination concerns. While there is an ongoing debate about the clinical utility of this information in the individual patient, there is no question that such information can predict risk on a population basis and is thus of great interest to life insurance companies and others that are in the business of estimating and monetizing risk.*"

as a "genetic discrimination". According to the legislation in Florida (Title XXXVII, Chapter 627, Section 4301, 2017) "*genetic information means information derived from genetic testing to determine the presence or absence of variations or mutations, including carrier status, in an individual's genetic material or genes that are scientifically or medically believed to cause a disease, disorder, or syndrome, or are associated with a statistically increased risk of developing a disease, disorder, or syndrome, which is asymptomatic at the time of testing. Such testing does not include routine physical examinations or chemical, blood, or urine analysis unless conducted purposefully to obtain genetic information, or questions regarding family history*". In Box 6.5, an excerpt explaining what "genetic tests" are is given, from Bélisle-Pipon et al. (2019).

According to Rawls (1971), the starting point for each person in society is the result of a social lottery (the political, social, and economic circumstances in which each person is born) and a natural lottery (the biological potentials with which each person is born - recently, Harden (2023) revisits this genetic lottery and its social consequences). John Rawls argues that the outcome of each person's social and natural lottery is a matter of good or bad "fortune" or "chance" (like ordinary lotteries). And since it is impossible to deserve the outcome of this lottery, discriminations resulting from these lotteries should not exist. For egalitarians ("luck-egalitarians"), it is appropriate to eliminate the differential effects on people's interests that, from their perspective, are a matter of luck. Affirmative action in favor of women is a means to neutralize the effects of sexist discrimination. Stone (2007) revisits the idea that this ex ante equality is part of what makes lotteries fair and appealing. Abraham (1986) discusses the consequences of natural lotteries in insurance. Around the same time, Wortham (1986) stated, "*those suffering from disease, a genetic defect, or disability on the basis of a natural lottery should not be penalized in insurance.*"

As Natowicz et al. (1992) explained, "*People at risk for genetic discrimination are (1) those individuals who are asymptomatic but carry a gene(s) that increases the probability that they will develop some disease, (2) individuals who are heterozygotes (carriers) for some recessive or X-linked genetic condition but who are and will remain asymptomatic, (3) individuals who have one or more genetic polymorphisms that are not known to cause any medical condition, and (4) immediate relatives of individuals with known or presumed genetic conditions.*"

6.4.2 Social Identity

As discussed so far, we have seen that it is perceived as “unfair" or “discriminatory" if individuals (or more precisely group of individuals) face higher premiums, or limited coverage, due to characteristics they cannot control. That would be the case of generic or biological characteristics. But when we talk about “racism", or “gender-neutral pricing", we’re talking about criteria that aren’t biologically defined, but socially. We’re talking about social identity. Social identity refers to a person’s membership in a social group. The common groups that make up a person’s social identity are age, ability, ethnicity, race, gender, sexual orientation, socioeconomic status and religion, as discussed by Tajfel (1978) and Tajfel et al. (1986).

The notion of “identity" is paradoxical. It articulates similarity and difference, uniqueness and community. For the individual, as for the group, identity is the result of complex interactions between feeling and objective social determinants, between self-perception and the gaze of others, between the intimate and the social. These are not just external factors that influence self-representation, but the constitutive processes on which identity is founded. The second paradox is linked to the fact that the individual must think of others in order to think of herself, or himself. It is because of their relationship with others that they become self-aware and construct their own identity. An individual’s “social identity" is the sum of relations of inclusion and exclusion in relation to the sub-groups that make up a society.

In many cases, sensitive attributed are related to auto-identification. Affirmative action measures require applicants to self-identify as “*native*," disabled, female, racialized or gender minorities.

6.4.3 “Lookism", Obesity and Discrimination

Williams (2017) and Liu (2017) discussed discrimination based on people’s appearance, coined “*lookism*." To quote Liu (2015), “*“everybody deserves to be treated based on what kind of person he or she is, not based on what kind of person other people are*." It is probably a reason why discrimination against “overweighted people" is challenging. As explained in Loos and Yeo (2022), while it is undeniable that changes in the environment have played a significant role in the rapid rise of obesity, it is important to recognize that obesity is the outcome of a complex interplay between environmental factors and inherent biological traits. An it is the “social norm" that make overweighted people possibly discriminated.

The stereotype “*fat is bad*" has existed in the medical field for decades, as Nordholm (1980) reminds us. Further study is needed to ascertain how this affects practice. It appears that obese individuals, as a group, avoid seeking medical care because of their weight. As claimed by Czerniawski (2007), “*with the rise of actuarial science, weight became a criterion insurance companies used to assess risk. Used originally as a tool to facilitate the standardization of the medical selection process throughout the life insurance industry, these tables later operationalized the notion of ideal weight and became recommended guidelines for body weights*." At the end of the 1950s, 26 insurance companies cooperated to determine the mortality of policyholders according to their body weight, as shown in Wiehl (1960). The conclusion is clear with regard to mortality: “*Studies bring out the clear-cut disadvantage of overweight-mortality ratios rising in every instance with increase in degree of overweight*." This is also what was said by Baird (1994), forty years after “*obesity is regarded by insurance companies as a substantial risk for both life and disability policies*." This risk increases proportionally with the degree of obesity (the same conclusion is found in Lew and Garfinkel (1979) or Must et al. (1999)).

6.5 Statistical Discrimination by Proxy

In statistics, as explained by Upton and Cook (2014), a “proxy" (we will also use “substitute variable") is a variable which, in a predictive model, replaces a useful but unobservable, unmeasurable variable, or, in our case, one that cannot be used because it is judged “discriminatory". And for a variable to be a good proxy, it must have a good correlation with the interest variable. A relatively popular example is the fact that in elementary school, shoe size is often a good proxy for reading ability. In reality, foot size has little to do with cognitive ability, but in children, foot size is highly correlated with age, which in turn is highly correlated with cognitive ability. This concept is quite related to notions of causality, discussed in the next Chapter.

We mentioned earlier economists' vision on discrimination, such as Gary Becker, and the link with a form of rationality and efficiency. And indeed, for many authors, what matters is that the association between variables is strong enough to make up a reliable predictor. For Norman (2003), group membership provides reliable information on the group, and by extension on any individual who is a member of it, the systematic use of this information (the generalization and stereotyping discussed in Schauer (2006) and Puddifoot (2021)) can be economically efficient. Taking an ethical counterpoint, Greenland (2002) reminds us that some information sources should be excluded from our decision making because they are irrelevant, or non-causal, even though they may provide fairly reliable information because of their strong correlation with another indicator. The central argument is that if variables are non-causal, then they lack moral justification. And proxy discrimination raises complex ethical issues. As Birnbaum (2020) states, “*if discriminating intentionally on the basis of prohibited classes is prohibited - e.g., insurers are prohibited from using race, religion or national origin as underwriting, tier placement or rating factors - why would practices that have the same effect be permitted?*" In other words, it is not enough to compare paid premiums, but the narrative process of modelling (i.e. the notions of interpretability and explicability, or the “*fable*" mentioned in the introduction) is equally important in judging whether a model is discriminatory or not. And the difficulty, as Seicshnaydre (2007) said, is that it is not about looking for a proof of a racist intention, or motivation, but of establishing that an algorithm discriminates according to a prohibited (because protected) criterion.

Obermeyer et al. (2019) reports that in 2019, a large health-care company had used medical expenses as a proxy for medical condition severity. The use of this proxy resulted in a racially discriminatory algorithm, because while the medical condition may not be racially discriminatory, health care spending is (in the U.S. at least). More generally, all sorts of proxies are used, more or less correlated with the interest variable. For example, a person's (or household's) income will be estimated by income tax, or by living conditions (or the neighborhood where the person lives). The first version of this paper will speak of “*indirect risk factors*." We will return to the importance of causal graphs for understanding whether one variable causes another, or whether they are simply correlated, in the section 7.1.

There are also certain quantities that are essential for modelling decision making in an uncertain context, but that are difficult to measure. This is the case of the abstract concept of “*risk aversion*" (widely discussed by Menezes and Hanson (1970) and Slovic (1987)). Hofstede (1995) proposes an uncertainty avoidance index (UAI), calculated from survey data. The first two studies, Outreville (1990) and 1996, suggested using the education level to assess risk aversion. According to Outreville (1996), education promotes an understanding of risk and therefore an increase in the demand for insurance, for example (although an inverse relationship could exist, if one assumes that increased levels of education are associated with an increase in transferable human capital, which induces greater risk-taking).

Recently, the Court of Justice of the European Union (CJEU) issued a ruling on August 1, 2022 with potentially far-reaching implications about inferred sensitive data (Case C-184/20, ECLI:EU:C:2022:601). In essence, the question posed to the European court was whether the disclosure of information such as the name of a partner or spouse would constitute processing of sensitive data within the meaning of the GDPR,

even though such data is not in itself directly sensitive, but only allows the indirect inference of sensitive information, such as the sexual orientation of the data subject. More precisely, the question was "*The referring court asks, in substance, whether* (. . .) *the publication* (. . .) *of personal data likely to disclose indirectly the political opinions, trade-union membership or sexual orientation of a natural person constitutes processing of special categories of personal data* (. . .)." The Court of Justice of the European Union has made a clear ruling on this issue, stating that processing of sensitive data must be considered to be "*processing not only of intrinsically sensitive data, but also of data which indirectly reveals, by means of an intellectual deduction or cross-checking operation, information of that kind*." For example, location data indicating places of worship or health facilities visited by an individual could now be qualified as sensitive data. As well as the recording of the order of a vegetarian menu at a restaurant in a food delivery application. And typically, if the pages "liked" by a user of a social network or the groups to which he belongs are not technically sensitive data, membership in a support group for pregnant women, or the placing of "likes" on the pages of politically oriented newspapers, allow the deduction of perfectly precise sensitive information relating to the state of health of a person or his political positions.

"*Humans think in stories rather than facts, numbers or equations - and the simpler the story, the better*," said Harari (2018), but for insurers, it is often a mixture of both. For Glenn (2000), like the Roman god Janus, an insurer's risk selection process has two sides: the one presented to regulators and policyholders, and the other presented to underwriters. On the one hand, there is the face of numbers, statistics and objectivity. On the other, there is the face of stories, character and subjective judgment. The rhetoric of insurance exclusion - numbers, objectivity and statistics - forms what Brian Glenn calls "*the myth of the actuary*," "*a powerful rhetorical situation in which decisions appear to be based on objectively determined criteria when they are also largely based on subjective ones*" or "*the subjective nature of a seemingly objective process*". And for Daston (1992), this alleged objectivity of the process is false, and dangerous, as also pointed out by Desrosières (2016). Glenn (2003) claimed that there are many ways to rate accurately. Insurers can rate risks in many different ways depending on the stories they tell on which characteristics are important and which are not. "*The fact that the selection of risk factors is subjective and contingent upon narratives of risk and responsibility has in the past played a far larger role than whether or not someone with a wood stove is charged higher premiums*." Going further, "*virtually every aspect of the insurance industry is predicated on stories first and then numbers*." We remember Box et al. (2011)'s "*all models are wrong but some models are useful*," in other words, any model is at best a useful fable.

6.5.1 Stereotypes and Generalization

As Bernstein (2013) reminds us, the word "stereotype" merges a Greek adjective meaning solid, στερές, with a noun meaning a mold, τυπος. Combining the two terms, the word refers to a hard molding, something that can leave a mark, which gave, a printing term, namely the printing form used for letterpress printing. In 1802, the dictionary of the French Academy, mentions, for the word "*stereotype*," "*a new word which is said of stereotyped books, or printed with solid forms or plates*," meaning that an image perpetuated without change. The American journalist and public intellectual Walter Lippmann gave the word its contemporary meaning in Lippmann (1922). For him, it was a description of how human beings fit "*the world outside*" into "*the pictures in our heads*," which will form simplified descriptive categories by which we seek to locate others or groups of individuals. Walter Lippmann tried to explain how images that spontaneously arise in people's minds become concrete. Stereotypes, he observed, are "*the subtlest and most pervasive of all influences*." A few years later, he began the first experiments to better understand this concept. One could observe that Lippmann (1922) was one of the first books on public opinion, manipulation and storytelling. Therefore, it is natural to see connections between the word "*stereotype*" and storytelling, as well as explanations and

interpretations.

The importance of stereotypes in understanding many decision-making processes will be analyzed in detail in Kahneman (2011), inspired in large part by Bruner (1957), and more recently, Hamilton and Gifford (1976) and especially Devine (1989). For Daniel Kahneman, schematically, two types of mechanisms are used to make a decision. System 1 is used for rapid decision making: it allows us to recognize people and objects, helps us direct our attention, and encourages us to fear spiders. It is based on knowledge stored in memory and accessible without intention, and without effort. It can be contrasted with system 2, which allows decision making in a more complex context, requiring discipline and sequential thinking. In the first case, our brain uses the stereotypes that govern representativeness judgments, and uses this heuristic to make decisions. If I cook a fish for some friends to eat, I open a bottle of white wine. The stereotype "fish goes well with white wine" allows me to make a decision quickly, without having to think. Stereotypes are statements about a group that are accepted (at least provisionally) as facts on each member. Whether correct or not, stereotypes are the basic tool for thinking about categories in System 1. But often, further, more constructed thinking - corresponding to System 2 - will lead to a better, even optimal decision. Without choosing just any red wine, perhaps a pinot noir would also be perfectly suited to grilled mullet. As Fricker (2007) asserted, "*stereotypes are [only] widely held associations between a given social group and one or more attributes*." Isn't this what actuaries do every day?

6.5.2 Generalization and Actuarial Science

In the "Ten Oever" judgement (*Gerardus Cornelis Ten Oever v Stichting Bedrijfspensioenfonds voor het Glazenwassers – en Schoonmaakbedrijf*, in April 1993), the Advocate General vangerven1993TenOever argued that "*the fact that women generally live longer than men has no significance at all for the life expectancy of a specific individual and it is not acceptable for an individual to be penalized on account of assumptions which are not certain to be true in his specific case*," as mentioned in De Baere and Goessens (2011). Schanze (2013) used the term "*injustice by generalization*." But at the same time, as explained by Schauer (2006), this "*generalization*" is probably the reason for the actuary's existence: "*To be an actuary is to be a specialist in generalization, and actuaries engage in a form of decision-making that is sometimes called actuarial*." This idea can be found in actuarial justice, for example in Harcourt (2008). Schauer (2006) reported that we might be led to believe that it is better to have airline pilots with good vision than bad ones (this point is also raised in the context of driving, and car insurance, Owsley and McGwin Jr (2010)). This criterion could be used in hiring, and would, of course, make up a kind of discrimination, distinguishing "good" pilots from "bad" ones, pilots with good vision from others. Some airlines might impose a maximum age for airline pilots (55 or 60, for example), age being a reliable, if imperfect, indicator of poor vision (or more generally of poor health, with impaired hearing or slower reflexes). If we exclude the elderly from being commercial airline pilots we will end up, *ceteris paribus*, with a cohort of airline pilots who have better vision, better hearing and faster reflexes. The explanation given here is clearly causal, and the underlying goals of the discrimination then seem clearly legitimate, so even the use of age becomes proxy discrimination, in the sense of Hellman (1998), which Schauer (2017) calls statistical discrimination, for want of testing the health status of pilots.

For Thiery and Van Schoubroeck (2006) (but also Worham (1985)), lawyers and actuaries have fundamentally different conceptions of discrimination and segmentation in insurance, one being individual, the other collective, as stigmatized in the United States by the Manhart and Norris cases (Hager and Zimpleman (1982), Bayer (1986)). In the Manhart case in 1978, the Court ruled that an annuity plan in which men and women received equal benefits at retirement, even though women made larger contributions, was illegal. In 1983, the Supreme Court ruled in the Norris case that the use of gender-differentiated actuarial factors

for benefits in pension plans was illegal because it fell within the prohibition against discrimination. These two decisions are a legal affirmation that insurance technique could not always be used as a guarantee to justify differential treatment of members of certain groups in the context of insurance premium segmentation. Indeed, legally, the right to equal treatment is one that is granted to a person in his or her capacity as an individual, not as a member of a racial, sexual, religious or ethnic group. Therefore, an individual cannot be treated differently because of his or her membership in such a group, particularly one to which he or she has not chosen to belong. These orders emphasize that individuals cannot be treated as mere components of a racial, religious, or sexual class, asserting that fairness to individuals trumps fairness to classes. But this view is fundamentally opposed to the actuarial approach, which historically analyzes risks and calculates premiums in terms of groups. As explained in Chapter 3, and Barry (2020b), until recently, actuaries considered individuals only as members of a group.

The actuarial approach is the one mentioned in the first paragraph. An individual belonging to a group with a higher statistical risk of survival or death ends up paying a higher premium or receiving fewer benefits. In automobile insurance, an individual in a group with a higher statistical risk of accident must pay higher premiums In the case of car insurance, an individual belonging to a group with a higher statistical risk of accident has to pay higher premiums, Brilmayer et al. (1983) recalled that it is the differences between the probabilities of having an accident according to gender (and not individual differences) that are taken into account to justify the difference in premiums, to explain the difference in benefits, or to base a selection mechanism. Insurance classification systems are based on the assumption that individuals meet the average characteristics, stereotyped as we say, of a group to which they belong. Insurers argue that current statistics indicate that, on average, more women than men drive without an accident and that, as a result, the average woman has a lower loss expectancy than the average man. Based on this data, women have to pay a lower premium than men. Insurance companies aim to preserve equality between groups, not individuals, and the reasons why insurers think in terms of the average woman and the average man.

An fundamental foundation of insurance is the idea of risk pooling, that is, the formation of groups. Risk in insurance cannot be considered without this notion of mutualization, and this is the major difference with financial mathematics, where there is a fundamental value of a risk (in a market). Mutualization is intrinsic to the segmentation of insurance risks, and imposes a form of solidarity within the group, since all the premiums of a group must be statistically entirely compensated by all the reimbursements of this same group. The insurer then imposes solidarity between policyholders who have the same risk profile (with a comparable probability of loss and size of loss). This is known as "*pure luck solidarity*," as coined in Barry (2020a). Without segmentation, or if the groups are not composed of members with a comparable risk profile, we will observe a phenomenon of subsidizing solidarity, in the sense that a person with a certain risk profile pays for the amount of the loss of persons with a higher loss expectancy.

We find again the opposition (using the terminology of Hume (1739)) between *is* and *ought*, between what is and what ought to be, between the statistical norm of the actuary and the norm of the legislator, which we mentioned in the introduction. From an empirical, descriptive point of view, to be within the norm means nothing more than to be within the average, not to be too far removed from this average. We can then define the norm as the frequency of what occurs most often, as the most frequently encountered attitude or the most regularly manifested preference. But this normality is not normativity, and "*to be in the norm*", to be exemplary, then comes under a different dimension, which this time is no longer linked to a description of reality but to an identification of what it must tend towards. Therefore, we move from the register of being to that of ought-to-be, from *is* to *ought*. It is indeed difficult to envisage the model (or normality) without slipping towards this second meaning that can be found in the concept of norm, and which deploys a dimension that is properly normative. This vision leads to a confusion between norms and laws, even if all normativity is not expressed in the form of laws. Therefore, David Hume notes that, in all moral systems, the

authors move from statements of fact, i.e. enunciative statements of the type "*there is*," to propositions that include a normative expression, such as "*it is necessary*," "*we must*." What he contests is the passage from one type of assertion to another: for him, these are two types of statements that have nothing to do with one another, and therefore that cannot be logically linked to each other, in particular from an empirical norm to a normative rule. For Hume, a non-normative statement cannot give rise to a normative conclusion. This assertion of David Hume has triggered many comments and interpretations, in particular because stated as it is, it seems to be an obstacle to any attempt at naturalizing morality (as detailed in MacIntyre (1969) or Rescher (2013)). In this sense, we find the strong distinction between the norm in regularity (normality) and the rule (normativity).

The principle of equality of human beings, recognized as a fundamental right, imposes the corresponding obligation not to discriminate. Therefore, to try to define discrimination is to try to specify what this principle of equality of all people means in concrete terms. Discrimination is defined as unequal and unfavorable treatment applied to certain people because of a criterion prohibited by law (i.e. race, origin, sex, etc.). According to article 225-1 of the French Penal Code, "*any distinction made between individuals on the basis of their origin, their gender, ... constitutes discrimination*." In the United States, the Prohibit Auto Insurance Discrimination Act (PAID), introduced in July 2019 in Congress, prohibits an automobile insurer from taking certain factors into account when determining the insurance premium or its eligibility. More explicitly, these prohibited factors include a driver's gender, employment status, zip code, census tract, marital status, and credit score.

Epstein and King (2002) points out that, unlike traditional statistical models, an artificial intelligence algorithm does not do this by relying on a human's initial intuition about the causal explanations for the statistical relationships between the input data and the target variable. Instead, AIs use training data to discover on their own which features can be used to predict the target variable. Although this process completely ignores causality, it inevitably leads AIs to "seek out" proxies for directly predictive features when data on those features are not made available to the AI due to legal prohibitions Mittelstadt et al. (2016). As pointed out by Barocas and Selbst (2016), "*therefore, a data mining model with a large number of variables will determine the extent to which membership in a protected class is relevant to the sought-after trait whether or not that information is an input*."

6.5.3 Massive Data and Proxy

The mutualization of individual risks is the founding principle of the insurance business. It consists, for a group of economic agents, in pooling a certain fraction of their resources to compensate the members of the group who suffer damage, as recalled by Henriet and Rochet (1987). In the early forms of insurance activity, this principle implied equal participation by each member of the community, or possibly participation proportional to individual resources, but certainly not participation proportional to individual risks. As competition in the insurance markets of several countries (especially in motor insurance) has become more and more active, a new trend has emerged: personalization of premiums. This antagonistic vision of mutualization considers that each insured should pay a premium proportional to his individual risk. Many insurers are strongly opposed to this trend towards personalization, some even rejecting the concept of "measurable individual risk" as statistical nonsense. However, personalization is also defended by others for equity reasons (each member participates proportionally to the burden he or she imposes on the community) and can even be reconciled with mutualization: if the market is large enough, individuals can be grouped into a very large number of mutual funds, each being homogeneous from the risk point of view. And the arrival of massive data and the use of machine learning techniques in the insurance world would seem to make this personalization possible, as Barry and Charpentier (2020) notes, with the hypothetical prospect of a "*pool of*

one risk," as stated by McFall et al. (2020).

Automobile insurance covers risks that are linked to the habits and behaviours of the driver, as Landes (2015) reminds us. Fire insurance compensates for risks that arise as a result of carelessness (regarding electrical appliances or furnaces). Home theft insurance covers risks that can usually be avoided with proper care (extra door locks, camera surveillance system, watch dogs, alarm, etc.). And actuaries are trying to capture this information, enriching their data. For Daniels (1990), we have morally to make sure that premiums reflect the risks of the insured.

And in the context of risk personalization, the idea that insurance products must first and foremost be "fair" to the insured is increasingly expressed by commentators, as discussed by Meyers and Van Hoyweghen (2018). Yet for McFall (2019), if individual behaviour rather than group membership were to become the basis for risk assessment, the social, economic, and political consequences would be considerable. It would disrupt the distributive and supportive character that is expressed in all health insurance plans, even those nominally designated as private or commercial. Personalized risk pricing is at odds with the infrastructure that currently defines, regulates and delivers health insurance. Going further, Van Lancker (2020) sees the widespread use of digital technology as changing the organization of the postwar welfare state in ways that affect its potential to ensure a decent standard of living for all. Moor and Lury (2018) explores how pricing has historically been implicated in the constitution of people and how the ability to "personalize" pricing reconfigures the ability of markets to discriminate. For many recent pricing techniques make it more difficult for consumers to identify themselves as part of a recognized group. She spoke of this when she wrote "*now, with the evolution of data science and network computers, insurance is facing fundamental change. With ever more information available (...) insurers will increasingly calculate risk for the individual and free themselves from the generalities of the larger pool*."

And while big data may present a danger, insurers like to create synthetic scores. Beniger (2009) points out that the volume of data is drastically reduced, with a single variable seemingly containing all the information related to all kinds of risks for a given individual. In automobile insurance, the vehicle may be associated with dozens of variables (make, model, maximum speed, weight, fuel, colour, number of seats, presence of safety features, etc.) but most insurers have a "vehiculier" that summarizes all the information related to the vehicle in about ten classes. In home insurance, the address of the dwelling, which can be associated with dozens of variables (flood zone, distance to the nearest fire station, building materials, etc.), is associated with a "zonier" made up of a few classes.

Prince and Schwarcz (2019) used the term "*discrimination by proxy*" to describe this incidental impact. Proxy discrimination occurs when a facially neutral feature is used as a substitute – or proxy – for a prohibited feature. Austin (1983), quoting Works (1977), claimed that "*although the core concern of the underwriter is the human characteristics of the risk, cheap screening indicators are adopted as surrogates for solid information about the attitudes and values of the prospective insured*." The invitations to underwriters to introduce prejudgments and biases and to indulge amateur psychological stereotypes are apparent. Even generalized underwriting texts include occupational, ethnic, racial, geographic, and cultural characterizations certain to give offence if publicly stated. Therefore, Prince and Schwarcz (2019) considers three types of proxy discrimination: causal proxy discrimination, opaque proxy discrimination, and indirect proxy discrimination. Opaque proxy discrimination occurs when one is unable to formally establish a causal link between the sensitive variable p and the target variable y. In the genetic context, even for many pathogenic genetic variants, it is often not known why a particular sequence of a gene leads to increased risk. In both causal and opaque proxy discrimination, prohibited characteristics are "*directly predictive*" of legitimate outcomes of interest. In indirect proxy discrimination, a variable x has significant predictive power simply because it is correlated with the target variable y, and the true causal variable is not present in the database. A typical example is that in school, shoe size is indirectly predictive of the number of spelling mistakes in a

dictation.

This unintended discrimination by proxy by AIs cannot be avoided simply by depriving the AI of information about the membership of individuals in legally suspect classes or obvious proxies for such group membership. However, the exclusion of the forbidden input alone may not be enough when there are other characteristics that are correlated with the forbidden input-an issue that is exacerbated in the context of big data. Bornstein (2018) discusses the "*stereotype theory of liability*," see also Selbst and Barocas (2018).

6.6 Names, Text and Language

The first name is a personal name used to designate a person, in addition to his patronymic, or family name. In most Indo-European languages, this first name precedes the family name. That is also called the Western order, opposed to the Eastern order, where the family name comes before the first name. Unlike the family name (usually inherited from the fathers in patriarchal societies), the first name is chosen by the parents at birth (or before), according to criteria influenced by the law and/or social conventions, pressures and/or trends. More precisely, at birth (and/or baptism), each person is usually given one or more first names, of which only one (which can be made up) will be used afterwards: the usual first name.

6.6.1 Last Name and Origin or Gender

The use of family names appeared in Venice in the IX-th century, Brunet and Bideau (2000) and Ahmed (2010), and came into use across Europe in the later Middle Ages (beginning roughly in the XI-th century), according to the 'Family names' entry of the Encyclopedia Britannica. Family names seems to have originated in aristocratic families and in big cities, since having hereditary surname that develops into a family name preserves the continuity of the family and facilitates official property records and other matters. Sources of family names are original nicknames (Biggs, Little, Grant), occupations came (Archer, Clark), place-names (Wallace, Murray, Hardes, Whitney, Fields, Holmes, Brookes, Woods), as mentioned in McKinley (2014).

In English, the suffixation of –**son** has been also very popular (Richardson, Dickson, Harrison, Gibson) but also using the prefix **Fitz**– (Fitzgerald), which goes back to Norman French *fis* (*fils* in French), meaning "son", as explained in McKinley (2014). In Russian, as discussed in Plakans and Wetherell (2000), the suffix –**ov** (ов, "son of") was also used, such as 'Ivan Petrov (Петров)', for "Ivan (Иван), the son of Piotr (Пиотр, orПетр+ов)", with the possibility of designating the successive fathers, with the use of patronymics: Vasily Ivanovich Petrov (Василиıу Иванович Петров), is Vasily (Василиıу) son of Ivan (Иван+ ович), born from the ancestor Piotr (Пиотр+ов).

Icelandic names are names used by people from Iceland. Icelandic surnames are different from most other naming systems in the modern Western world by being patronymic or occasionally matronymic, as mentioned in Willson (2009) and Johannesson (2013): they indicate the father (or mother) of the child and not the historic family lineage. Generally, with few exceptions, a person's last name indicates the first name of their father (patronymic) or in some cases mother (matronymic) in the genitive, followed by –**son** ("son") or –**dóttir** ("daughter"). For instance, in 2017, Iceland's national Women's soccer team players were Agla Maria Albert**sdóttir**, Sigridur Gardar**sdóttir**, Ingibjorg Sigurdar**dóttir**, Glodis Viggos**dóttir**, Dagny Brynjars**dóttir**, Sara Bjork Gunnars**dóttir**, Fanndis Fridriks**dóttir**, Hallbera Gisla**dóttir**, Gudbjorg Gunnars**dóttir**, Sif Atla**dóttir** or Gunnhildur Jons**dóttir**. In the national Men's soccer team, players were Hákon Rafn Valdimars**son**, Patrik Gunnars**son**, Höskuldur Gunnlaugs**son**, Júlíus Magnús**son**, Viktor Örlygur Andra**son** or Kristall Máni Inga**son**.

The development is slightly different among Jews. They chose, or were given, family names only at the end of the 18th and beginning of the XIX-th century, and most derived from religious vocations, as explained

in Kaganoff (1996): Cantor, Canterini, Kantorowicz (lower priest); Kohn, Cohen, Cahen, Kaan, Kahane (priest); Levi, Halévy, Löwy (name of the tribe of priests), etc. In China, Liu et al. (2012) mentioned that there are about 1,000 names, and only 60 are commonly used. Most Chinese surnames are monosyllabic and originate from a physical characteristic or description: Wang (yellow), Wong (wild water field), Chan (old) or Chu (mountain). In Japan, until the XIX-th century, only members of the nobility had a family name, as explained by Tanaka (2012). But everything changed when the Emperor declared that everyone should have a second name. Entire villages took the same name. That's why there are only about 10000 names in Japan and most of them are geographical place names: Arakawa (rough, river), Yamada (mountain, rice field), Hata (farm) and Shishido (flesh, door).

Using birth record data from California, Fryer Jr and Levitt (2004) find that "*Blacker names are associated with lower-income zip codes [and] lower levels of parental education*." In Table 6.1 we have names, and proportion of people who identify themselves as White, Black or Hispanic, per geographic area. In many cases, the last name can be a proxy of a racial attribute.

Name	Rank	White (%)	Black (%)	Hispanic (%)
Washington	138	5.2%	89.9%	1.5%
Jefferson	594	18.7%	75.2%	1.6%
Booker	902	30.0%	65.6%	1.5%
Banks	278	41.3%	54.2%	1.5%
Jackson	18	41.9%	53.0%	1.5%
Mosley	699	42.7%	52.8%	1.5%
Becker	315	96.4%	0.5%	1.4%
Meyer	163	96.1%	0.5%	1.6%
Walsh	265	95.9%	1.0%	1.4%
Larsen	572	95.6%	0.4%	1.5%
Nielsen	765	95.6%	0.3%	1.7%
McGrath	943	95.9%	0.6%	1.6%
Stein	720	95.6%	0.9%	1.6%
Decker	555	95.4%	0.8%	1.7%
Andersen	954	95.5%	0.6%	1.7%
Hartman	470	95.4%	1.5%	1.2%
Orozco	690	3.9%	0.1%	95.1%
Velazquez	789	4.0%	0.5%	94.9%
Gonzalez	23	4.8%	0.4%	94.0%
Hernandez	15	4.6%	0.4%	93.8%

Table 6.1: Last name, and racial proportions in the U.S., from Gaddis (2017).
(data from US Census (2012))

6.6.2 First Name and Age or Gender

As Bosmajian (1974) states, "*an individual has no definition, no validity for himself, without a name. His name is his badge of individuality, the means whereby he identifies himself and enters upon a truly subjective existence*." Names are often the first information people have in a social interaction. Sometimes we know individuals by name even before we meet them in person, as Erwin (1995) reminds us. First and last names

can carry a lot of information, as shown by Hargreaves et al. (1983), Dinur et al. (1996) or Daniel and Daniel (1998). To quote Young et al. (1993), "*the name Bertha might be judged as belonging to an older Caucasian women of lower-middle class social status, with attitudes common to those of an older generation* $(\cdots)$ *a person with a name such as Fred, Frank, Edith, or Norma is likely to be judged, at least in the absence of other information, to be either less intelligent, less popular, or less creative than would a person with a name such as Kevin, Patrick, Michelle, or Jennifer*."

In the course of testing operations, three profiles are mentioned: the first corresponds to the candidate whose name and surname are North African-sounding (for example, Abdallah Kaïdi, Soufiane Aazouz or Medji Ben Chargui), the second corresponds to the candidate whose first name is French-sounding and whose surname is North African-sounding (for example, François El Hadj, Olivier Ait Ourab or Nicolas Mekhloufi), and finally the one whose first name and last name are French sounding (for example Julien Dubois, Bruno Martin or Thomas Lecomte). Amadieu (2008) mentions that the first names (male) Sébastien, Mohammed and Charles-Henri are used for tests. The Table 6.2 shows the main first names on three generations of immigrants. For the first names in France, Coulmont (2011) goes back to the information from the first name.

		immigrants	children of immigrants	grandchildren of immigrants
Southern Europe	men	José, Antonio, Manuel	Jean, David, Alexandre	Thomas, Lucas, Enzo
	women	Maria, Marie, Ana	Marie, Sandrine, Sandra	Laura, Léa, Camille
Maghreb	men	Mohamed, Ahmed, Rachid	Mohamed, Karim, Mehdi	Yanis, Nicolas, Mehdi
	women	Fatima, Fatiha, Khaduja	Sarah, Nadia, Myriam	Sarah, Ines, Lina

Table 6.2: Top 3 first names by sex and generations in France, according to the origin (Southern Europe or Maghreb) of grandparents, Coulmont and Simon (2019).

In Box 6.6, Baptiste Coulmont discusses the use of the first name as a proxy for some sensitive attribute.

Similarly, in "*are Emily and Greg more employable than Lakisha and Jamal* ?", Bertrand and Mullainathan (2004) randomly assigned African-American or white-sounding names in resumes to manipulate the perception of race. "*White names*" received 50% more callbacks for interviews. Voicu (2018) presents the BIFSG ("Bayesian Improved First Name Surname Geocoding") model to use first name to improve the classification of race and ethnicity in a mortgage lending context, drawing on Coldman et al. (1988), Fiscella and Fremont (2006) and Analyzing data from the German Socio-Economic Panel, Tuppat and Gerhards (2021) shows that immigrants with first names considered uncommon in the host country disproportionately complain of discrimination. And if name is a marker indicating ethnicity, more highly educated immigrants more frequently report that they perceive discrimination in the host country than less educated immigrants ("discrimination paradox"). Rubinstein and Brenner (2014) shows that the Israeli labor market discriminates on the basis of perceived ethnicity (between Sephardic and Ashkenazi-sounding surnames). Carpusor and Loges (2006) analyzes the impact of first and last names on the rental market, while Sweeney (2013) analyzes their impact on online advertising.

6.6.3 Text and Biases

Chatbots will both raise critical ethical challenges and hold implications for democratization of technology, and implementing research addressing these directions is important. Chatbots permit users to interact through natural language, and consequently are a potential low threshold means to access information and services and promote inclusion. However, due to technological limitations and design choices, they can be the means

[5]Professor at the École Normale Supérieure Paris-Saclay

Box 6.6 **The first name as proxy**, by Baptiste Coulmont[5]

As early as the 1930s, historians such as Marc Bloch in France felt that first names could be used for historical research. "*The very choice of baptismal names, their nature, relative frequency, are all traits which, when properly interpreted, reveal currents of thought or feeling*," Bloch (1932). The first name of a person is indeed on the one hand the intimate choice of the parental couple (and their close relations), and on the other hand an information very often present in population registers. This connection between the private and the public spheres makes it possible to conceive the first name as a gateway to the study of a group's culture, because it says things about the givers of first names and the bearers of first names. What does it say? Marc Bloch is careful enough to insist on the necessity of a "*proper interpretation*." With the gender of the first name and the sex of the civil status, the associations are strong enough to guide the interpretation: if epicene first names exist, they name only a small part of the population. And it is even possible, for a large number of first names, from this simple first name, to infer with great certainty the gender of the person who bears it. This is how, after a time when it was forbidden to use married people's sex, it was possible to reconstitute the missing information, Carrasco (2007). The fashion phenomena around first names, visible with the craze then the disappearance during the years, make it possible to estimate the "average age" of the carriers of first name. Today, in France, Nathalie's are older than Emma's. But this does not make the first name a transparent signal of a person's age. An interesting survey, which asked participants to estimate a person's age from their first name, indicates that "*the perceived age of a first name corresponds only weakly with the true average year of birth of people with that name*." Interpretation must be undertaken with caution when the study focuses on other characteristics of first names. Parents in different locations in the social circles choose different first names, Besnard and Grange (1993). The frequency of the first names therefore varies with the social environment. Today, the Anouk, Adèle or Joséphine taken as a group have parents with more education than the Anissa, Mégane or Deborah. But this relationship will depend on time, as some of the first names spread - following fashion - from one environment to another. Finally, when the characteristics associated with first names are linked to fluid, contextual identities, or assigned administratively from outside, it is the addition of other variables that gives meaning to first names. Many studies (Tzioumis (2018) or Mazieres and Roth (2018)) seek to exploit the information on ethnic, national or geographical origin contained in the names and surnames, but they need, to start the investigation, a link between the first name and the variable studied. Directories from different countries, for example. The generalization of correlations between origin and first name to other countries or other populations must be done with caution.

to perpetuate and even reinforce existing biases in society, exclude or discriminate against some user groups, as discussed in Harwell et al. (2018) or Feine et al. (2019) and over-represent or enshrine specific values.

If we translate Turkish sentences that use the gender-neutral 'o' to English, we obtain outputs such as the ones in Table 6.3.

		2017	2023
o bir öğretmen	›	she is a teacher	he is a teacher
o bir hemşire	›	she is a nurse	she is a nurse
o bir doktor	›	he is a doctor	she is a doctor
o bir Şarkıcı	›	she is a singer	he is a singer
o bir sekreter	›	she is a secretary	she is a secretary
o bir dişçi	›	he is a dentist	he is a dentist
o bir çiçekçi	›	she is a florist	she is a florist
o çalışkan	›	he is hard working	he is hard working
o tembel	›	she is lazy	he is lazy
o güzel	›	she is beautiful	she is beautiful
o çirkin	›	he is ugly	he is ugly

Table 6.3: Translation based on https://translate.google.com/, from Turkish to English.

To make such decisions, it clearly means that corpuses used to train those models are sexists (possibly also agist or racist). Therefore, the words we use can be the proxy for some sensitive attributes. This is related to "Social Norm Bias" as defined in Cheng et al. (2023), inspired by Antoniak and Mimno (2021). Social Norm Bias is characterized by the associations between an algorithm's predictions and individuals' adherence to inferred social norms, that can be a a source of algorithmic unfairness, since penalizing individuals for their adherence or deviation to social norms is a form of algorithmic. More precisely, "Social Norm Bias" occurs when an algorithm is more likely to correctly classify the women in this occupation who adhere to these norms over the women in the same occupation who do not adhere to them bias. Tang et al. (2017) used manually-compiled lists of gendered words and using only the frequency of these words. "*Occupations are socially and culturally 'gendered'*" wrote Stark et al. (2020); many jobs (for instance in science, technology, and engineering) are traditionally masculine, Light (1999) and Ensmenger (2015). Different English words have gender and age related connotations, as shown in Moon (2014), inspired by Williams and Bennett (1975). Based on a large reference corpus (the 450-million-word Bank of English, `BoE`), Moon (2014) observed that the most frequent adjectives co-occurring with "young" are: inexperienced, beautiful, fresh, attractive, healthy, vulnerable, pretty, naive, talented, impressionable, energetic, crazy, single, dynamic, fit, strong, trendy, innocent, foolish, handsome, hip, stupid, ambitious, free, full (of life/ideas/hope/etc.), lovely, enthusiastic, eager, small, vibrant, gifted, immature, slim, good-looking. In contrast, the most frequent adjectives co-occurring with "old" are: sick, tired, infirm, frail, grey, fat, worn(-out), decrepit, disabled, wrinkly/wrinkled, slow, poor, weak, wise, beautiful, rare, ugly. The following associations were observed,

- precocious, shy: teens, tailing off in twenties
- pretty, promising: peaking with teens, twenties, tailing off in thirties
- beautiful, fresh-faced, stunning: mainly teens and twenties
- blonde: strongly associated with women in their twenties; to a lesser extent teens and thirties
- ambitious, brilliant, talented: peaking in twenties
- attractive: peaking in twenties, tailing off in thirties
- handsome: peaking in twenties, tailing off in forties
- balding, dapper, formidable, genial, portly, paunchy: mainly forties and older
- sprightly/spritely: beginning to appear in sixties, stronger in seventies/eighties
- frail: mainly seventies and older

Crawford et al. (2004) provide a corpus of 600 words and human-labeled gender scores, as scored on a 1–5 scale (1 is the most feminine, 5 is the most masculine) by undergraduates at U.S. universities. They find that words referring to explicitly gendered roles such as wife, husband, princess, and prince are the most strongly gendered, while words such as pretty and handsome also skew strongly in the feminine and masculine directions, respectively.

6.6.4 Language and Voice

The voice is also an important element, often very subjective during meetings with agents, in person, but today it is analyzed by algorithms, in particular when an insurer uses conversational robots (we will sometimes

speak of a "*chat bot*"), as analyzed by Hunt (2016), Koetter et al. (2019), Nuruzzaman and Hussain (2020), Oza et al. (2020) or Rodríguez Cardona et al. (2021). We can note, in France, the experience of the virtual assistant launched by Axa Legal Protection, called Maxime, described by Chardenon (2019).

In the Fall of 2020, in France, a law proposal repressing discrimination based on accent was presented to the parliament, as reported by Le Monde (2021). Linguists would speak of phonostyle discrimination, as in Léon (1993), or of "diastratic variation", with differences between usages by gender, age and social background (in the broad sense), in Gadet (2007), or of "*glottophobia*", a term introduced by Blanchet (2017). Glottophobia can be defined as "*contempt, hatred, aggression, rejection, exclusion, negative discrimination actually or allegedly based on the fact of considering incorrect, inferior, bad certain linguistic forms (perceived as languages, dialects or uses of languages) used by these people, in general by focusing on the linguistic forms (and without always being fully aware of the extent of the effects produced on the people)*." If Van Parijs (2002) is mainly interested in people discriminated against because French is not their mother tongue ("*the mother tongue, in this perspective, is as illegitimate a basis for discrimination as race, gender or faith*"), Blanchet (2017) emphasizes above all cultural differences (the lengthening of vowels, the distribution of pauses, the rate of speech, the accentuation of certain syllables, the richness of vocabulary, etc.). One can think of the "suburban accents" as Fagyal (2010) called them.

Blodgett and O'Connor (2017) has highlighted an important frontier in algorithmic fairness: the disparity in the quality of natural language processing algorithms when applied to the language of authors from different social groups. For example, current systems sometimes analyze the language of women and minorities more poorly than the language of whites and men. Finally, it is worth noting that in Canada, the expression "*speak white*" was a racist slur used by English speakers to characterize those who did not speak English in public places. Historically, this is what was said to Liberal Party MP Henri Bourassa in 1889 when he attempted to speak French during debates in the Canadian House of Commons. As Michèle Lalonde, author of the 1968 poem "*Speak White*" (reported by Dostie (1974)) puts it, "*Speak White is the protest of white Negroes in America. Language here is the equivalent of colour for the American Negro. The French language is our black colour*."

For Squires and Chadwick (2006), "linguistic profiling", i.e. the identification of a person's race from the sound of their voice and the use of this information to discriminate on the basis of race, has been documented in the home rental market, and Squires and Chadwick (2006) analyse its impact in the home insurance sector. Based on an analysis of paired tests conducted by fair housing organisations, they show that home insurance agents are generally able to detect the race of a person who contacts them by phone and that this information affects the services provided to people inquiring about purchasing a home insurance policy. Finally, as an anecdote, Durry (2001) reports that in 1982 an insurance company in France was sued for refusing to provide car insurance to people who could not read or write. On the legal basis that "*everything that is not* [explicitly] *prohibited is permitted*," the court ruled that the insurance company could not be held liable. However, the chosen criterion arguably led to the elimination of foreigners more often than French citizens, Durry (2001) pointed out. More recently, to return to our introductory discussion, the chat bot has shown itself capable of reproducing discrimination, mostly in relation to the speaker's gender, as explained by Feine et al. (2019), Aran et al. (2019), McDonnell and Baxter (2019) or Maedche (2020), but also be clearly racist, as shown by Schlesinger et al. (2018). One should then be careful when using voice in a "*black box*" algorithm.

6.7 Pictures

6.7.1 Pictures and Facial Information

More than a century ago, first Lombroso (1876), and then Bertillon and Chervin (1909), laid the foundations of phrenology and the "born criminal" theory, which assumes that physical characteristics are correlated with psychological traits and criminal inclinations. The idea was to build classifications of human types on the basis of morphological characteristics, in order to explain and predict differences in morals and behaviour. One could speak of the invention of a "*prima facie*". We can also mention "*ugly laws*," in the sense of Schweik (2009), taking up a term used by Burgdorf and Burgdorf Jr (1974) to describe laws in force in several cities in the United States until the 1920s, but some of which lasted until 1974. These laws allowed people with "*unsightly*" marks and scars on their bodies to be banned from public places, especially parks. In New York, in the XVIII-th century, Browne (2015) recalls that "lantern laws" demanded that Black, mixed-race and Indigenous enslaved people carry candle lanterns with them if they walked about the city after sunset, and not in the company of a white person. The law prescribed various punishments for those that didn't carry this supervisory device.

These debates are resurfacing as the number of applications of facial recognition technology increases, thanks to improvements in the quality of the images, the algorithms used, and the processing power of computers. The potential of these facial recognition tools to perform health assessment is demonstrated in Boczar et al. (2021). Historically, Anguraj and Padma (2012) had proposed a diagnostic tool for facial paralysis, and recently, Hong et al. (2021) uses the fact that many genetic syndromes have facial dysmorphism or and facial gestures that can be used as a diagnostic tool to recognize a syndrome. As shown on Figure 6.5, many applications online can, for free, label pictures, and extract information, personal if not sensitive, such as the gender (with a "confidence" value), the age, and also some sort of emotion.

In Chapter 3, we explain that in the context of risk modeling, a "classifier" that associate some pictures to a class is not extremely interesting, and having the probability to belong to classes is more interesting. On Figure 6.6, we challenge the "confidence" value given by Picpurify, using pictures generate by a GAN (used in Hill and White (2020) to generate faces), with a "continuous" transformation from a picture (top left) to another one (bottom right). Based on the terminology we will use later, when using barycenters in Chapter 12, we have here some sort of "geodesic" between the picture of a woman and a picture of a man. Surprisingly, we would expect the "confidence" (for identifying a "man") to increase continuously from a low value to a higher one. But it is not the case, the algorithm predicting with very high confidence that the person on the top right is a "female" and also with very high confidence that the person on the bottom left is a "male" (with only very few changes in the pixels between the two pictures).

More generally, beyond medical considerations, Wolffhechel et al. (2014) reminds us that several personality traits can be read from a face, and that facial features influence first impressions. That said, the prediction model considered fails to reliably predict personality traits from facial features. However, recent technical developments, accompanied by the development of large image banks, have made it possible, as claimed by Kachur et al. (2020), to predict multidimensional personality profiles from static facial images, using neural networks trained on large labeled data sets. About ten years ago, Cao et al. (2011) proposed to predict the gender of a person from facial images (Rattani et al. (2017) or Rattani et al. (2018)), and recently Kosinski (2021) used a facial recognition algorithm to predict political orientation (in a binary context, opposing liberals and conservatives, in the spirit of Rule and Ambady (2010)). Wang and Kosinski (2018) proposed to use these algorithms to predict sexual orientation (see also Leuner (2019)).

In Shikhare (2021), the idea of using a facial score model for life insurance underwriting was evoked (in connection with the concept of AUW – accelerated underwriting), with images like in Figure 6.5. The idea is

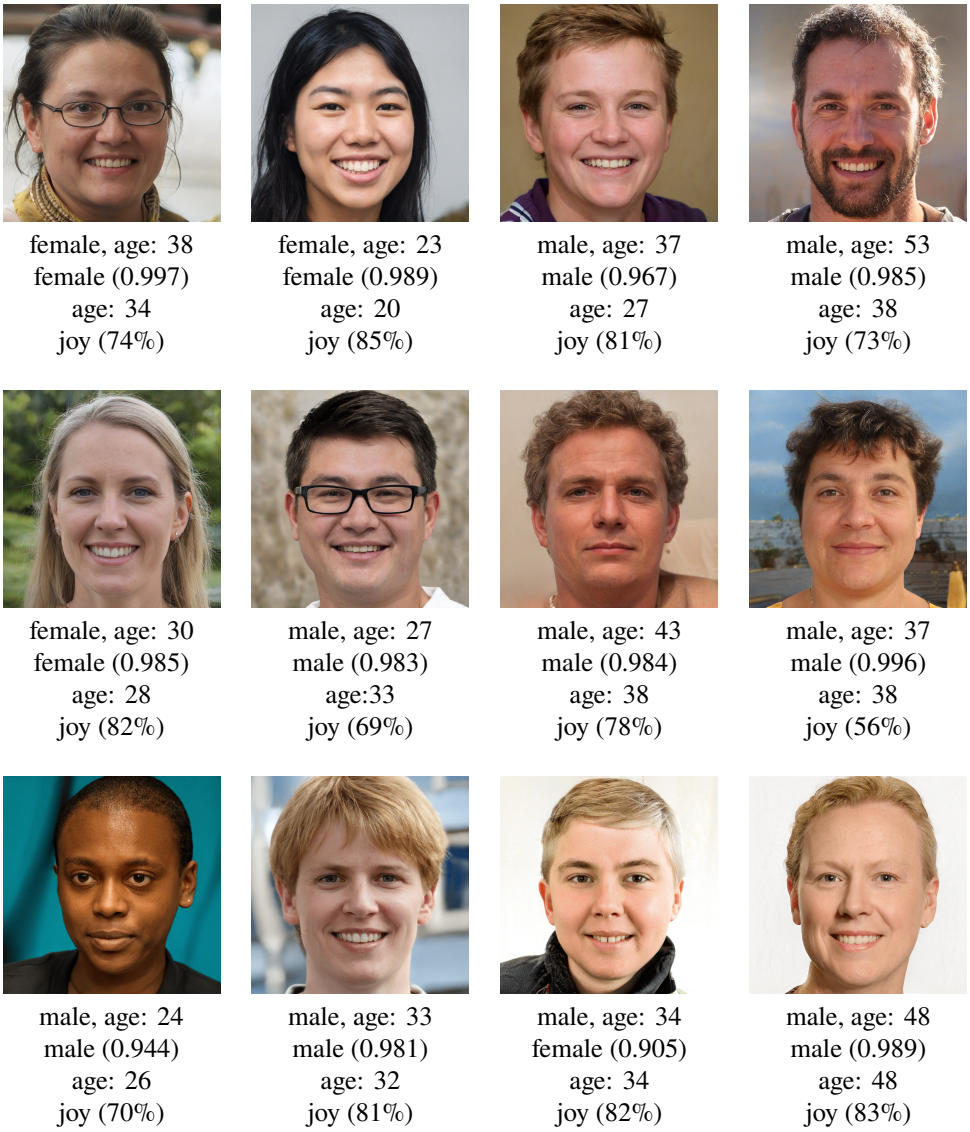


Figure 6.5: Faces generated by Karras et al. (2020). Gender and age were provided by https://gender.toolpie.com/ and facelytics, https://www.picpurify.com/demo-face-gender-age.html with a "confidence", and https://cloud.google.com/vision/, https://howolddoyoulook.com/ and https://www.facialage.com/ (accessed in January 2023).

to look for "*abnormal characteristics*" like related to a particular condition (Down's syndrome, Cornelia de Lange, Cushing's, acromegaly, etc.), including abnormal skin color to detect bronchial asthma or hepatitis, or to infer gender, as mentioned in Shikhare (2021).

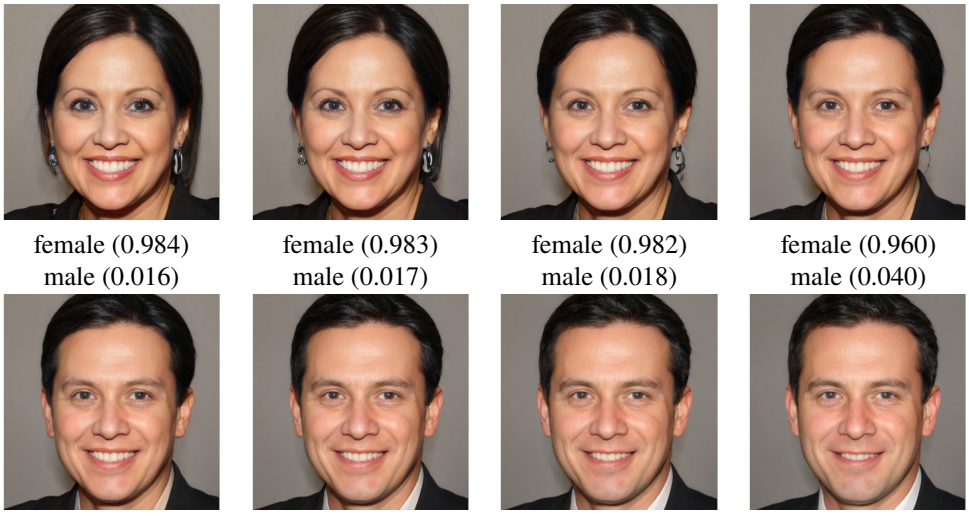


Figure 6.6: GAN used in Hill and White (2020) to generate faces, with a “continuous" transformation from a picture (top left) to another one (bottom right), and then gender predicted using https://www.picpurify.com/demo-face-gender-age.html with a “confidence" (accessed in March 2023).

6.7.2 Pictures of houses

Using deep learning algorithms, it is possible to extract information from pictures taken from cars after an accident, as discussed in Dwivedi et al. (2021), and several companies are using this techniques through various applications. After an accident, the app directs the policyholder to take photos of the damaged cars at certain angles, and in certain light. And using just those photos, a (black-box) model estimates how much it will cost to fix the car. “*Garage operators say the numbers can be wildly inaccurate,*" mentioned Marshall (2021). “*You can’t diagnose suspension damage or a bent wheel or frame misalignment from a photograph*" .

Some companies try to sell models that could estimate how much it will cost to fix a water damage or a fire, in a house. For example, on Figure 6.7, we can visualize three photos of damages to a house, tagged by Google API Cloud Vision.

It is also possible to use pictures in the underwriting process. On Figure 6.8, we have a variety of images associated with building search, with street photos on the left, with Google Street View (project launched in 2007), and aerial imagery, with Google Earth (project launched in 2001), on the right. It’s possible to spot a neighbor’s swimming pool close to the house, or the proximity of public electrical equipment, we could be a good thing to assess the true risk of a water damage, or a fire.

On Figure 6.9, we have examples of images associated with building search, with street photos on the left, with Google Street View (project launched in 2007), with different views, 2012 and 2018. It is possible to observe house extensions that were not declared by a policyholder, or that a tree is growing faster that expect, very close to a load-bearing wall.

On Figure 6.10, aerial imagery, with Google Earth, with four buildings. Some models are also used (see Galindo et al. (2009), Rodríguez-Cuenca and Alonso (2014) and Cunha et al. (2021)) to detect swimming

windows (91%)
wood (86%)
flooring (82%)

brown (98%)
wood (88%)
hardwood (70%)

wood (83%)
tints and shades (75%)
ceiling (72%)

Figure 6.7: Labels from Google API Cloud Vision https://cloud.google.com/vision/, (accessed in January 2023). (source: personal collection).

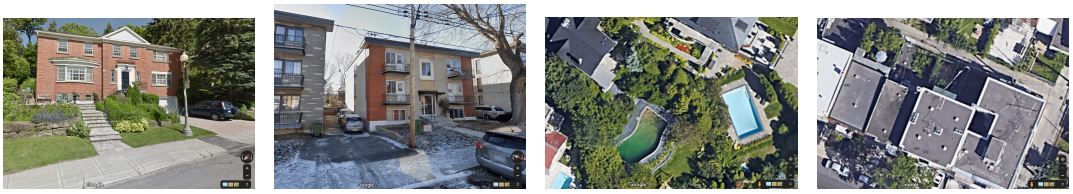


Figure 6.8: Examples of images associated with building search, with street photos on the left, with Google Street View, and aerial imagery, with Google Earth, on the right.

Figure 6.9: Examples of images associated with building search, with street photos on the left, with Google Street View, with different views from 2012 until 2018 on top, and from 2009 until 2018 below.

pools using aerial imagery. It is also possible to use those picture to have a better estimate of the size of the house. But when a house has a black tile roof, it's more difficult to separate the shadow of the house from the roof, which can lead to overestimates of the total surface area of the building, as we can see.

Figure 6.10: Examples of aerial imagery, with Google Earth.

6.8 Spatial Information

"*Geographic location is a well-established variable in many lines of insurance* ", recalls Bender et al. (2022). "*Geographic information is crucial for estimating the future costs of an insurance contract*" claimed Blier-Wong et al. (2021). For weather-related perils, such as flooding, geographic information is crucial because location is a key piece of information for predicting loss frequency. In motor insurance, policyholders living in rural areas are less likely to have accidents because they use roads with little traffic. When they do have accidents, they tend to generate higher claims costs because they are more severe. For this reason, insurance companies define territory levels, which serve as unique coding in their pricing models.

It is possible to use open data to extract a lot of information, such as OpenStreetMap. On Figure 6.11, we extracted buildings in Montréal (Canada), but we could also extract street and roads, or rivers and lakes.

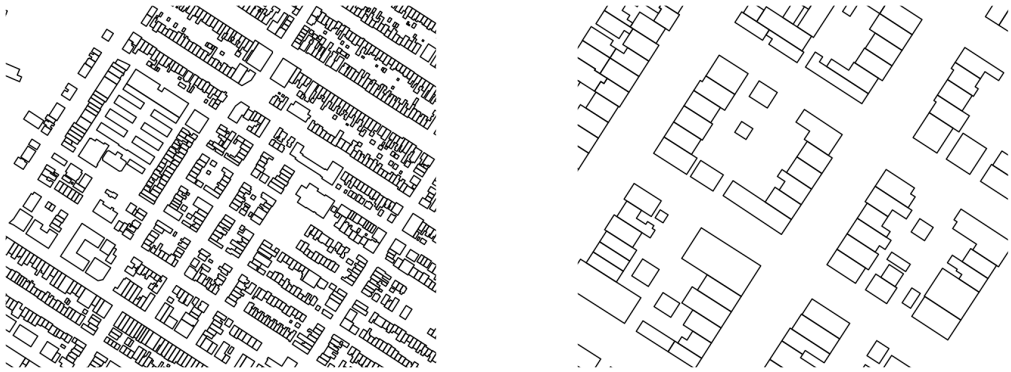


Figure 6.11: Some polygons of building contours, in Montréal (Canada), extracted from OpenStreetMap (`getbb` function of `osmdata` package).

Since locations of fire hydrants in Montréal is information publicly available, it is possible to visualize

it, and compute the distance between the house of a policyholder, and the closest fire hydrants, as on Figure 6.12.

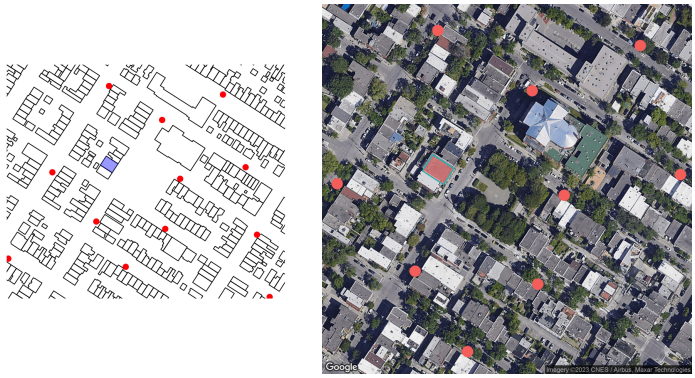


Figure 6.12: Locations of fire hydrants in a neighborhood in Montréal (Canada), with some satellite picture from GoogleView on the right (`qmap` function of `ggmap` package).

Figure 6.13, we can also visualize buildings with a three dimensional perspective, the colors are based on thermal diagnostics.

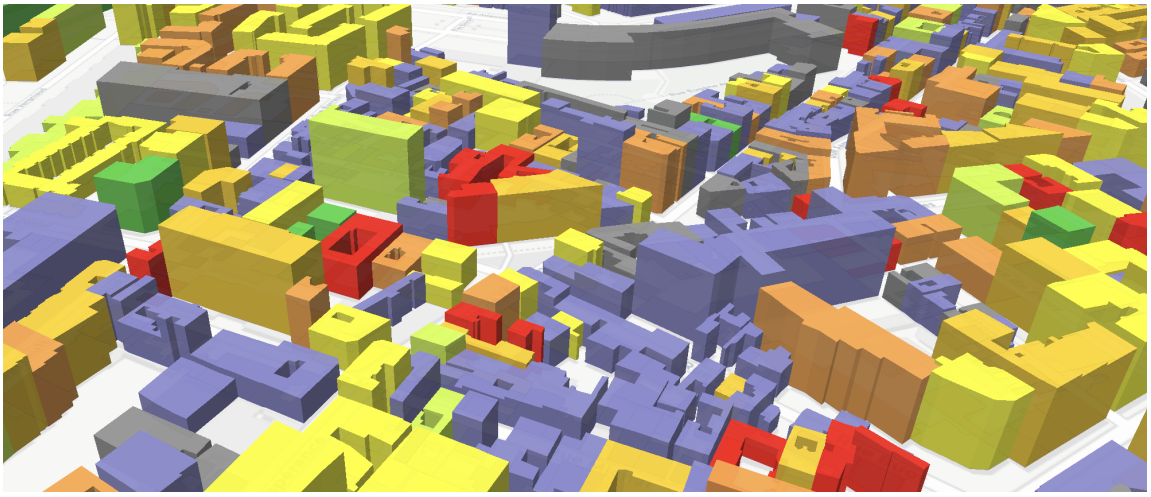


Figure 6.13: Three dimensional perspective of building in Paris (France). (source: https://particulier.gorenove.fr/)

6.8.1 Redlining

If geographic information is important to assess risk, it has been seen as a problematic information, because of “redlining" (see Squires and DeWolfe (1981) and Baker and McElrath (1997)). In the United States, the federal government created the Home Owners’ Loan Corporation (HOLC) during the Depression, in 1933, to slow down the dramatic increase in the rate of housing foreclosures. Redlining refers to lending (or insurance)

discrimination that bases credit decisions on the location of a property to the exclusion of characteristics of the borrower or property. Usually, it means that lenders will not make loans to areas with African Americans or other perceived risks to real estate investments. Jackson argued that the Federal Housing Administration (FHA) and private lenders obtained copies of the HOLC maps and that the grades on the maps impacted their lending decisions. Community groups in Chicago's Austin neighborhood coined the word "redlining" in the late 1960s, referring literally to red lines lenders and insurance providers admitted drawing around areas they would not service. "*The practice of redlining was first identified and named in the Chicago neighborhood of Austin in the late 1960s. Saving and loan associations, at the time the primary source of residential mortgages, drew red lines around neighborhood they thought were susceptible to racial change and refused to make mortgages in those neighborhoods*," explained Squires (2011). And because those areas were also with a higher proportion of African American people, "redlining" started to be perceived as a discriminatory practice. More generally, the use of geographic attributes may hide (intentionally or not) the fact that some neighborhoods are populated mainly by people of a specific race, or minority.

6.8.2 Geography and Wealth

Unfortunately, it is difficult to assess whether people are discriminated because of their ethnic origin, or because of their wealth. As observed in Institute and Faculty of Actuaries (2021), "people in poverty pay more for a range of products, including energy, through standard variable tariffs; credit, through high interest loans and credit cards; insurance, through postcodes considered higher risk; and payments, through not being able to benefit from direct debits as they are presently structured."

Historically, the address was used to associate an insured with a urban area. On Figure 6.14 we can visualize, Paris by IRIS zone[6], the median income (per household) of the neighborhood, the proportion of people over 65 years old, the rate of dwellings of less than 40^2 and of more than 100^2. We could have statistics, by IRIS, on the dilapidation of buildings, the burglary rate, etc.

Initially, Jean et al. (2016) had noted that at a fairly coarse level, night lighting was a rough indicator of economic wealth. Indeed, world nighttime maps show that many developing countries are poorly lit. Jean et al. (2016) combined nighttime maps with high-resolution daytime satellite imagery, and the combined images allowed – "*with a bit of machine-learning wizardry*" – to obtain accurate estimates of household consumption and assets (two quantities often difficult to measure in poorer countries). Subsequently, Seresinhe et al. (2017) suggested to estimate the amount of green space in different locations based on satellite images. Taking this a step further, Gebru et al. (2017) claimed to be able to quantify socioeconomic attributes such as income, ethnicity, education, and voting patterns from cars detected in Google Street View images. For example, in the United States, if the number of Sedans in a neighbourhood is greater than the number of pickup trucks, that neighbourhood is likely to vote Democrat in the next presidential election (88% chance); otherwise, the city is likely to vote Republican (82% chance).

As Law et al. (2019) puts it, when an individual buys a home, they are simultaneously buying its structural characteristics, its accessibility to work, and the neighbourhood's amenities. Some amenities, such as air quality, are measurable, while others, such as the prestige or visual impression of a neighbourhood, are difficult to quantify. Rundle et al. (2011) notes that Google Street View provides a sense of neighbourhood safety (related to traffic safety), if looking for crosswalks, the presence of parks and green spaces, etc. Using street and satellite image data, Law et al. (2019) shows that it is impossible to capture these un-quantifiable features and improve the estimation of housing prices, in London, U.K. A neural network, with input from

[6]IRIS = *Ilôts Regroupés pour l'Information Statistique*, a division of the French territory into grids of homogeneous size, with a reference size of 2,000 inhabitants per elementary grid. France has 16,100 IRIS, including 650 in the overseas departments, and Paris (*intra muros*) has 992 IRIS.

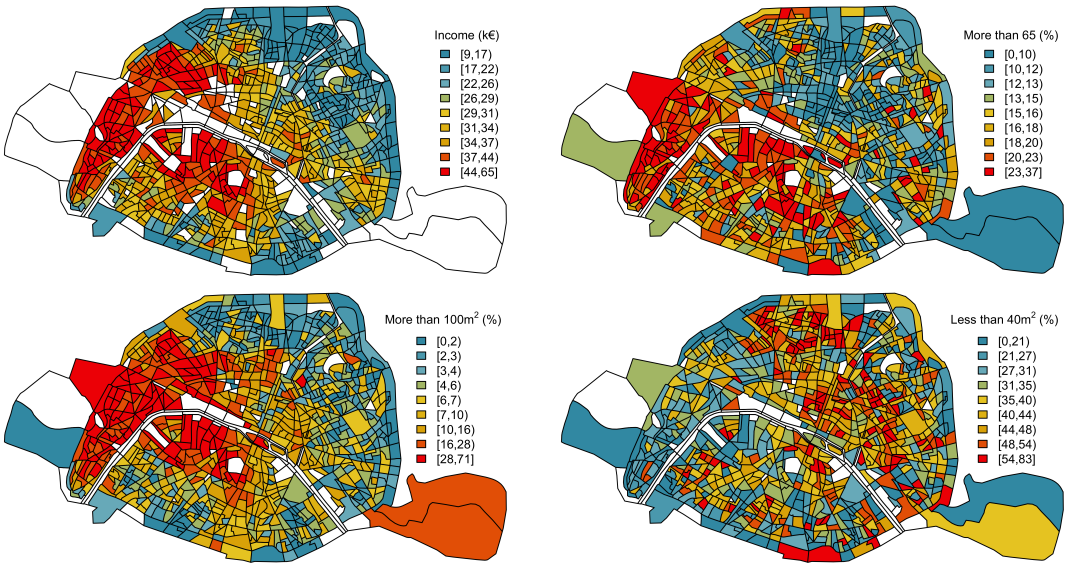


Figure 6.14: Median income per household, proportion of elderly people, proportion of dwellings according to their size, by neighborhood (IRIS), in Paris, France. This statistical information on the neighborhood (and not the insured) can be used to rate a home insurance policy, for example.
(source: INSEE open data).

traditional housing characteristics such as age, size and accessibility, as well as visual features from Google Street View and aerial images, is trained to estimate house prices. It is also possible to infer some possibly sensitive personal information, such as a possible disability with the presence of a ramp for the house (the methodology is detailed in Hara et al. (2014)), or sexual orientation with the presence of a rainbow flag in the window, or political with a Confederate flag (as mentioned in Mas (2020)). Ilic et al. (2019) also evokes this "*deep mapping*" of environmental attributes. A Siamese Convolutional Neural Network (SCNN), trained on temporal sequences of Google Street View (GSV) images is trained. The evolution over time confirms some urban areas known to be undergoing gentrification, while revealing areas undergoing gentrification that were previously unknown. Finally, Kita and Kidziński (2019) discusses possible direct applications in insurance.

6.9 Credit Scores

REP As some authors have historically suggested that discrimination based on wealth (or income) should be considered, as Brudno (1976), Gino and Pierce (2010), or more recently Paugam et al. (2017), but that criterion will be discarded here. The reflections of these authors raise however important questions on the links between discrimination, (economic) inequalities and meritocracy, as also underlined by Dubet (2014, 2016).

6.9.1 Credit Scoring

"*Credit scoring is one of the most successful applications of statistical and operations research modeling in finance and banking*" claimed Thomas et al. (2002). "Credit scoring" describes statistical models used to assist credit institution and banks in the process of running credit granting operations, including bank loans, credit cards, mortgages, etc. These techniques were developed to provide a more quantitative and supposedly more objective approach to historical approaches, which were essentially based on personal judgement. Credit cards provided a guaranteed source of credit, so that the customer did not have to go back and negotiate a loan for each individual purchase. In 1949, the "Diner's Club" card was offered, before banks offered the first cards in the 1950s and 1960s. In 1958, the Bank of America introduced the BankAmericard, a statewide card in California that offered customers a debit, and especially a credit, card service, before spreading nationwide in 1965. In 1966, a group of Illinois banks got together and formed the Interbank Card Association (ITC). Together they created Mastercard (originally called Master Charge, the current name dates from 1979). In England, Barclays launched Barclaycard (which had a monopoly until the arrival of the Access card in 1972). In 1973, electronic payment was introduced with faster authorisation systems (the waiting time went from 5 minutes to 56 seconds per authorisation) and clearing. The Bank of America card was renamed Visa in 1977 to facilitate the roll-out of the card worldwide. With the development of these credit cards, which were to allow credit to be granted almost instantaneously, it became necessary to quantify the associated risk. In the 60's, several articles discussed the use of predictive models, such as Chatterjee and Barcun (1970) or Churchill et al. (1977).

In the brief section "*how insurers determine your premium*," in the National Association of Insurance Commissioners (2011, 2022) reports, it is explained that "*most insurers use the information in your credit report to calculate a credit-based insurance score. They do this because studies show a correlation between this score and the likelihood of filing a claim. Credit-based insurance scores are different from other credit scores.*" And as mentioned in Kuhn (2020), credit scoring is allowed in 47 states in the U.S. (all except California, Massachusetts, and Hawaii), and it is used by the 15 largest auto insurers in the country and over 90% of all U.S. auto insurers.

More generally, credit scores are an important piece of individual information in the United States or Canada, widely used in many lines of insurance, not just loan insurance. As noted, a (negative) credit event, such as a default (or late) payment on a mortgage, or a bankruptcy, can impact an individual for a considerable period of time. These credit scores are numbers that represent an assessment of a person's creditworthiness, or the likelihood that he or she will pay back their debts. But increasingly, these scores are being used in quite different contexts, such as insurance. As mentioned by Kiviat (2019), "*the field of property and casualty insurance is dominated by the idea that it is fair to use data for pricing if the data actuarially relate to loss insurers expect to incur*." And as she explains, actuaries, underwriters and policymakers seem to have gone along with that, being conform with their sense of "moral deservingness" (using the term introduced by Watkins-Hayes and Kovalsky (2016)). Guseva and Rona-Tas (2001) recalls that in North America, Experian, Equifax and TransUnion keep records of a person's borrowing and repayment activities. And FICO (Fair Isaac Corporation) has developed a formula (not known) that calculates, based on these records, a score, based on debt and available credit, income, or rather their variations (along with payment history, number of recent credit applications and negative events – such as bankruptcy or foreclosure – as well as changes in income due to changes in employment or family situation). The FICO score starts at 300 and goes up to 850, with a poor score below 600, and an "exceptional" score above 800. The average score is around 711. This score, created for banking institutions, is now used in pre-employment screening, as Bartik and Nelson (2016) reminds us. For O'Neil (2016), this use of credit scores in hiring and promotion creates a dangerous vicious cycle in terms of poverty. The Credit-Based Insurance Scores cannot use any personal information

to determine the score, other than that on the credit report, specifically race and ethnicity, religion, gender, marital status, age, employment history, occupation, place of residence, child/family support obligations or rental agreements. On the Table 6.4, via Solution (2020) we can see the evolution of the proposed credit rate according to the credit score, over 30 years, for a “good risk" (3.2%) and a “bad risk" (4.8%), with the total amount of credit (a “bad risk" will have an extra cost of 20%), and on an insurance premium, for the same profile, a “bad risk" having an extra premium of about 70%.

In Table 6.4, we have on top, the cost of a 30-year credit, for an amount of $150,000, according to the credit score (with 5 categories, going from the most risky on the left to the least risky on the right), with the average rate proposed. Below, insurance premium charged to a 30 year old driver, with no convictions, driving an average car, 12,000 miles per year, in the city, based on the same credit score categories

	300-619	620-659	660-699	700-759	760-850
Total credit cost	$283,319	$251,617	$245,508	$239,479	$233,532
Rates	4.8%	3.8%	3.6%	3.4%	3.2%
Motor Insurance Premium	$2580	$2400	$2150	$2000	$1500

Table 6.4: Top, cost of a 30-year credit, for an amount of $150,000, according to the credit score, with the average interest rate. Below, insurance premium charged to a 30 year old driver, with no convictions, driving an average car, 12,000 miles per year, in the city
(source: InCharge Debt Solutions).

As noted by Lauer (2017), credit institutions have long compiled data on individuals' financial histories, such as timeliness of loan repayment's, total amount borrowed, frequency of applying for new loans, among others. Insurers quickly realized that this information was highly predictive of all kinds of risks. As early as the 1970s, companies in the United States were required by law to notify people when credit reports resulted in adverse actions, such as an increase in insurance premium or credit. Decades later, this rule let people know that credit scores were causing their auto insurance rates to suddenly rise. For Kiviat (2019), it is only recently that it has become clear how important credit score is in insurance product pricing, especially in auto insurance. Miller et al. (2003) illustrate credit reports contain a wide variety of credit information on an individual consumer. In addition to information that identifies a particular person, the report contains data on credit card and loan balances, types of credit, the status of each account, judgments, liens, collections, bankruptcies and credit inquiries.

These credit scores have long been used by lending institutions to predict the risk associated with repaying a loan or meeting another financial responsibility. Insurance score modellers have begun to combine and weight selected credit attributes to develop a single insurance score. These “*insurance scores*" are added as a risk factor to create risk classification schemes in order to obtain more accurate results. Although both the credit score and the insurance score are derived from a person's credit report, the two scores are different. There is no reason to believe that a credit score measuring the likelihood of loan repayment will be based on the same attributes (or that each attribute will be given the same weight) as those used to calculate an insurance score, and vice versa. Unfortunately, some in the insurance industry have come to call credit-based insurance scores simply “*credit score*." This abuse of language may have led some to conclude that the advantages and disadvantages of using credit scores in the lending industry are directly related to the advantages and disadvantages of insurance. It may also have led some to attempt to apply the results of studies of credit scores by lending institutions to the use of insurance scores by insurers. Morris et al. (2017) points out that credit-based insurance scores are perhaps the most important example of insured's characteristics that is regulated because it may yield potentially suspect classifications. Any correlation between insurance

scores and race or income is potentially troubling. First, insurance scoring can have a disparate impact on racial minorities racial minorities and low-income households, causing members of these groups to pay higher premiums on average. Credit information, for example, has been used for auto and home insurance underwriting for several years, despite its potential for proxy discrimination (as classified by Kiviat (2019) and Prince and Schwarcz (2019)).

In particular, Brockett and Golden (2007) questioned the use of credit scores to price insurance contracts. As noted by Kiviat (2019), inspired by Morris et al. (2017), policyholder income, in particular, could be predictive of future insurance claims if low-income policyholders are more likely to file claims even when losses are only slightly above their deductible. Some insurers may not even be aware that there is a correlation between the proxy variable (credit scores) and the suspect variable (race and income). Even if insurers are aware of this correlation, they may not believe that this correlation helps explain the power of credit information to predict claims. Instead, they may believe, as in fact much available evidence indicates, that credit information is predictive of claims because it measures policyholder care levels. Some insurers, such as Root Insurance explain that they refuse to use these credit scores because they consider them discriminatory: "*for decades, the car insurance industry has used credit score as a major factor in calculating rates. By basing rates on demographic factors like occupation, education, and credit score, the traditional car insurance industry has long relied on unfair, discriminatory biases in its pricing practices. These practices unfairly penalize historically under-resourced communities, immigrants, and those struggling to pay large medical expenses*."

Credit scores, or more precisely variations of credit scores can indicate changes in protected attribute. As explained in Avery et al. (2004), individuals who recently experienced a divorce might be expected to have a higher likelihood of payment problems on accounts, that will translate into their credit score. But marital status is usually a protected attribute. And as shown in Dean and Nicholas (2018); Dean et al. (2018), "*credit scores are increasingly used to understand health outcomes*," and more and more paper describe this so-called "out-of-pocket health care" in the U.S. (see Pisu et al. (2010), Bernard et al. (2011) or Zafar and Abernethy (2013)).

6.9.2 Discrimination Against the Poor

In 2013, Martin Hirsch (former director of Emmaüs – a charitable organization – and Assistance Publique - Hôpitaux de Paris – a university hospital trust – in France) claimed that "*it's getting expensive to be poor*." To understand why there could be discrimination against poor people, it is important to get back to "merit", and try to understand on what criteria do we admire people. For the Greeks, excellence, or αρετη (arete), was a major virtue. This excellence went beyond moral excellence: in the Greco-Roman world, the term evoked a form of nobility, recognizable by the beauty, strength, courage, or intelligence of the person. Now this excellence had little to do with wealth: thus Herodotus is astonished that the winners of the Olympian games were content with an olive wreath and a "glorious renown," (περι αρετης). In the Greek ethical vision, especially among the Stoics, a "good life" does not depend on material wealth – a precept pushed to its height by Diogenes who, seeing a child drinking from his hands at the fountain, throws away the bowl he had for all crockery, telling himself that it is again useless wealth.

Greek society was nevertheless a deeply hierarchical society, even if it was organized around values other than material wealth. We can then ask ourselves at what point in Western history wealth became the measure of all things. One thinks then of Max Weber's theory (Weber (1904)): the ethics of Protestantism pushes for work and earthly success as a revelation of a divine election to come: the rich of this world would be the chosen of the next. In the same way Adam Smith, taking a critical look at the birth of capitalism in the society of his time, titles a chapter of The Theory of Moral Sentiments, Smith (1759), "*of the corruption of*

our moral feelings occasioned by that disposition to admire the rich and great, and to despise or neglect the poor and lowly." Today, the cult of wealth seems to have never been so strong and material success is almost elevated to the rank of virtue. On the other hand, poverty becomes a stigma that is hard to get rid of; but history shows us that this is not natural.

Indeed, the poor have not always been "bad". In Europe, the Church has largely contributed to disseminating the image of the "good poor", as it appears in the Gospels: "*happy are you poor, the kingdom of God is yours,*" or "*God could have made all men rich, but he wanted there to be poor people in this world, so that the rich would have an opportunity to redeem their sins.*" Beyond this, the poor is seen as an image of Christ, Jesus having said "*whatever you do to the least of these, you will do to me*." Helping the poor, doing a work of mercy, is a means of salvation.

For Saint Thomas Aquinas, charity is thus essential to correct social inequalities by redistributing wealth through alms giving. In the Middle Ages, merchants were seen as useful, even virtuous, since they allowed wealth to circulate within the community. Priests played the role of social assistants, helping the sick, the elderly and the disabled. The hospices and "xenodochia" of the Middle Ages (ξενοδοχεῖον, the "place for strangers," ξένος) are the symbol of this care of the poor. And quite often, poverty is not limited to material capital, but also social and cultural, to use a more contemporary terminology.

Towards the end of the Middle Ages, the figure of the "bad poor", the parasitic and dangerous vagabond, appeared. Brant (1494) denounced these welfare recipients, "*some become beggars at an age when, young and strong, and in full health, one could work: why bother*." This mistrust was reinforced with the great pandemic of the Black Death. The hygienic theories of the end of the XIX-th century added the final touch: if fevers and diseases were caused by insalubrity and poor living conditions, then by keeping the poor out, they were protected from disease.

In the words of Mollat (2006), "*the poor are those who, permanently or temporarily, find themselves in a situation of weakness, dependence, humiliation, characterized by the deprivation of means, variable according to the times and the societies, of power and social consideration*." Recently, Cortina (2022) proposed the term "aporophobia", or "pauvrophobia", to describe a whole set of prejudices that exist towards the poor. The unemployed are said to be welfare recipients and lazy. These prejudices, which stigmatize a group, "*the poor*", lead to fear or hatred, generating an important cleavage, and finally a form of discrimination. Cortina (2022)'s "pauvrophobia" is a discrimination against social precariousness, which would be almost more important than standard forms of discrimination, such as racism or xenophobia. Cortina (2022) ironically notes that rich foreigners are often not rejected.

But these prejudices also turn into accusations. Szalavitz (2017) thus abruptly asks the question, "*why do we think poor people are poor because of their own bad choices?*." The "actor-observer" bias provides one element of an answer: we often think that it is circumstances, which constrain our own choices, but that it is the behavior of others that changes theirs. In other words, others are poor because they made bad choices, but if I am poor, it is because of an unfair system. This bias is also valid for the rich: winners often tend to believe that they got where they are by their own hard work, and that they therefore deserve what they have.

Social science studies show, however, that the poor are rarely poor by choice, and increasing inequality and geographic segregation do not help. The lack of empathy then leads to more polarization, more rejection and, in a vicious circle, even less empathy.

To discriminate is to distinguish (exclude or prefer) a person because of his/her "personal characteristics". Can we then speak of discrimination against the poor? Is poverty (like gender or skin color) a personal characteristic? In Québec, "social condition" (which explicitly includes poverty) is one of the protected attribute and therefore prohibited discrimination.

And there correlation between wealth and risk. In France, for deaths due to road accidents, there 3% of "executives" and 15% of "workers", while they represent nearly 20% of the working population each.

Blanpain (2018) points out that the gap in life expectancy at birth is 13 years between the most affluent and the most modest men, as discussed in Chapter 2.

6.10 Networks

From a mathematical perspective, a "graph" $\mathcal{G}$ is made up of "nodes" (also called "vertices", denoted V) which are connected by "edges" (denoted E). And those mathematical structures are use to define networks, that are graphs whose the nodes or edges possess attributes. In a social network, nodes are individuals (or policyholders) and edges will denote some sort of "connections" between individuals. On social media such as Facebook or Linkedin, an edge indicates that two individuals are friends or colleague, that will be interconnected, through some reciprocal connection. On Twitter, connections are directional, in the sense that one account is following another one (but of course, some edges could be bidirectional). In genealogical trees, nodes are individuals, and connections will usually denote parent-children connections. Other popular network structures are "bipartite graphs", where nodes are divided into two disjoint and independent sets V_1 and V_2, and edges connect nodes between V_1 and V_2 (and not within). Classical examples are "affiliate networks" where employees and employers are connected, or policyholders and brokers, car accident and experts, disability claims and medical doctors, etc. It is also possible have two groups of nodes, and within and between edges, such as disease and drug networks, and edges denote associations.

"*Villagers experience health problems, crop failures, wildly changing employment, in addition to a variety of needs for cash, such as dowries. They don't have insurance or much, if any, savings: they rely on each other for help,*" said Jackson (2019). More and more, insurers try to extract information from various networks. And as Bernstein (2007) claimed, "*network and data analyses compound and reflect discrimination embedded within society*."

6.10.1 On the Use of Networks

Scism (2019) presented a series of "life hacks", including tips on how to behave on social media in order to bypass insurers' profile evaluations. For example "*do not post photos of yourself smoking*," "*post pictures of yourself exercising (but not while engaging in a risky sport)*," "*use fitness tracking devices to show you are concerned about your health*," "*purchase food from healthy online meal-preparation services*," and "*visit the gym with mobile location-tracking enabled (while leaving your phone at home when you go to a bar)*."

Social networks are also important to analyse fraud. Fraud is often committed through illegal set-ups with many accomplices. When traditional analytical techniques fail to detect fraud due to a lack of evidence, social network analysis might give new insights by investigating how people influence each other. These are the so-called guilt-by-associations, where we assume that fraudulent influences run through the network. For example, insurance companies often have to deal with groups of fraudsters, trying to swindle by resubmitting the same claim using different people. Suspicious claims often involve the same claimers, claimees, vehicles, witnesses, and so on. By creating and analyzing an appropriate network, inspectors might gain new insight in the suspiciousness of the claim and can prevent pursuit of the claim. In many applications, it might be useful to integrate a second node type in the network. Affiliation or bipartite networks represent the reason why people connect to each other, and include the events that network objects—like people or companies—attend or share. An event can for example refer to a paper (scientific fraud), a resource (social security fraud), an insurance claim (insurance fraud), a store (credit card fraud), and so on. Adding a new type of nodes to the network does not only enrich the imaginative power of graphs, but also creates new insights in the network structure and provides additional information neglected before. On the other hand, including a second type of node results in an increasing complexity of the analysis.

As mentioned by the National Association of Insurance Commissioners (2011, 2022), "*insurance companies can base premiums on all insured drivers in your household, including those not related by blood, such as roommates.*" And Boyd et al. (2014) asserted that there is a new kind of discrimination associated not with personal characteristics (like those discussed in the previous section) but with personal networks. Beyond their personal characteristics (such as race or gender), an important source of information is "*who they know*." In the context of usage-based auto insurance, the nuance is that personal networks are not those represented by driving behavior (strictly speaking), but those defined by the places to which people physically go.

In many countries, when it comes to employment, most companies are required to respect equal opportunity: discrimination on the basis of race, gender, beliefs, religion, color, and national origin is prohibited. Additional regulations prohibit many employers from discriminating on the basis of age, disability, genetic information, military history and sexual orientation. However, nothing prevents an employer from discriminating based on a person's personal network. And increasingly, as Boyd et al. (2014) reminds us, technical decision-making tools are providing new mechanisms by which this can happen. Some employers use LinkedIn (and other social networking sites) to determine a candidate's "*cultural fit*" for hire, including whether or not a candidate knows people already known to the company. While hiring on the basis of personal relationships is by no means new, it takes on new meaning when it becomes automated and occurs on a large scale. Algorithms that identify our networks, or predict our behaviour based on them, offer new opportunities for discrimination and unfair treatment.

6.10.2 Mathematics of Networks, and Paradoxes

Feld (1991) has shown that in any network the average degree (i.e. the number of neighbours, or connections) of the neighbour of a node is strictly greater than the average degree of nodes in the network as a whole. Applied to networks of friendship, this translates simply as "*on average your friends have more friends than you do.*" This phenomenon is known as the "friendship paradox." A related phenomenon, the generalized friendship paradox, describes similar behaviour with respect to other attributes of network nodes, Jo and Eom (2014). Are your friends richer than you? Generalized friendship paradoxes arise when such attributes are correlated with node degree. If richer people are on average also more popular, then wealth and popularity will be positively correlated and hence the tendency for your friends to be more popular than you could mean they are also richer. Heuristically, the friendship paradox states that people's friends tend to be more popular than they themselves are. Stated a little more precisely, nodes in a network tend to have a lower degree than their neighbours do. Consider an undirected network of n individuals (nodes). Let $\boldsymbol{A} = [A_{i,j}]$ denote the $n \times n$ adjacency matrix, and $\boldsymbol{d} = (d_i)$ denote the n-vector of degrees, in the sense that $d_i = A_i^\top \mathbf{1}$.

Propoition 6.10.1 *The average number of friends of the collection of friends of individuals in a social network will be higher than the average number of friends of the collection of the individuals themselves. More formally*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{d_i}\sum_{j=1}^{n}A_{ij}d_j\right) \geq \frac{1}{n}\sum_{i=1}^{n}d_i.$$

Define differences Δ_i's between the average of its neighbours' degrees and its own degree, in the sense that

$$\Delta_i = \frac{1}{d_i}\sum_{j=1}^{n}A_{ij}d_j - d_i.$$

where we suppose here that all nodes with non-zero degrees (at least one neighbour). The friendship paradox states that the average of Δ_i across all nodes is greater than zero. In order to prove it, write the average as

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{d_i}\sum_{j=1}^{n}A_{ij}d_j - d_i\right) = \frac{1}{n}\sum_{ij=1}^{n}\left(A_{ij}\frac{d_j}{d_i} - A_{ij}\right),$$

that yields

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i = \frac{1}{n}\sum_{ij=1}^{n}A_{ij}\left(\frac{d_j}{d_i} - 1\right) \text{ but also } \frac{1}{n}\sum_{ij=1}^{n}A_{ij}\left(\frac{d_i}{d_j} - 1\right)$$

by exchanging the summation indices, and because $\boldsymbol{A}$ is a symmetric matrix. By adding the two, we can write

$$\frac{2}{n}\sum_{i=1}^{n}\Delta_i = \frac{1}{n}\sum_{ij}A_{ij}\left(\frac{d_j}{d_i} + \frac{d_i}{d_j} - 2\right) = \frac{1}{2n}\sum_{ij}A_{ij}\left(\sqrt{\frac{d_j}{d_i}} - \sqrt{\frac{d_i}{d_j}}\right)^2 \geq 0.$$

Observe that the exact equality holds only when $d_i = d_j$ for all pairs of neighbors, corresponding to the case where the network is a regular graph (or possibly the reunion of disjoint regular graphs).

REP For the generalized friendship paradox, which considers attributes other than degree, as in Cantwell et al. (2021), one can define an analogous quantity, $\Delta_i^{(x)}$, for some attribute x (such as the wealth) is defined as

$$\Delta_i^{(x)} = \frac{1}{d_i}\sum_{j}A_{ij}x_j - x_i,$$

which measures the difference between the average of the attribute for node i's neighbours and the value for i itself. When the average of this quantity over all nodes is positive one may say that the generalized friendship paradox holds. In contrast to the case of degree, this is not always true—the value of $\Delta_i^{(x)}$ can be zero or negative—but we can write the average as

$$\frac{1}{n}\sum_{i}\Delta_i^{(x)} = \frac{1}{n}\sum_{i}\left(\frac{1}{d_i}\sum_{j}A_{ij}x_j - x_i\right) = \frac{1}{n}\sum_{i}\left(x_i\sum_{j}\frac{A_{ij}}{d_j} - x_i\right),$$

where the second line again follows from interchanging summation indices. Defining the new quantity

$$\delta_i = \sum_{j}\frac{A_{ij}}{d_j}$$

and noting that

$$\frac{1}{n}\sum_{i}\delta_i = \frac{1}{n}\sum_{ij}\frac{A_{ij}}{d_j} = \frac{1}{n}\sum_{j}\frac{1}{d_j}\sum_{i}A_{ij} = 1,$$

we can then write

$$\frac{1}{n}\sum_{i}\Delta_i^{(x)} = \frac{1}{n}\sum_{i}x_i\delta_i - \frac{1}{n}\sum_{i}x_i\,\frac{1}{n}\sum_{i}\boldsymbol{\delta}_i = \mathrm{Cov}(\boldsymbol{x},\boldsymbol{\delta}).$$

Thus, we will have a generalized friendship paradox in the sense defined here if (and only if) $\boldsymbol{x}$ and $\boldsymbol{\delta}$ are positively correlated. But this is not always the case

$$\left.\begin{array}{l}\mathrm{Cov}(\boldsymbol{d},\boldsymbol{\delta}) \geq 0\\ \mathrm{Cov}(\boldsymbol{x},\boldsymbol{\delta}) \geq 0\end{array}\right\} \not\Longrightarrow \mathrm{Cov}(\boldsymbol{d},\boldsymbol{x}) \geq 0.$$

French poet Paul Verlaine warned us[7], "*il ne faut jamais juger les gens sur leurs fréquentations. Tenez, Judas, par exemple, il avait des amis irréprochables*." Nevertheless, several organizations have proposed to use information about our "friends" in order to learn more about us, following an homophily principle (in the sense of McPherson et al. (2001)), because as popular saying goes, "*birds of a feather flock together*." Therefore Bhattacharya (2015) noted that "*you apply for a loan and your would-be lender somehow examines the credit ratings of your Facebook friends. If the average credit rating of these members is at least a minimum credit score, the lender continues to process the loan application. Otherwise, the loan application is rejected*." As we can see, mathematical guarantees are not strong here, and it is very likeliy that such strategy will create more biases.

[7] "*you should never judge people by who they associate with. Take Judas, for example, he had friends who were beyond reproach*".

Chapter 7

Observations or Experiments: Data in Insurance

In important challenge for actuaries is that they need to answer causal questions with observational data. After a brief discussion about correlation and causality, we will describe the “causation ladder", and the three rungs: association or correlation (“*what if I see...*"), intervention (“*what if I do...*"), and counterfactuals (“*what if I had done...*"). Counterfactuals will be important to quantify discrimination.

To take up the classification of Rosenbaum (2018), it is important to distinguish “experimental" and “observational" data. In the latter case, we use data also called “administrative data", such as financial transaction or medical visit records, which show what people actually do, and not what they say they do (as in surveys, for example). Such data allows us to see what people buy, eat, where they travel, etc., from what they leave behind. There were 185 billion transactions using Visa cards in 2019, and 85 billion using Mastercard. This massive data gives us information on the holder’s behaviour, without their knowledge. Looking at transactions probably allows us to better quantify the number of air travels for a given person than a survey or a quick poll would, we also suspect that these records are biased and do not reflect all transactions, and probably give a biased view of their behaviour. If we want to understand the impact of prevention on risk, we take data we collected from doctors, and compare the outcome of those who go for a routine check, and those who don’t. Those are “observational data" in the sense that agents are free to act as they please, and we are content to observe from a distance, as explained in Rosenbaum (2005). But of course, there might be a strong bias, as discussed previously: people might be doing a routine exam *because* they know that they are at risk. With “experimental data", we consider some controlled randomized assignment to some treatment, as explained in Shadish and Luellen (2005). In our example, we will ask some people, randomly chosen, to do routine check.

7.1 Correlation and Causation

In section 3.4, we have described various predictive models, where some variables, called “explanatory variables", $\boldsymbol{x}$ are used to “predict" a variable y, through some function m. We needed variables $\boldsymbol{x}$ to be

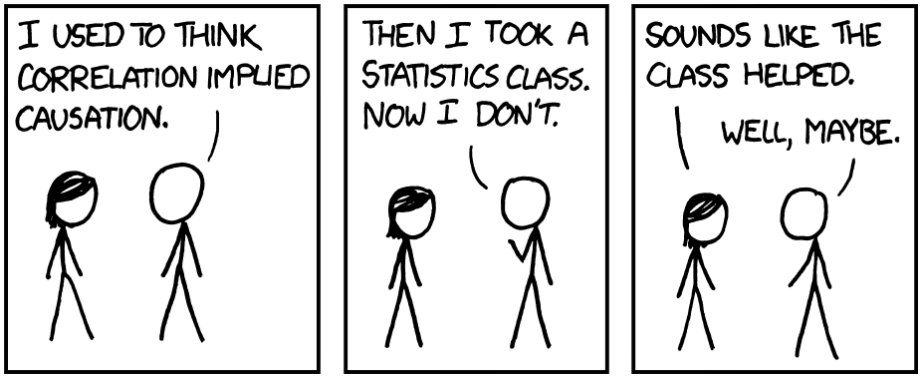


Figure 7.1: Correlation, by Randall Munroe, 2009 (source: https://xkcd.com/552/).

as correlated as possible with y, so that $m(\boldsymbol{x}$ is also, hopefully, correlated with y. This correlation make them appear as valid predictors, and that is usually a sufficient motivation to use them as pricing variables. "*Databases about people are full of correlations, only some of which meaningfully reflect the individual's actual capacity or needs or merit, and even fewer of which reflect relationships that are causal in nature*," claimed Williams et al. (2018). In section 4.1, we discussed interpretability, and explainability, we try to explain "*why*" a variable is legitimate in a pricing model, and "*why*" a specific individual get a large prediction. But as mentioned by Kiviat (2019), the causal effect from $\boldsymbol{x}$ to y, is usually not a worthwhile question for insurers, and the concept of "actuarial fairness" allows to sweep this problem under the rug. But more and more "*policymakers want(ed) to understand why*" some variable "*predict(ed) insurance loss in order to determine if any links in the causal chain held the wrong people to account, and as a consequence gave them prices they did not deserve*" recalls Kiviat (2019). That is actually simply a regulatory requirement across many countries, states and provinces, if one wants to prove that rates are not "unfairly discriminatory".

7.1.1 Correlation is (Probably) Not Causation

Everyone is familiar with the adage "*correlation does not imply causation*." And indeed, a classical question for researcher, but also policy makers, is whether a significant correlation between two variables represents a causal effect. As mentioned in Traag and Waltman (2022), a similar question arises when we observe a difference in outcomes y between two groups: does the difference represent a bias or disparity? If the difference is based on a causal effect, there may be a disparity, or a bias. But if the difference does not represent a causal effect, there is probably no bias or disparity.

To illustrate that issue, in an article published in the Wall Street Journal, entitled *Medicaid is worse than no coverage at all*, Scott Gottlieb stated that "*dozens of recent medical studies show that Medicaid patients suffer for it. In some cases, they'd do just as well without health insurance.*" Specifically, Gottlieb (2011) relied on LaPar et al. (2010), which showed, statistically, that uninsured patients were about 25% less likely than those with Medicaid to have an "in-hospital death'", see Table 7.1.

Several studies have found that Medicaid patients with specific conditions (e.g. cancer) or who have undergone specific treatments (e.g. lung transplantation or coronary angioplasty) had significantly poorer health outcomes than patients with private insurance, the same conditions, or same procedures. For example, for major surgery, uninsured patients were about 25% less likely than Medicaid patients to die in the hospital. Being insured by Medicaid would almost seem to make the risk worse! The potential problem with this type

	Medicare	Medicaid	Uninsured	Insurance
In-hospital mortality	4.4%	3.7%	3.2%	1.3%
Pulmonary resection	4.3%	4.3%	6.2%	2.0%
Esophagectomy	8.7%	7.5%	6.5%	3.0%
Colectomy	7.5%	5.4%	3.9%	1.8%
Pancreatectomy	6.1%	5.8%	8.4%	2.7%
Gastrectomy	10.8%	5.4%	5.0%	3.5%
Aortic aneurysm	12.4	14.5	14.8	7.0
Hip replacement	0.4%	0.2%	0.1%	0.1%
Coronary artery bypass grafting	4.0%	2.8%	2.3%	1.4%
Number of cases	491,829	40,259	24,035	337,535
Age (years)	73.5 ± 8.6	49.8 ± 16.4	51.8 ± 12.8	55.5 ± 11.4
Women	49.6%	48.8%	35.8%	39.7%
Length of stay (days)	9.5 ± 0.1	12.7 ± 0.4	10.1 ± 0.3	7.4 ± 0.1
Total Cost ($)	76, 374 ± 53.1	93, 567 ± 251.4	78, 279 ± 231.0	63, 057 ± 53.0
Rural location	10.1%	8.5%	9.8%	6.6%

Table 7.1: In-hospital mortality for all patients undergoing major surgery, by major payer group. (source LaPar et al. (2010), Tables 4 and 5).

of study is that the comparison groups, Medicaid patients and privately insured or uninsured patients, may be subject to self-selection. It is likely that many patients do not enroll in Medicaid until after, sometimes long after, the onset of a serious medical problem. In this case, those who choose to enroll in Medicaid may be sicker than those with private insurance or those who are uninsured. In a way, this self-selection effect is a form of reverse causation. Being sick drives people to enroll in Medicaid, not the other way around. This is the concern with “observational data".

In 2008, due to severe budget constraints, Oregon found itself with a Medicaid waiting list of 90,000 people and only enough money to cover 10,000 of them. So the state created a lottery to randomly select people who would qualify for Medicaid, therefore recreating the necessary preconditions. The reality, however, was a bit more complex, as many of the lottery winners were not eligible for Medicaid or chose not to submit their paperwork to enroll in the program. Compared to the control group (people who did not have access to Medicaid), Finkelstein et al. (2012) observed that the treatment group had substantially and statistically significantly higher health care use (including primary and preventive care as well as hospitalizations), lower out-of-pocket medical expenditures and medical debt (including fewer bills sent to collection), and better self-reported physical and mental health. These “experiments", which are often difficult to implement (for financial and sometimes ethical reasons), make it possible to bypass the bias of administrative data. Having non-null correlation is quite easy, but proving a causal effect is difficult.

Definition 7.1.1 (Common cause) *Reichenbach (1956). If X and Y are non-independent, $X \not\perp\!\!\!\perp Y$, then, either*

$$\begin{cases} X \text{ causes } Y \\ Y \text{ causes } X \\ \text{there exists } Z \text{ such that } Z \text{ causes both } X \text{ and } Y \end{cases}$$

This concept of “common cause" is ill-defined and might be seen as tautological, since “cause" has not been defined properly. Heuristically, the (probabilistic) causation is not that the occurrence of a single

event “causes" another to happen, but rather that the occurrence increases the likelihood of the other event to happen, all else being equal (see Hitchcock (1997)). A starting point could be the case of sequential variables, where the dynamics could make causality easier to define.

7.1.2 Causality in a Dynamic Context

Before defining causality in the context of individual data, let us recall that, in a context of temporal data, Granger (1969) introduced a concept of “causality" which takes a relatively simple form. Sequences of observations will be useful to properly capture this “causal" effect. Consider here a standard bivariate autoregressive time series, where a regression of variables at time $t+1$ on the same varibles at time t is performed

$$\begin{cases} x_{1,t+1} = c_1 + a_{1,1}x_{1,t} + a_{1,2}x_{2,t} + \varepsilon_{1,t} \\ x_{2,t+1} = c_2 + a_{2,1}x_{1,t} + a_{2,2}x_{2,t} + \varepsilon_{2,t}, \end{cases}$$

also noted $\boldsymbol{x}_{t+1} = \boldsymbol{c} + \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{\varepsilon}_{t+1}$ where the off-diagonal terms of the autoregression matrix $\boldsymbol{A}$ allow to quantify the lagged causality, i.e. a lagged causal effect (between t and $t+1$) with respectively $x_1 \rightarrow x_2$ or $x_1 \leftarrow x_2$ (see Hamilton (1994) or Reinsel (2003) for more details). For example Figure 7.2 shows the scatterplot $(x_{1,t}, x_{2,t+1})$ and $(x_{2,t}, x_{1,t+1})$, left and right, where respectively x_1 denotes the number of cyclists in Helsinki, per day, in 2014 (at a given road intersection) and x_2 denotes the (average) temperature on the same day. The graph on the left is equivalent to asking whether the temperature “causes" the number of cyclists (if the temperature rises, the number of cyclists on the roads rises) and the graph on the right is equivalent to asking whether the number of cyclists “causes" the temperature (if the number of cyclists on the roads rises, the temperature rises). In both cases, if we estimate

$$\begin{cases} x_1 \rightarrow x_2 : & x_{2,t+1} = \gamma_1 + \alpha_{2,1}x_{1,t} + \eta_{1,t}, \\ x_1 \leftarrow x_2 : & x_{1,t+1} = \gamma_2 + \alpha_{1,2}x_{2,t} + \eta_{2,t}, \end{cases}$$

we observe significant regressions (but this is not a causal test)

$$\begin{cases} x_1 \rightarrow x_2 : & x_{2,t+1} = \underset{(234)}{4320} + \underset{(23)}{757}\, x_{1,t} + \eta_{1,t},\ R^2 = 75.72\% \\ x_1 \leftarrow x_2 : & x_{1,t+1} = \underset{(0.04)}{-1.98} + \underset{(0.00003)}{0.001}\, x_{2,t} + \eta_{2,t},\ R^2 = 72.41\% \end{cases}$$

We can use the Granger test (see Hamilton (1994)), on the data of Figure 7.2, on the two causal hypotheses (not on the levels but on the daily variations, of the number of cyclists, and of the temperature)

$$\begin{cases} x_1 \rightarrow x_2 : & H_0 : a_{2,1} = 0,\ p-\text{value} = 56.66\% \\ x_1 \leftarrow x_2 : & H_0 : a_{1,2} = 0,\ p-\text{value} = 0.004\% \end{cases}$$

In other words, temperature is causally related to the presence of cyclists on the road (the temperature “causes" the number of cyclists), but not vice versa.

In a non-dynamic context, defining causality will be a more perilous exercise. In the next sections, we will get back to the “*causation ladder*," introduced in Pearl (2009b) and more recently in Pearl and Mackenzie (2018). The first stage (section 7.2), “*association*," corresponds to the most basic level, we see that two or more quantities are somehow related, or correlated. The second stage (section 7.3), “*intervention*," corresponds to the case that we can see association, but we can also change the world through appropriate interventions (or experiments). The third stage (section 7.4), is about “*counterfactuals*." The “*fundamental*

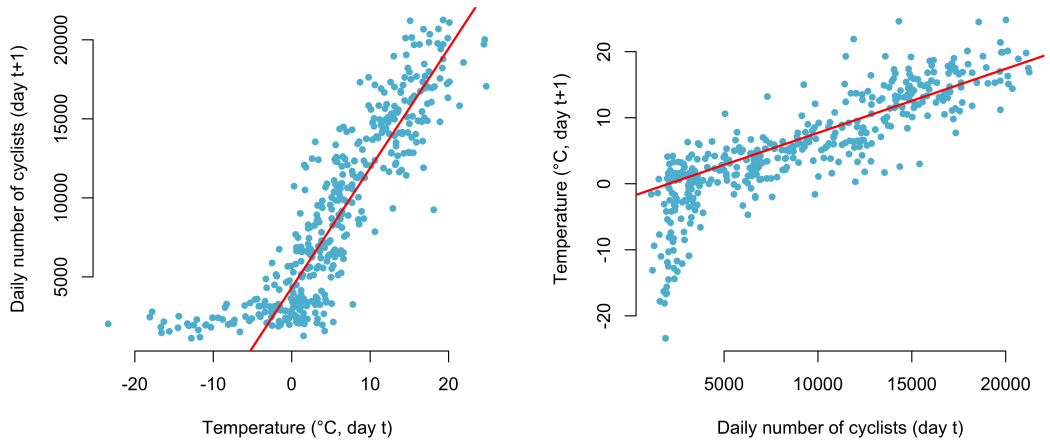


Figure 7.2: Number of cyclists per day (x_2), in 2014, in Helsinki (Finland), and average daily temperature (x_1), respectively at time t on the x-axis and $t+1$ on the y-axis. The regression lines are estimated only on days when the temperature exceeded 0°C.
(source, NaytaData, https://www.reddit.com/r/dataisbeautiful/comments/8k40wl)

problem of causal inference" as Holland (1986) claimed, is that we we only ever observe one realization, even if there could have been, at some point, several alternatives. In that section, we will use counterfactuals for quantity the causal effect, which would be the difference between what we observed, and the "potential outcome" if the individual had had the treatment.

7.2 Rung 1, Association (Seeing, "*what if I see...*")

Two variables are independent when the value of one gives no information about the value of the other. In everyday language, dependence, association and correlation are used interchangeably. Technically, however, association is synonymous with dependence (or non independence) and is different from correlation. "*Association is a very general relationship: one variable provides information about another*, as explained in Altman and Krzywinski (2015), in the sense that there could be association even if the "correlation" is statistically null.

7.2.1 Independence and Correlation

Here, we will have to study the joint distribution of some variables, and their conditional distribution.

Propoition 7.2.1 (Chain Rule) *In dimension 2, for all sets $\mathcal{A}$ and $\mathcal{B}$,*

$$\mathbb{P}\big[X \in \mathcal{A}, Y \in \mathcal{B}\big] = \mathbb{P}\big[Y \in \mathcal{B} \big| X \in \mathcal{A}\big] \cdot \mathbb{P}\big[X \in \mathcal{A}\big]$$

and in dimension 3, for all sets $\mathcal{A}$, $\mathcal{B}$ and C,

$$\mathbb{P}\big[X \in \mathcal{A}, Y \in \mathcal{B}, Z \in C\big] = \mathbb{P}\big[Z \in C \big| Y \in \mathcal{B} X \in \mathcal{A}\big] \cdot \mathbb{P}\big[Y \in \mathcal{B} \big| X \in \mathcal{A}\big] \cdot \mathbb{P}\big[X \in \mathcal{A}\big]$$

This chain rule is a simple consequence of the definition of the conditional probability.

Definition 7.2.1 (Independence (dimension 2)) *X and Y are independent, denoted $X \perp\!\!\!\perp Y$, if for any sets $\mathcal{A}, \mathcal{B} \subset \mathbb{R}$,*

$$\mathbb{P}\big[X \in \mathcal{A}, Y \in \mathcal{B}\big] = \mathbb{P}\big[X \in \mathcal{A}\big] \cdot \mathbb{P}\big[Y \in \mathcal{B}\big]$$

An equivalent expression is, from the Chain Rule, that

$$\mathbb{P}\big[Y \in \mathcal{B} \big| X \in \mathcal{A}\big] = \mathbb{P}\big[Y \in \mathcal{B}\big].$$

Definition 7.2.2 (Linear Independence) *Consider two random variables X and Y. $X \perp Y$ if and only if for any $Cov[X, Y] = 0$.*

Propoition 7.2.2 *Consider two random variables X and Y. $X \perp\!\!\!\perp Y$ if and only if for any functions $\varphi : \mathbb{R} \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}$ (such that the expected values below exist and are well-defined) $Cov[\varphi(\boldsymbol{X}), \psi(\boldsymbol{Y})] = 0$, i.e.*

$$\mathbb{E}[\varphi(X)\psi(Y)] = \mathbb{E}[\varphi(X)] \cdot \mathbb{E}[\psi(Y)].$$

Hirschfeld (1935), Gebelein (1941), Rényi (1959) and Sarmanov (1963) introduced the concept of "maximal correlation", defined as

$$r^\star(X, Y) = \max_{\varphi, \psi} \big\{ \text{Corr}[\varphi(\boldsymbol{X}), \psi(\boldsymbol{Y})] \big\},$$

(provided that expected values exist and are well-defined). And those authors proved that $X \perp Y$ if and only if $r^\star(X, Y) = 0$.

Definition 7.2.3 (Linear Independence) *In a general context, consider two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, in $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$, respectively. $\boldsymbol{X} \perp \boldsymbol{Y}$ if and only if for any $\boldsymbol{a} \in \mathbb{R}^{d_x}$ and $\boldsymbol{b} \in \mathbb{R}^{d_y}$*

$$Cov[\boldsymbol{a}^\top \boldsymbol{X}, \boldsymbol{b}^\top \boldsymbol{Y}] = 0.$$

Definition 7.2.4 (Independence) *In a general context, consider two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$ if and only if for any $\mathcal{A} \subset \mathbb{R}^{d_x}$ and $\mathcal{B} \subset \mathbb{R}^{d_y}$,*

$$\mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\} \cap \{\boldsymbol{Y} \in \mathcal{B}\}] = \mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\}] \cdot \mathbb{P}[\{\boldsymbol{Y} \in \mathcal{B}\}].$$

Propoition 7.2.3 *Consider two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$ if and only if for any functions $\varphi : \mathbb{R}^{d_x} \to \mathbb{R}$ and $\psi : \mathbb{R}^{d_y} \to \mathbb{R}$ (such that the expected values below exist and are well-defined)*

$$\mathbb{E}[\varphi(\boldsymbol{X})\psi(\boldsymbol{Y})] = \mathbb{E}[\varphi(\boldsymbol{X})] \cdot \mathbb{E}[\psi(\boldsymbol{Y})],$$

or equivalently

$$Cov[\varphi(\boldsymbol{X}), \psi(\boldsymbol{Y})] = 0.$$

This is the extension in higher dimension of Proposition 7.2.2.

Definition 7.2.5 (Mutual Independence) *Let $\boldsymbol{Y} = (Y_1, \cdots, Y_d)$ denote some random vector. All components of $\boldsymbol{Y}$ are (mutually) independent if for any $\mathcal{A}_1, \cdots, \mathcal{A}_d \subset \mathbb{R}$*

$$\mathbb{P}\left[\{(Y_1, \cdots, Y_d) \in \bigcap_{i=1}^{d} \mathcal{A}_i\}\right] = \prod_{i=1}^{d} \mathbb{P}[\{Y_i \in \mathcal{A}_i\}].$$

Definition 7.2.6 (Independent Version of Some Random Vector) *Let $\boldsymbol{Y} = (Y_1, \cdots, Y_d)$ denote some random vector. $\boldsymbol{Y}^\perp = (Y_1^\perp, \cdots, Y_d^\perp)$ is an independent version of $\boldsymbol{Y}$ if*

$$\begin{cases} (Y_1^\perp, \cdots, Y_d^\perp) \text{ are mutually independent random variables} \\ Y_i^\perp \overset{\mathcal{L}}{=} Y_i, \ \forall i = 1, \cdots, d. \end{cases}$$

In section 4.1, we have discussed the difference between *ceteris paribus* ('all other things being equal" or "other things held constant") and *mutatis mutandis* ("once the necessary changes have been made"). As discussed, *ceteris paribus* has to do with versions of some random vector with some independent components (we consider some explanatory variable as if they were independent of the other ones),

Definition 7.2.7 (Version of Some Random Vector with Independent Margin) *Let $\boldsymbol{Y} = (Y_1, \cdots, Y_d)$ denote some random vector. $(Y_1^\perp, Y_2, \cdots, Y_d)$ is a version of $\boldsymbol{Y}$ with independent first margin if*

$$\begin{cases} Y_1^\perp \perp\!\!\!\perp \boldsymbol{Y}_{-1} = (Y_2, \cdots, Y_d) \\ Y_1^\perp \overset{\mathcal{L}}{=} Y_1. \end{cases}$$

One can easily extend the previous concept for some subset of indices $J \subset \{1, \cdots, d\}$. All those concepts can be extended to the case of condition independence,

Definition 7.2.8 (Conditional Independence (dimension 2)) *X and Y are independent conditionally on Z, denoted $X \perp\!\!\!\perp Y \mid Z$, if for any sets $\mathcal{A}, \mathcal{B}, \mathcal{C} \subset \mathbb{R}$,*

$$\mathbb{P}\big[X \in \mathcal{A}, Y \in \mathcal{B} \big| Z \in \mathcal{C}\big] = \mathbb{P}\big[X \in \mathcal{A} \big| Z \in \mathcal{C}\big] \cdot \mathbb{P}\big[Y \in \mathcal{B} \big| Z \in \mathcal{C}\big].$$

Again, an alternative characterization is

$$\mathbb{P}\big[Y \in \mathcal{B} \big| X \in \mathcal{A}, Z \in \mathcal{C}\big] = \mathbb{P}\big[Y \in \mathcal{B} \big| Z \in \mathcal{C}\big].$$

Again, this notion can be extended in higher dimension,

Definition 7.2.9 (Conditional Independence) *In a general context, consider three random vectors $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$. $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y})|\boldsymbol{Z}$ if and only if for any $\mathcal{A} \subset \mathbb{R}^{d_x}$, $\mathcal{B} \subset \mathbb{R}^{d_y}$ and $\mathcal{C} \subset \mathbb{R}^{d_z}$,*

$$\mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\} \cap \{\boldsymbol{Y} \in \mathcal{B}\}|\boldsymbol{Z} \in \mathcal{C}] = \mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\}|\boldsymbol{Z} \in \mathcal{C}] \cdot \mathbb{P}[\{\boldsymbol{Y} \in \mathcal{B}\}|\boldsymbol{Z} \in \mathcal{C}]$$

for all $\boldsymbol{z} \in \mathbb{R}^{d_z}$.

See Dawid (1979) for various properties on conditional independence. All those concepts are well-known in actuarial science, but there still are several pitfalls. So let us recall some valid properties, and some

Propoition 7.2.4 *Consider three random variables X, Y, and Z. If $X \perp Z$ and $Y \perp Z$, then $aX + bY \perp Z$, for any $a, b \in \mathbb{R}$.*

By the linearity of the covariance,

$$\text{Cov}[aX+bY,Z] = a\underbrace{\text{Cov}[X,Z]}_{=0\ (X\perp Z)} + b\underbrace{\text{Cov}[Y,Z]}_{=0\ (Y\perp Z)} = 0.$$

Propoition 7.2.5 *Consider three random variables X, Y, and Z. If $X \perp Z$ and $Y \perp Z$, it does not imply that $\psi(X,Y) \perp Z$, for any $\psi : \mathbb{R}^2 \to \mathbb{R}$.*

$$(X,Y,Z) = \begin{cases} (0,0,0) & \text{with probability } 1/4, \\ (0,1,1) & \text{with probability } 1/4, \\ (1,0,1) & \text{with probability } 1/4, \\ (1,1,0) & \text{with probability } 1/4. \end{cases}$$

One can easily get that

$$\begin{cases} X,Y,Z \sim \mathcal{B}(1/2) \\ XY,YZ,XZ \sim \mathcal{B}(1/4) \\ XYZ = 0, \text{ a.s.} \end{cases}$$

Thus, on the one hand

$$\mathbb{E}[XZ] = \frac{1}{4} = \frac{1}{2}\cdot\frac{1}{2} = \underbrace{\mathbb{E}[X]}_{=1/2}\cdot\underbrace{\mathbb{E}[Z]}_{=1/2},$$

and therefore $\text{Cov}(X,Z) = 0$, and similarly for the pair (Y,Z), so $X \perp Z$ and $Y \perp Z$. On the other hand, $XY \not\perp Z$ since $\text{Cov}(XY,Z) \neq 0$, because

$$\mathbb{E}[XY\cdot Z] = 0 \neq \underbrace{\mathbb{E}[XY]}_{=1/4}\cdot\underbrace{\mathbb{E}[Z]}_{=1/2}.$$

Propoition 7.2.6 *Consider a random vector $\boldsymbol{X}$ in $\mathbb{R}^k$, and a random variable Z.*
$\boldsymbol{X} \perp Z$ does not imply that $\psi(\boldsymbol{X}) \perp Z$, for any $\psi : \mathbb{R}^k \to \mathbb{R}$.

In the context of fairness, if we get even if $\boldsymbol{X}^\perp$ orthogonal to S (by construction), we can still have $\widehat{Y} = m(\boldsymbol{X}^\perp) \not\perp S$.

Propoition 7.2.7 *Consider three random variables X, Y, and Z. Even if $X \perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z$, it does not imply either that $\psi(X,Y) \perp Z$ or that $\psi(X,Y) \perp\!\!\!\perp Z$, for any $\psi : \mathbb{R}^2 \to \mathbb{R}$.*

A counter-example can be obtained using variables that are pairwise independent, but not mutually independent. The example of the previous proof works (with $\psi(x,y) = xy$). Pairs are more than non-correlated, they are pairwise independent, so $X \perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z$. But

$$(XY,Z) = \begin{cases} (0,0) & \text{with probability } 1/4, \\ (0,1) & \text{with probability } 1/2, \\ (1,0) & \text{with probability } 1/4, \end{cases}$$

and therefore, $XY \not\perp\!\!\!\perp Z$ since

$$\mathbb{P}\big[XY=1, Z=0\big] = \frac{1}{4} \neq \frac{1}{4}\cdot\frac{1}{2} = \mathbb{P}\big[XY=1\big]\cdot\mathbb{P}\big[Z=0\big].$$

Propoition 7.2.8 *Consider a random vector* $\boldsymbol{X}$ *in* $\mathbb{R}^k$*, and a random variable* Z*.*
$\boldsymbol{X} \perp\!\!\!\perp Z$ *does not imply either that* $\psi(\boldsymbol{X}) \perp\!\!\!\perp Z$ **or** $\psi(\boldsymbol{X}) \perp\!\!\!\perp Z$*, for any* $\psi : \mathbb{R}^k \to \mathbb{R}$*.*

In the context of fairness, if we were able to insure that $\boldsymbol{X}^{\perp} \perp\!\!\!\perp S$, we can still have $\widehat{Y} = m(\boldsymbol{X}^{\perp}) \not\perp\!\!\!\perp S$ (even $\widehat{Y} = m(\boldsymbol{X}^{\perp}) \not\perp\!\!\!\perp S$).

7.2.2 Dependence with Graphs

If we have introduced various concepts related to independence, we do not have any proper definition of a "*causal effect*." Spirtes et al. (1993) noted that it would be difficult to "define" what causality is, and that it would probably be simpler to "axiomatize" it. In other words, we will start by determining the attributes that would be considered necessary for a relation to be considered as "causal," we formalize these properties mathematically, and we look to see if these axioms translate into an interpretable characterization of a causal relation.

For example, it seems legitimate that these relations are transitive: if x_1 causes x_2 and if x_2 causes x_3, then it must also be true that x_1 causes x_3. We could then talk about global causality. But a local version seems to exist: if x_1 causes x_3 only through x_2, then it is possible to block the influence of x_1 on x_3 if we prevent x_2 from being influenced by x_1. One could also ask that the causal relation be irreflexive, in the sense that x_1 cannot cause itself. The danger of this property is that it tends to seek a causal explanation for any variable. Finally, an asymmetry property of the relation is often desired, in the sense that x_1 causes x_2 implies that x_2 cannot cause x_1. Here again, a precision is necessary, especially if the variables are dynamic: this property does not prevent a lagged feedback effect: It is indeed possible for $x_{1,t}$ to cause $x_{2,t+1}$ and for $x_{2,t+1}$ to cause $x_{1,t+2}$, but not for $x_{1,t}$ to cause $x_{2,t}$ and for $x_{2,t}$ to cause $x_{1,t}$, with this approach. As noted by Wright (1921), well before Pearl (1988), the most natural tool to describe visually and simply these causal relations is probably that of directed graphs.

A variable will be here a node of the network (for example x_1 or x_2), and a causal relation, in the sense "x_1 causes x_2" will be translated by an arrow directed from x_1 to x_2 (as we did on time series).

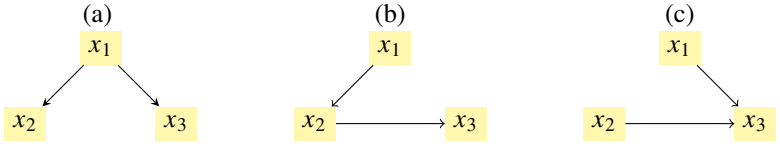


Figure 7.3: Some examples of directed (acyclical) graphs, with 3 nodes, and 2 connections. (a) corresponds to the case where x_1 is a confounder for x_2 and x_3, corresponding to a common shock or mutual dependence (also called "fork"), (b) corresponds to the case where x_2 is a mediator for x_1 and x_3 (also called "chain"), and (c) corresponds to the case where x_3 is a collider for x_1 and x_2, corresponding to a mutual causation case

.

Definition 7.2.10 (Confounder) *A variable is a confounder when it influences both the dependent variable and independent variable, causing a spurious association.*

The existence of confounders is an important quantitative explanation why "correlation does not imply causation". For example, on Figure 7.3 (a), x_1 is a confounder for x_2 and x_3. The term "fork" is also used.

Definition 7.2.11 (Collider) *A variable is a collider when it is causally influenced by two or more variables.*

For example, on Figure 7.3 (c), x_3 is a collider for x_1 and x_2. The term “inverted fork" is also used.

Definition 7.2.12 (Mediator) *A mediation model proposes that the independent variable influences the mediator variable, which in turn influences the dependent variable.*

For example, on Figure 7.3 (b), x_2 is a mediator for x_1 (the independent variable) and x_3 (the dependent variable). The term “chain" is also used.

In example (b), we will say that x_2 and x_3 are causally dependent on x_1, x_2 will be directly and x_3 indirectly. We will say that x_1 is a causal parent of x_2 (a cause), and conversely that x_2 is a causal child of x_1 (a consequence). This parent/child relationship is associated with the existence of a link between the two variables. We will say that x_1 is a causal ancestor of x_3, and conversely that x_3 is a causal descendant of x_1. This ancestor / descendant relation is associated with the existence of a (directed) path between the two variables, i.e. a succession of links. If there is no path between two nodes, we say that the two variables are causally independent. And if there is a directed path from node x_1 to node x_2, then x_1 causally affects x_2: if ${}_x1$ had been different, x_2 would also have been different too. Therefore, causality has a direction. A collider is a variable which is the consequence of two or more variables, like x_3 on (c). This type of variable is very much related to Berkson’s paradox, discussed earlier. A non-collider is a variable influenced by only one, and it allows a consequence to be causally transmitted along a path. The causal variables that influence the collider are themselves not necessarily associated. If this is the case, the collider is said to be shielded and the variable is the vertex of a triangle. For (a) $x_2 \not\perp\!\!\!\perp x_3$ while $x_2 \perp\!\!\!\perp x_3 \mid x_1$, for (b) $x_1 \not\perp\!\!\!\perp x_3$ while $x_1 \perp\!\!\!\perp x_3 \mid x_2$, and for (c) $X_1 \perp\!\!\!\perp X_2$.

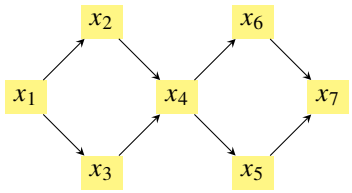


Figure 7.4: An example of directed graph, with 7 nodes, $x_1, \cdots, x_7$, and 8 edges, $\{1 \to 2\}, \cdots, \{6 \to 7\}$.

In Figure 7.4, x_4 is a “descendant" of x_1, a child of x_2 (and x_4), a parent of x_5 (and x_6) and an “ancestor" of x_7. The variables x_3 and x_5 are not causally independent. x_4 is a collider, but x_6 is not. x_4 is an unshielded collider because x_2 and x_3 (the two parents) are not connected (but they are not independent)

Consider a causal graph $\mathcal{G}$ (i.e. a collection of variables $\boldsymbol{x}$ and arrows). Let x_1 and x_2 be two nodes, and $\boldsymbol{y}$ be a set of nodes of $\mathcal{G}$ containing neither x_1 nor x_2. We will say that x_1 and x_2 are separated (or d-separated, “*directed separation*") if there is no undirected path π between x_1 and x_2 such that (a) every collider (in π) is either in $\boldsymbol{y}$, or has a descendant in $\boldsymbol{y}$ and (b) no other node on π is in $\boldsymbol{y}$. We will then note $x_1 \perp\!\!\!\perp x_2 | \boldsymbol{y}$. We have here necessary and sufficient conditions for two vertices of a causal graph to be independent from a probabilistic point of view, after conditioning on another set of vertices. For it is possible to translate this causal terminology, presented here in terms of graphs, into a probabilistic form. If Kiiveri and Speed (1982) introduced most of the concepts, Koller and Friedman (2009) and Peters et al. (2017) provided recent overviews. Let us formalize here the concepts described above.

Definition 7.2.13 (Path) *A path π from a node x_i to another node x_j is a sequence of nodes and edges starting at x_i and ending at x_j.*

On the causal graph of Figure 7.4, $\pi = \{x_1 \to x_2 \to x_4 \to x_5\}$ is a path from node x_1 to x_5. To go further, a conditioning set $\boldsymbol{x}_c$ is simply a collection of nodes.

Definition 7.2.14 (Blocking path) *A path π from a node x_i to another node x_j is blocked by $\boldsymbol{x}_c$ whenever there is a node x_k such that either $x_k \in \boldsymbol{x}_c$ and*

$$\{x_{k^-} \to x_k \to x_{k^+}\} \text{ or } \{x_{k^-} \leftarrow x_k \leftarrow x_{k^+}\} \text{ or } \{x_{k^-} \leftarrow x_k \to x_{k^+}\}$$

or $x_k \notin \boldsymbol{x}_c$, as well as all descents of x_k, and $\{x_{k^-} \to x_k \leftarrow x_{k^+}\}$ In that case, write $x_i \perp_{\mathcal{G}-\pi} x_j \mid \boldsymbol{x}_c$.

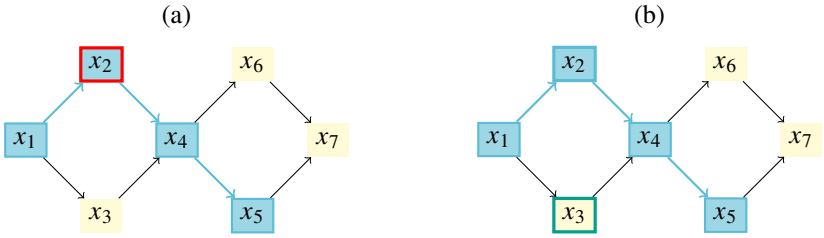


Figure 7.5: On the directed graph of Figure 7.4, examples of blocking path. Path $\pi = \{x_1 \to x_2 \to x_4 \to x_5\}$ is blocked by x_2 (on the left, (a)), and not blocked by x_3 (on the right, (b))

On the causal graph of Figure 7.5, $\pi = \{x_1 \to x_2 \to x_4 \to x_5\}$ is blocked by x_2 (on the left, (a)), and not blocked by x_3 (on the right, (b)).

Definition 7.2.15 (d-separation (nodes)) *A node x_i is said to be d-separated with another node x_j by $\boldsymbol{x}_c$ whenever every path from x_i to x_j is blocked by $\boldsymbol{x}_c$. We will simply denote $x_i \perp_{\mathcal{G}} x_j \mid \boldsymbol{x}_c$.*

On the causal graph of Figure 7.6, nodes x_1 and x_5 are d-separated by x_4 (a), by (x_2, x_3) (b), by (x_2, x_3, x_6), and not blocked by x_3 (c) and (x_3, x_6) (d).

Definition 7.2.16 (d-separation (sets)) *A set of nodes $\boldsymbol{x}_i$ is said to be d-separated with another set of nodes $\boldsymbol{x}_j$ by $\boldsymbol{x}_c$ whenever every path from any $x_i \in \boldsymbol{x}_i$ to any $x_j \in \boldsymbol{x}_j$ is blocked by $\boldsymbol{x}_c$. We will simply denote $\boldsymbol{x}_i \perp_{\mathcal{G}} \boldsymbol{x}_j \mid \boldsymbol{x}_c$.*

Propoition 7.2.9 *Two nodes x_i and x_j are d-separated by $\boldsymbol{x}_c$ if and only members of $\boldsymbol{x}_c$ block all paths from x_i to x_j.*

Propoition 7.2.10 *Two nodes x_i and x_j are d-separated by $\boldsymbol{x}_c$ if and only members of $\boldsymbol{x}_c$ block all paths from x_i to x_j.*

Definition 7.2.17 (Markov Property) *Given a causal graph $\mathcal{G}$ with nodes $\boldsymbol{x}$, the joint distribution of $\boldsymbol{X}$ satisfies the (global) Markov property with respect to $\mathcal{G}$ if, for any disjoints $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_c$*

$$\boldsymbol{x}_1 \perp_{\mathcal{G}} \boldsymbol{x}_2 \mid \boldsymbol{x}_c \implies \boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2 \mid \boldsymbol{X}_c.$$

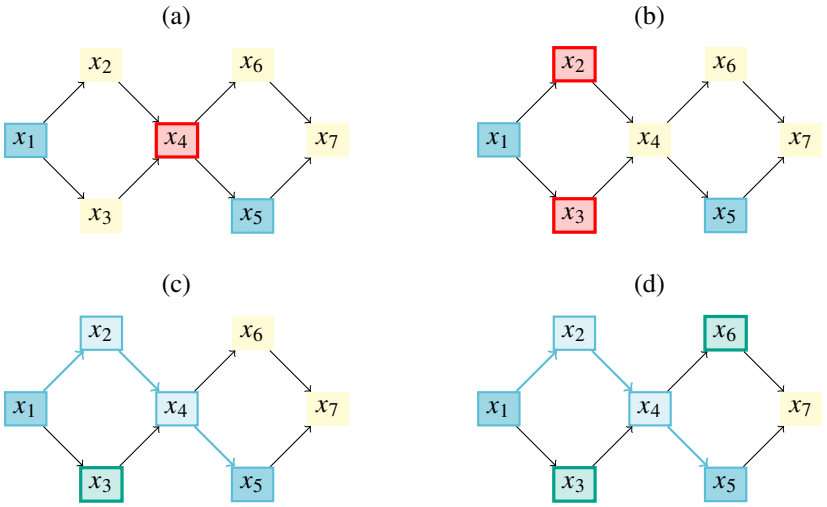


Figure 7.6: On the directed graph of Figure 7.4, examples of d-separation. Nodes x_1 and x_5 are d-separated by x_4 (top left (a)), by (x_2, x_3) (top right (a)), by (x_2, x_3, x_6), and not blocked by x_3 (bottom left (c)) and (x_3, x_6) (bottom right (d)).

The probability chain rule allows us to calculate the probability of an intersection of events using conditional probabilities, since (if $\mathbb{P}[\mathcal{B}|\mathcal{A}]$ is denoted $\mathbb{P}_{\mathcal{A}}[\mathcal{B}]$)

$$\mathbb{P}[\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_n] = \mathbb{P}[\mathcal{A}_1] \times \mathbb{P}_{\mathcal{A}_1}(\mathcal{A}_2) \times \mathbb{P}_{\mathcal{A}_1 \cap \mathcal{A}_2}[\mathcal{A}_3] \times \cdots \times \mathbb{P}_{\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_{n-1}}[\mathcal{A}_n],$$

which we could write

$$\mathbb{P}[x_1, \cdots, x_n] = \mathbb{P}[x_1] \times \mathbb{P}[x_2|x_1] \times \mathbb{P}[x_3|x_1, x_2] \times \cdots \times \mathbb{P}[x_n|x_1, \cdots, x_{n-1}].$$

But this writing is not unique, since we could also write (for example)

$$\mathbb{P}[x_1, \cdots, x_n] = \mathbb{P}[x_n] \times \mathbb{P}[x_{n-1}|x_n] \times \mathbb{P}[x_{n-2}|x_n, x_{n-1}] \times \cdots \times \mathbb{P}[x_1|x_n, \cdots, x_2].$$

since $\mathbb{P}[x_n, x_{n-1}] = \mathbb{P}[x_n] \times \mathbb{P}[x_{n-1}|x_n]$ as well as $\mathbb{P}[x_{n-1}] \times \mathbb{P}[x_n|x_{n-1}]$.

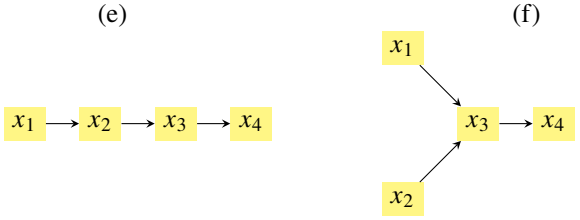


Figure 7.7: Other examples of directed causal graphs.

The idea here will be to write conditional probabilities involving only the variables and their causal parents. For example, the graph (e) in Figure 7.7 would correspond to

$$\mathbb{P}[x_1, x_2, x_3, x_4] = \mathbb{P}[x_1] \times \mathbb{P}[x_2|x_1] \times \mathbb{P}[x_3|x_1, x_2] \times \mathbb{P}[x_4|x_1, x_2, x_3].$$

whereas graph (f) of Figure 7.7 would be associated with the writing

$$\mathbb{P}[x_1, x_2, x_3, x_4] = \mathbb{P}[x_1, x_2] \times \mathbb{P}[x_3|x_1, x_2] \times \mathbb{P}[x_4|x_1, x_2, x_3].$$

which can also be written as

$$\mathbb{P}[x_1, x_2, x_3, x_4] = \mathbb{P}[x_1] \times \mathbb{P}[x_2] \times \mathbb{P}[x_3|x_1, x_2] \times \mathbb{P}[x_4|x_1, x_2, x_3].$$

because x_1 and x_2 are assumed to be independent. It is not uncommon to add a Markovian assumption, corresponding to the case where each variable is independent of all its ancestors conditional on its parents. For example, on the graph (e) of Figure 7.7, the Markov hypothesis allows to write

$$\mathbb{P}[x_3|x_1, x_2] = \mathbb{P}[x_3|x_2]$$

Also, the graph (e) of Figure 7.7 would correspond to

$$\mathbb{P}[x_1, x_2, x_3, x_4] = \mathbb{P}[x_1] \times \mathbb{P}[x_2|x_1] \times \mathbb{P}[x_3|x_2] \times \mathbb{P}[x_4|x_3].$$

whereas graph (f) of Figure 7.7 would be associated with the writing

$$\mathbb{P}[x_1, x_2, x_3, x_4] = \mathbb{P}[x_1] \times \mathbb{P}[x_2] \times \mathbb{P}[x_3|x_1, x_2] \times \mathbb{P}[x_4|x_3].$$

To go further in the examples, on the diagram (a) of Figure 7.3

$$\mathbb{P}[x_1, x_2, x_3] = \mathbb{P}[x_2|x_1] \cdot \mathbb{P}[x_3|x_1]$$

such that

$$\mathbb{P}[x_2, x_3|x_1] = \frac{\mathbb{P}[x_1, x_2, x_3]}{\mathbb{P}[x_1]} = \mathbb{P}[x_2|x_1] \cdot \mathbb{P}[x_3|x_1]$$

and therefore $x_2 \perp\!\!\!\perp x_3$ conditionally to x_1. In the diagram (b) of Figure 7.3

$$\mathbb{P}[x_1, x_2, x_3] = \mathbb{P}[x_1]\mathbb{P}[x_2|x_1] \cdot \mathbb{P}[x_3|x_2]$$

such that

$$\mathbb{P}[x_1, x_3|x_2] = \frac{\mathbb{P}[x_1, x_2, x_3]}{\mathbb{P}[x_2]} = \frac{\mathbb{P}[x_1]\mathbb{P}[x_2|x_1]}{\mathbb{P}[x_2]}\mathbb{P}[x_3|x_2] = \mathbb{P}[x_1|x_2] \cdot \mathbb{P}[x_3|x_2]$$

and therefore $x_1 \perp\!\!\!\perp x_3$ conditionally to x_2.

In the diagram (c) of Figure 7.3

$$\mathbb{P}[x_1, x_2, x_3] = \mathbb{P}[x_1] \cdot \mathbb{P}[x_2] \cdot \mathbb{P}[x_3|x_1, x_2]$$

such that

$$\mathbb{P}[x_1, x_2] = \mathbb{P}[x_1] \cdot \mathbb{P}[x_2]$$

in other words $x_1 \perp\!\!\!\perp x_2$, but

$$\mathbb{P}[x_1, x_2|x_3] = \frac{\mathbb{P}[x_1, x_2, x_3]}{\mathbb{P}[x_3]} = \frac{\mathbb{P}[x_1] \cdot \mathbb{P}[x_2] \cdot \mathbb{P}[x_3|x_1, x_2]}{\mathbb{P}[x_3]}$$

and therefore x_1 and x_2 are not independent, conditional on x_3.

7.3 Rung 2, Intervention (Doing, "*what if I do...*")

7.3.1 The do() Operator and Computing Causal Effects

To formalize the concept of "intervention", we will simply note that $\mathbb{P}[Y \in \mathcal{A}|X = x]$ describes how $Y \in \mathcal{A}$ is likely to occur if X happened to be equal to x. Therefore, it is an observational statement. We will denote $P[Y \in \mathcal{A}|\text{do}(X = x)]$ to describe how $Y \in \mathcal{A}$ is likely to occur if X is set to x (to avoid confusion, we will use P and not $\mathbb{P}$, in this introduction). It is here an intervention statement. Using causal graphs, the intervention $\text{do}(X = x)$ means that all incoming edges to x are cut. Hence, $P[Y \in \mathcal{A}|\text{do}(X = x)]$ can be seen as a $\mathbb{Q}[Y \in \mathcal{A}|X = x]$ where the causal graph has been manipulated. On the two graphs on the left of Figure 7.8, (a) and (b), $P[Y \in \mathcal{A}|\text{do}(X = x)] = \mathbb{P}[Y \in \mathcal{A}|X = x]$. On the two graphs on the right of Figure 7.8, (c) and (d),

$$P[Y \in \mathcal{A}|\text{do}(X = x)] = \mathbb{Q}[Y \in \mathcal{A}|X = x] = \sum_z \mathbb{Q}[Y \in \mathcal{A}, Z = z|X = x]$$

by the law of total probability. Using Bayes' rule

$$P[Y \in \mathcal{A}|\text{do}(X = x)] = \sum_z \mathbb{Q}[Y \in \mathcal{A}|X = x, Z = z] \cdot \mathbb{Q}[Z = z]$$

and since $\mathbb{Q}$ is the probability on the cut graph,

$$P[Y \in \mathcal{A}|\text{do}(X = x)] = \sum_z \mathbb{P}[Y \in \mathcal{A}|X = x, Z = z] \cdot \mathbb{P}[Z = z] \neq \mathbb{P}[Y \in \mathcal{A}|X = x].$$

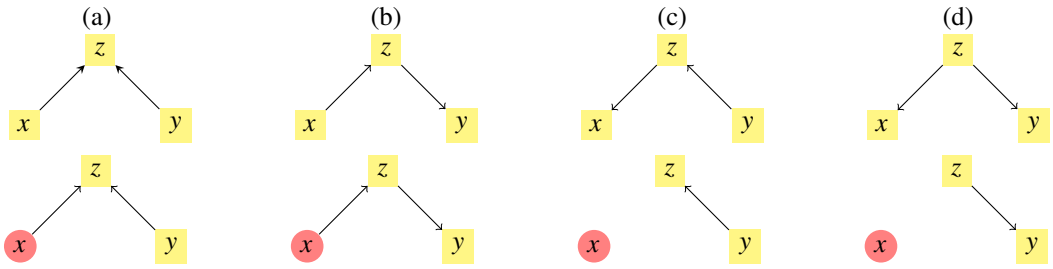


Figure 7.8: Illustration of the do operator, with two forks, (a) and (d), z being a collider on (a) and a cofounder on (d), and two chains, (b) and (c). On top, the causal graphs, and below, the implied graphs when an intervention on x, corresponding to "$\text{do}(x)$" is considered.

If $P[Y \in \mathcal{A}|\text{do}(X = x)] \neq \mathbb{P}[Y \in \mathcal{A}|X = x]$, it means that X and Y are confounded. And now that we have a better understanding, as in most textbooks, let us denote $P[Y \in \mathcal{A}|\text{do}(X = x)]$ this expression (having in mind that it is not equal to $\mathbb{P}[Y \in \mathcal{A}|X = x]$). To get a better understanding this idea of interventions via this do() operator, let us introduce "*structural causal models*."

7.3.2 Structural Causal Models

For simplicity here, we will present here on the linear Gaussian Stuctural Causal Models, associated with an acyclic causal graph. The definition is very close to the "algorithmic" definition of the Markov property, well

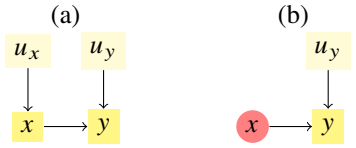


Figure 7.9: Causal diagram of a Structural Causal Model, with an intervention on x, on the right.

know in the context of homogeneous Markov chains (see Rolski et al. (2009)): if (X_t) is a Markov chain, then

$$(X_t|X_{t-1} = x_{t-1}, X_{t-2} = X_{2-2}, \cdots, X_1 = X_1, X_0 = x_0) \overset{\mathcal{L}}{=} (X_t|X_{t-1} = x_{t-1}),$$

and equivalently, there are independent variable (U_t) and a measurable function h such that $X_t = h(X_{t-1}, U_t)$. In the case of causal graph, quite naturally, if C is a cause, and E the effect, we should expect to have $E = h(C, U)$ for some measurable function h and some random noise U. This is the idea of structural models,

Definition 7.3.1 (Structural Causal Models (SCM)) *Pearl (2009b) In a simple causal graph, with two nodes C (the cause) and E (the effect), the causal graph is $C \longrightarrow E$, and the mathematical interpretation can be summarized in two assignments*

$$\begin{cases} C = h_c(U_C) \\ E = h_e(C, U_E) \end{cases}$$

where U_C and U_E are two independent random variables, $U_C \perp\!\!\!\perp U_E$.

More generally, a structural causal model is a triplet $(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{h})$, as in Pearl (2010) or Halpern (2016). The variables in $\boldsymbol{U}$ are called exogenous variables, in other words, they are external to the model (we do not have to explain how they are caused). The variables in $\boldsymbol{V}$ are called endogenous. Each endogenous variable is a descendant of at least one exogenous variable. Exogenous variables cannot be descendants of any other variable and, in particular, cannot be descendants of an endogenous variable. Also, they have no ancestors and are represented as roots in causal graphs. Finally, if we know the value of each exogenous variable, we can, using $\boldsymbol{h}$ functions, determine with perfect certainty the value of each endogenous variable. The causal graphs we have described consist of a set of n nodes representing the variables in $\boldsymbol{U}$ and $\boldsymbol{V}$, and a set of edges between the n nodes representing the functions in $\boldsymbol{h}$. Observe that we consider acyclical graphs, not only for a mathematical reason (to insurance that the model is solvable) but also for interpretation: a cycle between x, y and z would mean that x causes y, y causes z and z causes x. In a static setting (as the one we consider here), that is not possible.

In the causal diagram (a) in Figure 7.9, we have two endogenous variables, x and y, and two exogenous variables, u_x and u_y. The diagram (a) is a representation of the real world, but we assume here that it is possible to make interventions, and to change the value of x, assuming that all things remain equal. We will use here the notation $Y^\star$ to write the "potential" outcome if an intervention was to be consired,

$$\begin{array}{cc} \text{real world} & \text{with intervention (do}(x)) \\ \begin{cases} X = h_x(U_x) \\ Y = h_y(X, U_y) \end{cases} & \begin{cases} X = x \\ Y_x^\star = h_y(x, U_y) \end{cases} \end{array}$$

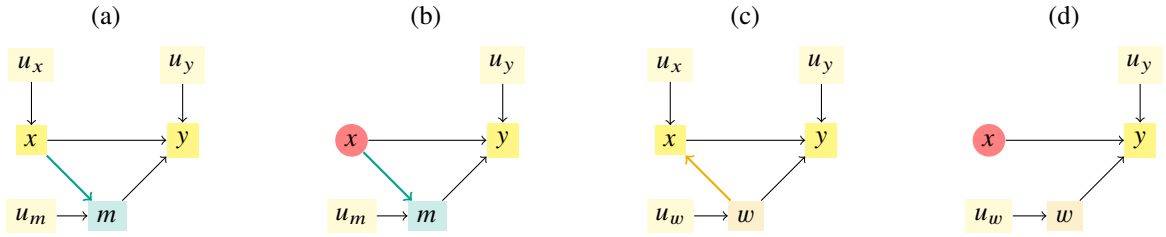


Figure 7.10: Two causal diagrams, $x \longrightarrow y$, with a mediator m on the left ((a) and (b)) with and an intervention on x and with a confounding factor w on the right ((c) and (d)) with and an intervention on x.

In the causal diagram (a) in Figure 7.10, we have three endogenous variables, x, y, and a mediator m, and three exogenous variables, u_x, u_y and u_m. The diagram (a) is a representation of the real world, but as before, it is assumed here that it is possible to make interventions on X.

In other words,

$$\text{in the presence of a mediator } (m)$$

$$\begin{array}{ll} \text{real world} & \text{with intervention } (\text{do}(x)) \\ \begin{cases} X = h_x(U_x) \\ M = h_m(X, U_m) \\ Y = h_y(X, M, U_y) \end{cases} & \begin{cases} X = x \\ M_x = h_m(x, U_m) \\ Y_x^\star = h_y(x, M_x, U_y) \end{cases} \end{array}$$

In the causal diagram (b) of Figure 7.10, we have three endogenous variables, x, y, and a confounding factor w, and three exogenous variables, u_x, u_y and u_w. Diagram (c) is a representation of the real world, but as before, it is assumed here that it is possible to make interventions on X. In other words,

$$\text{in the presence of a confusion factor } (w)$$

$$\begin{array}{ll} \text{real world} & \text{with intervention } (\text{do}(x)) \\ \begin{cases} X = h_x(W, U_x) \\ W = h_w(U_w) \\ Y = h_y(X, W, U_y) \end{cases} & \begin{cases} X = x \\ W = h_w(U_w) \\ Y_x^\star = h_y(x, W, U_y) \end{cases} \end{array}$$

$$\begin{cases} \text{mediator}: & \mathbb{P}[Y_x^\star = 1] = \mathbb{P}[Y = 1|\text{do}(X = x)] = \mathbb{P}[Y = 1|X = x] \\ \text{confusion}: & \mathbb{P}[Y_x^\star = 1] = \mathbb{P}[Y = 1|\text{do}(X = x)] \neq \mathbb{P}[Y = 1|X = x] \end{cases}$$

In fact, in the presence of a confounding factor, $\mathbb{P}[Y_x^\star = 1]$ which corresponds to $\mathbb{P}[Y = 1|\text{do}(X = x)]$ should be written

$$\sum_w \mathbb{P}[Y = 1|W = w, X = x] \cdot \mathbb{P}[W = w] = \mathbb{E}(\mathbb{P}[Y = 1|W, X = x]).$$

For example, we can suppose that $\mathbb{P}[Y = 1|W = w, X = x]$ is obtained using a logistic model: if $\mu_x(w) = \mathbb{P}[Y = 1|W = w, X = x]$

$$\widehat{\mu}_x(w) = \frac{\exp[\widehat{\beta}_0 + \widehat{\beta}_x x + \widehat{\beta}_w w]}{1 + \exp[\widehat{\beta}_0 + \widehat{\beta}_x x + \widehat{\beta}_w w]}$$

the average causal effect, ACE = $\mathbb{E}(\mu_1(W) - \mu_0(W))$ will be estimated by

$$\widehat{\text{ACE}} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}_1(w_i) - \widehat{\mu}_0(w_i))$$

As explained in Pearl (1998), the structural equation $y = h_y(x, u_y)$ represents a causal mechanism that specifies the value taken by the variable y in response to each pair of values taken by the variable x and the (exogeneous) factor u_y. $y = f_y(x, u_y)$ is computed under the formal intervention "*X is set to x*", which Pearl (1998) notes "do$(X = x)$" (or simply do(x)), then assigned to y (historically, from Wright (1921) to Holland (1986), passing by Neyman et al. (1923) or Rubin (1974), various notations have been proposed). To use the notation and interpretation of Pearl (2010), Y would be $y^{\star}_{x}$ had X been x in situation u_y will be written $Y^{\star}_{x}(u_y) = y$, the structural error u_y not being impacted by an intervention on x.

In probabilistic terminology, $\mathbb{P}[Y = y|X = x]$ denotes the population distribution of Y among individuals whose X value is x. Here, $\mathbb{P}[Y = y|\text{do}(X = x)]$ represents the population distribution of Y if all individuals in the population had their X value set to x. And more generally, $\mathbb{P}(Y = y|\text{do}(X = x), Z = z)$ will denote the conditional probability that $Y = y$, given $Z = z$, in the distribution created by the intervention do$(X = x)$. Also, in the literature, the average causal effect (ACE) corresponds to $\mathbb{E}[Y|\text{do}(X = 1)] - \mathbb{E}[Y|\text{do}(X = 0)]$, or $\widehat{Y}_1 - \widehat{Y}_0$ if $\widehat{Y}_x = \mathbb{E}[Y|\text{do}(X = x)]$ (which we will also note hereafter $Y^{\star}_{X\leftarrow 1} - Y^{\star}_{X\leftarrow 0}$, as in Russell et al. (2017)). To calculate this quantity, given a causal graph,

$$\mathbb{P}[Y = y|\text{do}(X = x)] = \sum_{z}\mathbb{P}[Y = y|X = x, PA = z] \cdot \mathbb{P}[PA = z]$$

where PA denotes the parents of x, and z covers all combinations of values that the variables in PA can take. A sufficient condition for identifying the causal effect $\mathbb{P}(y|\text{do}(x))$ is that each path between X and one of its children traces at least one arrow emanating from the measured variable, as in Tian and Pearl (2002).

To illustrate this difference between the intervention (via the do operator) and conditioning, consider the causal graph (c) of Figure 7.10, discussed in de Lara (2023), based on the following Structural Causal Model, with additive functions,

$$\begin{cases} W = h_w(u_w) = u_w \\ X = h_x(w, u_x) = w + u_x \\ Y = h_y(x, w, u_y) = x + w + u_y \end{cases}$$

A solution (X, W, Y) is such that

$$\begin{cases} X = W + u_x \\ W = u_w \\ Y = X + W + u_y \end{cases} \quad \text{or} \quad \begin{cases} X = u_w + u_x \\ W = u_w \\ Y = 2u_w + u_x + u_y \end{cases}$$

As mentioned in Bongers et al. (2021), Structural Causal Models are not always solvable, this is why the "acyclicity" assumption is important, since it insures unique solvability. If we consider now a "do intervention", with do$(x = 0)$, we now have

$$\begin{cases} X = 0 \\ W^{\star}_{x\leftarrow 0} = u_w \\ Y^{\star}_{x\leftarrow 0} = x + w + u_y \end{cases}, \text{ or } \begin{cases} X = 0 \\ W^{\star}_{x\leftarrow 0} = u_w \\ Y^{\star}_{x\leftarrow 0} = u_w + u_y \end{cases}$$

Thus, on the one hand, observe that $W|X=0$ as the same distribution as U_w conditional on $U_x+U_w=0$, i.e. $W|X=0$ as the same distribution as $-U_x$. On the other hand, the distribution of $W^\star$ is U_w. Thus, generally, $(W|X=0) \overset{\mathcal{L}}{\neq} W^\star_{x\leftarrow 0}$ which corresponds to $(W|\text{do}(x=0))$.

Before concluding, and reaching rung 3 in the "*ladder of causation,*" observe that it is possible to use causal graphs to defined more precisely "bias" and "disparity" (or "discrimination").

Definition 7.3.2 (Bias) *Traag and Waltman (2022). A bias is a direct causal effect of s on y that is considered unjustified (from a moral or ethical sense).*

As stressed in Traag and Waltman (2022), the fact that a direct causal effect is justified, or not is "*an ethical question that cannot be determined empirically using data*."

Definition 7.3.3 (Disparity) *Traag and Waltman (2022). A disparity as a causal effect of s on y that includes a bias.*

Therefore, there is a disparity of s on y if at least one causal causal from s on y represents a bias. Each bias is a disparity, but a disparity does not need to be a bias. And a disparity is seen as unfair. It is possible to discuss redlining, mentioned in the early parts of Chapter 1, and in Section 6.1.2 with respect to those definitions. In its strict (historical) definition, the practice of redlining in the United States, corresponds to the case where organisations, such as financial institutions and insurers, deny people services based on the area in which they live (no loans, or no insurance). And as we've been, due to geographic racial segregation, intentionally or not, this yields to not serving people of a certain races. Here, there could be multiple "biases". The use of ZIP codes, to determine whom to insure or not, may be considered "unjustified", in which case there is a location bias, and a racial disparity, when insuring. But it is also possible that there is a geographic racial bias, where people of a certain race could be denied access to certain neighbourhoods. This geographic racial bias consequently produces a racial disparity in insuring, even if using ZIP codes was no longer "unjustified". If insurers use race to determine whom to insure, there is a racial bias in insuring, and not only a racial disparity. Therefore, even if there is no racial bias in insuring, it would not imply that there is no problem. A racial disparity in insurance indicates that the outcome is unfair with respect to race, as discussed in Traag and Waltman (2022).

7.4 Rung 3, Counterfactuals (Imagining, "*what if I had done...*")

In the previous section, we considered Pearl's causal modeling approach, based on Pearl (2009a). According to Holland (1986), there is "*no causation without manipulation*," and therefore, there should be only two rungs on the "*ladder of causation*" (and gender or race should not be understood as having a causal effect, as discussed in Holland (2003)). Nevertheless, it is possible to go one step further, with the "*potential-outcome framework*," also known as Neyman-Rubin causal modeling, from Rubin (1974), based on the concept of "counterfactuals". Rather than considering whether some variables (such as the gender, or the race) are manipulable, we will consider here hypothetical possibilities (see Kohler-Hausmann (2018) for further discussions on counterfactual causal thinking in the context of racial discrimination).

7.4.1 Counterfactuals

Pearl and Mackenzie (2018) noted that causal inference was intended to answer the question "*what would have happened if...*?" This question is central in epidemiology (what would have happened if this person

had received the treatment?) or as soon as we try to evaluate the impact of a public policy (what would have happened if we had not removed this tax?). But we note that this is the question we ask as soon as we talk about discrimination (what would have happened if this woman had been a man?). In causal inference, in order to quantify the effect of a drug or a public policy measure, two groups are constituted, one that will receive the treatment and another that will not, and will therefore serve as a counterfactual, in order to answer the question "*what would have happened if the same person had had access to the treatment?*" When analysing discrimination, similar questions are asked, for example, "*would the price of risk be different if the same person had been a man and not a woman?*" except that here gender is not a matter of choice, of an arbitrary assignment to a treatment (random in so-called randomized experiments). In fact, this parallel between discrimination analysis and causal inference was initially criticised: changing treatment is possible, whereas changing sex is a figment of the imagination. One can also think of the questions regarding the links between smoking and certain cancers: seeing smoking as a "treatment" may make sense mathematically, but ethically, one could not force someone to smoke just to quantify the probability of getting cancer a few years later[1] (whereas in a clinical experiment, one could imagine that a patient is given the blue pills, instead of the red pills). We enter here in the category of the so-called "quasi-experimental" approaches, in the sense of Cook et al. (2002) and DiNardo (2016).

In the data, y (often called "outcome") is the variable that we seek to model and predict, and which will serve as a measure of treatment effectiveness. The potential outcomes are the outcomes that would be observed under each possible each possible treatment, and we note $y_t^\star$ the outcome that would be observed if the treatment T had taken the value t. And the counterfactual outcomes are what would have been observed if the treatment had been different, in other words, for a person of type t, his counterfactual outcome is $y_{1-t}^\star$ (because t takes the values $\{0, 1\}$). The typical example is that of a person who received a vaccine ($t = 1$), who did not get sick ($y = 0$), whose counterfactual outcome would be $y_0^\star$, sometimes noted $y_{t\leftarrow 0}^\star$. Before launching the vaccine efficacy study, the two outcomes are potential, $y_0^\star$ and $y_1^\star$. Once the study is launched, the observed outcome will be y and the counterfactual outcome will be $y_{1-t}^\star$. Note that different notations are used in the literature, $y(1)$ and $y(0)$ in Imbens and Rubin (2015), y^1 and y^0 in Cunningham (2021), or $y_{t=1}$ and $y_{t=0}$ in Pearl and Mackenzie (2018). Here we will use $y_{t\leftarrow 1}^\star$ and $y_{t\leftarrow 0}^\star$, the star being a reminder that those quantities are potential outcomes, as reminded in Table 7.2.

	Treatment	Outcome			Features	···
	t_i	y_i	$y_{i,T\leftarrow 0}^\star$	$y_{i,T\leftarrow 1}^\star$	x_i	···
1	0	75	75	?	172	···
2	1	52	?	52	161	···
3	1	57	?	57	163	···
4	0	78	78	?	183	···

Table 7.2: Excerpt of a standard table, with observed data t_i, x_i, y_i, and potential outcomes $y_{i,T\leftarrow 0}^\star$ and $y_{i,T\leftarrow 1}^\star$, respectively when treatment (t) is either 0 and 1. The question mark ? corresponds to the unobserved outcome, and will correspond to the counterfactual value of the observed one.

The treatment formally corresponds, in our vaccine example, to an intervention, which is formally a shot given to a person, or a pill that the person must swallow. In this section, it will not be possible to manipulate the variable whose causal effect we want to measure. In the introduction, we mentioned the idea that body

[1] In a humorous article, Smith and Pell (2003) asked the question of setting up randomized experiments to prove the causal link between having a parachute and surviving a plane crash.

mass index (BMI) could have an impact on health status, but BMI is not a pill, it is an observed quantity. It could be possible to manipulate variables that will have an impact on the index (by forcing a person to practice sports regularly, change their eating habits, etc.), so that one is not measuring *strictly speaking* the causal effect of the body mass index, but rather that of the interventions that influence the index. In the same way, it is impossible to intervene on certain variables, said to be immutable, such as gender or racial origin. The counterfactual is then purely hypothetical. Dawid (2000) was very critical of the idea that we can create (or observe) a counterfactual, because "*by definition, we can never observe such [counterfactual] quantities, nor can we assess empirically the validity of any modelling assumption we may make about them, even though our conclusions may be sensitive to these assumptions*."

We will say that there is a causal effect (or "*identified causal effect*") of t on y if $y_0^\star$ and $y_1^\star$ are significantly different. And since we cannot observe these variables at the individual level, we will compare the effect on sub-populations, as shown by Rubin (1974), Hernán and Robins (2010) or Imai (2018). Quite naturally, one might want to measure the causal effect as the difference in $\overline{y}$ between the two groups, the treated ($t = 1$) and the untreated ($t = 0$), but unless additional assumptions are made, this difference will not correspond to the average causal effect (ATE, "average treatment effect"). But let us formalize a little be more the different concepts used here.

Definition 7.4.1 (Average Treatment Effect) *Holland (1986). Given a treatment T, the average treatment effect on outcome Y is*

$$\tau = ATE = \mathbb{E}\left[Y^\star_{t\leftarrow 1} - Y^\star_{t\leftarrow 0}\right].$$

Definition 7.4.2 (Conditional Average Treatment Effect) *Wager and Athey (2018). Given a treatment T, the conditional average treatment effect on outcome Y, given some covariates $\boldsymbol{X}$,is*

$$\tau(\boldsymbol{x}) = CATE(\boldsymbol{x}) = \mathbb{E}\left[Y^\star_{t\leftarrow 1} - Y^\star_{t\leftarrow 0}\middle|\boldsymbol{X} = \boldsymbol{x}\right].$$

Definition 7.4.3 (Individual Average Treatment Effect) *Given a treatment T, the conditional average treatment effect on outcome Y, for individual i, given covariates $\boldsymbol{X}_i$, is*

$$IATE(i) = \mathbb{E}\left[Y^\star_{i,t\leftarrow(1-t_i)} - Y^\star_{i,t\leftarrow t_i}\right].$$

A naïve attempt to estimate the average treatment effect is to consider

$$\widehat{\tau}_{\text{naive}} = \underbrace{\frac{\sum_{i=1}^n y_i \mathbf{1}(t_i = 1)}{\sum_{i=1}^n \mathbf{1}(t_i = 1)}}_{\overline{y}_1} - \underbrace{\frac{\sum_{i=1}^n y_i \mathbf{1}(t_i = 0)}{\sum_{i=1}^n \mathbf{1}(t_i = 0)}}_{\overline{y}_0},$$

where $\overline{y}_1$ is the average outcome of treated observations ($t_i = 1$), and $\overline{y}_0$ is the average outcome of individuals in the control group ($t_i = 0$). Observe that $\overline{y}_1$ and $\overline{y}_0$ are unbiased estimates of $\mathbb{E}[Y|T = 1]$ and $\mathbb{E}[Y|T = 0]$, respectively. Therefore, $\widehat{\tau}_{\text{naive}}$ is unbiased estimate of

$$\mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]. \tag{7.1}$$

Equation (7.1) is called "*association effect*" of T on Y (first level on the ladder of causality). This is not yet the average treatment effect, since $\mathbb{E}[Y^\star_{T\leftarrow t}]$ and $\mathbb{E}[Y|T = t]$ are different quantities (unless data were obtained in a randomized experiment, without any selection bias). It is possible to "identify" causal effect, when adding some properties.

Definition 7.4.4 (unconfoundedness / Ignorability) *The treatment indicator T is said to be unconfounded, or ignorable, if*

$$(Y^\star_{T\leftarrow 1}, Y^\star_{T\leftarrow 0}) \perp\!\!\!\perp T.$$

This property is a classical consequence of randomization, but not only. It implies that knowing T gives no information for predicting $(Y^\star_{T\leftarrow 1}, Y^\star_{T\leftarrow 0})$ (the distribution of potential outcomes in the different treatment groups are the same). Of course, it is necessary to take into account heterogeneity, through covariates $\boldsymbol{x}$. A stronger condition is then needed

Definition 7.4.5 (Conditional unconfoundedness / Strong Ignorability) *The treatment indicator T is said to be conditionally unconfounded, if*

$$(Y^\star_{T\leftarrow 1}, Y^\star_{T\leftarrow 0}) \perp\!\!\!\perp T \,\big|\, \boldsymbol{X}.$$

For simplicity, assume that all confounders have been identified, and are categorical variables. Within a class (or stratum) $\boldsymbol{x}$, there is no confounding effect, and therefore, the causal effect can be identified by naive estimation, and the overall average treatment effect is identified by aggregation: by the law of total probability, $\mathbb{P}[Y^\star_{T\leftarrow 1} = y]$ is equal to

$$\sum_{\boldsymbol{x}} \mathbb{P}[Y^\star_{T\leftarrow 1} = y|\boldsymbol{X}=\boldsymbol{x}]\mathbb{P}[\boldsymbol{X}=\boldsymbol{x}] = \sum_{\boldsymbol{x}} \mathbb{P}[Y^\star_{T\leftarrow 1} = y|\boldsymbol{X}=\boldsymbol{x}, T=1]\mathbb{P}[\boldsymbol{X}=\boldsymbol{x}],$$

that we can write

$$\mathbb{P}[Y^\star_{T\leftarrow 1} = y] = \sum_{\boldsymbol{x}} \mathbb{P}[Y = y|\boldsymbol{X}=\boldsymbol{x}, T=1]\mathbb{P}[\boldsymbol{X}=\boldsymbol{x}],$$

and similarly for $\mathbb{P}[Y^\star_{T\leftarrow 0} = y]$, so that

$$\mathbb{E}[Y^\star_{T\leftarrow 1} - Y^\star_{T\leftarrow 0}] = \sum_{\boldsymbol{x}} \big(\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}, T=1] - \mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}, T=0]\big)\mathbb{P}[\boldsymbol{X}=\boldsymbol{x}],$$

and the estimate will be

$$\widehat{\tau}_{\text{strata}}(\boldsymbol{x}) = \sum_{\boldsymbol{x}} \widehat{\tau}_{\text{naive}}(\boldsymbol{x})\widehat{p}_n(\boldsymbol{X}=\boldsymbol{x}),$$

also called "exact matching estimate." Here, $\widehat{p}_n(\boldsymbol{X}=\boldsymbol{x})$ is the proportion of stata $\boldsymbol{x}$ is the training dataset.

Unfortunately, strata might be sparse, and that will generate a lot of variability. So, instead of considering all possible strata, it is possible to create a score, such that conditional on the score, we will have independence. This is the idea of a "balancing score,"

Definition 7.4.6 (Balancing score) *Rosenbaum and Rubin (1983) A balancing score is a function $b(\boldsymbol{X})$ such that*

$$\boldsymbol{X} \perp\!\!\!\perp T \,\big|\, b(\boldsymbol{X}).$$

An obvious scoring function is $b(\boldsymbol{X}) = \boldsymbol{X}$, but the idea here is to have $b : \mathcal{X} \to \mathbb{R}^d$ where d is small (ideally $d = 1$). It could be seen as the equivalent of the concept of "sufficient statistic" in parametric inference, in the sense that all the information contained in a sample $\boldsymbol{X}$ can be summarized into that statistic.

Propoition 7.4.1 *If $b(\boldsymbol{X})$ is a balancing score is a function, and if conditional unconfoundedness (as in Definition 7.4.5) is satisfied, then*

$$(Y^\star_{T\leftarrow 1}, Y^\star_{T\leftarrow 0}) \perp\!\!\!\perp T \,\big|\, b(\boldsymbol{X}).$$

The complete proof can be found in Rosenbaum and Rubin (1983) and Borgelt et al. (2009). It is a direct consequence of the so called "contraction" property, in the sense that $Y \perp\!\!\!\perp T \mid B$ and $Y \perp\!\!\!\perp B$ imply $Y \perp\!\!\!\perp (T, B)$. See Zenere et al. (2022) for more details about balancing scores and conditional independence properties.

A popular balancing score is the "propensity score". With our previous notations, if $n(\boldsymbol{x})$ is the number of observations (out of n) within stratum $\boldsymbol{x}$, and $\widehat{p}_n(\boldsymbol{x}) = n(\boldsymbol{x})/n$. The propensity score will be $\widehat{e}_n(\boldsymbol{x}) = n_1(\boldsymbol{x})/n(\boldsymbol{x})$, where $n_1(\boldsymbol{x})$ is the number of treated individuals in stratum $\boldsymbol{x}$. And one can write

$$\widehat{\tau}_{\text{strata}}(\boldsymbol{x}) = \sum_{\boldsymbol{x}} \widehat{\tau}_{\text{naive}}(\boldsymbol{x}) \frac{n(\boldsymbol{x})}{n} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i \mathbf{1}(t_i = 1)}{e(\boldsymbol{x}_i)} - \frac{y_i \mathbf{1}(t_i = 0)}{1 - e(\boldsymbol{x}_i)} \right).$$

With a little bit a calculations (see in the next section for the probabilistic version), we can actually write the later as

$$\widehat{\tau}_{\text{strata}}(\boldsymbol{x}) = \frac{1}{n} \sum_{e} \sum_{e(\boldsymbol{x}_i)=e} \left(\frac{y_i \mathbf{1}(t_i = 1)}{e} - \frac{y_i \mathbf{1}(t_i = 0)}{1 - e} \right) = \sum_{e} \widehat{\tau}_{\text{naive}}(e) \frac{n(e)}{n},$$

where we recognize a matching estimator, not on strata $\boldsymbol{x}$, but on the score. The interpretation, is that, conditional on the score, we can pretend that data were collected through some randomization process.

Definition 7.4.7 (Propensity Score) *Rosenbaum and Rubin (1983). The propensity score is the probability to be assigned to a particular treatment given a set of observed covariates. For a binary treatment ($t \in \{0, 1\}$)*

$$e(\boldsymbol{x}) = \mathbb{P}[T = 1 | \boldsymbol{X} = \boldsymbol{x}].$$

As mentioned earlier, and as proved in Rosenbaum and Rubin (1983), the propensity score is a balancing score. Even more, a score $b(\boldsymbol{x})$ is balanced if and only if it is a function of the propensity score $e(\boldsymbol{x})$. Since T is binary, this comes from the fact that $\mathbb{P}[T = 1 | \boldsymbol{X}, e(\boldsymbol{X})] = \mathbb{P}[T = 1 | e(\boldsymbol{X})]$.

7.4.2 Weights and Importance Sampling

A classical technique in surveys to correct biases is to use properly chosen weights, as explained in Pfeffermann (1993) and Biemer and Christ (2012). Inverse probability weighting leverages the propensity score $e(\boldsymbol{x})$ to balance covariates between groups. Intuitively, if we divide the weight of each unit by the probability that it will be treated, each unit has an equal probability of being processed. Mathematically, this corresponds to a change in probability, from $\mathbb{P}$ to $\mathbb{Q}$, where

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \frac{1}{2} \left(\frac{T}{e(\boldsymbol{X})} + \frac{1 - T}{1 - e(\boldsymbol{X})} \right),$$

so that $\mathbb{Q}$ is such that $\mathbb{Q}(T = 0) = \mathbb{Q}(T = 1)$, and under $\mathbb{Q}$, $T \perp\!\!\!\perp \boldsymbol{X}$. This means that the pseudo population (obtained by re-weighting) looks as if the treatment was randomly allocated by tossing an unbiased coin, and

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}[Y \mid T = 1] = \frac{\mathbb{E}_{\mathbb{Q}}[TY]}{\mathbb{Q}(T = 1)} &= 2 \cdot \frac{1}{2} \cdot \mathbb{E}_{\mathbb{P}} \left[YT \left(\frac{T}{e(\boldsymbol{X})} + \frac{1 - T}{1 - e(\boldsymbol{X})} \right) \right] \\
&= \mathbb{E}_{\mathbb{P}} \left[\frac{T}{e(\boldsymbol{X})} \cdot Y \right] = \mathbb{E}_{\mathbb{P}} \left[\mathbb{E}_{\mathbb{P}} \left[\frac{T}{e(\boldsymbol{X})} \cdot Y \middle| \boldsymbol{X} \right] \right] = \mathbb{E}_{\mathbb{P}} \left[\frac{\mathbb{E}_{\mathbb{P}}[TY \mid \boldsymbol{X}]}{e(\boldsymbol{X})} \right] \\
&= \mathbb{E}_{\mathbb{P}} \left[\frac{\mathbb{E}_{\mathbb{P}} \left[TY^{\star}_{T \leftarrow 1} \mid \boldsymbol{X} \right]}{e(\boldsymbol{X})} \right] = \mathbb{E}_{\mathbb{P}} \left[\frac{e(\boldsymbol{X}) \cdot \mathbb{E}_{\mathbb{P}} \left[Y^{\star}_{T \leftarrow 1} \mid \boldsymbol{X} \right]}{e(\boldsymbol{X})} \right] \\
&= \mathbb{E}_{\mathbb{P}} \left[\mathbb{E}_{\mathbb{P}} \left[Y^{\star}_{T \leftarrow 1} \mid \boldsymbol{X} \right] \right] = \mathbb{E}_{\mathbb{P}} \left[Y^{\star}_{T \leftarrow 1} \right],
\end{aligned}$$

and

$$\mathbb{E}_{\mathbb{Q}}[Y \mid T=0] = \mathbb{E}_{\mathbb{P}}\left[\frac{1-T}{1-e(\boldsymbol{X})}\cdot Y\right] = \mathbb{E}_{\mathbb{P}}\left[Y^{\star}_{T\leftarrow 0}\right].$$

Thus, if we combine,

$$\mathbb{E}_{\mathbb{P}}\left[Y^{\star}_{T\leftarrow 1} - Y^{\star}_{T\leftarrow 0}\right] = \mathbb{E}_{\mathbb{P}}\left[\frac{T}{e(\boldsymbol{X})}\cdot Y\right] - \mathbb{E}_{\mathbb{P}}\left[\frac{1-T}{1-e(\boldsymbol{X})}\cdot Y\right].$$

The price to pay to get to be able to identify the average treatment effect under $\mathbb{P}$ is that we need to estimate the propensity score e (see Kang and Schafer (2007) or Imai and Ratkovic (2014)). We will get back to Radon–Nikodym derivative and weights in Section 12.2.

Importance sampling is a classical techniques, popular in Monte Carlo techniques. Recall that Monte Carlo is simply based on the law of large numbers: if we can draw i.i.d. copies of a random variable X_i's, under probability $\mathbb{P}$, then

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i) \to \mathbb{E}_{\mathbb{P}}[h(X)], \text{ as } n\to\infty.$$

And much more can be obtained, since the empirical distribution $\mathbb{P}_n$ (associated with sample $\{x_1,\cdots,x_n\}$) converges to $\mathbb{P}$ as $n\to\infty$ (see e.g. Van der Vaart (2000)).

Now, assume that we can draw i.i.d. copies of a random variable X_i's, under probability $\mathbb{Q}$, and we still want to compute $\mathbb{E}_{\mathbb{P}}[h(X)]$. The idea of importance sampling is to use some weights,

$$\frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{\mathrm{d}\mathbb{P}(x_i)}{\mathrm{d}\mathbb{Q}(x_i)}}_{\omega_i} h(x_i) \to \mathbb{E}_{\mathbb{P}}[h(X)], \text{ as } n\to\infty,$$

where weights are simply based on the likelihood ratio of $\mathbb{P}$ over $\mathbb{Q}$. To introduce notations that we will use afterward, define

$$\widehat{\mu}_{\mathsf{is}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}\mathbb{P}(x_i)}{\mathrm{d}\mathbb{Q}(x_i)}h(x_i)$$

and if the likelihood ratio is known only up to a multiplicative constant, define a "*self-normalized importance sampling*" estimate

$$\widehat{\mu}'_{\mathsf{is}} = \frac{\sum_{i=1}^{n}\omega_i h(x_i)}{\sum_{i=1}^{n}\omega_i} \text{ with } \omega_i \propto \frac{\mathrm{d}\mathbb{P}(x_i)}{\mathrm{d}\mathbb{Q}(x_i)}.$$

On top of Figure 7.11, we supposed that we had a nice code to generate a Poisson distribution $\mathcal{P}(8)$, unfortunately, we want to generate some Poisson $\mathcal{P}(5)$. Below, we consider the opposite, we can generate some $\mathcal{P}(5)$ variable, but we want a $\mathcal{P}(8)$. On the left, the values of the weights, $\mathrm{d}\mathbb{P}(x)/\mathrm{d}\mathbb{Q}(x)$, with $x\in\mathbb{N}$. In the center, the histogram of $n=500$ observations from the algorithm we have ($\mathcal{P}(8)$ on top, $\mathcal{P}(5)$ below) and on the right, a weighted histogram for observations we wish we had, mixing the first sample and appropriate weights ($\mathcal{P}(5)$ on top, $\mathcal{P}(8)$ below). Below, we generate data from $\mathcal{P}(5)$, and the largest observation was here 13 (before, all values from 0 to 11 were obtained). As we can see on the right, it is not possible to get data outside the range of data initially obtained. Clearly, this approaches works well only when the supports are close. The weighted histogram was obtained using `wtd.hist`, in package `weights`.

In our context, one can define the importance sampling estimator of $\mathbb{E}[Y^{\star}_{T\leftarrow 1}]$, as

$$\widehat{\mu}_{\mathsf{is}}(Y^{\star}_{T\leftarrow 1}) = \frac{1}{n_1}\sum_{t_i=1}\frac{y_i}{e(\boldsymbol{x}_i)}\frac{n_T}{n} = \frac{1}{n}\sum_{t_i=1}\frac{y_i}{e(\boldsymbol{x}_i)},$$

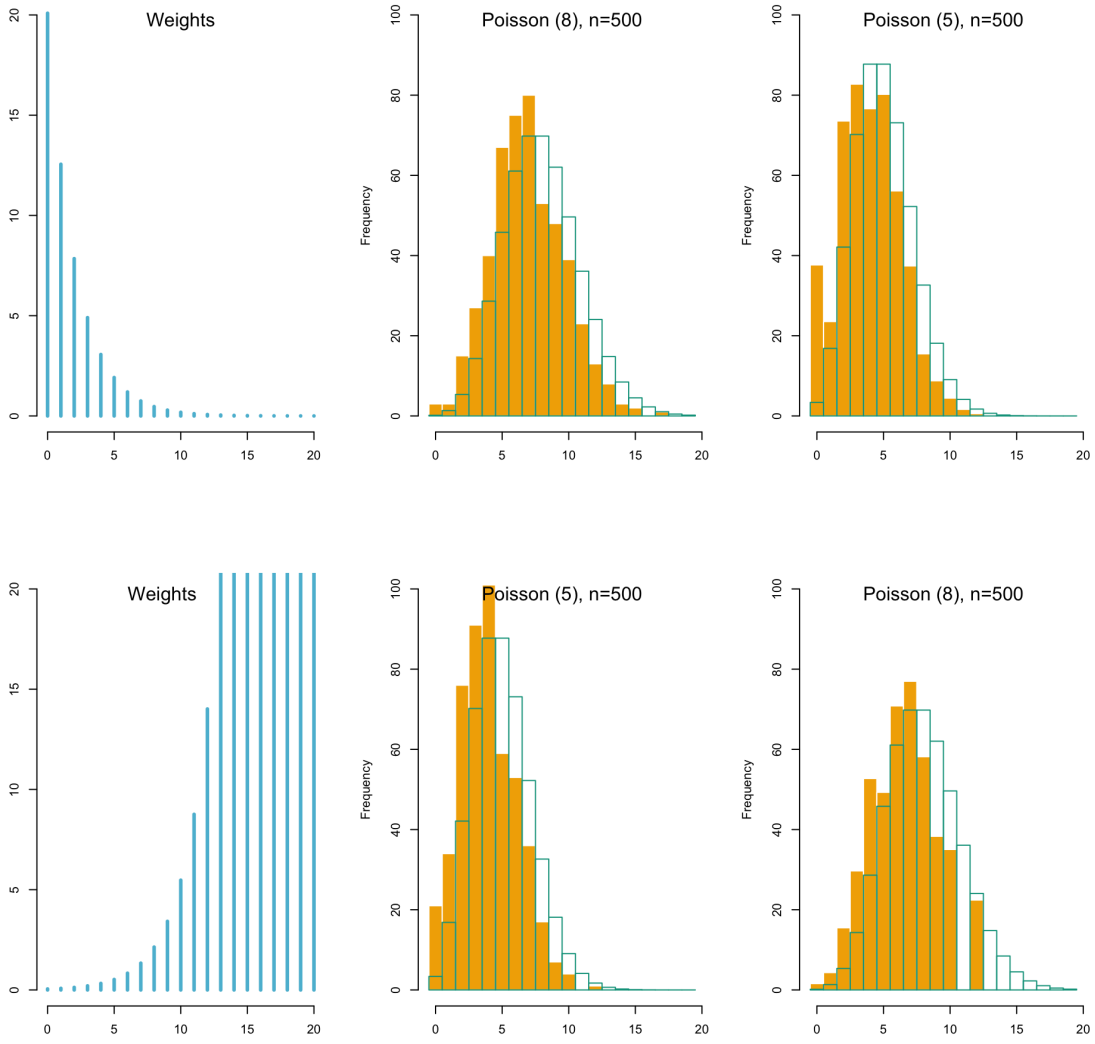


Figure 7.11: Illustration of the importance sampling procedure, and the use of weights ($\mathrm{d}\mathbb{P}(x)/\mathrm{d}\mathbb{Q}(x)$). On top, we have an algorithm to generate a Poisson distribution $\mathcal{P}(8)$ (in the middle), unfortunately, we want to generate some Poisson $\mathcal{P}(5)$ (on the right, obtained from the initial sample, and the weights on the left). Below, we have an algorithm to generate a Poisson distribution $\mathcal{P}(5)$ (in the middle), unfortunately, we want to generate some Poisson $\mathcal{P}(8)$ (on the right, obtained from the initial sample, and the weights on the left).

and a "*self-normalized importance sampling*" estimate for $\mathbb{E}[Y^\star_{T\leftarrow 1}]$ is

$$\widehat{\mu}'_{\mathsf{is}}(Y^\star_{T\leftarrow 1}) = \frac{\sum_{t_i=1}\omega_i y_i}{\sum_{t_i=1}\omega_i}, \text{ where } \omega_i = \frac{1}{e(\boldsymbol{x}_i)}.$$

The "*self-normalized importance sampling*" estimate for the conditional average treatment effect is

$$\widehat{\tau}'_{\text{is}} = \frac{\sum_{t_i=1} \omega_i y_i}{\sum_{t_i=1} \omega_i} - \frac{\sum_{t_i=0} \omega'_i y_i}{\sum_{t_i=0} \omega'_i} -, \text{ where } \omega_i = \frac{1}{e(\boldsymbol{x}_i)} \text{ and } \omega'_i = \frac{1}{1-e(\boldsymbol{x}_i)}.$$

7.5 Causal Techniques in Insurance

We will use these ideas of counterfactuals in Section 9.3, to quantify "individual fairness", but observe that causal inference is important in insurance at least in two applications, prevention and uplift modeling.

Heuristically, "prevention" means that measures are taken to prevent risk, such as diseases or injuries rather than curing them or treating their symptoms, in health insurance. If there is a causal relationship, an intervention could actually be effective.

Farbmacher et al. (2022) investigate the causal effect of health insurance coverage (T) on general health (Y) and decompose it into an indirect pathway via the incidence of a regular medical checkup (X) and a direct effect entailing any other causal mechanisms. Whether or not an individual undergoes routine checkups appears to be an interesting mediator, as it is likely to be affected by health insurance coverage and may itself have an impact on the individual's health (simply because checkups can help to identify medical conditions before they get serious).

Another classical application of causal inference and predictive modeling could be "*uplift modeling*." The idea is to model the impact of a treatment (that would be as a direct marketing action) on an individual's behaviour. Those ideas were formalized more than twenty years ago, in Hansotia and Rukstales (2002) or Hanssens et al. (2003). In Radcliffe and Surry (1999), the term "*true response modelling*" was used, Lo (2002) used "*true lift*," and finally Radcliffe (2007) suggested techniques for "*building and assessing uplift models*." More specifically, Hansotia and Rukstales (2002) used two models, estimated separately (namely two logistic regressions), one for the treated individuals, and one for the non-treated ones and Lo (2002) suggested an interaction model, where interaction terms between predictive variables $\boldsymbol{x}$ and the treatment t are added. Over the past twenty years, several papers appeared to apply those techniques, in personalised medicine, such as Nassif et al. (2013), but also in insurance, with Guelman et al. (2012) Guelman and Guillén (2014) and Guelman et al. (2014).

Part III

Fairness

"*For leaders today – both in business and regulation – the dominant theme of 21st century financial services is fast turning out to be a complicated question of fairness*", Wheatley (2013), Chief Executive of the FCA, at Mansion House, London

"*When you can measure what you are speaking about, and express it in numbers, you know something about it; but* **when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind**[2]*: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be,*" Lord Kevin, Thomson (1883)

"*Voulez-vous croire au réel, mesurez-le,*"

"*Do you want to believe in reality? Measure it,*" Bachelard (1927)

[2]Sociologist William Ogburn, a onetime head of the Social Sciences Division at the University of Chicago, was responsible for perhaps the most contentious carving on campus. Curving around an oriel window facing 59th Street is the quote from Lord Kelvin: "**when you cannot measure, your knowledge is meager and unsatisfactory**". Leu (2015) mentioned that in a 1939 symposium, economist Frank Knight Jr. snarkily suggested that the quote should be changed to "**if you cannot measure, measure anyhow**".

Chapter 8

Group Fairness

Assessing whether a model is discriminatory, or not, is a complex problem. As in Chapter 3, where we discussed global and local interpretability of predictive models, we will start with some global approaches (the local ones will be discussed in Chapter 9), also called "*group fairness*", comparing quantities between groups, usually identified by sensitive attributes (e.g., gender, ethnicity, age, etc.). Using the formalism introduced in the previous chapters, we will note y the variable of interest, $\widehat{y}$ or $m(\boldsymbol{x})$ the prediction given by the model, and s the sensitive attribute. Most concepts are derived from three main principles: independence ($\widehat{y} \perp\!\!\!\perp s$), separation ($\widehat{y} \perp\!\!\!\perp s$ conditional on y) and sufficiency ($y \perp\!\!\!\perp s$ conditional on $\widehat{y}$). We will review these approaches here, linking them while opposing them, and we will implement metrics related to those notions on various datasets.

Before starting formally to define fairness principles, remember that we have a simple dataset, `toydataset2` dataset, with only three admissible features, and several models were considered in Chapter 3. On Figure 8.1, we can visualize those models on a subset with $n = 40$ individuals. As in Kearns and Roth (2019), the shape of the points reflects the value of the sensitive attribute s, with here two types of individuals, circles and squares (as in Ruillier (2004)), corresponding to group `A` and group `B`, respectively. In this dataset y is binary, taking values in $\{0, 1\}$, corresponding to a good or a bad outcome. The probability of a bad outcome is a quantify of interest since the premium will be proportional to that quantity. If y is the numerical version of the categorical output (here the indicator $\mathbf{1}(y = 1)$) the probability $\mathbb{P}[Y = 1]$ will correspond to $\mathbb{E}[Y]$. The sensitive attribute is also binary, taking values in $\{\texttt{A}, \texttt{B}\}$, that we will correspond to a circle • or a square ■. There are also three legitimate continuous covariates, x_1, x_2 and x_3. Those variables $\boldsymbol{x}$ were used to build a score $m(\boldsymbol{x}) \in [0, 1]$, that will be interpreted as an estimation of both $\mathbb{P}[Y = 1|\boldsymbol{X} = \boldsymbol{x}]$ or $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$. This predictor is said to be "*fair through unawareness*" (discussed in section 8.1) when s was not used (also denoted "without sensitive" in the applications). But it is also possible to include s, and the score is $m(\boldsymbol{x}, s) \in [0, 1]$. Four different models are considered here (and trained on the complete `toydataset2` dataset): a plain logistic regression (GLM, fitted with `glm`), an additive logistic regression (GAM, with splines, for the three continuous variables, using `gam` from the `mgcv` package), a classification tree (CART, fitted using `rpart`) and a random forest (fitted using `randomForest`). Details on those $n = 40$ individuals are given in Table 8.1. On Figure 8.1, on top, the x-value (from the left to the right) of the $n = 40$ points correspond to values of $m(\boldsymbol{x}_i)$'s in $[0, 1]$. Colors correspond to the value of y, with $\{0, 1\}$ and the

shape corresponds to the value of s (■ and • correspond to A and B, respectively. As expected, individuals associated with $y_i = 0$ are more likely to be on the left (small $m(\boldsymbol{x}_i)$'s).

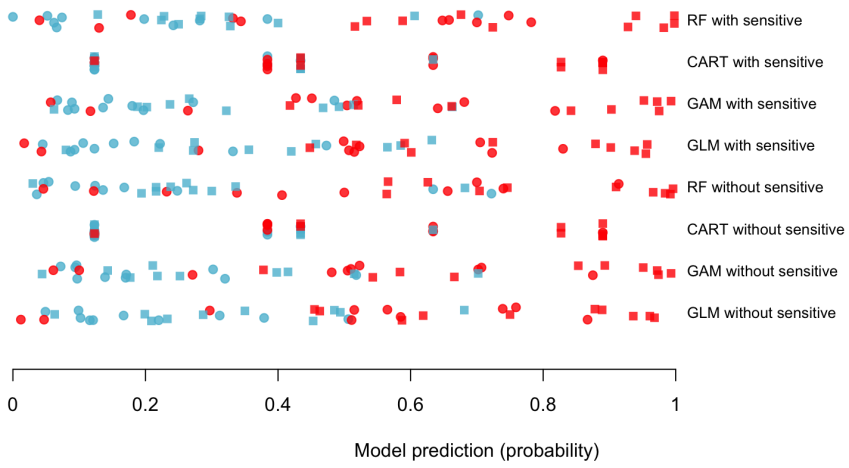


Figure 8.1: Scores $m(\boldsymbol{x}_i, s_i)$, on top ("with sensitive" attribute) or $m(\boldsymbol{x}_i) \in [0, 1]$ below ("without sensitive" attribute), for the different models, fitted on the `toydataset2` dataset. Colors correspond to the value of y, with $\{0, 1\}$ and the shape corresponds to the value of s (■ and • correspond to A and B, respectively).

On Figure 8.2, instead of visualizing $n = 40$ individuals on a scatter-plot (as on Figure 8.1), we can visualize the distribution of scores, that is the distribution of $m(\boldsymbol{x}_i, s_i)$ ("with sensitive" attribute) or $m(\boldsymbol{x}_i) \in [0, 1]$ ("without sensitive" attribute) using box plots, respectively when $y_i = 0$ and $y_i = 1$ on top, and respectively when $s_i = A$ and $s_i = B$ below. For example at the bottom of the two graphs, the two boxes correspond to prediction $m(\boldsymbol{x}_i)$ when a logistic regression is considered. On top, we have the distinction $y_i = 0$ and $y_i = 1$: for individuals $y_i = 0$, the median value for $m(\boldsymbol{x}_i)$ is 23% with the logistic regression, while it is 60% for individuals $y_i = 1$. All models here were able to "discriminate" according to the risk of individuals. Below, we have the distinction $s_i = A$ and $s_i = B$: for individuals $s_i = A$, the median value for $m(\boldsymbol{x}_i)$ is 30%, while it is 47% for individuals $s_i = B$. All models here were able to "discriminate" according to the risk of individuals. Based on the discussion we had in the introduction, in section 1.1.6 (and Figure 1.2 on the COMPAS dataset with Dieterich et al. (2016) and Feller et al. (2016) interpretations), if $m(\boldsymbol{x}_i)$ is a strong component of the premium, individuals in group B would be asked, on average, a higher premium than individuals in group A, and that difference could be perceived as discriminatory. When comparing box plots on top and below, at least, observe that all models "discriminate" more, between groups, based on the true risk y than on the sensitive attribute s. On Figure 8.3 we can visualize survival functions (on the small dataset, with $n = 40$ individuals) of those scores, since, as discussed in Chapter 3, classical quantities used on classifiers (true positive rates, etc) can be visualized on such a graph. Even if each curve is obtained on 20 individuals (as shown in Table 8.1, the dataset is well balanced, with 10 individuals in each group (y, s)). In table 8.2, Kolmogorov-Smirnov test is performed, to assess weather the difference between the survival functions is significant. Here H_0 would be "there is no discrimination with respect to s" (for the lower part of the Table, it would be y on top), while H_1 would be "there is no discrimination with respect to s". Here,

x_1	-1.220	-1.280	-0.930	-0.330	0.050	1.000	-0.070	-1.340	-1.800	-1.890
x_2	3.700	9.600	4.500	5.800	2.800	1.200	1.400	3.800	5.900	3.200
x_3	-1.190	-0.410	-1.050	-0.660	-0.890	0.250	-1.530	-0.300	-1.780	-0.570
s	A	A	A	A	A	A	A	A	A	A
y	0	0	0	0	0	0	0	0	0	0
p	0.090	0.500	0.130	0.270	0.14	0.220	0.070	0.090	0.160	0.060
$m(\boldsymbol{x})$	0.099	0.506	0.167	0.379	0.22	0.312	0.121	0.102	0.116	0.049

x_1	-0.330	0.800	-1.040	-0.160	-0.990	-0.440	-0.220	-3.200	-1.720	-1.090
x_2	7.900	9.200	9.400	9.800	6.700	7.900	9.700	3.700	2.700	9.600
x_3	-0.230	-0.650	-1.530	-1.650	-1.150	-0.030	-0.040	-3.440	-0.700	-1.620
s	A	A	A	A	A	A	A	A	A	A
y	1	1	1	1	1	1	1	1	1	1
p	0.460	0.840	0.500	0.690	0.260	0.440	0.660	0.100	0.050	0.510
$m(\boldsymbol{x})$	0.585	0.867	0.511	0.739	0.297	0.565	0.759	0.012	0.047	0.515

x_1	1.480	0.720	0.400	-0.470	-0.230	-0.820	0.740	1.440	0.200	-0.610
x_1	3.000	0.000	4.800	5.300	6.500	0.500	0.700	1.200	3.000	3.900
x_1	3.520	1.690	0.080	0.910	0.220	1.220	0.560	1.620	0.190	-0.010
s	B	B	B	B	B	B	B	B	B	B
y	0	0	0	0	0	0	0	0	0	0
p	0.670	0.160	0.460	0.310	0.470	0.050	0.210	0.480	0.260	0.190
$m(\boldsymbol{x})$	0.681	0.209	0.485	0.350	0.494	0.063	0.233	0.453	0.287	0.199

x_1	0.820	2.040	1.740	1.580	1.550	1.030	0.440	3.090	2.180	2.300
x_1	4.200	2.800	0.400	2.300	9.600	8.500	4.200	5.900	7.100	4.900
x_3	2.920	2.480	2.010	1.570	2.830	1.240	0.570	3.510	1.260	1.200
s	B	B	B	B	B	B	B	B	B	B
y	1	1	1	1	1	1	1	1	1	1
p	0.540	0.840	0.540	0.650	0.970	0.900	0.420	1.000	0.980	0.960
$m(\boldsymbol{x})$	0.619	0.75	0.463	0.587	0.961	0.889	0.455	0.968	0.936	0.878

Table 8.1: The small version of `toydata2`, with $n = 40$ observations. p corresponds to the "true probability" (used to generate y), and $m(\boldsymbol{x})$ is the predicted probability, from a plain logistic regression.

we do not consider some favored-disfavored distinction between the two groups, discrimination could be in both direction.

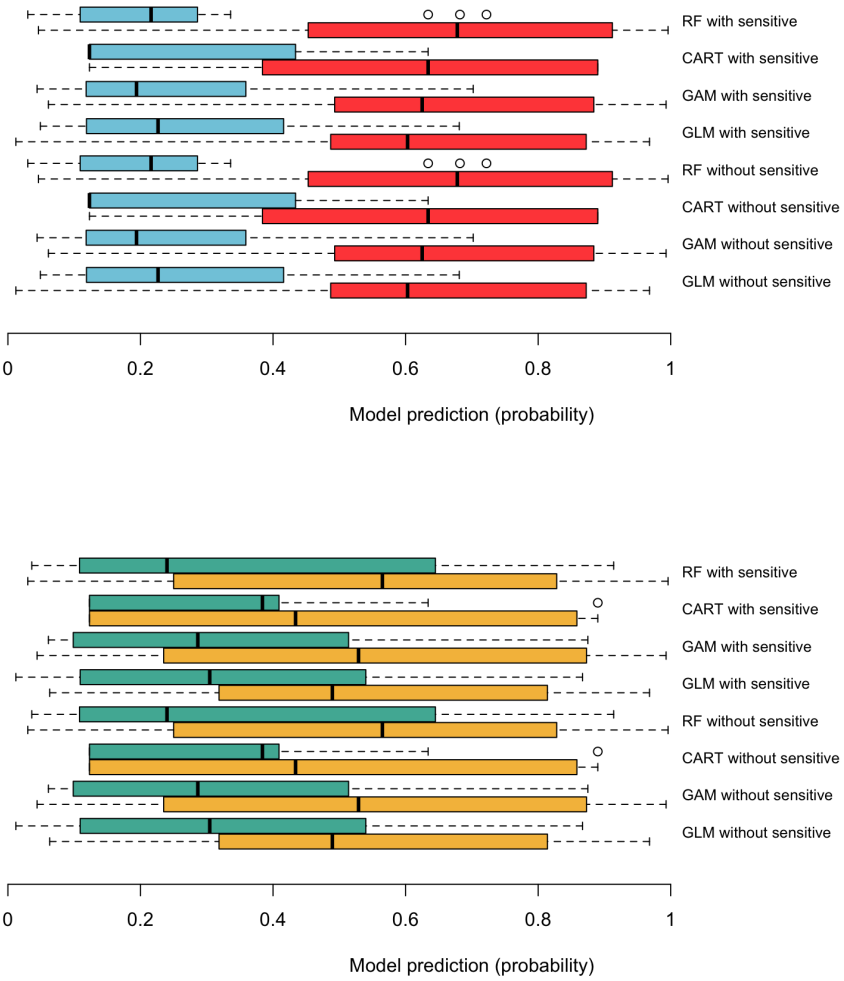


Figure 8.2: Box plot of the scores $m(\boldsymbol{x}_i, s_i)$ ("with sensitive" attribute) or $m(\boldsymbol{x}_i) \in [0, 1]$ ("without sensitive" attribute), for the different models, conditional on $y \in \{0, 1\}$ on top and conditional on $s \in \{\texttt{A}, \texttt{B}\}$ below.

8.1 Fairness Through Unawareness

The very first concept we will discuss is based on "*blindness*", also coined "*fairness through unawareness*" in Dwork et al. (2012), and is based on the (naïve) idea that a model fitted on the subpart of the dataset that contains only legitimate attributes (and not sensitive ones) will be "*fair*",

	unaware (without s)				aware (with s)			
	GLM	GAM	CART	RF	GLM	GAM	CART	RF
Difference between score distributions, $y \in \{0, 1\}$								
distance	0.700	0.650	0.500	0.700	0.650	0.600	0.500	0.650
p-value	0.01%	0.03%	0.29%	0.01%	0.03%	0.11%	0.29%	0.03%
Difference between score distributions, $s \in \{\texttt{A}, \texttt{B}\}$								
distance	0.350	0.350	0.450	0.400	0.400	0.400	0.450	0.400
p-value	17.45%	17.13%	1.66%	7.87%	8.11%	8.11%	1.66%	8.11%

Table 8.2: Kolmogorov-Smirnov test, to compare the conditional distribution of $m(\boldsymbol{x}_i, s_i)$ ("with sensitive" attribute) or $m(\boldsymbol{x}_i) \in [0, 1]$ ("without sensitive" attribute), conditional on the value of y on top, and s below. The distance is the maximum distance between the two survival functions, and the p-value is the one obtained with a two-sided alternative hypothesis. H_0 corresponds to the hypothesis that both distributions are identical ("no discrimination").

Definition 8.1.1 (Fairness through unawareness) *Dwork et al. (2012). A model $m(\cdot)$ satisfies the fairness through unawareness criteria, with respect to sensitive attribute $s \in \mathcal{S}$ if $m : \mathcal{X} \to \mathcal{Y}$.*

Based on that idea, we will extend the notion of "*regression function*" (from Definition 3.3.1), and distinguish "*aware*" and "*unaware regression functions*"

Definition 8.1.2 (Aware and unaware regression functions μ) *The aware regression function is $\mu(\boldsymbol{X}, S) = \mathbb{E}[Y|\boldsymbol{X}, S]$ and the unaware regression function is $\mu(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$.*

In Lindholm et al. (2022a,b), the "*aware regression function*" is named "*best-estimate price (given full information)*". On Figure 8.4, we can visualize empirical versions of those two regression functions, on the `toydata2` dataset, with scatterplots $\{\widehat{\mu}(\boldsymbol{x}_i, s_i), \widehat{\mu}(\boldsymbol{x}_i)\}$, where three kinds of models are considered.

This principle prescribes not to explicitly employ sensitive features when making decisions, and assessing premiums. It is not *per se* a group fairness principle, but it is the first one we will mention in this chapter. Apfelbaum et al. (2010) reminds us that "*the color-blind approach to managing diversity has become a leading institutional strategy for promoting racial equality, across domains and scales of practice*", making this principle extremely important. Goodwin and Voola (2013) discussed the "*the gender-blind approach*", asking if "*gender neutral*" or "*gender blind*" are equivalent. To go further, a simple way to achieve this is simply not to ask for such information (gender, age, race, etc). It is precisely why most resumes do not mention the age of candidates, or their gender. The main drawback of such a strategy is that not collecting sensitive information (on training data as well as testing or validation data) will not allow us to assess weather a model is fair, or not, as mentioned in Lewis (2004). But as stated in Apfelbaum et al. (2010), "*institutional messages of color blindness may therefore artificially depress formal reporting of racial injustice. Color-blind messages may thus appear to function effectively on the surface even as they allow explicit forms of bias to persist*". As discussed in the previous chapters, proxy based discrimination is still possible, and this approach of fairness "*function effectively on the surface*", only. So we need other definition to assess whether a model is fair, or not.

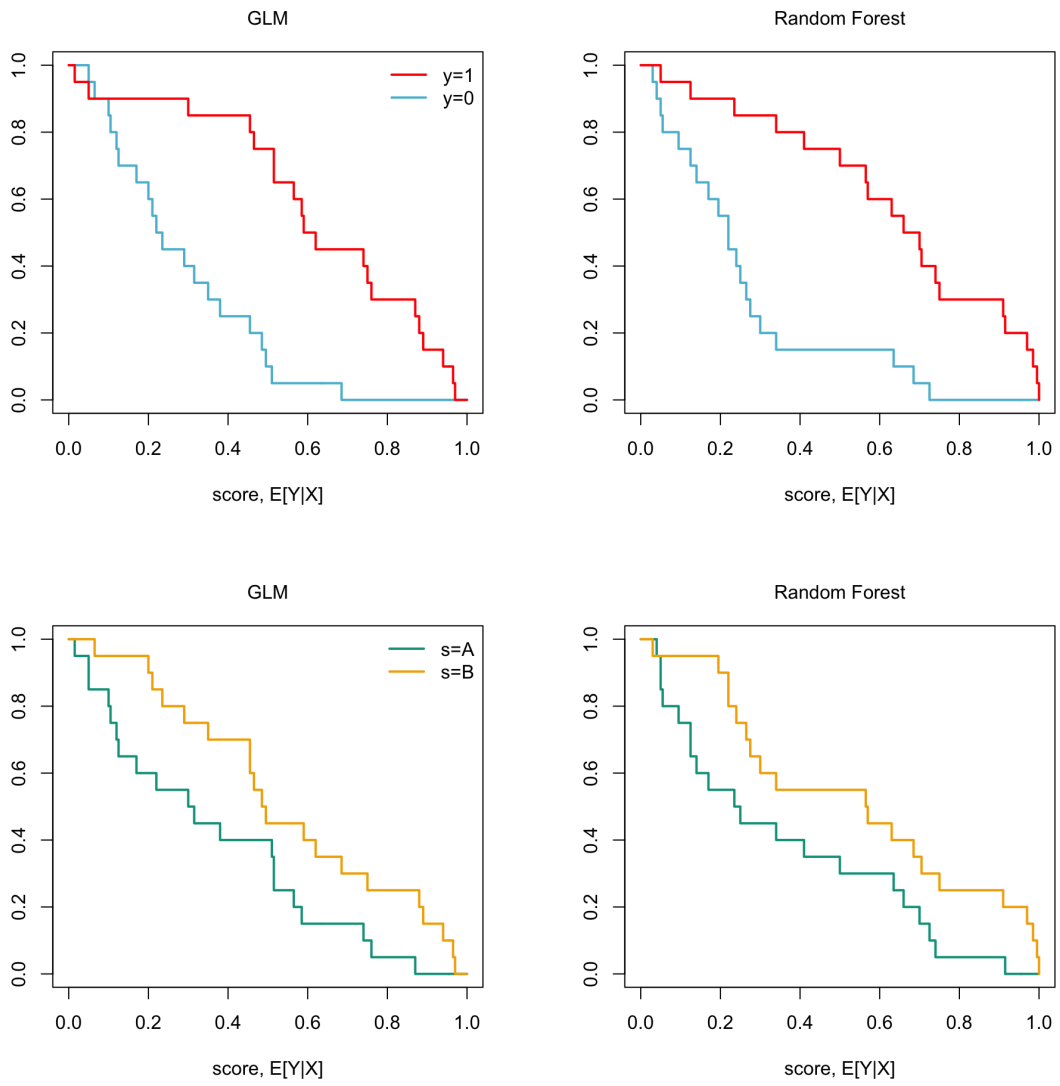


Figure 8.3: Survival distribution of the scores $m(\boldsymbol{x}_i) \in [0, 1]$ ("without sensitive" attribute), for the different models (plain logistic regression on the left and random forest on the right), conditional on $y \in \{0, 1\}$ on top and conditional on $s \in \{\texttt{A}, \texttt{B}\}$ below.

8.2 Independence and Demographic Parity

As pointed out by Caton and Haas (2020), there are at least a dozen ways to define (formally) the fairness of a classifier, or more generally of a model. For example, one can wish for independence between the score[1] and

[1] As discussed in Chapter 3, we will use here $\boldsymbol{z}$ as a generic notation, to denote either $\boldsymbol{x}$, or $(\boldsymbol{x}, s)$.

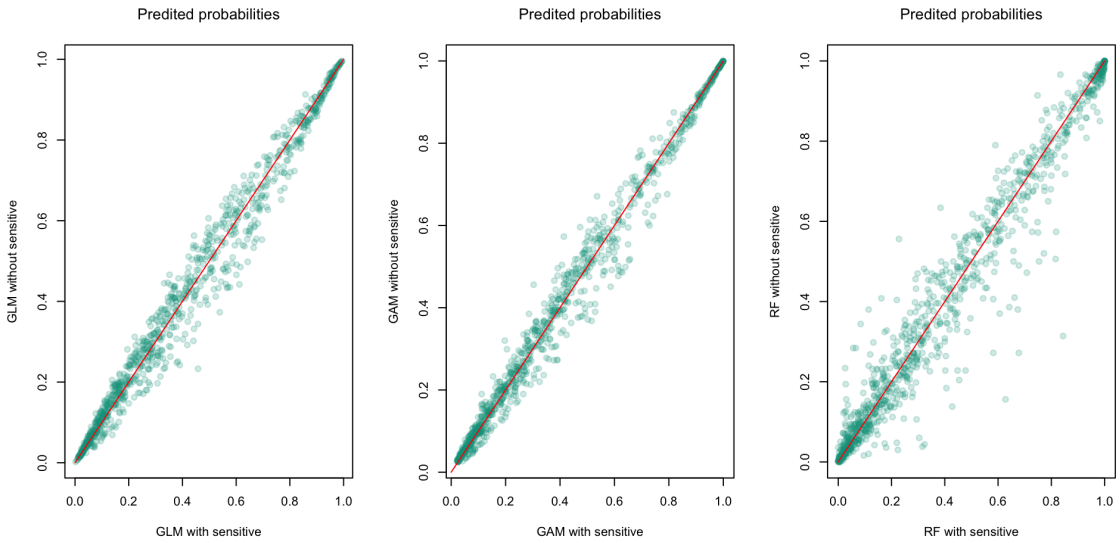


Figure 8.4: Scatterplot with aware (with the sensitive attribute s) on the x-axis and unaware model (without the sensitive attribute s) on the y-axis, or $\{\widehat{\mu}(\boldsymbol{x}_i, s_i), \widehat{\mu}(\boldsymbol{x}_i)\}$, for the GLM, the GAM and the Random Forest models, on the `toydata2` dataset, $n = 1000$ individuals.

the group membership, $m(\boldsymbol{Z}) \perp\!\!\!\perp S$, or between the prediction (as a class) and the protected variable $\widehat{Y} \perp\!\!\!\perp S$.

Definition 8.2.1 (Independence) *Barocas et al. (2017) A model m satisfies the independence property if $m(\boldsymbol{Z}) \perp\!\!\!\perp S$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$.*

Observe that this is implicitly in the introduction of this chapter, for example in Table 8.2 when we compared conditional distributions of the score $m(\boldsymbol{x}_i)$ in the two groups, $s = \texttt{A}$ and $s = \texttt{B}$. Inspired by Darlington (1971) define as follows, for a classifier m_t

Definition 8.2.2 (Demographic Parity) *Calders and Verwer (2010), Corbett-Davies et al. (2017). A decision function $\widehat{y}$ – or a classifier $m_t(\cdot)$, taking values in $\{0, 1\}$ – satisfies demographic parity, with respect to some sensitive attribute S if (equivalently)*

$$\begin{cases} \mathbb{P}[\widehat{Y} = 1|S = A] = \mathbb{P}[\widehat{Y} = 1|S = B] = \mathbb{P}[\widehat{Y} = 1] \\ \mathbb{E}[\widehat{Y}|S = A] = \mathbb{E}[\widehat{Y}|S = B] = \mathbb{E}[\widehat{Y}] \\ \mathbb{P}[m_t(\boldsymbol{Z}) = 1|S = A] = \mathbb{P}[m_t(\boldsymbol{Z}) = 1|S = B] = \mathbb{P}[m_t(\boldsymbol{Z}) = 1]. \end{cases}$$

In the regression case, when y is continuous, there are two possible definitions of demographic parity. If $\widehat{y} = m(\boldsymbol{z})$, we ask only for the equality of the conditional expectation (weak notion) or for the equality of the conditional distributions (strong notion)

Definition 8.2.3 (Weak Demographic Parity) *A decision function $\widehat{y}$ satisfies weak demographic parity if*

$$\mathbb{E}[\widehat{Y}|S = A] = \mathbb{E}[\widehat{Y}|S = B].$$

Definition 8.2.4 (Strong Demographic Parity) *A decision function $\widehat{y}$ satisfies demographic parity if $\widehat{Y} \perp\!\!\!\perp S$, i.e. for all A,*

$$\mathbb{P}[\widehat{Y} \in \mathcal{A}|S = \texttt{A}] = \mathbb{P}[\widehat{Y} \in \mathcal{A}|S = \texttt{B}], \ \forall \mathcal{A} \subset \mathcal{Y}.$$

If y and $\widehat{y}$ are binary, the two definitions are equivalent, but it is usually not the case. When the score is used to select clients, as to authorize the granting of a loan by a bank or a financial institution, this "*demographic parity*" concept (also called "*statistical fairness*", "*equal parity*", "*equal acceptance rate*" or simply "*independence*", as mentioned in Calders and Verwer (2010)) simply requires that the fraction of applicants in group `A` who are granted credit be approximately the same as the fraction of applicants in group `B` who are granted credit. And by symmetry, the rejection proportions must be identical. Using the same threshold t on the scores, to grant a loan, we get values of Table 8.3. For example, with a plain logistic regression

$$\mathbb{P}[\widehat{Y} = 1|S = \texttt{A}] = \frac{8}{20} = 40\% \text{ while } \mathbb{P}[\widehat{Y} = 1|S = \texttt{B}] = \frac{9}{20} = 45\%,$$

so that, strictly speaking,

$$\mathbb{P}[\widehat{Y} = 1|S = \texttt{A}] \neq \mathbb{P}[\widehat{Y} = 1|S = \texttt{B}].$$

In Table 8.3, the small dataset with $n = 40$ is on top, and with the entire dataset ($n = 1000$) below. One can observe an important difference between the ratio, for identical thresholds t. This is simply because the small dataset $n = 40$ is a distorted version of the entire one, with selection bias (as discussed in Section 5.7). Recall that in the entire dataset, $\mathbb{P}[Y = 0, S = \texttt{A}] \sim 47\%$, $\mathbb{P}[Y = 1, S = \texttt{B}] = 27\%$ and $\mathbb{P}[Y = 0, S = \texttt{B}] \sim \mathbb{P}[Y = 1, S = \texttt{A}] \sim 13\%$, while in our small dataset, all probabilities are 25% (in order to have exactly 10 individuals in each group). So clearly, selection bias has an impact on discrimination assessment.

On Figure 8.5 we can visualize $t \mapsto \mathbb{P}[m_t(\boldsymbol{Z}) = 0|S = \texttt{A}]/\mathbb{P}[m_t(\boldsymbol{Z}) = 0|S = \texttt{B}]$ for aware and unaware models (here a plain logistic regression) on the left, and more specifically $t \mapsto \mathbb{P}[m_t(\boldsymbol{Z}) = 0|S = \texttt{A}]$ and $t \mapsto \mathbb{P}[m_t(\boldsymbol{Z}) = 0|S = \texttt{B}]$ in the middle and on the right. The model without the sensitive attribute is more fair, with respect to the demographic parity criteria, than the one with the sensitive attribute. Points on the dots are the ones obtained on the small dataset, with $n = 40$. As mentioned previously, on that dataset, it seems that there is less discrimination (with respect to s) probably because of some selection bias. When $t = 30\%$ for the plain unaware logistic regression, out of 300,000 individuals in group `A`, there are 100,644 "positive", which is a 33.55% proportion ($\mathbb{P}[\widehat{Y} = 1|S = \texttt{A}]$), and out of 200,000 individuals in group `B`, there are 163,805 "positive", which is a 81.9% proportion ($\mathbb{P}[\widehat{Y} = 1|S = \texttt{B}]$). Thus, the ratio $\mathbb{P}[\widehat{Y} = 1|S = \texttt{B}]/\mathbb{P}[\widehat{Y} = 1|S = \texttt{A}]$ is 2.44. As t increases, both proportions (of "positive") decrease, but the ratio $\mathbb{P}[\widehat{Y} = 1|S = \texttt{B}]/\mathbb{P}[\widehat{Y} = 1|S = \texttt{A}]$ increases. On Figure 8.6, we visualize on the right $\mathbb{P}[\widehat{Y} = 1|S = \texttt{B}]$ and $\mathbb{P}[\widehat{Y} = 1|S = \texttt{A}]$, and on the left, $t \mapsto \mathbb{P}[m_t(\boldsymbol{X}) \leq t|S = \texttt{A}]$ and $t \mapsto \mathbb{P}[m_t(\boldsymbol{X}) \leq t|S = \texttt{B}]$.

As we can see in Definition 8.2.2, this measure is base on m_t, and not m. A classical choice for t is 50% (close the majority rule in bagging approaches), but other choices are possible. An alternative is to choose t so that (at least on the training dataset) $\mathbb{P}[m_t(\boldsymbol{Z} > t]$ should be close to $\mathbb{P}[Y = 1]$. In section 8.9, we will consider the probability to claim a loss in motor insurance (on the `frenchmotor` dataset). In the training subset 8.72% of the policyholders claimed a loss, 8.55% in the validation dataset. With the logistic regression model, on the validation dataset, the average prediction ($\overline{m}(\boldsymbol{z})$) is 9% and the median one is 8%. With a threshold $t = 16\%$, and a logistic regression about 10% of the policyholders get $\widehat{y} = 1$ (which is close to the claim frequency in the dataset) and 9% with a classification tree. Therefore, another natural threshold is the quantile (associated with $\overline{m}(\boldsymbol{z}_i)$'s) when probability is the proportion of 0's among y_i's. Finally, as discussed in Chapter 9 in Krzanowski and Hand (2009), several approaches can be considered, based on the

	unaware (without s)				aware (with s)			
	GLM	GAM	CART	RF	GLM	GAM	CART	RF
$n = 40$, $t = 50\%$, ratio $\mathbb{P}[\widehat{Y} = 1 \vert S = \mathtt{B}]/\mathbb{P}[\widehat{Y} = 1 \vert S = \mathtt{A}]$								
$\mathbb{P}[\widehat{Y} = 1 \vert S = \mathtt{A}]$	40%	35%	20%	30%	30%	25%	20%	30%
$\mathbb{P}[\widehat{Y} = 1 \vert S = \mathtt{B}]$	45%	55%	40%	55%	60%	55%	40%	55%
ratio	1.125	1.571	2.000	1.833	2.000	2.200	2.000	1.833
$n = 40$, various t, ratio $\mathbb{P}[\widehat{Y} = 1 \vert S = \mathtt{B}]/\mathbb{P}[\widehat{Y} = 1 \vert S = \mathtt{A}]$								
$t = 30\%$	1.500	1.400	1.273	1.333	1.778	1.875	1.273	1.556
$t = 50\%$	1.125	1.571	2.000	1.833	2.000	2.200	2.000	1.833
$t = 70\%$	2.000	2.333	7.000	2.333	2.000	6.000	7.000	2.000
$n = 40$, various t, ratio $\mathbb{P}[\widehat{Y} = 0 \vert S = \mathtt{A}]/\mathbb{P}[\widehat{Y} = 0 \vert S = \mathtt{B}]$								
$t = 30\%$	2.000	1.667	1.500	1.375	2.750	2.400	1.500	1.833
$t = 50\%$	1.091	1.444	1.333	1.556	1.750	1.667	1.333	1.556
$t = 70\%$	1.214	1.308	1.462	1.308	1.214	1.357	1.462	1.214
$n = 40$, ratio $\mathbb{E}[m(\boldsymbol{X}) \vert S = \mathtt{B}]/\mathbb{E}[m(\boldsymbol{X}) \vert S = \mathtt{A}]$								
$\mathbb{E}[m(\boldsymbol{X}) \vert S = \mathtt{A}]$	34.840%	33.570%	33.185%	34.580%	32.290%	31.125%	33.185%	33.110%
$\mathbb{E}[m(\boldsymbol{X}) \vert S = \mathtt{B}]$	54.800%	54.410%	50.400%	54.000%	56.870%	56.155%	50.400%	54.700%
ratio	1.125	1.571	2.000	1.833	2.000	2.200	2.000	1.833
$n = 1000$, various t, ratio $\mathbb{P}[\widehat{Y} = 1 \vert S = \mathtt{B}]/\mathbb{P}[\widehat{Y} = 1 \vert S = \mathtt{A}]$								
$t = 30\%$	1.652	1.519	1.235	1.559	1.918	1.714	1.235	1.798
$t = 50\%$	1.877	2.451	2.918	2.404	2.944	3.457	2.918	2.180
$t = 70\%$	6.033	8.711	26.000	4.621	7.917	19.333	26.000	4.578
$n = 1000$, various t, ratio $\mathbb{P}[\widehat{Y} = 0 \vert S = \mathtt{A}]/\mathbb{P}[\widehat{Y} = 0 \vert S = \mathtt{B}]$								
$t = 30\%$	5.507	4.667	4.059	4.682	8.510	5.746	4.059	5.873
$t = 50\%$	3.603	3.806	3.256	4.159	4.735	6.825	4.884	6.332
$t = 70\%$	2.648	2.868	2.902	2.869	2.781	3.010	2.902	2.938
$n = 1000$, ratio $\mathbb{E}[m(\boldsymbol{X}) \vert S = \mathtt{B}]/\mathbb{E}[m(\boldsymbol{X}) \vert S = \mathtt{A}]$								
$\mathbb{E}[m(\boldsymbol{X}) \vert S = \mathtt{A}]$	34.840%	33.570%	33.185%	34.580%	32.290%	31.125%	33.185%	33.110%
$\mathbb{E}[m(\boldsymbol{X}) \vert S = \mathtt{B}]$	54.800%	54.410%	50.400%	54.000%	56.870%	56.155%	50.400%	54.700%
ratio	×1.573	×1.621	×1.519	×1.562	×1.761	×1.804	×1.519	×1.652

Table 8.3: Quantifying demographic parity on `toydata2`, using `dem_parity` from R package `fairness`. Ratio $\mathbb{P}[\widehat{Y} = 1 \vert S = \mathtt{B}]/\mathbb{P}[\widehat{Y} = 1 \vert S = \mathtt{A}]$ (and $\mathbb{P}[\widehat{Y} = 0 \vert S = \mathtt{A}]/\mathbb{P}[\widehat{Y} = 0 \vert S = \mathtt{B}]$) should be equal to 1 to satisfy the demographic parity criteria.

ROC curve or the rate of error. On Figure 8.7, we can see the “optimal" threshold, in the sense of maximum predictive power on the left, or minimizing the rate of error committed, on the right.

In this definition of fairness, observe that we consider the conditional distribution of $\widehat{y}$ on s, but other variables are not considered. If we have a lot of heterogeneity based on a specific legitimate attribute x, it could interesting to add it. Therefore, we can consider the following extension,

Definition 8.2.5 (Conditional demographic parity) *Corbett-Davies et al. (2017). We will have a condi-*

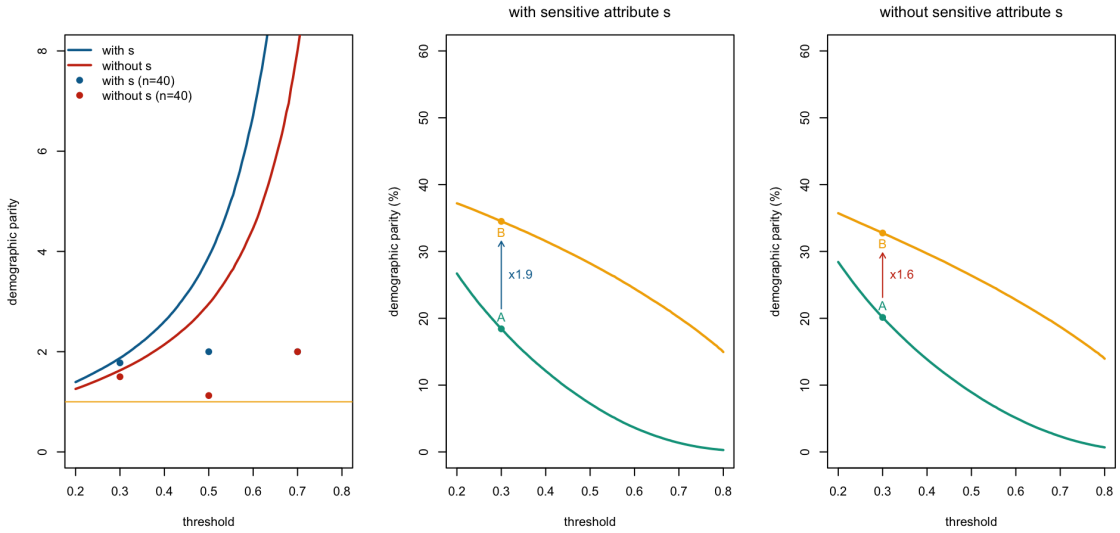


Figure 8.5: Demographic parity as a function of the threshold t, for classifier $m_t(\boldsymbol{x})$, when m is a plain logistic regression – with and without the sensitive attribute s – with groups A and B, on `toydata2`, using `dem_parity` from R package `fairness`. Here $n = 500,000$ simulated data are considered for plain lines, while dots on the left are empirical values obtained on the smaller subset, with $n = 40$, as in Table 8.3. In the middle $t \mapsto \mathbb{P}[m_t(\boldsymbol{X}) > t | S = \mathtt{B}]$ and $t \mapsto \mathbb{P}[m_t(\boldsymbol{X}) > t | S = \mathtt{A}]$ on the model with s, and on the right without s. On the left, evolution of the ratio ratio $\mathbb{P}[\widehat{Y} = 1 | S = \mathtt{B}]/\mathbb{P}[\widehat{Y} = 1 | S = \mathtt{A}]$.

tional demographic parity if (at choice) for all $\boldsymbol{x}$,

$$\begin{cases} \mathbb{P}[\widehat{Y} = 1 | \boldsymbol{X}_L = \boldsymbol{x}, S = \mathtt{A}] = \mathbb{P}[\widehat{Y} = 1 | \boldsymbol{X}_L = \boldsymbol{x}, S = \mathtt{B}], \ \forall y \in \{0, 1\} \\ \mathbb{E}[\widehat{Y} | \boldsymbol{X}_L = \boldsymbol{x}, S = \mathtt{A}] = \mathbb{E}[\widehat{Y} | \boldsymbol{X}_L = \boldsymbol{x}, S = \mathtt{B}], \\ \mathbb{P}[\widehat{Y} \in \mathcal{A} | \boldsymbol{X}_L = \boldsymbol{x}, S = \mathtt{A}] = \mathbb{P}[\widehat{Y} \in \mathcal{A} | \boldsymbol{X}_L = \boldsymbol{x}, S = \mathtt{B}], \ \forall \mathcal{A} \subset \mathcal{Y} \end{cases}$$

where L denotes a "legitimate" subset of unprotected covariates.

Finally, "demographic parity" claims that a model is "fair" if the prediction $\widehat{y}$ is independent of the protected attribute s,

$$\widehat{Y} \perp\!\!\!\perp S.$$

An alternative formulation would be (in a very general setting, such as a regression problem) to use distances between conditional distribution, $\widehat{Y}|S = \mathtt{A}$ and $\widehat{Y}|S = \mathtt{B}$, as defined in Section 3.3.1. If the conditional distribution of $m(\boldsymbol{X})$ in the two groups is too different, e.g. a large population stability index (PSI) as defined in Definition 3.3.1, the empirical estimation of $\mathbb{P}(\widehat{Y} = 1 | S = \mathtt{A}) - \mathbb{P}(\widehat{Y} = 1 | S = \mathtt{B})$ might be not robust, as discussed in Siddiqi (2012). Therefore, Szepannek and Lübke (2021) suggested to use a "*group unfairness index*" (GUI) defined as

$$\text{GUI}(\widehat{y}, s) = \sum_{i \in \{0,1\}} \big(\mathbb{P}(\widehat{Y} = i | s = \mathtt{B}) - \mathbb{P}(\widehat{Y} = i | s = \mathtt{A})\big) \log \frac{\mathbb{P}(\widehat{Y} = i | s = \mathtt{A})}{\mathbb{P}(\widehat{Y} = i | s = \mathtt{B})}.$$

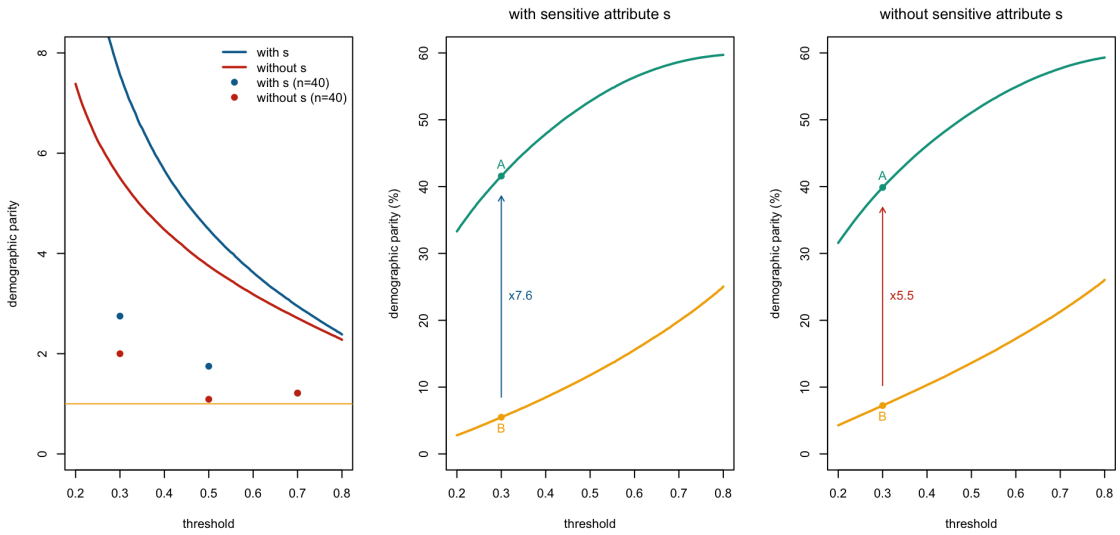


Figure 8.6: Alternative demographic parity graphs (compared with Figure 8.5), with ratio In the middle and on the right $t \mapsto \mathbb{P}[m_t(\boldsymbol{X}) \leq t | S = \texttt{B}]$ and $t \mapsto \mathbb{P}[m_t(\boldsymbol{X}) \leq t | S = \texttt{A}]$, and on the left, the ratio of the two, $\mathbb{P}[\widehat{Y} = 0 | S = \texttt{A}] / \mathbb{P}[\widehat{Y} = 0 | S = \texttt{B}]$.

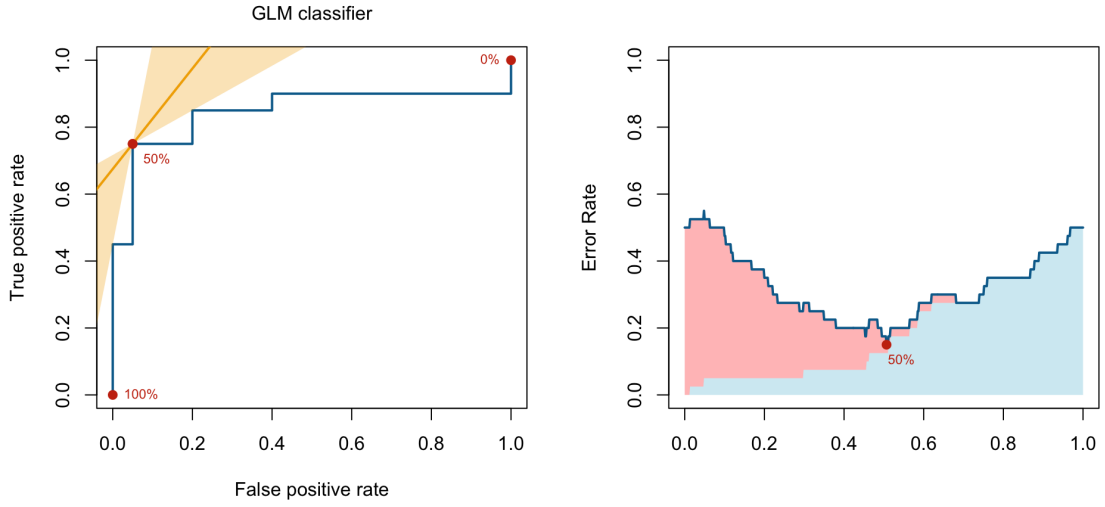


Figure 8.7: ROC curve on the plain logistic regression on the left, on the small dataset with $n = 40$ individuals, with the optimal threshold ($t = 50\%$) and the evolution of the rate of error on the right, and the optimal threshold (also $t = 50\%$).

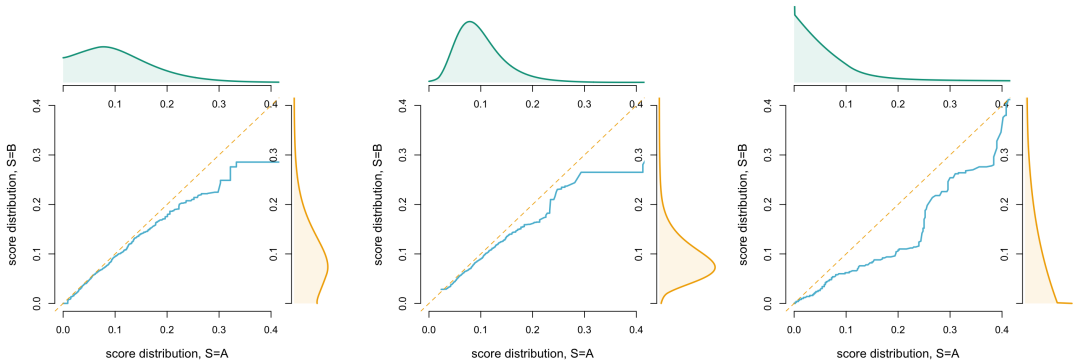


Figure 8.8: Matching between $m(\boldsymbol{x}, s = \texttt{A})$ (distribution on the x-axis) and $m(\boldsymbol{x}, s = \texttt{B})$ (distribution on the y-axis), where m is (from the left to the right) GLM, GBM and RF. The plain line is the (monontonic) optimal transport $\mathcal{T}^\star$ for different models.

And inspired by Shannon and Weaver (1949), it is possible also to define some mutual information, based on Kullback–Leibler divergence between the joint distribution and the independent version (see Definition 3.3.7).

$$\text{IM}(\widehat{y}, s) = \sum_{i\in\{0,1\}} \sum_{s\in\{\texttt{A},\texttt{B}\}} \mathbb{P}(\widehat{Y} = i, S = s) \log \frac{\mathbb{P}(\widehat{Y} = i, S = s)}{\mathbb{P}(\widehat{Y} = i)\mathbb{P}(S = s)}.$$

"Strong demographic parity" can be related to total distance (as in Definition 3.3.6), since a decision function $\widehat{y}$ satisfies strong demographic parity if $\widehat{Y} \perp\!\!\!\perp S$, i.e. for all $\mathcal{A} \subset \mathbb{R}$, $d_{\text{TV}}(\mathbb{P}_{\texttt{A}}, \mathbb{P}_{\texttt{B}}) = 0$, where $\mathbb{P}_{\texttt{A}}$ and $\mathbb{P}_{\texttt{B}}$ denote the conditional distributions of the score $m(\boldsymbol{X}, S)$. Quite naturally, one could consider another distance, for instance the Wassertein distance, as defined in Definition 3.3.11,

Propoition 8.2.1 *A model m satisfies the strong demographic parity property if and only if $W_2(\mathbb{P}_{\texttt{A}}, \mathbb{P}_{\texttt{B}}) = 0$.*

or Kullback-Leibler

Propoition 8.2.2 *A model m satisfies the strong demographic parity property if and only if $D_{\text{KL}}(\mathbb{P}_{\texttt{A}} \| \mathbb{P}) = D_{\text{KL}}(\mathbb{P}_{\texttt{B}} \| \mathbb{P}) = 0$.*

It is possible to visualize that property on Figure 8.8, with three models estimated on `frenchmotor`, with on the x-axis, the distribution of the score in group A, and on the y-axis the distribution of the score in group B. The plain line is the (monontonic) optimal transport $\mathcal{T}^\star$. If that line is on the diagonal, m is fair (for the "strong demographic parity" criteria).

One shortcoming with those approaches is that "*demographic parity*" is simply based on the independence between the protected variable s and the prediction $\widehat{y}$. And this does not take into account the fact that the outcome y may be correlated with the sensitive variable s. In other words, if the groups induced by the sensitive attribute s have different underlying distributions for y, ignoring these dependencies may lead to results that would be considered as fair. Therefore, quite naturally, an extension of the independence property is the "*separation*" criterion that adds the value of the outcome y. More precisely, we will require independence between the prediction $\widehat{y}$ and the sensitive variable s, conditional on the value of the outcome variable y, or formally $\widehat{Y} \perp\!\!\!\perp S$ conditional on Y.

8.3 Separation and Equalized Odds

In a general context, define "separation" as follows,

Definition 8.3.1 (Separation) *Barocas et al. (2017) A model $m : \mathcal{Z} \to \mathcal{Y}$ satisfies the separation property if $m(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$.*

Based on that principle, several notions can be introduced, when y is a binary variable ($y \in \{0, 1\}$). The first one, "*equal opportunity*" is achieved when the predicted target variable of a $\widehat{y}$ model and the label of a protected category s are statistically independent of each other, conditional on the actual value of the target variable y (either 0 or 1). In this binary classification problem, this corresponds to the fact that true positive rates and false positive rates be equal between groups, where the groups are determined by the protected category. A slightly less demanding fairness criterion is equal opportunity, in which only the probability of the true positive is equalized across groups in a protected category. Formally, we have the following definitions, where we require parity of false or true positives, in the two groups, A and B (see Figure 8.9).

Definition 8.3.2 (True positive equality, (Weak) Equal Opportunity) *Hardt et al. (2016) A decision function $\widehat{y}$ – or a classifier $m_t(\cdot)$, taking values in $\{0, 1\}$ – satisfies equal opportunity, with respect to some sensitive attribute S if*

$$\begin{cases} \mathbb{P}[\widehat{Y} = 1 | S = A, Y = 1] = \mathbb{P}[\widehat{Y} = 1 | S = B, Y = 1] = \mathbb{P}[\widehat{Y} = 1 | Y = 1] \\ \mathbb{P}[m_t(\boldsymbol{Z}) = 1 | S = A, Y = 1] = \mathbb{P}[m_t(\boldsymbol{Z}) = 1 | S = B, Y = 1] = \mathbb{P}[m_t(\boldsymbol{Z}) = 1 | Y = 1], \end{cases}$$

which corresponds to parity of true positives, in the two groups, $\{A, \}$.

The previous property can also be named "weak equal opportunity", and as for demographic parity, a stronger property can be defined,

Definition 8.3.3 (Strong Equal Opportunity) *A classifier $m(\cdot)$, taking values in $\{0, 1\}$, satisfies equal opportunity, with respect to some sensitive attribute S if*

$$\mathbb{P}[m(\boldsymbol{X}, S) \in \mathcal{A} | S = A, Y = 1] = \mathbb{P}[m(\boldsymbol{X}, S) \in \mathcal{A} | S = B, Y = 1] = \mathbb{P}[m(\boldsymbol{X}, S) \in \mathcal{A} | Y = 1],$$

for all $\mathcal{A} \subset [0, 1]$.

Definition 8.3.4 (False positive equality) *Hardt et al. (2016) A decision function $\widehat{y}$ – or a classifier $m_t(\cdot)$, taking values in $\{0, 1\}$ – satisfies parity of false positives, with respect to some sensitive attribute s, if*

$$\begin{cases} \mathbb{P}[\widehat{Y} = 1 | S = A, Y = 0] = \mathbb{P}[\widehat{Y} = 1 | S = B, Y = 0] = \mathbb{P}[\widehat{Y} = 1 | Y = 0] \\ \mathbb{P}[m_t(\boldsymbol{Z}) = 1 | S = A, Y = 0] = \mathbb{P}[m_t(\boldsymbol{Z}) = 1 | S = B, Y = 0] = \mathbb{P}[m_t(\boldsymbol{Z}) = 1 | Y = 0], \end{cases}$$

On Figure 8.9, point • in the top left corner corresponds to the case where the threshold t is 50% in group B, on the left, and point ◦ in the bottom left corner corresponds to the case where the threshold t is 50% in group A. Since those two points are not on same vertical or horizontal line, m_t satisfies neither "true positive equality" nor "false positive equality", when $t = 50\%$. Nevertheless if we suppose that t can be different, it is possible to achieve both (but never together). If we use a threshold $t = 24.1\%$ in group A, we have "false positive equality" with the classifier obtained when using a threshold $t = 50\%$ in group B. And if we use a threshold $t = 15.2\%$ in group A, we have "true positive equality" with the classifier obtained when using a

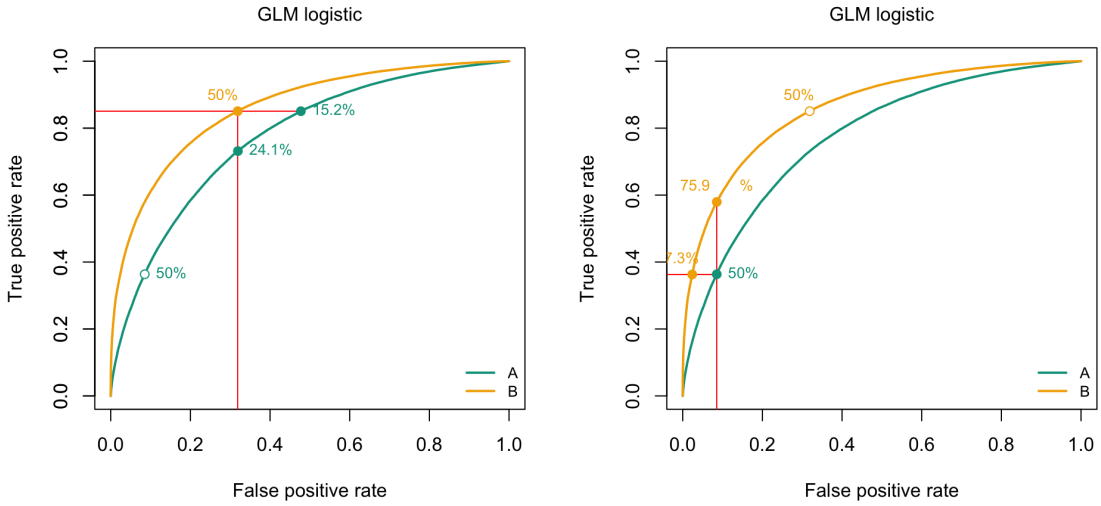


Figure 8.9: ROC curves (true positive rates against false positive rates) for the plain logistic regression m, on the `toydata2` dataset. Percentages are thresholds t that should be used, in each group (A and B), with the false positive rate (on the x-axis) and the true positive rate (on the y-axis).

threshold $t = 50\%$ in group B. If we consider the graph on the right, we have thresholds that should be used in group B to achieve either "true positive equality" or "false positive equality" when threshold $t = 50\%$ is used in group A.

If the two properties are satisfied at the same time, we have "*equalized odds*".

Definition 8.3.5 (Equalized Odds) *Hardt et al. (2016). parity of false positives A decision function* $\widehat{y}$ *– or a classifier* $m_t(\cdot)$ *taking values in* $\{0, 1\}$ *– satisfies equal odds constraint, with respect to some sensitive attribute S, if*

$$\begin{cases} \mathbb{P}[\widehat{Y} = 1|S = A, Y = y] = \mathbb{P}[\widehat{Y} = 1|S = B, Y = y] = \mathbb{P}[\widehat{Y} = 1|Y = y], \ \forall y \in \{0, 1\} \\ \mathbb{P}[m_t(\boldsymbol{Z}) = 1|S = A, Y = y] = \mathbb{P}[m_t(\boldsymbol{Z}) = 1|S = B, Y = y], \ \forall y \in \{0, 1\}, \end{cases},$$

which corresponds to parity of true positive and false positive, in the two groups.

Note that instead of using the value of $\widehat{y}$ (conditional on y and s), "*equalized odds*" means, for some specific threshold t,

$$\mathbb{E}[m_t(\boldsymbol{Z})|Y = y, S = A] = \mathbb{E}[m_t(\boldsymbol{Z})|Y = y, S = B], \ \forall y \in \{0, 1\}.$$

The separation criterion implies that the prediction $\widehat{y}$ should show the same error rate for each value of s, which explains why it is coined "*equalized odds*".

Therefore, as in Kleinberg et al. (2016), we can consider a so-called "*class balance*" property, where we consider m instead of m_t,

Definition 8.3.6 (Class balance) *(Kleinberg et al. (2016) We will have class balance in the weak sense if*

$$\mathbb{E}[m(\boldsymbol{X})|Y=y,S=A]=\mathbb{E}[m(\boldsymbol{X})|Y=y,S=B],\ \forall y\in\{0,1\}$$

or in the strong sense if

$$\mathbb{P}[m(\boldsymbol{X})\in\mathcal{A}|Y=y,S=A]=\mathbb{P}[m(\boldsymbol{X})\in\mathcal{A}|Y=y,S=B],\ \forall\mathcal{A}\subset[0,1],\ \forall y\in\{0,1\}.$$

As in the previous section, it is possible to compute for a threshold t the ratios $t\mapsto\mathbb{P}[m_t(\boldsymbol{Z})=1|S=A,Y=y]/\mathbb{P}[m_t(\boldsymbol{Z})=1|S=B,Y=y]=\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=y]$ or $t\mapsto\mathbb{P}[m_t(\boldsymbol{Z})=0|S=A,Y=y]/\mathbb{P}[m_t(\boldsymbol{Z})=0|S=B,Y=y]=\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=y]$, as in Table 8.4, with $y=1$ and $y=0$ on top and below, respectively.

	unaware (without s)				aware (with s)			
	GLM	GAM	CART	RF	GLM	GAM	CART	RF
False positive rate ratio, various t								
$t=50\%$	1.50	4.00	2.00	-	4.00	5.00	2.00	Inf
$t=70\%$	1.75	1.75	3.00	2.67	1.75	2.25	3.00	2.00
False negative rate ratio, various t								
$t=30\%$	0.60	0.75	0.60	1.00	0.33	0.20	0.60	0.50
$t=50\%$	1.00	0.50	0.00	2.00	0.00	0.00	0.00	1.00

Table 8.4: False positive and false negative metrics, on our $n=40$ individuals from dataset `toydata2`, using `fpr_parity` and `fnr_parity` from R package `fairness`. The ratio (group B against group A) should be close to 1 to have either false positive fairness (on topc) or false negative fairness (below).

On top of Table 8.4, we can visualize the False Positive Rate (FPR) parity metric as described by Chouldechova (2017). For example, on the aware logistic regression, and a threshold of $t=50\%$, in groups `A` and `B`, the false positive rate was 30% and 20%, respectively (on the small dataset with $n=40$ observation). With a basis 1 in group `B`, it would have been 1.5 in `A`, which is the value given in the Table. This value should be close to 1 if the false positive rate equality criteria was satisfied. On Figure 8.10, we can look at this metric, as a function of the threshold t, when n is very large (to have a smooth function). At the bottom of Table 8.4, False Negative Rate (FNR) parity metric is used, as described by Chouldechova (2017). Here, on the aware logistic regression, and a threshold of $t=50\%$, on both groups `A` and `B`, the false positive rate is the same, 10%. With a basis 1 in group `B`, it would have been 1 in `A`. On Figure 8.11, we can look at this metric, as a function of the threshold t, when n is very large.

The metric used for "*equalized odds*" (also known as "*equal opportunity*", "*positive rate parity*" or simply "*separation*") in function `equal_odds` from R package `fairness` is simply the "*sensitivity*" (or true positive rate). In Table 8.5, we can see them on our $n=40$ database, and on Figure 8.12, we can look at this metric, as a function of the threshold t, when n is very large.

A concept of "lack of disparate mistretement" has been introduced simulatneously in Zafar et al. (2019) and Berk et al. (2017) (and coined "ccuracy equality),

Definition 8.3.7 (Similar mistreatement) *Zafar et al. (2019) We will have similar mistreatement, or* "lack of disparate mistreatment*", if*

$$\begin{cases}\mathbb{P}[\widehat{Y}=Y|S=A]=\mathbb{P}[\widehat{Y}=Y|S=B]=\mathbb{P}[\widehat{Y}=Y]\\ \mathbb{P}[m_t(\boldsymbol{X})=Y|S=A]=\mathbb{P}[m_t(\boldsymbol{X})=Y|S=B]=\mathbb{P}[m_t(\boldsymbol{X})=Y].\end{cases}$$

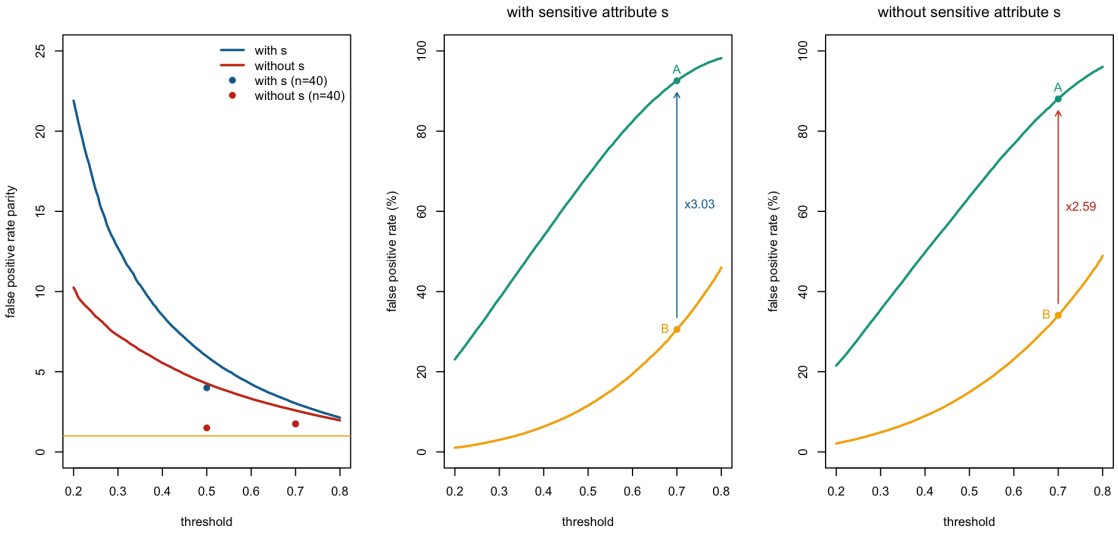


Figure 8.10: On the right, evolution of the false positive rates, in groups A and B, for $m_t(\boldsymbol{x})$ – without the sensitive attribute s – as a function of threshold t (on a plain logistic regression), on `toydata2`, using `fpr_parity` from R package `fairness`.

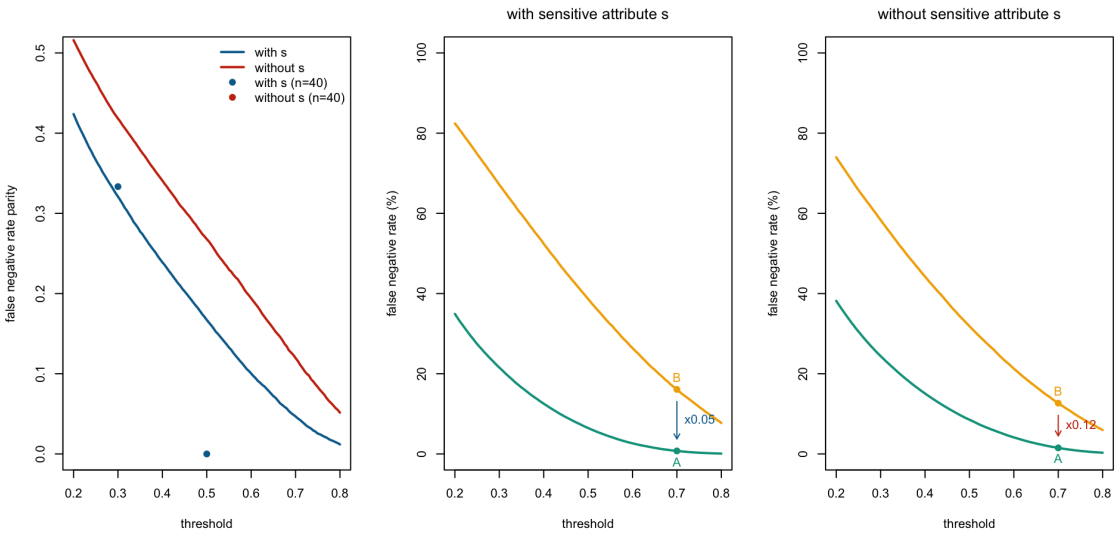


Figure 8.11: On the right, evolution of the false negative rates, in groups A and B, for $m_t(\boldsymbol{x})$ – without the sensitive attribute s – as a function of threshold t (on a plain logistic regression), on `toydata2`, using `fnr_parity` from R package `fairness`. Here $n = 500,000$ simulated data are considered.

	unaware (without s)				aware (with s)			
	GLM	GAM	CART	RF	GLM	GAM	CART	RF
$t = 30\%$	1.400	1.167	1.400	1.000	2.000	1.800	1.400	1.333
$t = 50\%$	1.000	1.125	1.111	0.889	1.429	1.250	1.111	1.000
$t = 70\%$	1.000	1.111	1.000	0.900	1.000	1.000	1.000	0.900

Table 8.5: "*equalized odds*" on the $n = 40$ subset of `toydata2`, using `equal_odds` from R package `fairness`.

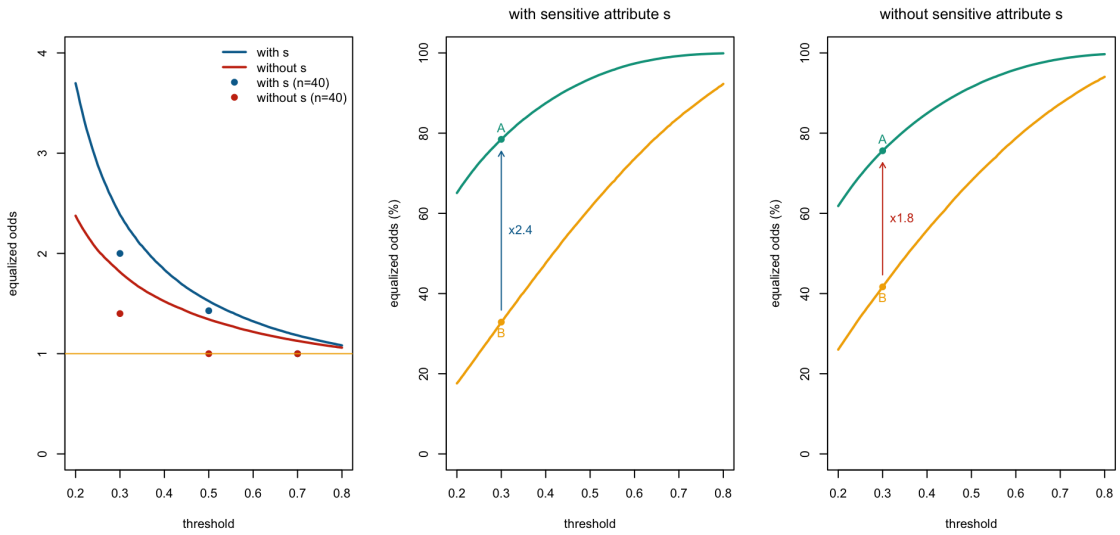


Figure 8.12: On the right, evolution of the equalized odds metrics, in groups A and B, for $m_t(\boldsymbol{x})$ – without the sensitive attribute s – as a function of threshold t (on a plain logistic regression), on `toydata2`, using `equal_odds` from R package `fairness`.

In the context of a regression model, we could consider a strong property on the distribution of residuals,

$$\mathbb{P}[Y - \widehat{Y} \in \mathcal{A} | S = \mathtt{A}] = \mathbb{P}[Y - \widehat{Y} \in \mathcal{A} | S = \mathtt{B}] = \mathbb{P}[Y - \widehat{Y} \in \mathcal{A}], \ \forall \mathcal{A}.$$

Because we as for equality of the distribution of residuals, we could consider a weaker condition (but easier to test), written using moments

$$\mathbb{E}\big[|Y - \widehat{Y}|^a | S = \mathtt{A}\big] = \mathbb{E}\big[|Y - \widehat{Y}|^a | S = \mathtt{B}\big] = \mathbb{E}\big[|Y - \widehat{Y}|^a\big], \ \forall a > 0.$$

Several extensions of this notion, introduced in Zafar et al. (2019), could be considered, such as lack of disparate mistreatment for "*false positive signals*"

$$\mathbb{P}[Y \neq \widehat{Y} | S = \mathtt{A}, Y = 0] = \mathbb{P}[Y \neq \widehat{Y} | S = \mathtt{B}, Y = 0],$$

or for "*false discovery signals*",

$$\mathbb{P}[Y \neq \widehat{Y} | S = \mathtt{A}, \widehat{Y} = 1] = \mathbb{P}[Y \neq \widehat{Y} | S = \mathtt{B}, \widehat{Y} = 1],$$

for example.

One can also use any metrics based on confusion matrices, such as ϕ, introduced by Matthews (1975), also denoted MCC, for "*Matthew's correlation coefficient*" in Baldi et al. (2000) or Tharwat (2021),

Definition 8.3.8 (ϕ-fairness) *Chicco and Jurman (2020) We will have ϕ-fairness if $\phi_1 = \phi_0$, where ϕ_s denotes Matthews correlation coefficient for the s group,*

$$\phi_s = \frac{TP_s \cdot TN_s - FP_s \cdot FN_s}{\sqrt{(TP_s + FP_s)(TP_s + FN_s) \cdot (TN_s + FP_s)(TN_s + FN_s)}}, \ s \in \{\texttt{A}, \texttt{B}\}.$$

In table 8.6, can visualize the evolution of $\phi_{\texttt{A}}/\phi_{\texttt{B}}$ as a function of the treshold. For example, with a 50% threshold t, in group `A`, MCC is 0.612 with a plain logistic regression, while it is 0.704 in group `A`, with the same model. So if the later was 1 (basis 1 in group `B`, it would have been 0.870 in group `A` (see Table 3.2 for details on computations, based on confusion matrices). On Figure 8.13, we can look at this metric, as a function of the threshold t, when n is very large.

	unaware (without s)				aware (with s)			
	GLM	GAM	CART	RF	GLM	GAM	CART	RF
$t = 30\%$	0.693	0.611	1.151	0.615	1.005	1.061	1.151	0.768
$t = 50\%$	0.870	0.745	0.816	0.241	1.069	0.821	0.816	0.482
$t = 70\%$	0.642	0.801	0.313	0.191	0.642	0.350	0.313	0.214

Table 8.6: Evolution of ϕ-fairness, as a function of threshold t, on the small version of `toydata2` (40 observations), using `mcc_parity` from R package `fairness`.

Another extension could be, inspired by Definition 8.3.6, where we introduced "*balance class*" conditionally on some covariates. Here also, in all the conditions based on conditional probabilities or conditional expectations, it is possible to add "$\boldsymbol{X}_L = \boldsymbol{x}$" in the conditioning.

Finally, observe that instead of asking for true positive rates and false positive rates to be equal, one can ask to have identical ROC curves, in the two groups. As a reminder, we have defined (see Definition 4.2.4) the ROC curve as $C(t) = \text{TPR} \circ \text{FPR}^{-1}(t)$, where $\text{FPR}(t) = \mathbb{P}[m(\boldsymbol{X}) > t|Y = 0]$ and $\text{TPR}(t) = \mathbb{P}[m(\boldsymbol{X}) > t|Y = 1]$.

Definition 8.3.9 (Equality of ROC curves) *Vogel et al. (2021) Let $FRP_s(t) = \mathbb{P}[m(\boldsymbol{X}) > t|Y = 0, S = s]$ and $TPR_s(t) = \mathbb{P}[m(\boldsymbol{X}) > t|Y = 1, S = s]$, where $s \in \{\texttt{A}, \texttt{B}\}$. Set $\Delta_{TPR}(t) = TPR_{\texttt{B}} \circ TPR_{\texttt{A}}^{-1}(t) - t$ et $\Delta_{FRP}(t) = FPR_{\texttt{B}} \circ FPR_{\texttt{A}}^{-1}(t) - t$. We will have an fairness of ROC curves if $\|\Delta_{TPR}\|_\infty = \|\Delta_{FPR}\|_\infty = 0$.*

And a weaker conditional can be asked, as in in Beutel et al. (2019) and Borkan et al. (2019), using not the entier ROC curve, but only the area under that curve,

Definition 8.3.10 (AUC fairness) *Borkan et al. (2019) We will have AUC fairness if $AUC_{\texttt{A}} = AUC_{\texttt{B}}$, where AUC_s is the AUC associated with model m within the s group.*

In table 8.7, we can visualize the ratio $\text{AUC}_{\texttt{B}}/\text{AUC}_{\texttt{A}}$. For example, with a logistic regression without the sensitive attribute s, AUC in groups `A` and `B` are respectively 77% and 92%. Thus, a basis 1 in group `B`, we would have had 0.837 in group `A`.

Recall that equal opportunity is satisfied if the "*separation*" property is satisfied, in the sense that the prediction $\widehat{Y}$ is conditionally independent of the protected attribute S, given the actual value Y,

$$\forall y : \ \widehat{Y} \perp\!\!\!\perp S \mid Y = y.$$

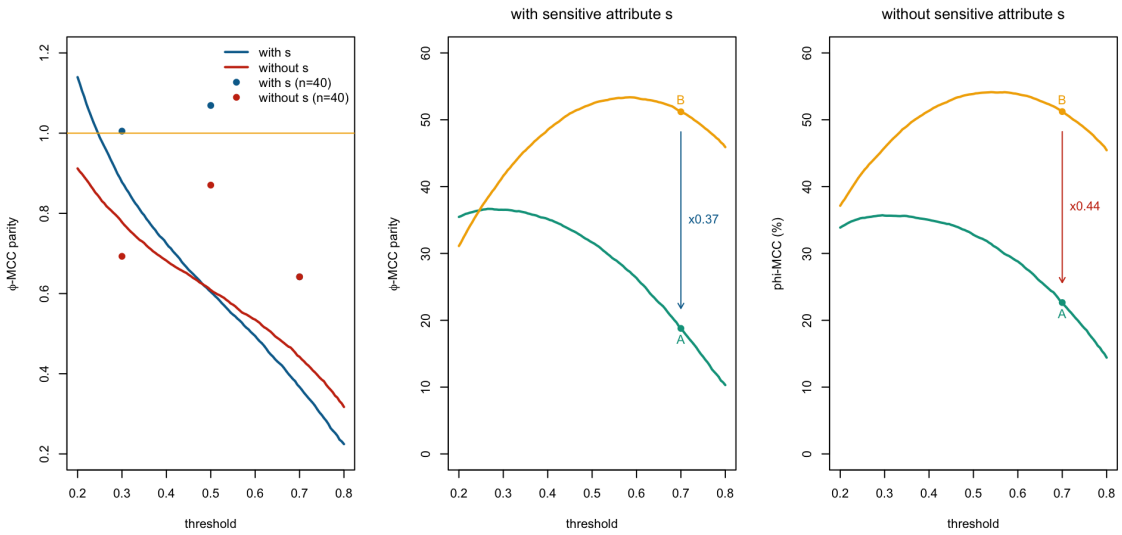


Figure 8.13: On the right, evolution of the ϕ-fairness metric, in groups A and B, for $m_t(\boldsymbol{x})$ – without the sensitive attribute s – as a function of threshold t (on a plain logistic regression), on `toydata2`, using `mcc_parity` from R package `fairness`. Here $n = 500,000$ simulated data are considered. In the middle evolution of MCC, in groups A and B, for $m_t(\boldsymbol{x}, s)$ – with the sensitive attribute s, as function of t. On the left, evolution of the ratio between groups A and B, respectively with and without the use of the sensitive attribute in m_t, as function of t. Dots on the lefts are empirical values obtained on a smaller subset, as in Table 8.6.

	unaware (without s)				aware (with s)			
	GLM	GAM	CART	RF	GLM	GAM	CART	RF
ratio of AUC	0.837	0.839	0.913	0.768	0.857	0.860	0.913	0.763

Table 8.7: AUC fairness on the small version of `toydata2`, using `roc_parity` (that actually compared AUC and not ROC curves) based on the ratio of AUC in the two groups, from R package `fairness`.

In our illustration, with $n = 40$ individuals, equality of opportunity is impossible to achieve. Indeed, this definition of fairness suggests that the false positive and true positive rates be the same for both populations. This may be reasonable, but in the illustrative example, it is impossible because the two ROC curves do not intersect, as visualized on Figure 8.9. Note that if the curves did cross, this could impose threshold choices that would be unattractive in practice (with acceptance rates potentially much too low, or too high). Nevertheless, on some applications, it is possible to transform

$$\mathbb{P}[m_t(\boldsymbol{X}) = 1|S = \mathtt{A}, Y = y] = \mathbb{P}[m_t(\boldsymbol{X}) = 1|S = \mathtt{B}, Y = y], \ \forall y \in \{0, 1\},$$

into

$$\mathbb{P}[m_{t_A}(\boldsymbol{X}) = 1|S = \mathtt{A}, Y = y] = \mathbb{P}[m_{t_B}(\boldsymbol{X}) = 1|S = \mathtt{B}, Y = y], \ \forall y \in \{0, 1\},$$

for some appropriate threshold t_A and t_B. We can visualize this on Figure 8.14. On the left side of Figure 8.14, the thresholds are chosen so that the rate of false positives is the same for both populations (A and B). In

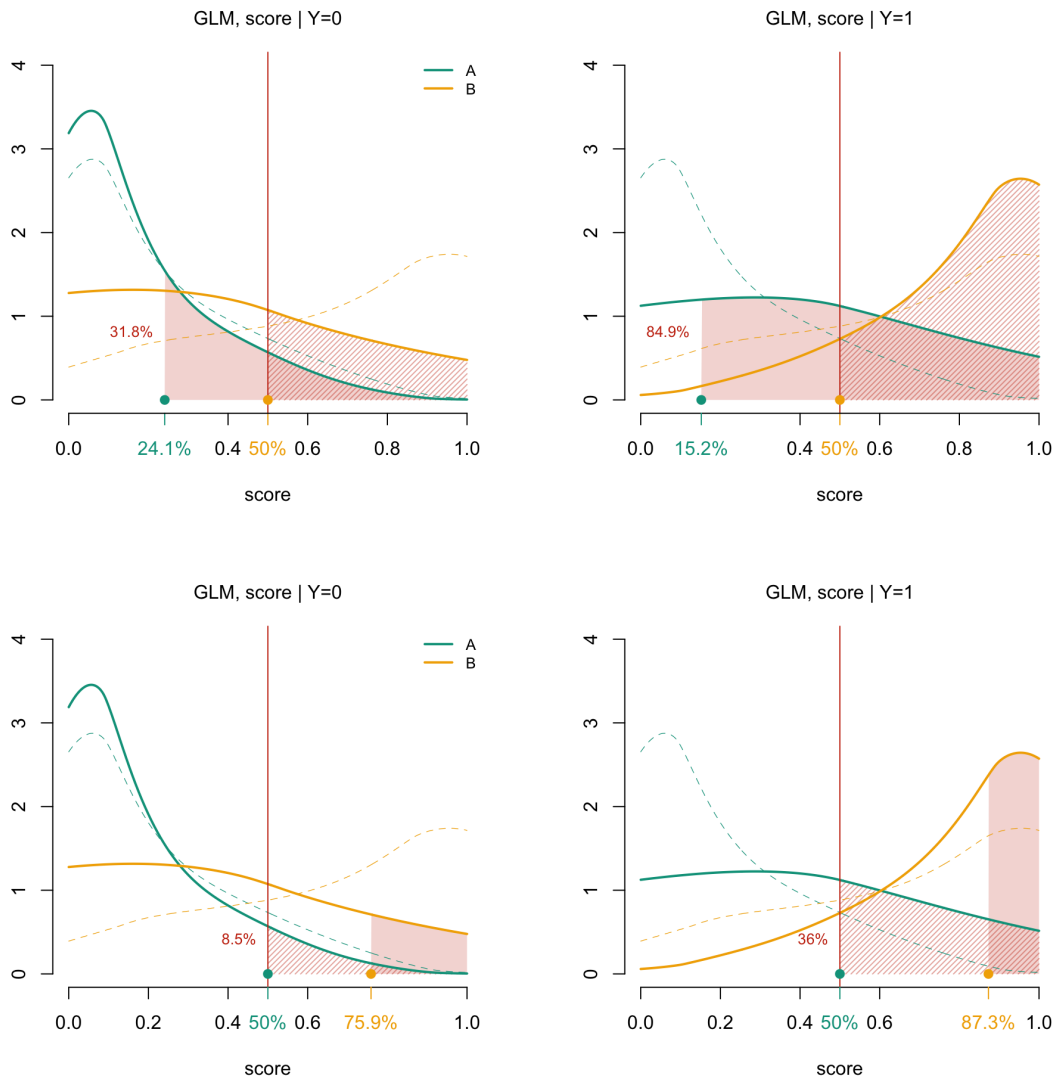


Figure 8.14: Densities of scores in plain strong lines, conditional of $y = 0$ on the left and $y = 1$ on the right, in groups A and B on the large `toydata2` dataset. Dotted lines are unconditional on y. We consider here threshold $t = 50\%$, and positive predictions $\widehat{y} = 1$ or $m_t(\boldsymbol{x}) > t$.

other words, if our model is related to load acceptance, we must have the same proportion of individuals who got offered a loan in each group (advantaged and disadvantaged). For example, here we keep the threshold of 50% on the score for group B (corresponding to a true positive rate of about 31.8%), and we must use a threshold of about 24.1% on the score for group A. On the right, we consider true positive.

Definition 8.3.11 (Equal treatment) *Berk et al. (2021a). We have equality of treatment, the rate of false positives and false negatives are identical in the protected groups,*

$$\frac{\mathbb{P}[\widehat{Y}=1|S=A,Y=0]}{\mathbb{P}[\widehat{Y}=0|S=A,Y=1]}=\frac{\mathbb{P}[\widehat{Y}=1|S=B,Y=0]}{\mathbb{P}[\widehat{Y}=0|S=B,Y=1]}.$$

Berk et al. (2021a) uses the processing term in connection with causal inference which we will discuss next. If the classifier produces more false negatives than false positives for the supposedly privileged group, this means that more disadvantaged individuals receive a favorable outcome than the reverse. A slightly different version had been proposed by Jung et al. (2020),

Definition 8.3.12 (Equalizing Disincentives) *Jung et al. (2020). The difference between the true positive rate and the false positive rate must be the same in the protected groups,*

$$\begin{aligned}&\mathbb{P}[\widehat{Y}=1|S=A,Y=1]-\mathbb{P}[\widehat{Y}=1|S=A,Y=0]\\=\;&\mathbb{P}[\widehat{Y}=1|S=B,Y=1]-\mathbb{P}[\widehat{Y}=1|S=B,Y=0].\end{aligned}$$

Because moving to the third criteria, it should be stressed that "*equalized odds*" might not be a legitimate criteria, because equal error rates could simply reproduce existing biases. If there is a disparity of s in y, then equal error rates could simply reproduce this disparity in the prediction $\widehat{y}$. And therefore, correcting the disparity of s in y actually requires different error rates for different values of s.

8.4 Sufficiency and Calibration

A third commonly used criterion is sometimes called "*sufficiency*", which requires independence between the target Y and the sensitive variable S, conditional either on a given score $m(\boldsymbol{X})$ or prediction $\widehat{Y}$, introduced by Sokolova et al. (2006), and later taken up by Kleinberg et al. (2016), Zafar et al. (2019) and Pleiss et al. (2017). Formally, it means that

$$Y \perp\!\!\!\perp S \mid \widehat{Y} \text{ or } m(\boldsymbol{X}).$$

Definition 8.4.1 (Sufficiency) *Barocas et al. (2017)A model $m : \mathcal{Z} \to \mathcal{Y}$ satisfies the sufficiency property if $Y \perp\!\!\!\perp S \mid m(\boldsymbol{Z})$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$.*

As discussed in Section 4.2 (and Definition 4.3.3) this property is closely related to calibration of the model. For Hedden (2021), it is the only interesting criteria to define fairness with solid philosophical grounds, and Baumann and Loi (2023) relates this criteria to the concept of "actuarial fairness" discussed earlier,

Definition 8.4.2 (Calibration parity, accuracy parity) *Kleinberg et al. (2016), Zafar et al. (2019). We have calibration parity if*

$$\mathbb{P}[Y=1|m(\boldsymbol{X})=t,S=A]=\mathbb{P}[Y=1|m(\boldsymbol{X})=t,S=B],\ \forall t\in[0,1].$$

We can go further by asking for a little more, by asking not only for parity, but also for a good calibration,

Definition 8.4.3 (Good calibration) *Kleinberg et al. (2017), Verma and Rubin (2018) We have fairness of good calibration if*

$$\mathbb{P}[Y=1|m(\boldsymbol{X})=t,S=A]=\mathbb{P}[Y=1|m(\boldsymbol{X})=t,S=B]=t,\ \forall t\in[0,1].$$

In the context of a classifier, instead of conditioning on $m(\boldsymbol{X})$, we can simply use $\widehat{Y}$, as suggested in Chouldechova (2017),

Definition 8.4.4 (Predictive parity (1) - outcome test) *Chouldechova (2017) We have a predictive parity if*

$$\mathbb{P}[Y = 1|\widehat{Y} = 1, S = \texttt{A}] = \mathbb{P}[Y = 1|\widehat{Y} = 1, S = \texttt{B}].$$

Note that if $\widehat{y}$ is not a perfect classifier ($\mathbb{P}[\widehat{Y} \neq Y] > 0$), and if the two groups are not balanced ($\mathbb{P}[S = \texttt{A}] \neq \mathbb{P}[S = \texttt{B}]$), then it is impossible to have predictive parity and equal opportunity at the same time. Note that positive predictive value is

$$\text{PPV}_s = \frac{\text{TPR} \cdot \mathbb{P}[S = s]}{\text{TPR} \cdot \mathbb{P}[S = s] + \text{FPR} \cdot (1 - \mathbb{P}[S = s])}, \ \forall s \in \{\texttt{A}, \texttt{B}\},$$

such that $\text{PPV}_0 = \text{PPV}_1$ implies that either TPR or FPR is zero, and since negative predictive value can be written

$$\text{NPV}_s = \frac{(1 - \text{FPR}) \cdot (1 - \mathbb{P}[S = s])}{(1 - \text{TPR}) \cdot \mathbb{P}[S = s] + (1 - \text{FPR}) \cdot (1 - \mathbb{P}[S = s])}, \ \forall s \in \{\texttt{A}, \texttt{B}\},$$

such that $\text{NPV}_{\texttt{A}} \neq \text{NPV}_{\texttt{B}}$, and predictive parity cannot be achieved.

Continuing the formalism of Chouldechova (2017), Barocas et al. (2019) proposed an extension to predictive parity, with a distinction,

Definition 8.4.5 (Predictive parity (2)) *Barocas et al. (2019)*

$$\begin{cases} \mathbb{P}[Y = 1|S = \texttt{A}, \widehat{Y} = 1] = \mathbb{P}[Y = 1|S = \texttt{B}, \widehat{Y} = 1] & \textit{positive prediction} \\ \mathbb{P}[Y = 1|S = \texttt{A}, \widehat{Y} = 0] = \mathbb{P}[Y = 1|S = \texttt{B}, \widehat{Y} = 0] & \textit{negative prediction} \end{cases}$$

or

$$\mathbb{P}[Y = 1|S = \texttt{A}, \widehat{Y} = \widehat{y}] = \mathbb{P}[Y = 1|S = \texttt{B}, \widehat{Y} = \widehat{y}], \ \forall \widehat{y} \in \{0, 1\}.$$

Finally, let us note that Kleinberg et al. (2017) introduced a notion of balance for positive / negative class.

$$\begin{cases} \mathbb{E}(m(\boldsymbol{X})|Y = 1, S = \texttt{B}) = \mathbb{E}(M|Y = 1, S = \texttt{A}), \text{ balance for positive class} \\ \mathbb{E}(m(\boldsymbol{X})|Y = 0, S = \texttt{B}) = \mathbb{E}(M|Y = 0, S = \texttt{A}), \text{ equilibrium for the negative class.} \end{cases}$$

In Table 8.8, we use the "*accuracy parity metric*" as described by Kleinberg et al. (2016) and Friedler et al. (2019). In groups A and B, accuracy metrics are 80% and 85%, in the $n = 40$ dataset. Therefore, in basis 1 in B, A would have been 0.941. On Figure 8.15, we can look at this metric, as a function of the threshold, when n is very large.

Another approach can be inspired by Kim (2017), for whom, another way to define if a classification is fair, or not, is to say that we cannot tell from the result if the subject was member of a protected group or not. In other words, if an individual's score does not allow us to predict that individual's attributes better than guessing them without any information, we can say that the score was assigned fairly.

Definition 8.4.6 (Non-reconstruction of the protected attribute) *Kim (2017) If we cannot tell from the result ($\boldsymbol{x}$, $m(\boldsymbol{x})$, y and $\widehat{y}$) whether the subject was a member of a protected group or not, we will talk about fairness by non-reconstruction of the protected attribute*

$$\mathbb{P}[S = \texttt{A}|\boldsymbol{X}, m(\boldsymbol{X}), \widehat{Y}, Y] = \mathbb{P}[S = \texttt{B}|\boldsymbol{X}, m(\boldsymbol{X}), \widehat{Y}, Y].$$

	unaware (without s)				aware (with s)			
	GLM	GAM	CART	RF	GLM	GAM	CART	RF
$t = 30\%$	0.933	0.875	1.071	0.833	1.071	1.067	1.071	0.938
$t = 50\%$	0.941	0.882	0.875	0.632	1.000	0.882	0.875	0.737
$t = 70\%$	0.812	0.867	0.647	0.647	0.812	0.688	0.647	0.688

Table 8.8: Accuracy parity on the small subset of `toydata2`, using `acc_parity` from R package `fairness`. 1.071 means that accuracy is 7.1% higher in group `A` than in group `B`.

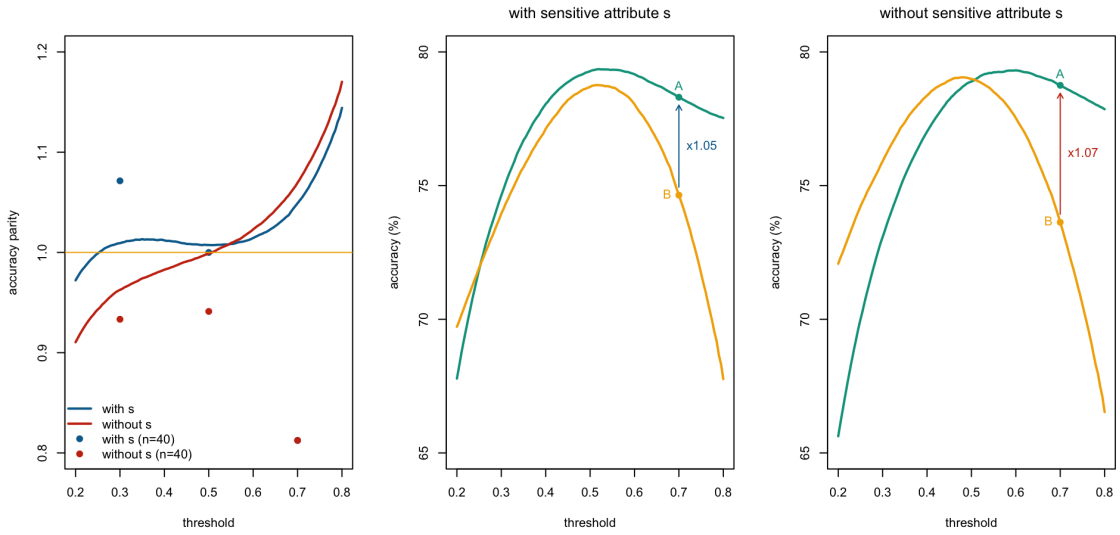


Figure 8.15: On the right, evolution of accuracy, in groups `A` and `B`, for $m_t(\boldsymbol{x})$ – without the sensitive attribute s – as a function of threshold t (on a plain logistic regression), on `toydata2`, using `acc_parity` from R package `fairness`. Here $n = 500,000$ simulated data are considered. In the middle evolution of accuracy, in groups `A` and `B`, for $m_t(\boldsymbol{x}, s)$ – with the sensitive attribute s, as function of t.

8.5 Comparisons and Impossibility Theorems

The different notions of "*group fairness*" can be summarized in the table 8.9. And as we will now see, those notions are incompatible.

Propoition 8.5.1 *Suppose that a model m satisfies the independence condition (8.2) and the sufficiency property (8.4), with respect to a sensitive attribute s, then necessarily, $Y \perp\!\!\!\perp S$.*

From the sufficiency property (8.4), $S \perp\!\!\!\perp Y \mid m(\boldsymbol{Z})$, then, for $s \in \mathcal{S}$ and $\mathcal{A} \subset \mathcal{Y}$,

$$\mathbb{P}[S = s, Y \in \mathcal{A}] = \mathbb{E}\Big[\mathbb{P}[S = s, Y \in \mathcal{A}|m(\boldsymbol{Z})]\Big],$$

can be written

$$\mathbb{P}[S = s, Y \in \mathcal{A}] = \mathbb{E}\Big[\mathbb{P}[S = s|m(\boldsymbol{Z})] \cdot \mathbb{P}[Y \in \mathcal{A}|m(\boldsymbol{Z})]\Big].$$

independence, $\widehat{Y} \perp\!\!\!\perp S$, (8.2.1)

statistical parity	Dwork et al. (2012)	$\mathbb{P}[\widehat{Y} = 1 \vert S = s] = \text{cst}, \ \forall s$	(8.2.2)
conditional statistical parity	Corbett-Davies et al. (2017)	$\mathbb{P}[\widehat{Y} = 1 \vert S = s, X = x] = \text{cst}_x, \ \forall s, y$	(8.2.5)

separation, $\widehat{Y} \perp\!\!\!\perp S \mid Y$, (8.3.1)

equalized odds	Hardt et al. (2016)	$\mathbb{P}[\widehat{Y} = 1 \vert S = s, Y = y] = \text{cst}_y, \ \forall s, y$	(8.3.5)
equalized opportunity	Hardt et al. (2016)	$\mathbb{P}[\widehat{Y} = 1 \vert S = s, Y = 1] = \text{cst}, \ \forall s$	(8.3.2)
predictive equality	Corbett-Davies et al. (2017)	$\mathbb{P}[\widehat{Y} = 1 \vert S = s, Y = 0] = \text{cst}, \ \forall s$	(8.3.4)
balance	Kleinberg et al. (2016)	$\mathbb{E}[M \vert S = s, Y = 1] = \text{cst}_y, \ \forall s$,y	(8.3.6)

sufficiency, $Y \perp\!\!\!\perp S \mid \widehat{Y}$, (8.4.1)

disparate mistreatment	Zafar et al. (2019)	$\mathbb{P}[Y = y \vert S = s, \widehat{Y} = y] = \text{cst}_y, \ \forall s, y$	(8.3.7)
predictive parity	Chouldechova (2017)	$\mathbb{P}[Y = 1 \vert S = s, \widehat{Y} = 1] = \text{cst}, \ \forall s$	(8.4.4)
calibration	Chouldechova (2017)	$\mathbb{P}[Y = 1 \vert M = m, S = s] = \text{cst}_m, \ \forall m, s$	(8.4.2)
well-calibration	Chouldechova (2017)	$\mathbb{P}[Y = 1 \vert M = m, S = s] = s, \ \forall m, s$	(8.4.3)

other

AUC fairness	Borkan et al. (2019)	same AUC in both groups	(8.3.10)
ROC fairness	Vogel et al. (2021)	same ROC curves in both groups	(8.3.9)
non-reconstruction	Kim (2017)	$\mathbb{P}[S = s \vert \boldsymbol{X}, M, \widehat{Y}, Y] = \text{cst}_s, \ \forall s$	(8.4.6)

Table 8.9: Group Fairness definitions, where $M = m(\boldsymbol{Z})$.

And from the independence property (8.4), $m(\boldsymbol{Z}) \perp\!\!\!\perp S$, we can write the first component $\mathbb{P}[S = s \vert m(\boldsymbol{Z})] = \mathbb{P}[S = s]$, almost surely, and therefore

$$\mathbb{P}[S = s, Y \in \mathcal{A}] = \mathbb{E}\big[\mathbb{P}[S = s] \cdot \mathbb{P}[Y \in \mathcal{A} \vert m(\boldsymbol{Z})]\big] = \mathbb{P}[S = s] \cdot \mathbb{P}[Y \in \mathcal{A}\big],$$

for all $s \in \mathcal{S}$ and $\mathcal{A} \subset \mathcal{Y}$, corresponding to the independence between S and Y.

Therefore, unless the sensitive attribute s has no impact on the outcome y, there is no model $m(\cdot)$ which satisfies independence and sufficiency simultaneously.

Propoition 8.5.2 *Consider a classifier $m_t(\cdot)$ taking values in $\mathcal{Y} = \{0, 1\}$. Suppose that m_t satisfies the independence condition (8.2) and the separation property (8.3), with respect to a sensitive attribute s, then necessarily either $m_t(\boldsymbol{Z}) \perp\!\!\!\perp Y$ or $Y \perp\!\!\!\perp S$ (possibly both).*

Because m_t satisfies the independence condition (8.2), $m_t(\boldsymbol{Z}) \perp\!\!\!\perp S$, and the separation property (8.3), $m_t(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y$, them, for $\widehat{y} \in \mathcal{Y}$ and for $s \in \mathcal{S}$,

$$\mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}] = \mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y} \vert S = s] = \mathbb{E}\big[\mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y} \vert Y, S = s]\big],$$

that we can write

$$\mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}] = \sum_y \mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y} \vert Y = y, S = s\big] \cdot \mathbb{P}\big[Y = y \big\vert S = s\big],$$

or

$$\mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}] = \sum_y \mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y} \vert Y = y\big] \cdot \mathbb{P}\big[Y = y \big\vert S = s\big],$$

almost surely. Furthermore, we can also write

$$\mathbb{P}[m_t(\boldsymbol{Z}) = \widehat{y}] = \sum_y \mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y}|Y = y\big] \cdot \mathbb{P}\big[Y = y\big],$$

so that, if we combine the two expressions, we get

$$\sum_y \mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y}|Y = y\big] \cdot \Big(\mathbb{P}\big[Y = y\big|S = s\big] - \mathbb{P}\big[Y = y\big]\Big) = 0,$$

almost surely. And since we assumed that y was a binary variable, $\mathbb{P}[Y = 0] = 1 - \mathbb{P}[Y = 1]$, as well as $\mathbb{P}[Y = 0|S = s] = 1 - \mathbb{P}[Y = 1|S = s]$, and therefore

$$\mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y}|Y = 1\big] \cdot \Big(\mathbb{P}\big[Y = 1\big|S = s\big] - \mathbb{P}\big[Y = 1\big]\Big)$$

or

$$-\mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y}|Y = 0\big] \cdot \Big(\mathbb{P}\big[Y = 0\big|S = s\big] - \mathbb{P}\big[Y = 0\big]\Big)$$

can be written

$$\mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y}|Y = 0\big] \cdot \Big(\mathbb{P}\big[Y = 1\big|S = s\big] - \mathbb{P}\big[Y = 1\big]\Big).$$

Thus, either $\mathbb{P}\big[Y = 1\big|S = s\big] - \mathbb{P}\big[Y = 1\big]$ almost surely, or $\mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y}|Y = 0\big] = \mathbb{P}\big[m_t(\boldsymbol{Z}) = \widehat{y}|Y = 1\big]$ (or both).

Of course, the previous proposition holds only when y is a binary variable.

Propoition 8.5.3 *Consider a classifier $m_t(\cdot)$ taking values in $\mathcal{Y} = \{0, 1\}$. Suppose that m_t satisfies the sufficiency condition (8.4) and the separation property (8.3), with respect to a sensitive attribute s, then necessarily either $\mathbb{P}[\gtrapprox(\boldsymbol{Z}) = \lessapprox|\mathbb{Y} = \lessapprox] = 0$ or $Y \perp\!\!\!\perp S$ or $m_t(\boldsymbol{Z} \perp\!\!\!\perp Y$.*

Suppose that m_t satisfies the sufficiency condition (8.4) and the separation property (8.3), respectively $Y \perp\!\!\!\perp S \mid m_t(\boldsymbol{Z})$ and $m_t(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y$. For all $s \in \mathcal{S}$, we can write, using Bayes formula

$$\mathbb{P}[Y = 1|S = s, m_t(\boldsymbol{Z}) = 1] = \frac{\mathbb{P}[m_t(\boldsymbol{Z}) = 1|Y = 1, S = s] \cdot \mathbb{P}[Y = 1|S = s]}{\mathbb{P}[m_t(\boldsymbol{Z}) = 1|S = s]},$$

i.e.

$$\mathbb{P}[Y = 1|S = s, m_t(\boldsymbol{Z}) = 1] = \frac{\mathbb{P}[m_t(\boldsymbol{Z}) = 1|Y = 1] \cdot \mathbb{P}[Y = 1|S = s]}{\displaystyle\sum_{y \in \{0,1\}} \mathbb{P}[m_t(\boldsymbol{Z}) = 1|Y = y] \cdot \mathbb{P}[Y = 1|S = s]},$$

that should not depend on s (from the sufficiency property). So a similar property holds if $S = s'$. Observe further that $\mathbb{P}[m_t(\boldsymbol{Z}) = 1|Y = 1]$ is the *true positive rate* (TPR) while $\mathbb{P}[m_t(\boldsymbol{Z}) = 1|Y = 0]$ is the *false positive rate* (TPR). Let $p_s = \mathbb{P}[Y = 1|S = s]$, so that

$$\mathbb{P}[Y = 1|S = s, m_t(\boldsymbol{Z}) = 1] = \frac{\text{TPR}}{p_s \cdot \text{TPR} + (1 - p_s) \cdot \text{FPR}}.$$

Supposes that Y and S are not independent (otherwise $Y \perp\!\!\!\perp S$ as stated in the proposition), i.e. there are s and s' such that $p_s = \mathbb{P}[Y = 1|S = s] \neq \mathbb{P}[Y = 1|S = s'] = p_{s'}$. Hence, $p_s \neq p_{s'}$, but at the same time

$$\frac{\text{TPR}}{p_s \cdot \text{TPR} + (1 - p_s) \cdot \text{FPR}} = \frac{\text{TPR}}{p_{s'} \cdot \text{TPR} + (1 - p_{s'}) \cdot \text{FPR}}.$$

Supposes that TPR $\neq 0$ (otherwise TPR $= \mathbb{P}[\succcurlyeq_{\approx}(\boldsymbol{Z}) = \nVdash | \mathbb{Y} = \nVdash] = 0$ as stated in the proposition), then

$$(p_s - p_{s'}) \cdot \text{TPR} = (p_s - p_{s'}) \cdot \text{FPR} \neq 0,$$

and therefore $m_t(\boldsymbol{Z}) \perp\!\!\!\perp Y$.

So, to summarize that section, unless very specific properties are assumed on $\mathbb{P}$, there is no prediction function $m(\cdot)$ that can satisfy at the same time two fairness criteria.

8.6 Relaxation and Confidence Intervals

We had seen that the demographic fairness is translated by the equality

$$\frac{\mathbb{P}[\widehat{Y} = 1 | S = \texttt{A}]}{\mathbb{P}[\widehat{Y} = 1 | S = \texttt{B}]} = 1 = \frac{\mathbb{P}[\widehat{Y} = 1 | S = \texttt{B}]}{\mathbb{P}[\widehat{Y} = 1 | S = \texttt{A}]}$$

If this approach is interesting, the statistical reality is that having a perfect equality between two (predictive) probabilities is usually impossible. It is actually possible to relax that equality, as follows,

Definition 8.6.1 (Disparate impact) *Feldman et al. (2015). A decision function $\widehat{Y}$ has a disparate impact, for a given threshold τ, if,*

$$\min\left\{ \frac{\mathbb{P}[\widehat{Y} = 1 | S = \texttt{A}]}{\mathbb{P}[\widehat{Y} = 1 | S = \texttt{B}]}, \frac{\mathbb{P}[\widehat{Y} = 1 | S = \texttt{B}]}{\mathbb{P}[\widehat{Y} = 1 | S = \texttt{A}]} \right\} < \tau \text{ (usually 80\%).}$$

This so-called "*four-fifths rule*", coupled with the $\tau = 80\%$ threshold, was originally defined by the State of California Fair Employment Practice Commission (FEPC) Technical Advisory Committee on Testing, which issued the California State Guidelines on Employee Selection Procedures in October 1972, as recalled in Feldman et al. (2015), Mercat-Bruns (2016), Biddle (2017) or Lipton et al. (2018). This standard was later adopted in the 1978 Uniform Guidelines on Employee Selection Procedures, used by the Equal Employment Opportunity Commission (EEOC), the U.S. Department of Labor, and the U.S. Department of Justice. An important point here is that this form of discrimination occurred even when the employer did not intend to discriminate, but by looking at employment statistics (on gender or racial grounds), it was possible to observe (and correct) discriminatory bias.

For example, on the `toydataset2` with $n = 1000$ individuals,

$$\frac{\mathbb{P}[\widehat{Y} = 1 | S = \texttt{A}]}{\mathbb{P}[\widehat{Y} = 1 | S = \texttt{B}]} = \frac{134}{270}\frac{400}{600} = \frac{134}{405} \sim 33.1\% \ll 80\%.$$

Another approach, suggested to relax the equality $\mathbb{P}(\widehat{Y} = 1 | S = \texttt{A}) = \mathbb{P}(\widehat{Y} = 1 | S = \texttt{B})$, consists in introducing a notion of "*approximate-fairness*" in Holzer and Neumark (2000), Collins (2007) and Feldman et al. (2015), or ε-fairness in Hu (2022)

$$\left| \mathbb{P}(\widehat{Y} = 1 | S = \texttt{A}) - \mathbb{P}(\widehat{Y} = 1 | S = \texttt{B}) \right| < \varepsilon.$$

The left deviation is sometimes called "*statistical parity difference*" (SPD). Žliobaite (2015) suggests normalizing the statistical parity difference,

$$\text{NSPD}\frac{\text{SPD}}{D_{\max}} \text{ where } D_{\max} = \min\left\{ \frac{\mathbb{P}(\widehat{Y} = 1)}{\mathbb{P}(S = \texttt{B})}, \frac{\mathbb{P}(\widehat{Y} = 0)}{\mathbb{P}(S = \texttt{A})} \right\},$$

so that NSPD = 1 for maximum discrimination (otherwise NSPD < 1).

For strong concepts of fairness, we can use distances (or divergences) between distributions, as in Proportion 8.2.1, using Wasserstein distance,

$$W_2(\mathbb{P}_{\mathtt{A}}, \mathbb{P}_{\mathtt{B}}) < \varepsilon,$$

or Proposition 8.2.2, using Kullback-Leibler divergence,

$$D_{\mathrm{KL}}(\mathbb{P}_{\mathtt{A}} \| \mathbb{P}) + D_{\mathrm{KL}}(\mathbb{P}_{\mathtt{B}} \| \mathbb{P}) < \varepsilon.$$

Besse et al. (2018) considered another approach, based on confidence intervals for fairness criteria. For example, for the disparate impact, we have seen that we should calculate

$$T = \frac{\mathbb{P}[\widehat{Y} = 1 | S = \mathtt{A}]}{\mathbb{P}[\widehat{Y} = 1 | S = \mathtt{B}]},$$

whose empirical version is

$$T_n = \frac{\sum_i \widehat{y}_i \mathbf{1}(s_i = \mathtt{A})}{\sum_i \widehat{y}_i \mathbf{1}(s_i = \mathtt{B})} \cdot \frac{\sum_i \mathbf{1}(s_i = \mathtt{B})}{\sum_i \mathbf{1}(s_i = \mathtt{A})},$$

which can be used to construct a confidence interval for T (Besse et al. (2018) proposes an asymptotic test, but using resampling techniques is also possible).

8.7 Using Decomposition and Regressions

If the three previous approaches are now quite popular in machine learning literature (independence, separation and sufficiency), other techniques to quantify discrimination have been introduced in econometrics. A classical case in labor economics is the gender wage gap. Such a gap has been observed for decades and economists have tried to explain the difference in average wages between men and women. In a nutshell, as in insurance, such as gap could be a "fair demographic discrimination" if there were group differences in wage determinants, that is in characteristics that are relevant for wages, such as education. This is called "*compositional differences*". But that gap can also be associated with "*unfair discrimination*", if there was a differential compensation for these determinants, such as different returns to education for men an women. This is called "*differential mechanisms*". In order to construct a counterfactual state – to answer a question "*what the wage would be if women had the same characteristics as men?*" – economists have considered decomposition method, to recovery the causal effect of a sensitive attribute. Cain (1986) and Fortin et al. (2011) provided state-of-the-art on those techniques.

If the seminal work by Kitagawa (1955) and Solow (1957) introduced the "decomposition method", Oaxaca (1973) and Blinder (1973) have laid the foundations of the decomposition approach to analyze mean wage differences between groups, based either on gender, or race. This approach is a naturally way to disentangle cause and effect in the context of linear models. Consider here a regression model,

$$y_i = \gamma \mathbf{1}_{\mathtt{B}}(s_i) + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,$$

where y is the salary, s denotes the binary sensitive attribute $\mathbf{1}_{\mathtt{B}}(s)$, $\boldsymbol{x}$ is a collection of predictive variable (or control variables, that might have an influence on the salary, such as education or work experience). In such a model, $\widehat{\gamma}$ can be use to answer the question we asked earlier "*what the wage y would be if women ($s = \mathtt{B}$)*

had the same characteristics $\boldsymbol{x}$ *as men* $(s = \texttt{A})$*?*". To introduce the Oaxaca-Blinder approach, suppose that we consider two (linear) models, for men and women respectively,

$$\begin{cases} y_{\texttt{A}:i} = \boldsymbol{x}_{\texttt{A}:i}^\top \boldsymbol{\beta}_{\texttt{A}} + \varepsilon_{\texttt{A}:i} \ \text{(group A)} \\ y_{\texttt{B}:i} = \boldsymbol{x}_{\texttt{B}:i}^\top \boldsymbol{\beta}_{\texttt{B}} + \varepsilon_{\texttt{B}:i} \ \text{(group B)}. \end{cases}$$

Using ordinary least squares estimates (and standard properties of linear models),

$$\begin{cases} \overline{y}_{\texttt{A}} = \overline{\boldsymbol{x}}_{\texttt{A}}^\top \widehat{\boldsymbol{\beta}}_{\texttt{A}} \ \text{(group A)} \\ \overline{y}_{\texttt{B}} = \overline{\boldsymbol{x}}_{\texttt{B}}^\top \widehat{\boldsymbol{\beta}}_{\texttt{B}} \ \text{(group B)} \end{cases},$$

so that $\overline{y}_{\texttt{A}} - \overline{y}_{\texttt{B}} = \overline{\boldsymbol{x}}_{\texttt{A}}^\top \widehat{\boldsymbol{\beta}}_{\texttt{A}} - \overline{\boldsymbol{x}}_{\texttt{B}}^\top \widehat{\boldsymbol{\beta}}_{\texttt{B}}$, that we can write (by adding and removing $\overline{\boldsymbol{x}}_{\texttt{A}}^\top \widehat{\boldsymbol{\beta}}_{\texttt{B}}$)

$$\overline{y}_{\texttt{A}} - \overline{y}_{\texttt{B}} = \underbrace{\left(\overline{\boldsymbol{x}}_{\texttt{A}} - \overline{\boldsymbol{x}}_{\texttt{B}}\right)^\top \widehat{\boldsymbol{\beta}}_{\texttt{B}}}_{\text{characteristics}} + \underbrace{\overline{\boldsymbol{x}}_{\texttt{A}}^\top \left(\widehat{\boldsymbol{\beta}}_{\texttt{A}} - \widehat{\boldsymbol{\beta}}_{\texttt{B}}\right)}_{\text{coefficients}}, \tag{8.1}$$

where the first term is the "characteristics effect", which describes how much the difference of outcome y (on average) is due to the differences in the levels of explanatory variables $\boldsymbol{x}$, while the second term is the "coefficients effect" which describes how much the difference of outcome y (on average) is due to differences in the magnitude of regression coefficients $\boldsymbol{\beta}$. The first one is the legitimate component, also called "*endowment effect*" in Woodhams et al. (2021) or "composition effect" in Hsee and Li (2022), while the second one can be interpreted as some illegitimate discrimination, is called "returns effect" in Agrawal (2013) or "structure effect" in Firpo (2017). For the first component,

$$\left(\overline{\boldsymbol{x}}_{\texttt{A}} - \overline{\boldsymbol{x}}_{\texttt{B}}\right)^\top \widehat{\boldsymbol{\beta}}_{\texttt{B}} = \left(\overline{x}_{\texttt{A}:1} - \overline{x}_{\texttt{B}:1}\right)^\top \widehat{\beta}_{\texttt{B}:1} + \cdots + \left(\overline{x}_{\texttt{A}:j} - \overline{x}_{\texttt{B}:j}\right)^\top \widehat{\beta}_{\texttt{B}:j} + \cdots + \left(\overline{x}_{\texttt{A}:p} - \overline{x}_{\texttt{B}:p}\right)^\top \widehat{\beta}_{\texttt{B}:p}$$

were, for the j-th term, we explicitly see the average difference in the two groups, $\left(\overline{x}_{\texttt{A}:j} - \overline{x}_{\texttt{B}:j}\right)$, while for the second component

$$\overline{\boldsymbol{x}}_{\texttt{A}}^\top \left(\widehat{\boldsymbol{\beta}}_{\texttt{A}} - \widehat{\boldsymbol{\beta}}_{\texttt{B}}\right) = \overline{x}_{\texttt{A}:1}^\top \left(\widehat{\beta}_{\texttt{A}:1} - \widehat{\beta}_{\texttt{B}:1}\right) + \cdots + \overline{x}_{\texttt{A}:j}^\top \left(\widehat{\beta}_{\texttt{A}:j} - \widehat{\beta}_{\texttt{B}:j}\right) + \cdots + \overline{x}_{\texttt{A}:p}^\top \left(\widehat{\beta}_{\texttt{A}:p} - \widehat{\beta}_{\texttt{B}:p}\right)$$

But similarly, we could have written,

$$\overline{y}_{\texttt{A}} - \overline{y}_{\texttt{B}} = \underbrace{\left(\overline{\boldsymbol{x}}_{\texttt{A}} - \overline{\boldsymbol{x}}_{\texttt{B}}\right)^\top \widehat{\boldsymbol{\beta}}_{\texttt{A}}}_{\text{characteristics}} + \underbrace{\overline{\boldsymbol{x}}_{\texttt{B}}^\top \left(\widehat{\boldsymbol{\beta}}_{\texttt{A}} - \widehat{\boldsymbol{\beta}}_{\texttt{B}}\right)}_{\text{coefficients}} \tag{8.2}$$

which is the analogous of the previous decomposition, in Equation 8.1, but can be rather different. This is more or less the same as the regression on categorical variables: changing the reference does not change the prediction, only the interpretation of coefficient. To visualize it, consider the case were the is only one single characteristics x, as on Figure 8.16.

In the context of categorical variable, it is rather common to use a contrast approach where all quantities are expressed with respect to some "*average benchmark*". We will do the same here, except that the "*average benchmark*" is now a "*fair benchmark*". So suppose that it could be possible to have a nondiscriminatory (potential) outcome $y^\star$, so that

$$y^\star = \boldsymbol{x}^\top \boldsymbol{\beta}^\star + \varepsilon,$$

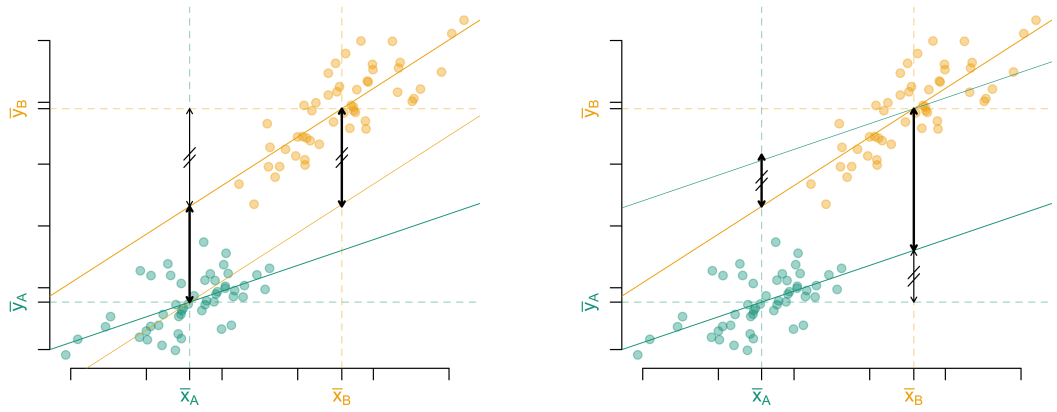


Figure 8.16: Visualization of Oaxaca-Blinder decomposition of $\bar{y}_A - \bar{y}_B$, as $x_A(\widehat{\beta}_A - \widehat{\beta}_B)$ and $(\bar{x}_A - \bar{x}_B)\widehat{\beta}_B$ (as in Equation 8.1) on the left, and $x_B(\widehat{\beta}_A - \widehat{\beta}_B)$ and $(\bar{x}_A - \bar{x}_B)\widehat{\beta}_A$ (as in Equation 8.2) on the right. (fictitious data)

where $\mathbb{E}[\varepsilon|\boldsymbol{X}] = 0$. Then we could write, non ambiguously

$$\bar{y}_A - \bar{y}_B = \underbrace{(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_B)^\top \widehat{\boldsymbol{\beta}}^\star}_{\text{characteristics}} + \Big(\underbrace{\bar{\boldsymbol{x}}_A^\top (\widehat{\boldsymbol{\beta}}_A - \widehat{\boldsymbol{\beta}}^\star)}_{\text{coefficients (A)}} + \underbrace{\bar{\boldsymbol{x}}_B^\top (\widehat{\boldsymbol{\beta}}^\star - \widehat{\boldsymbol{\beta}}_B)}_{\text{coefficients (B)}} \Big), \tag{8.3}$$

where the "coefficient effect" component is now decomposed in two part: an illegitimate discrimination *in favor of* group A, and an illegitimate discrimination *against* group B. This approach can be used to get a better interpretation of the first two models. In fact, if we assume that there is only discrimination against group B, and no discrimination on favor of group 0, then $\boldsymbol{\beta}^\star = \boldsymbol{\beta}_A$ and we obtain Equation (8.1), while if we assume that there is only discrimination in favor of group A, and no discrimination on against group 1, then $\boldsymbol{\beta}^\star = \boldsymbol{\beta}_B$ and we obtain Equation (8.2).

As for the contrast approach, one can consider an average approach for the fair benchmark. Reimers (1983) suggested to consider the average coefficient between the two groups,

$$\widehat{\boldsymbol{\beta}}^\star = \frac{1}{2}\widehat{\boldsymbol{\beta}}_A + \frac{1}{2}\widehat{\boldsymbol{\beta}}_B,$$

while Cotton (1988) suggested a weighted average, based on population sizes in the two group,

$$\widehat{\boldsymbol{\beta}}^\star = \frac{n_A}{n_A + n_B}\widehat{\boldsymbol{\beta}}_A + \frac{n_B}{n_A + n_B}\widehat{\boldsymbol{\beta}}_B.$$

And quite naturally, consider

$$\widehat{\boldsymbol{\beta}}^\star_\omega = \omega\widehat{\boldsymbol{\beta}}_A + (1-\omega)\widehat{\boldsymbol{\beta}}_B, \quad \text{where } \omega \in [0,1].$$

Then we can write Equation (8.3) as

$$\overline{y}_{\mathtt{A}} - \overline{y}_{\mathtt{B}} = \left(\overline{\boldsymbol{x}}_{\mathtt{A}} - \overline{\boldsymbol{x}}_{\mathtt{B}}\right)^\top \left(\omega\widehat{\boldsymbol{\beta}}_{\mathtt{A}} + (1-\omega)\widehat{\boldsymbol{\beta}}_{\mathtt{B}}\right) + \left((1-\omega)\overline{\boldsymbol{x}}_{\mathtt{A}} + \omega\overline{\boldsymbol{x}}_{\mathtt{B}}\right)^\top \left(\widehat{\boldsymbol{\beta}}_{\mathtt{A}} - \widehat{\boldsymbol{\beta}}_{\mathtt{B}}\right) \tag{8.4}$$

or more generally

$$\overline{y}_{\mathtt{A}} - \overline{y}_{\mathtt{B}} = \left(\overline{\boldsymbol{x}}_{\mathtt{A}} - \overline{\boldsymbol{x}}_{\mathtt{B}}\right)^\top \left(\Omega\widehat{\boldsymbol{\beta}}_{\mathtt{A}} + (\boldsymbol{I}-\Omega)\widehat{\boldsymbol{\beta}}_{\mathtt{B}}\right) + \left((\boldsymbol{I}-\Omega)\overline{\boldsymbol{x}}_{\mathtt{A}} + \Omega\overline{\boldsymbol{x}}_{\mathtt{B}}\right)^\top \left(\widehat{\boldsymbol{\beta}}_{\mathtt{A}} - \widehat{\boldsymbol{\beta}}_{\mathtt{B}}\right) \tag{8.5}$$

for some $p \times p$ matrix Ω, that corresponds to relative weights given to the coefficients of group 0. Oaxaca and Ransom (1994) suggested to use

$$\Omega = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}_{\mathtt{A}}^\top\boldsymbol{X}_{\mathtt{A}}).$$

But this approach suffers several drawbacks, related toe the "*omitted variable bias*" discussed in section 5.6, see also Jargowsky (2005) or more recently Cinelli and Hazlett (2020). Therefore, Jann (2008) suggested simply to use a pooled regression over the two groups, controlled by the group membership, that is

$$y_i = \gamma\mathbf{1}_{\mathtt{B}}(s_i) + \boldsymbol{x}_i^\top\boldsymbol{\beta}^\star + \varepsilon_i$$

so that, with heuristically standard matrix notations,

$$\widehat{\boldsymbol{\beta}}^\star = \left([\boldsymbol{X},\boldsymbol{s}]^\top[\boldsymbol{X},\boldsymbol{s}]\right)^{-1}[\boldsymbol{X},\boldsymbol{s}]^\top\boldsymbol{y}.$$

As in Brown et al. (1980), consider now some binary sensitive attribute s, and possible some class x, corresponding to the industry. This corresponds to the framework of Simpson's paradox discussed in Section 5.6, where y is the acceptance in a graduate program, s the gender of the candidate, and x the program. Here, y is the wage, s the gender, and x the industry. Let $p_{s:j}$ denote the probability for a person of gender s to enter industry j. Then

$$\overline{y}_s = \sum_j p_{s:j}\overline{y}_{s:j},$$

so that the average wage gap between men and women in the labor market is

$$\overline{y}_{\mathtt{B}} - \overline{y}_{\mathtt{A}} = \sum_j \left(p_{\mathtt{B}:j}\overline{y}_{\mathtt{B}:j} - p_{\mathtt{A}:j}\overline{y}_{\mathtt{A}:j}\right),$$

and we will decompose this wage gap into industry wage differentials, and the probability of entering a certain industry. With personal data, one can consider some multinomial logit-model, so that the probability for individual with characteristics $\boldsymbol{x}_i$ joins industry j would be

$$p_{j,i} = \frac{\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}.$$

Such a model can be estimated for men and women, independently,

$$p_{s:j,i} = \frac{\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}_s)}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}_s)}.$$

For a woman ($s = \mathtt{B}$) with characteristics $\boldsymbol{x}_i$, define the "non-discriminatory" probability (or probability of working in industry j if that person had been treated as a men) as

$$p^\star_{\mathtt{B}:j,i} = \frac{\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{\mathtt{A}})}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}_{\mathtt{A}})},$$

and decompose the average wage gap

$$\overline{y}_{\mathtt{B}} - \overline{y}_{\mathtt{A}} = \sum_j \left(p_{\mathtt{B}:j} \overline{y}_{\mathtt{B}:j} - p_{\mathtt{A}:j} \overline{y}_{\mathtt{A}:j} \right) = \sum_j d_j,$$

in two parts,

$$d_j = \left([p_{\mathtt{B}:j} - p_{\mathtt{A}:j}] \overline{y}_{\mathtt{B}:j} + p_{\mathtt{A}:j} [\overline{y}_{\mathtt{B}:j} - \overline{y}_{\mathtt{A}:j}] \right).$$

The term can also be decomposed as

$$\left([p_{\mathtt{B}:j} - p_{\mathtt{A}:j}] \overline{y}_{\mathtt{B}:j} \right) = \left[(p_{\mathtt{B}:j} - p^{\star}_{\mathtt{A}:j}) + (p^{\star}_{\mathtt{A}:j} - p_{\mathtt{A}:j}) \right] \overline{y}_{\mathtt{B}:j}.$$

And similarly, it is possible to use Oaxaca-Blinder decomposition for the second term, since

$$[\overline{y}_{\mathtt{B}:j} - \overline{y}_{\mathtt{A}:j}] = [\overline{\boldsymbol{x}}^{\top}_{\mathtt{B}:j} \widehat{\boldsymbol{\beta}}_{\mathtt{B}:j} - \overline{\boldsymbol{x}}^{\top}_{\mathtt{A}:j} \widehat{\boldsymbol{\beta}}_{\mathtt{A}:j}],$$

i.e.

$$[\overline{\boldsymbol{x}}_{\mathtt{B}:j} - \overline{\boldsymbol{x}}_{\mathtt{A}:j}]^{\top} \widehat{\boldsymbol{\beta}}_{\mathtt{B}:j} + \overline{\boldsymbol{x}}^{\top}_{\mathtt{A}:j} [\widehat{\boldsymbol{\beta}}_{\mathtt{B}:j} - \widehat{\boldsymbol{\beta}}_{\mathtt{A}:j}].$$

And therefore, d_j (and $\overline{y}_{\mathtt{B}} - \overline{y}_{\mathtt{A}}$) can be decomposed intro four components

$$\begin{aligned} d_j &= p_{\mathtt{A}:j} [\overline{\boldsymbol{x}}_{\mathtt{B}:j} - \overline{\boldsymbol{x}}_{\mathtt{A}:j}]^{\top} \widehat{\boldsymbol{\beta}}_{\mathtt{B}:j} \text{ legitimate difference within industries} \\ &+ p_{\mathtt{A}:j} \overline{\boldsymbol{x}}^{\top}_{\mathtt{A}:j} [\widehat{\boldsymbol{\beta}}_{\mathtt{B}:j} - \widehat{\boldsymbol{\beta}}_{\mathtt{A}:j}] \text{ illegitimate within industries} \\ &+ [(p_{\mathtt{B}:j} - p^{\star}_{\mathtt{A}:j})] \left] \overline{y}_{\mathtt{B}:j} \right. \text{ legitimate across industries} \\ &+ [(p^{\star}_{\mathtt{A}:j} - p_{\mathtt{A}:j}] \overline{y}_{\mathtt{B}:j} \text{ illegitimate across industries.} \end{aligned}$$

If x is no-longer defined on classes, DiNardo et al. (1995) suggested a continuous version, where the vector of probabilities $p_{s:j}$ has now a continuous density f_s that can be estimated using kernel density estimates.

An alternative is to consider so-called "*direct and reverse regressions*", to quantify discrimination. In the context of labor market discrimination, Conway and Roberts (1983) considered some data, where y denotes the income, a job offer or a promotion. The protected variable p is either the gender or the indicator associated with racial minorities (and will be binary). And $\boldsymbol{x}$ denotes some information about job qualifications, that will be univariate here, and is considered to be an imperfect measure of actual productivity ("*we shall mainly think of x as a job qualification rather than as a proxy for productivity*"). When performing a regression of y on x and s, the regression coefficient of s estimates mean salary differences between females and males after statistical allowance for the measured qualifications, x. Such a regression corresponds to the idea that discrimination has to do with disparity in mean salaries, for given measured job qualification. Conway and Roberts (1983) considered another type of possible discrimination, called is "*placement discrimination* », which refers to the "*shunting*" or "*steering*" of females (or minorities) into lower job levels than their qualifications warrant. Therefore, Conway and Roberts (1983) claims that it is more natural to compare the average qualifications of males and females within each job group, that is, to regress x on y and s, so that discrimination will correspond to disparities in mean measured qualifications for given entering job groups. Hence, the conditional distribution of y given x is named "*direct regression method*", while the one based on the conditional distribution of x given y is the "*reverse regression method*" . The first one was considered in Finkelstein (1980), Fisher (1980), or Weisberg and Tomberlin (1982), that discuss the use of linear regression

models in legal cases of employment discrimination. And later, the idea was discussed in Ferber and Green (1982a,b), Kamalich and Polachek (1982), Goldberger (1984), Greene (1984), Michelson and Blattenberger (1984), in the early 80's.

Those two approaches provide two different perspectives on fairness: fairness in the sense that the distributions of male and female incomes are the same at given qualifications will be obtained using the direct regression, while fairness in the sense that the distributions of male and female qualifications are the same at given incomes will corresponds to reverse regression.

$$\begin{cases}\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}] = \beta_0 + \boldsymbol{x}^\top\boldsymbol{\beta}_1 + \beta_2 p \\ \mathbb{E}[X|Y=y] = \alpha_0 + \alpha_1 y + \alpha_2 p\end{cases}$$

or

$$\begin{cases}y_i = \beta_0 + \boldsymbol{x}^\top\boldsymbol{\beta}_1 + \beta_2 p + \varepsilon_i \\ x_i = \alpha_0 + \alpha_1 y_i + \alpha_2 p_i + \eta_i\end{cases}$$

The idea underlying both proposals is the intuitively appealing one that if the protected class is underpaid, that is, $\beta_2 < 0$, then we should expect its members to be overqualified in the jobs they hold, that is, $\alpha_2 > 0$. Kamalich and Polachek (1982) suggested a multivariate extension, with

$$\begin{cases}x_{1,i} = \alpha_{1,0} + \alpha_{1,1}y_i + \alpha_{1,2}p_i + \alpha_{1,3,2}x_{2,i} + \alpha_{1,3,3}x_{3,i} + \cdots + \alpha_{1,3,k}x_{k,i} + \eta_{1,i} \\ x_{2,i} = \alpha_{2,0} + \alpha_{2,1}y_i + \alpha_{2,2}p_i + \alpha_{2,3,1}x_{1,i} + \alpha_{2,3,3}x_{3,i} + \cdots + \alpha_{2,3,k}x_{k,i} + \eta_{2,i} \\ \quad\vdots \\ x_{p,i} = \alpha_{p,0} + \alpha_{k,1}y_i + \alpha_{k,2}p_i + \alpha_{k,3,1}x_{1,i} + \alpha_{k,3,2}x_{2,i} + \cdots + \alpha_{k,3,k-1}x_{k-1,i} + \eta_{k,i}\end{cases}$$

while Conway and Roberts (1983) suggested some aggregated index of $\boldsymbol{x}$. More precisely,

$$\begin{cases}y_i = \overbrace{\beta_0 + \boldsymbol{x}^\top\boldsymbol{\beta}_1}^{z_i} + \beta_2 p + \varepsilon_i \\ z_i = \alpha_0 + \alpha_1 y_i + \alpha_2 p_i + \eta_i\end{cases}$$

Greene (1984) observed that

$$\widehat{\alpha}_2 = \frac{n}{n - n_{\mathtt{A}}} \cdot (\overline{y}_{\mathtt{A}} - \overline{y})(1 - R^2) - \widehat{\beta}_2,$$

As mentioned by Conway and Roberts (1983), "*because of the conditioning on salary, reverse regression exposes features of the data that might go undetected by the use of direct regression alone. Comparison of employee qualifications for given jobs or salaries is especially pertinent when one is trying to detect shunting, which is usually thought of as placement of members of protected classes into lower job groups than would be suggested by their qualifications*".

Racine and Rilstone (1995) suggested to use nonparametric regression to avoid possible misspecifications. Indeed, if

$$\begin{cases}y_i = g(x_i) + \beta_2 s \\ x_i = h(y_i) + \alpha_2 s\end{cases}$$

that can be written

$$\begin{cases}y_i - \mathbb{E}[Y|X=x_i] = \beta_2(s_i - \mathbb{E}[S|X=x_i]) + \varepsilon_i \\ x_i - \mathbb{E}[X|Y=y_i] = \alpha_2(s_i - \mathbb{E}[S|Y=y_i]) + \eta_i\end{cases}$$

First, the unknown conditional means are non-parametrically estimated. Then they are substituted for the unknown functions, and least squares is used to estimate : and a, which in this case correspond to the appropriate partial derivatives. Robinson (1988) proved that these estimates are asymptotically equivalent to those obtained using the true conditional mean functions for estimation.

8.8 Application on the `germancredit` Dataset

In section 4.2, we presented, on the `germancredit` dataset four models (logistic regression, classification tree, boosting and bagging), with and without the sensitive variable (the gender). Cumulative distribution functions of the scores, for the plain logistic regression and the boosting algorithm can be visualized on Figure 8.17.

In the training subset of `frenchmotor` 30.1% of the person got a default (y = `BAD`), 29.7% in the validation dataset ($\overline{y}$). If we consider predictions from the logistic regression model, with the sensitive attribute, on the validation dataset, the average prediction ($\overline{m}(\boldsymbol{x})$) is 28.7% and the median one is 20.4%. With a threshold $t = 20\%$, we have a balanced dataset, with $\widehat{y} = 1$ for 50% of the risky person (see Figure 8.17). With a threshold $t = 40\%$, 30% of the policyholders get $\widehat{y} = 1$ (which is close to the default frequency in the dataset). In Table 8.10 and 8.11, respectively, we use thresholds $t = 20\%$ and 40%.

	with sensitive				without sensitive			
	GLM	tree	boosting	bagging	GLM	tree	boosting	bagging
$\mathbb{P}[m(\boldsymbol{X}) > t]$	51.7%	28.0%	54.7%	61.7%	50.7%	28.0%	56.0%	60.7%
Predictive Rate Parity	0.992	1.190	0.992	1.050	0.957	1.190	1.041	1.037
Demographic Parity	0.998	1.091	1.159	1.027	1.213	1.091	1.112	1.208
FNR Parity	1.398	0.740	1.078	1.124	1.075	0.740	1.064	0.970
Proportional Parity	0.922	1.008	1.071	0.949	1.121	1.008	1.027	1.116
Equalized odds	0.816	1.069	0.947	0.888	0.956	1.069	0.953	1.031
Accuracy Parity	0.843	1.181	0.912	0.904	0.896	1.181	0.943	0.966
FPR Parity	1.247	0.683	1.470	0.855	2.004	0.683	0.962	1.069
NPV Parity	0.676	1.141	0.763	0.772	0.735	1.141	0.799	0.823
Specificity Parity	0.941	1.439	0.930	1.028	0.851	1.439	1.007	0.990
ROC AUC Parity	0.928	1.162	0.997	1.108	0.926	1.162	1.004	1.090
MCC Parity	0.604	2.013	0.744	0.851	0.639	2.013	0.884	0.930

Table 8.10: Fairness metrics on the `germancredit` dataset, with the `fairness` R package, by Varga and Kozodoi (2021), for women, reference being men, with threshold at 20%.

8.9 Application on the `frenchmotor` Dataset

In section 4.2, we presented, on the `frenchmotor` dataset four models (logistic regression, classification tree, boosting and bagging), with and without the sensitive variable (the gender). In the training subset of `frenchmotor` 8.72% of the policyholders claimed a loss, 8.55% in the validation dataset ($\overline{y}$). If we consider predictions from the logistic regression model, with the sensitive attribute, on the validation dataset, the average prediction ($\overline{m}(\boldsymbol{x})$) is 9% and the median one is 8%. With a threshold $t = 8\%$, we have a balanced dataset, with $\widehat{y} = 1$ for 50% of the risky drivers (see Figure 8.18). With a threshold $t = 16\%$, 10% of

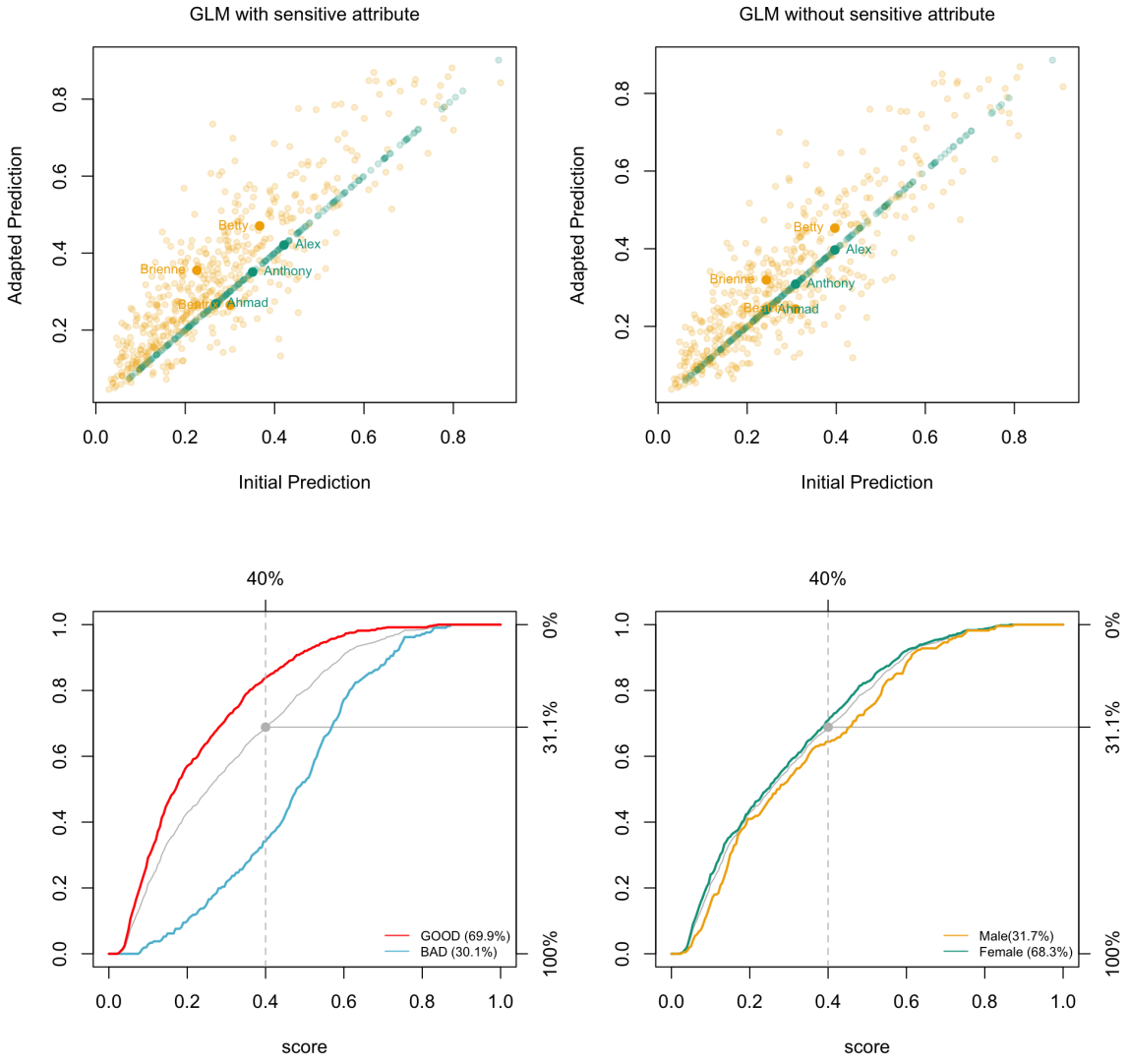


Figure 8.17: Distributions of the score $m(z)$, on the `germancredit` dataset, conditional on y on the left and s on the right, when m is a plain logistic regression without sensitive attribute s on top, and boosting without sensitive attribute s below, with threshold $t = 40\%$.

the policyholders get $\widehat{y} = 1$ (which is close to the claim frequency in the dataset). In Table 8.12 and 8.13, respectively, we use thresholds $t = 8\%$ and 16%.

	with sensitive				without sensitive			
	GLM	tree	boosting	bagging	GLM	tree	boosting	bagging
$\mathbb{P}[m(\boldsymbol{X}) > t]$	30.3%	26.0%	27.7%	25.7%	30.7%	26.0%	28.0%	27.0%
Predictive Rate Parity	1.030	1.179	1.110	1.182	1.034	1.179	1.111	1.200
Demographic Parity	1.090	1.062	1.074	1.069	1.108	1.062	1.044	1.019
FNR Parity	1.533	0.851	1.110	0.781	1.342	0.851	1.322	0.962
Proportional Parity	1.007	0.981	0.992	0.987	1.024	0.981	0.964	0.942
Equalized odds	0.925	1.032	0.982	1.041	0.944	1.032	0.955	1.008
Accuracy Parity	0.949	1.154	1.054	1.164	0.963	1.154	1.038	1.159
FPR Parity	1.118	0.703	0.820	0.653	1.118	0.703	0.784	0.641
NPV Parity	0.738	1.080	0.890	1.108	0.766	1.080	0.848	1.082
Specificity Parity	0.935	1.470	1.169	1.480	0.935	1.470	1.203	1.652
ROC AUC Parity	0.928	1.162	0.997	1.108	0.926	1.162	1.004	1.090
MCC Parity	0.745	1.817	1.105	1.754	0.779	1.817	1.056	2.055

Table 8.11: Fairness metrics on the `germancredit` dataset, with the `fairness` R package, by Varga and Kozodoi (2021), for women, reference being men, with threshold at 40%.

	with sensitive				without sensitive			
	GLM	tree	boosting	bagging	GLM	tree	boosting	bagging
$\mathbb{P}[m(\boldsymbol{X}) > t]$	51.1%	29.2%	49.6%	18.7%	50.8%	29.2%	51.6%	18.6%
Predictive Rate Parity	1.019	1.021	1.017	1.011	1.018	1.021	1.027	1.012
Demographic Parity	0.673	0.588	0.700	0.589	0.649	0.588	0.693	0.588
FNR Parity	0.833	0.900	0.789	0.813	0.865	0.900	0.806	0.818
Proportional Parity	1.182	1.034	1.231	1.035	1.141	1.034	1.217	1.033
Equalized odds	1.187	1.040	1.234	1.031	1.145	1.040	1.232	1.030
Accuracy Parity	1.161	1.051	1.198	1.037	1.125	1.051	1.205	1.036
FPR Parity	1.004	0.886	1.125	0.775	0.975	0.886	0.956	0.727
NPV Parity	1.004	1.054	0.986	1.071	0.982	1.054	1.060	1.074
Specificity Parity	0.998	1.141	0.927	1.079	1.012	1.141	1.026	1.091
ROC AUC Parity	1.023	1.098	1.027	1.059	1.023	1.098	1.046	1.063
MCC Parity	1.482	1.496	1.505	1.128	1.394	1.496	2.273	1.136

Table 8.12: Fairness metrics on the `frenchmotor` dataset, with the `fairness` R package, by Varga and Kozodoi (2021), for women, reference being men, with threshold at $t = 8\%$.

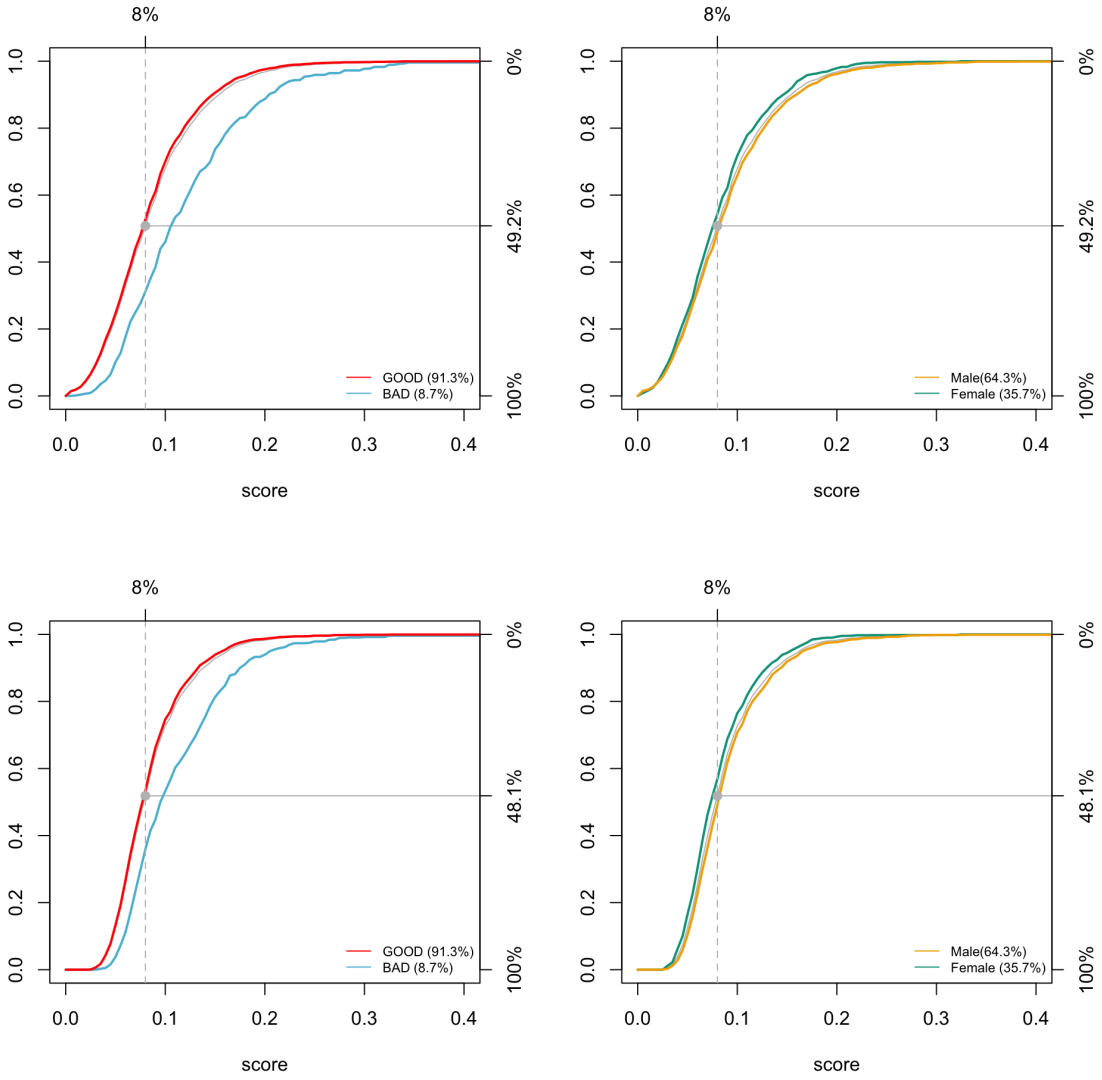


Figure 8.18: Distributions of the score $m(z)$, on the `frenchmotor` dataset, conditional on y on the left and s on the right, when m is a plain logistic regression without sensitive attribute s on top, and boosting without sensitive attribute s below, with threshold $t = 8\%$.

	with sensitive				without sensitive			
	GLM	tree	boosting	bagging	GLM	tree	boosting	bagging
$\mathbb{P}[m(\boldsymbol{X}) > t]$	10.0%	9.2%	6.6%	14.6%	10.2%	9.2%	5.6%	14.5%
Predictive Rate Parity	1.011	1.016	1.009	1.022	1.014	1.016	1.010	1.020
Demographic Parity	0.596	0.591	0.587	0.577	0.588	0.591	0.592	0.577
FNR Parity	0.618	0.620	0.642	0.819	0.710	0.620	0.478	0.827
Proportional Parity	1.048	1.039	1.032	1.014	1.034	1.039	1.040	1.014
Equalized odds	1.045	1.040	1.027	1.021	1.033	1.040	1.034	1.020
Accuracy Parity	1.050	1.050	1.032	1.038	1.043	1.050	1.040	1.036
FPR Parity	1.071	1.003	1.090	0.569	1.015	1.003	1.090	0.613
NPV Parity	1.011	1.259	0.652	1.160	1.092	1.259	0.847	1.143
Specificity Parity	0.748	0.987	0.467	1.256	0.944	0.987	0.467	1.234
ROC AUC Parity	1.023	1.098	1.027	1.059	1.023	1.098	1.046	1.063
MCC Parity	0.993	1.452	0.354	1.289	1.236	1.452	0.610	1.265

Table 8.13: Fairness metrics on the `frenchmotor` dataset, with the `fairness` R package, by Varga and Kozodoi (2021), for women, reference being men, with threshold at $t = 16\%$.

Chapter 9

Individual Fairness

Group fairness, as studied in Chapter 8, considered fairness from a global perspective, in the entire population, by attempting to answer the question "*are individuals in the advantaged group and in the disadvantaged group treated differently?* " Or more formally, are the predictions and the protected variable globally independent. Here, we focus on a specific individual, in the disadvantaged group, and we will talk about discrimination (in a broad sense, or inequity) by asking what the model would have predicted if this same person had been in the favored group. We will return here to the classical approaches, emphasizing the different approaches to constructing a counterfactual for this individual.

In the previous Chapter, we were interested in a notion of "group fairness" (with subgroups constituted from the values of y, s or $\widehat{y}$). It was probably first formalized in Dwork et al. (2012). The notion of individual fairness emphasizes that similar individuals (based on unprotected attributes only) should be treated similarly. "*Two actual cases may be cited as examples: Two neighbours, a man and woman, residing door to door in the city of Toronto, insured their Ford sedans through different agents with this insurer within a few days of each other, no change in rating policy having taken place in the meantime. A scrutiny of the written applications showed the risks as alike as two peas in a pod. The woman paid a thirteen per cent. higher rate than the man. The only explanation of the discrimination on the daily report was 'Premium arranged by A.H.B.'. 'A.H.B.' were the initials of a general agent of the insurer* ", Edwards (1932), cited in Barry (2020a).

Many definitions of individual fairness have been recently introduced in the literature (for instance, Jung et al. (2019a,b) introduced "*fairness elicitation*", while Salimi et al. (2020) defined "*justifiable fairness*", among many others). In this chapter, we will first discuss the popular idea that two similar individuals should have the same prediction (related to some "*Lipschitz property*"), and then, discuss ideas related to causal inference and counterfactual fairness, based on concepts introduced in section 7.4. As previously, consider a binary sensitive attribute s, taking values in $\{\texttt{A}, \texttt{B}\}$, $\texttt{A}$ being the favored group (or supposed to be), and $\texttt{B}$ the disfavored one.

9.1 Similarity Between Individuals (and Lipschitz Property)

The natural idea, formalized in Luong et al. (2011), is that two "close" individuals (in the sense of unprotected characteristics $\boldsymbol{x}$) must have the same prediction.

Definition 9.1.1 (Similarity Fairness) *Luong et al. (2011), Dwork et al. (2012). Consider two metrics, one on $\mathcal{Y} \times \mathcal{Y}$ (or for a classifier $[0,1]$ and not $\{0,1\}$) noted D_y, and one on $\mathcal{X}$ noted D_x, such that we will have similarity fairness on a database of size n if we have the following property (called Lipschitz property)*

$$D_y(m(\boldsymbol{x}_i, s_i), m(\boldsymbol{x}_j, s_j)) \leq L D_x(\boldsymbol{x}_i, \boldsymbol{x}_j), \ \forall i, j = 1, \cdots, n,$$

for some $L < \infty$.

Duivesteijn and Feelders (2008) defined a quite similar concept, called "*monotonic classification*". From a practical perspective, it is difficult to determine which metric to use to measure the similarity of two individuals (i.e. between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$), as explained by Kim et al. (2018). As Dwork et al. (2012) noticed, "*our approach is centered around the notion of a task-specific similarity metric describing the extent to which pairs of individuals should be regarded as similar for the classification task at hand. The similarity metric expresses ground truth. When ground truth is unavailable, the metric may reflect the "best" available approximation as agreed upon by society. Following established tradition – Rawls (2001) – the metric is assumed to be public and open to discussion and continual refinement. Indeed, we envision that, typically, the distance metric would be externally imposed, for example, by a regulatory body or externally proposed by a civil rights organization.*". The most usual ones for numeric variables are derived from Mahalanobis distance, to take into account the different scales between the variables (a normalized Euclidean distance), or some normalized ℓ_1 distance. For categorical variables, some similarty indices can be used (as in Jaccard (1901), Dice (1945) or Sorensen (1948)). Gower (1971) suggested to combined those two for mixed data. But one should keen in mind that this choice is not neutral. Nevertheless, because of the separation property of distances, whatever they are ($d(x_1, x_2) = 0$ if and only if $x_1 = x_2$), a consequence of the Lipschitz property is that a fair model m should satisfy $m(\boldsymbol{x}, \texttt{A}) = m(\boldsymbol{x}, \texttt{B})$.

Inspired from Definition 9.1.2,

$$\frac{D_y(m(\boldsymbol{x}_i, s_i), m(\boldsymbol{x}_j, s_j))}{D_x(\boldsymbol{x}_i, \boldsymbol{x}_j)} \leq L, \ \forall i, j = 1, \cdots, n,$$

for some L, Petersen et al. (2021) defined "local fairness" as follows,

Definition 9.1.2 (Local individual fairness) *Petersen et al. (2021). Consider two metrics, one on $\mathcal{Y} \times \mathcal{Y}$ (or for a classifier $[0,1]$ and not $\{0,1\}$) noted D_y, and one on $\mathcal{X}$ noted D_x, model m is locally individually fair if*

$$\mathbb{E}_{(\boldsymbol{X},S)}\left[\underset{\boldsymbol{x}': D_x(\boldsymbol{X},\boldsymbol{x}') \to 0}{limsup} \frac{D_y(m(\boldsymbol{X}, S), m(\boldsymbol{x}', S))}{D_x(\boldsymbol{X}, \boldsymbol{x}')} \right] \leq L < \infty.$$

Heidari and Krause (2018) and Gupta and Kamble (2021) considered some dynamic extension of that rule, to define fair rules in reinforcement learning, for example, inspired by Bolton et al. (2003) who found that customers' impression of fairness of prices critically relies on past prices that act as reference points.

9.2 Fairness with Causal Inference

As Loftus et al. (2018) wrote it, "*only by understanding and accurately modelling the mechanisms that propagate unfairness through society can we make informed decisions as to what should be done*". Galles and Pearl (1998) first introduced the idea of counterfactual reasoning in the context of fairness, formalized later on by Kusner et al. (2017) and Kilbertus et al. (2017), which states that the decision made should remain fixed, even if, hypothetically, the protected attribute (such as the race, or the gender) were to be changed.

As pointed out by Wu et al. (2019) and Carey and Wu (2022), there are a few concepts of fairness that are related to causal inference. In Chapter 7, we introduced the "*ladder of causation*" distinguishing "*intervention*" (in section 7.3) and "*counterfactuals*" (in section 7.4). Those two perspectives yield two definitions of individual fairness,

Definition 9.2.1 (Proxy Based Fairness) *Kilbertus et al. (2017). A decision making process $\widehat{y}$ exhibits no proxy discrimination with respect to proxy p if*

$$\mathbb{E}\left[\widehat{Y}\middle|do(S=\mathtt{A})\right]=\mathbb{E}\left[\widehat{Y}\middle|do(S=\mathtt{B})\right].$$

Definition 9.2.2 (Fairness on Average Treatment Effect) *Kusner et al. (2017). We achieve fairness on average treatment effect (counterfactual fairness on average)*

$$ATE=\mathbb{E}\left[Y^{\star}_{S\leftarrow\mathtt{A}}-Y^{\star}_{S\leftarrow\mathtt{B}}\right]=0.$$

Based on previous definition, quite natural extension is a local one,

Definition 9.2.3 (Counterfactual Fairness) *Kusner et al. (2017). We achieve counterfactual fairness for an individual with characteristics $\boldsymbol{x}$ if*

$$CATE(\boldsymbol{x})=\mathbb{E}\left[Y^{\star}_{S\leftarrow\mathtt{A}}-Y^{\star}_{S\leftarrow\mathtt{B}}\middle|\boldsymbol{X}=\boldsymbol{x}\right]=0.$$

Observe that there are several variations in the literature around those definitions. Kusner et al. (2017) defined formally counterfactual fairness "*conditional on a factual condition*", while Wu et al. (2019) considered "*path-specific causal fairness*". Zhang and Bareinboim (2018) distinguished "*counterfactual direct effect*", "*counterfactual indirect effect*" and "*counterfactual spurious effect*". To explain quickly the differences, following Avin et al. (2005), let us define the idea of "*path-specific causal effect*" (studied in Zhang et al. (2016) and Chiappa (2019)).

Definition 9.2.4 (Path-Specific Effect) *Avin et al. (2005). Given a causal diagram, and a path π some s to y ,the π-effect of a change of s from $\mathtt{B}$ to $\mathtt{A}$ on y is*

$$PE_{\pi}(\mathtt{B}\rightarrow\mathtt{A})=\mathbb{E}[Y|do_{\pi}(S=\mathtt{A})]-\mathbb{E}[Y|S=\mathtt{B}],$$

where "$do_{\pi}(S=\mathtt{A})$" denotes the intervention on s transmitted only along path π.

Transmission along a path (in a causal graph) was introduced with Definition 7.2.13. Then, following Wu et al. (2019), define the "*path-specific counterfactual effect*"

Definition 9.2.5 (Path-Specific Counterfactual Effect) *Wu et al. (2019). Given a causal diagram, a factual condition (denoted $\mathcal{F}$), and a path π some s to y, the π-effect of a change of s from $\mathtt{B}$ to $\mathtt{A}$ on y is*

$$PCE_{\pi}(\mathtt{B}\rightarrow\mathtt{A}|\mathcal{F})=\mathbb{E}[Y|do_{\pi}(S=\mathtt{A}),\mathcal{F}]-\mathbb{E}[Y|S=\mathtt{B},\mathcal{F}].$$

The "*factual condition*" $\mathcal{F}$ is a very general notation, that could be converted later on into $\{\boldsymbol{X} = \boldsymbol{x}\}$ or $\{Y = y\}$, for example. Based on those concepts, one can define "*path-specific causal fairness*" simply by asking that $\text{PCE}_\pi(\texttt{B} \to \texttt{A}|\mathcal{F})$ should be null. Based on those definitions, "*counterfactual fairness*" is obtained when $\mathcal{F} = \{(\boldsymbol{X}, S) = (\boldsymbol{x}, s)\}$, for any path π, while "*counterfactual indirect fairness*" is obtained when $\mathcal{F} = \{(Y, S) = (y, s)\}$, for any indirect path π (the effect of s on y should be transmitted through some $\boldsymbol{x}$'s) and "*indirect causal fairness*" is only for any indirect path π (and no factual condition). See Baer et al. (2019) for a review of fairness concepts based on causal graphs.

9.3 Counterfactuals and Optimal Transport

So far, we have used the general terminology in causal inference to assess if a model in discriminatory, or not. But we have not created, *per se* a counterfactual of an individual who may feel discriminated. Counterfactuals were introduced in section 7.4. Hume (1748) probably introduced the idea of counterfactuals: "*we may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, if the first object had not been, the second never had existed*". Unfortunately, it took some time to introduce the concept of having a "potential counterfactual", starting with Lewis (1973). Formally, having observed $(\boldsymbol{x}, s, y)$ we hope to create a counterfactual by considering $(\boldsymbol{x}, s^\star, y^\star)$, where $s^\star$ is another category than s, when the sensitive attribute is categorical (if we had observed $s = \texttt{A}$, we wish to create a protected class observation $s = \texttt{B}$, and vice versa), in order to compare the outcome y with the potential outcome $y^\star$ of the counterfactual. And a briefly mentioned in section 7.4, if s (hypothetically) changes, chances are that $\boldsymbol{x}$ will also change, as reminded by Gordaliza et al. (2019), Berk et al. (2021b) Torous et al. (2021), de Lara et al. (2021) and Charpentier et al. (2023). Formally, we simply say that the distributions of $\boldsymbol{x}$, conditionally to $s = \texttt{A}$ or $s = \texttt{B}$, are not identical, which will necessarily happen in the presence of a proxy of s among the explanatory variables $\boldsymbol{x}$. This can be related to the concept of "*pairwise fair representations*", as defined in Lahoti et al. (2019).

In section 4.2.1, we have mentioned Kullback–Leibler divergence (in Definition 3.3.7), a symmetric extension with Jensen–Shannon (in Definition 3.3.10) and Wasserstein distance (in Definition 3.3.11). Given two measures on p and q on $\mathbb{R}^d$, with a norm $\|\cdot\|$, then Wasserstein distance is defined a $W_k(p, q)$ where

$$W_k(p,q)^k = \inf_{\pi\in\Pi(p,q)} \left\{ \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\boldsymbol{x}-\boldsymbol{y}\|^k \mathrm{d}\pi(\boldsymbol{x},\boldsymbol{y}) \right\},$$

where $\Pi(p, q)$ is the set of all couplings of p and q. As mentioned in Definition 3.3.11, without any further specification, the Wasserstein distance will be W_2, where d is the Euclidean distance. In the univariate case, it is possible to simplify, and to see a connection with quantiles

$$W_k(p,q)^k = \int_0^1 \left|F_p^{-1}(u) - F_q^{-1}(u)\right|^k \,\mathrm{d}u.$$

9.3.1 Quantile Based Transport

Before defining properly this concept of "quantile based transport", let us recall a simple lemma,

Lemma 9.3.1 *Let X denote an random variable with cumulative distribution function F_p, and $h : \mathbb{R} \to \mathbb{R}$ an increasing one-to-one mapping, then the distribution of $Y = h(X)$ is $F_q(y) = F_p\big(h^{-1}(y)\big)$. And conversely, given two distributions F_p and F_q, let $h^{-1} = F_p^{-1} \circ F_q$ or $h = F_q^{-1} \circ F_p$, then if X has distribution F_p, $Y = h(Y)$ has distribution F_q (and in that case, X and Y are comonotone).*

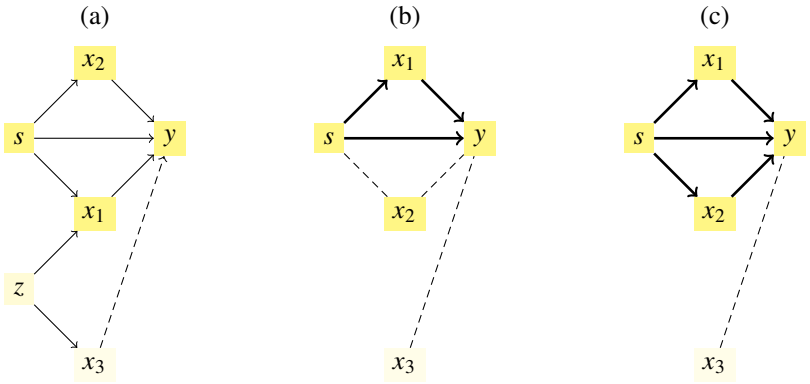


Figure 9.1: (a) Causal graph used to generate variables in `toydataset2`. (b) Simple causal graph that can be used on `toydataset2`, where the sensitive attribute might actually cause the outcome y, either directly (upper arrow), or indirectly, through x_1, a mediator variable. (c) Causal graph where the sensitive attribute s might cause the outcome y, either directly or indirectly, via with two possible paths and two mediator variables, x_1 and x_2.

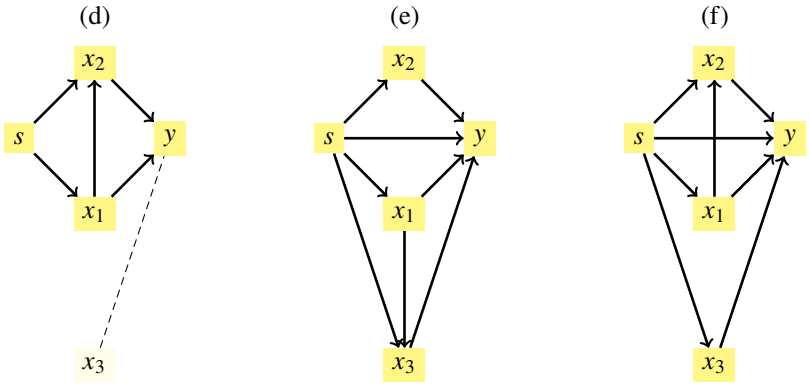


Figure 9.2: (d) Causal graph with no direct impact of s on y, but two mediators, and possibly, x_1 might cause x_2. (e) is similar to (c) with an additional indirect connection from x_1 to y, via mediator x_3. (f) is similar to (d) with an additional indirect connection from x_1 to y, via mediator x_3.

This is a consequence of the probability integral transform and the probability density function after transformation.

On Figure 9.1 (a), we have the causal graph used to generate the `toydataset2` dataset (see Section 1.4 for a description). To illustrate the "quantile based transformed", consider a simplified version, with the casual model of Figure 9.1 (b) : s, a binary variable (taking values in $\{\texttt{A}, \texttt{B}\}$), is generated; then x_1 is generated, conditionally on s, and finally; y will be generated, function of s and x_1.

Let m denote a model fitted on the data, so that $m(\boldsymbol{z})$ is an estimate of the regression function $\mu(\boldsymbol{z}) = \mathbb{E}[Y|\boldsymbol{Z} = \boldsymbol{z}]$, where $\boldsymbol{z}$ is either only $\boldsymbol{x}$ (fairness through unawareness, without the sensitive attribute) or

$(\boldsymbol{x}, s)$ (function of the sensitive attribute). Consider an individual with characteristics $\boldsymbol{x}$, in class $\mathtt{B}$, could be considered as discriminated, if a counterfactual version would get a different output. And the counterfactual of $(\boldsymbol{x}, \mathtt{B})$ would be $(\boldsymbol{x}_{s\leftarrow\mathtt{A}}, \mathtt{A})$, since an (hypothetical) intervention of s should have an impact on $\boldsymbol{x}$, from the model considered, Figure 9.1 (b).

In this section, $\boldsymbol{x}$ is a single variable, x_1. Heuristically, counterfactuals are obtained by requiring relative stability within the classes. It means that the counterfactual of $(x_1, \mathtt{B})$ will depend on the relative position of x_1 within group B. Let $F_{1\mathtt{A}}$ and $F_{1\mathtt{B}}$ denote the empirical distribution functions of X_1 in both groups, $F_{1s}(x) = \mathbb{P}[X_1 \leq x | S = s]$. If u is the probability associated with x_1 in group $\mathtt{B}$ – in the sense $u = F_{1\mathtt{B}}(x_1)$ – then the counterfactual should be associated with the quantile (in group $\mathtt{A}$) with the same probability u. Thus, the counterfactual of $(x_1, \mathtt{B})$ would be $(\mathcal{T}(x_1), \mathtt{A})$, where $\mathcal{T} = F_{1\mathtt{A}}^{-1} \circ F_{1\mathtt{B}}$. Following Berk et al. (2021b) and Charpentier et al. (2023) it is possible to define a "*quantile based counterfactual*" (or "*adaptation with quantile preservation*", as defined in Plečko et al. (2021)), as follows

Definition 9.3.1 (Quantile based counterfactual) *The counterfactual of* $(x, \mathtt{B})$ *is* $(\mathcal{T}(x), \mathtt{A})$*, where* $\mathcal{T}(x_1) = F_{1A}^{-1} \circ F_{1B}(x_1)$.

Definition 9.3.2 (Quantile based counterfactual discrimination) *There is counterfactual discrimination with model* m *for individual* (x_1, B) *if*

$$m(x_1, B) \neq m(\mathcal{T}(x_1), \mathtt{A}), \text{ where } \mathcal{T} = F_{1A}^{-1} \circ F_{1B}.$$

In real life applications, $\mathcal{T}$ is estimated, empirically, using $\widehat{\mathcal{T}}_n = \widehat{F}_{1\mathtt{A}}^{-1} \circ \widehat{F}_{1\mathtt{B}}$. And if X_1 has a Gaussian distribution, conditional on s, in the sense that,

$$X_{\mathtt{A}} \overset{\mathcal{L}}{=} X|S = \mathtt{A} \sim \mathcal{N}(\mu_{\mathtt{A}}, {\sigma_{\mathtt{A}}}^2) \text{ and } X_{\mathtt{B}} \overset{\mathcal{L}}{=} X|S = \mathtt{B} \sim \mathcal{N}(\mu_{\mathtt{B}}, {\sigma_{\mathtt{B}}}^2),$$

then

$$\underbrace{\mu_{\mathtt{A}} + \sigma_{\mathtt{A}} \cdot \frac{X_{\mathtt{B}} - \mu_{\mathtt{B}}}{\sigma_{\mathtt{B}}}}_{\mathcal{T}(X_{\mathtt{B}})} \overset{\mathcal{L}}{=} X_{\mathtt{A}}.$$

9.3.2 Optimal Transport (Discrete Setting)

The empirical version of the problem described in the previous section could be expressed as follows: consider two samples with identical size, denoted $\{x_1^{\mathtt{A}}, \cdots, x_n^{\mathtt{A}}\}$ and $\{x_1^{\mathtt{B}}, \cdots, x_n^{\mathtt{B}}\}$. For each individual $x_i^{\mathtt{B}}$, the counterfactual will be an individual in the other group $x_j^{\mathtt{A}}$, with two constraints: (1) it should be a one-to-one matching: each individual in group $\mathtt{B}$ should be associated with with a single observation in group $\mathtt{A}$, and conversely; (2) individuals should be matched with a "close" one, in the other group. Stuart (2010) used the name "*1:1 nearest neighbor matching*", to describe that matching procedure, see also Dehejia and Wahba (1999) or Ho et al. (2007). The first condition imposes that a matching is simply a permutation σ of $\{1, 2, \cdots, n\}$, so that, for all i, the counterfactual of $x_i^{\mathtt{B}}$ will be $x_{p(i)}^{\mathtt{A}}$. Recall that p can be characterized by a $n \times n$ permutation matrix $\boldsymbol{P}$ (with $P_{ij} = 1$ if $j = \sigma(i)$ and $P_{ij} = 0$ otherwise). $n \times n$ permutation matrix, with entries in $\{0, 1\}$, satisfy $\boldsymbol{P}\mathbf{1}_n = \mathbf{1}_n$ and $\boldsymbol{P}^\top \mathbf{1}_n = \mathbf{1}_n$, as defined in Brualdi (2006). If $\boldsymbol{C}$ denote the $n \times n$ matrix that quantifies the distance between individuals in the two groups, $C_{i,j} = d({x_i}^{\mathtt{B}}, {x_j}^{\mathtt{A}})^2 = (x_i^{\mathtt{B}} - x_j^{\mathtt{A}})^2$, the optimal matching is solution of

$$\min_{\boldsymbol{P}\in\mathcal{P}} \left\{ \langle \boldsymbol{P}, \boldsymbol{C} \rangle \right\} = \min_{\boldsymbol{P}\in\mathcal{P}} \left\{ \sum_{i,j} P_{i,j} C_{i,j} \right\}, \tag{9.1}$$

where $\mathcal{P}$ is the set of $n \times n$ permutation matrices. The solution is very intuitive, and based on the following rearrangement inequality,

Lemma 9.3.2 (Hardy–Littlewood–Pólya inequality (1)) *Given $x_1 \leq \cdots \leq x_n$ and $y_1 \leq \cdots \leq y_n$ n pairs of ordered real numbers, for every permutation σ of $\{1, 2, \cdots, n\}$,*

$$\sum_{i=1}^{n} x_i y_{n+1-i} \leq \sum_{i=1}^{n} x_i y_{\sigma(i)} \leq \sum_{i=1}^{n} x_i y_i.$$

See Hardy et al. (1952).

That previous inequality can be extended, from this product version (terms are products between x_i and some y_j) to more general function $\Phi(x_i, y_j)$.

Definition 9.3.3 (Supermodular) *Function $\Phi : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ is supermodular if for any $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^k$,*

$$\Phi(\boldsymbol{z} \wedge \boldsymbol{z}') + \Phi(\boldsymbol{z} \vee \boldsymbol{z}') \geq \Phi(\boldsymbol{z}) + \Phi(\boldsymbol{z}'),$$

where $\boldsymbol{z} \wedge \boldsymbol{z}'$ and $\boldsymbol{z} \vee \boldsymbol{z}'$ denote respectively the maximum and the minimum componentwise. If $-\Phi$ is supermodular, Φ is said to be submodular.

From Topkis' characterization theorem (see Topkis (1998)), if $\Phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is twice differentiable, Φ is supermodular if and only if $\partial^2\Phi/\partial x \partial y \geq 0$, for all $i \neq j$. And as mentioned in Galichon (2016), many popular functions in applied mathematics, and economics, satisfy this property, such as Cobb-Douglas functions, $\Phi(x, y) = x^a y^b$ when $a, b \geq 0$ (on $\mathbb{R}_+ \times \mathbb{R}_+$) or if $\Phi(x, y) = \gamma(x - y)$ for some concave function $\gamma : \mathbb{R} \to \mathbb{R}$, such as $\Phi(x, y) = -|x - y|^k$ with $k \geq 1$ or $\Phi(x, y) = -(x - y - k)_+$. Note that $-\Phi$ can be seen as a cost function.

Lemma 9.3.3 (Hardy–Littlewood–Pólya inequality (2)) *Given $x_1 \leq \cdots \leq x_n$ and $y_1 \leq \cdots \leq y_n$ n pairs of ordered real numbers, and some supermodular function $\Phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, for every permutation σ of $\{1, 2, \cdots, n\}$,*

$$\sum_{i=1}^{n} \Phi(x_i, y_{n+1-i}) \leq \sum_{i=1}^{n} \Phi(x_i, y_{\sigma(i)}) \leq \sum_{i=1}^{n} \Phi(x_i, y_i),$$

while if $\Phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is submodular,

$$\sum_{i=1}^{n} \Phi(x_i, y_i) \leq \sum_{i=1}^{n} \Phi(x_i, y_{\sigma(i)}) \leq \sum_{i=1}^{n} \Phi(x_i, y_{n+1-i}),$$

See Hardy et al. (1952).

Another way to write that inequality is that, given two sets of real numbers $\{x_1, \cdots, x_n\}$ and $\{y_1, \cdots, y_n\}$ with no ties, if Φ is submodular (such as $\Phi(x, y) = |x - y|^k$ with $k \geq 1$),

$$\sum_{i=1}^{n} \Phi(x_i, y_i) \geq \sum_{i=1}^{n} \Phi(x_i, y_{\sigma^\star(i)}),$$

where $\sigma^\star$ is the permutation such that the rank of $y_{\sigma^\star(i)}$ (among the y's) is equal to the rank of x_i (among the x's), as discussed in Chapter 2 of Santambrogio (2015). This corresponds to a "*monotone rearrangement*".

	7	8	9	10	11	12
1	0.41	0.55	0.22	0.64	0.04	0.25
2	0.28	0.24	0.73	0.22	0.64	0.80
3	0.28	0.47	0.32	0.52	0.16	0.37
4	0.28	0.62	0.81	0.25	0.64	0.85
5	0.41	0.37	0.89	0.25	0.81	0.97
6	0.66	0.76	0.21	0.89	0.22	0.14

	7	8	9	10	11	12	
1	·	·	·	·	1	·	1 → 11
2	·	1	·	·	·	·	2 → 8
3	·	·	1	·	·	·	3 → 9
4	1	·	·	·	·	·	4 → 7
5	·	·	·	1	·	·	5 → 10
6	·	·	·	·	·	1	6 → 12

Table 9.1: Optimal matching, $n = 6$ individuals in class B (per row) and in class A (per column). Table on the left is the distance matrix $\boldsymbol{C}$, while table on the right is the optimal permutation $\boldsymbol{P}^\star$, solution of Equation (9.1).

A numerical illustration is provided in Table 9.1. There are $n = 6$ individuals in class B (per row) and in class A (per column). Table on the left is the distance matrix $\boldsymbol{C}$, between x_i and x_j, while table on the right is the optimal permutation $\boldsymbol{P}$, solution of Equation (9.1). Here, individual $i = 3$, in group B, is matched with individual $j = 9$, in group A. Thus, on this very specific example, model m would be seen as fair for individual $i = 3$ if $m(\boldsymbol{x}_3, B) = m(\boldsymbol{x}_9, A)$. Observed that fairness can be assessed here only for individuals that belong to the training dataset (and not any fictional individual $(\boldsymbol{x}, B)$).

A more general setting could be considered, where the two groups do not longer have the same size n. Consider two groups, A and B. Given $\boldsymbol{\nu}_B \in \mathbb{R}_+^{n_B}$ and $\boldsymbol{\nu}_A \in \mathbb{R}_+^{n_A}$ such that $\boldsymbol{\nu}_B^\top \mathbf{1}_{n_B} = \boldsymbol{\nu}_A^\top \mathbf{1}_{n_A}$ (identical sums), define

$$U(\boldsymbol{\nu}_B, \boldsymbol{\nu}_A) = \left\{ M \in \mathbb{R}_+^{n_B \times n_A} : M\mathbf{1}_{n_A} = \boldsymbol{\nu}_B \text{ and } M^\top \mathbf{1}_{n_B} = \boldsymbol{\nu}_A \right\},$$

where $\mathbb{R}_+^{n_B \times n_A}$ is the set of $n_B \times n_A$ matrices with positive entries. This set of matrices is a convex polytope (see Brualdi (2006)).

In our case, let us denote $U\left(\mathbf{1}_B, \dfrac{n_B}{n_A}\mathbf{1}_A\right)$ as U_{n_B,n_A}. Then the problem we want to solve is simply

$$\boldsymbol{P}^\star \in \underset{P \in U_{n_B,n_A}}{\operatorname{argmin}} \left\{ \langle P, C \rangle \right\} \text{ or } \underset{P \in U_{n_B,n_A}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n_B} \sum_{j=1}^{n_A} P_{i,j} C_{i,j} \right\}. \tag{9.2}$$

Here $\boldsymbol{P}^\star$ is not longer a permutation matrix, but $\boldsymbol{P}^\star \in U_{n_B,n_A}$, so that sums per row are equal to one (with positive entries), and can be considered as weights (as permutation matrices), but here sums per column are equal, but not to one - they are equal to the ratio n_B/n_A. To illustrate, consider Table 9.2, with $n_B = 6$ individuals in class B (per row) and $n_A = 10$ individuals in class A (per column). Consider individual $i = 3$, in group B. She will be is matched with a weighted sum of individuals in group A, namely $j = 7$ and 13, with respective weights $^3/_5$ and $^2/_5$. Thus, on this very specific example, model m would be seen as fair for individual $i = 3$ if

$$m(\boldsymbol{x}_3, B) = \frac{3}{5} m(\boldsymbol{x}_7, A) + \frac{2}{5} m(\boldsymbol{x}_{13}, A).$$

9.3.3 Optimal Transport (General Setting)

A transport $\mathcal{T}$ is a deterministic function that couple x_0 and x_1 (or more generally vectors in the same space) in the sense that $x_1 = \mathcal{T}(x_0)$. And $(X_0, \mathcal{T}(X_0))$ is a coupling of two measures $\mathbb{P}_0$ and $\mathbb{P}_1$ (for a more general framework than the p and q used previously) if $X \sim \mathbb{P}_0$ and $\mathcal{T}(X_0) \sim \mathbb{P}_1$. We will denote $\mathcal{T}_\# \mathbb{P}_0 = \mathbb{P}_1$ the "*push-forward*" of $\mathbb{P}_0$ by mapping $\mathcal{T}$.

	7	8	9	10	11	12	13	14	15	16
1	0.41	0.55	0.22	0.64	0.04	0.25	0.24	0.77	0.74	0.55
2	0.28	0.24	0.73	0.22	0.64	0.80	0.76	0.76	0.12	0.10
3	0.28	0.47	0.32	0.52	0.16	0.37	0.27	0.68	0.63	0.45
4	0.28	0.62	0.81	0.25	0.64	0.85	0.58	0.32	0.51	0.48
5	0.41	0.37	0.89	0.25	0.81	0.97	0.91	0.81	0.05	0.25
6	0.66	0.76	0.21	0.89	0.22	0.14	0.33	0.96	0.99	0.79

	7	8	9	10	11	12	13	14	15
1	·	·	$1/5$	·	$3/5$	·	$1/5$	·	·
2	·	$2/5$	·	·	·	·	·	·	·
3	$3/5$	·	·	·	·	·	$2/5$	·	·
4	·	·	·	$2/5$	·	·	·	$3/5$	·
5	·	$1/5$	·	$1/5$	·	·	·	·	$3/5$
6	·	·	$2/5$	·	·	$3/5$	·	·	·

Table 9.2: Optimal matching, $n_B = 6$ individuals in class B (per row) and $n_A = 10$ individuals in class A (per column). Table on the left is the distance matrix $\boldsymbol{C}$, while table on the right is the optimal weight matrix $\boldsymbol{P}^\star$ in $U_{6,10}$, solution of Equation (9.2).

Lemma 9.3.4 *The distribution of X on $\mathbb{R}$ is the push-forward measure of the uniform measure on $[0,1]$, with $\mathcal{T} = F_X^{-1}$.*

This is a consequence of the probability integral transform.

There are many coupling, many transport functions, and we will seek an "optimal" one. An optimal transport $\mathcal{T}^\star$ (in Brenier's sense, from Brenier (1991), see Villani (2009) or Galichon (2016)) from $\mathbb{P}_0$ towards $\mathbb{P}_1$ will be solution of

$$\mathcal{T}^\star \in \underset{\mathcal{T}:\mathcal{T}_\#\mathbb{P}_0=\mathbb{P}_1}{\operatorname{arginf}} \left\{ \int_{\mathbb{R}^k} \gamma(\boldsymbol{x} - \mathcal{T}(\boldsymbol{x})) \mathrm{d}\mathbb{P}_0(\boldsymbol{x}) \right\},$$

that we can related to the equation below, which could be seen as the continuous (and more general) version of Equation (9.2),

$$\inf_{\nu \in \Pi(\mathbb{P}_0,\mathbb{P}_1)} \int_{\mathbb{R}^k \times \mathbb{R}^k} \underbrace{\Phi(\boldsymbol{x}_0, \boldsymbol{x}_1)}_{=C} \underbrace{\nu(\mathrm{d}\boldsymbol{x}_0, \mathrm{d}\boldsymbol{x}_1)}_{=P},$$

for some function $\Phi : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_+$, such that $\Phi(\boldsymbol{x}_0, \boldsymbol{x}_1) = \gamma(\boldsymbol{x}_0 - \boldsymbol{x}_1)$.

Definition 9.3.4 (Monge-Kantorovich Problem) *Given a submodular function $\Phi : \mathbb{R}^k \times \mathbb{R}^k$, and two measures $\mathbb{P}_0$ and $\mathbb{P}_1$ on $\mathbb{R}^k$, the coupling $\nu^\star \in \Pi(\mathbb{P}_0, \mathbb{P}_1)$ is optimal if*

$$\nu^\star \in \underset{\nu \in \Pi(\mathbb{P}_0,\mathbb{P}_1)}{\operatorname{arginf}} \left\{ \mathbb{E}\big[\Phi(\boldsymbol{X}_0, \boldsymbol{X}_1)\big] \right\}, \text{ where } (\boldsymbol{X}_0, \boldsymbol{X}_1) \sim \nu.$$

An optimal transport is a mapping $\mathcal{T}^\star : \mathbb{R}^k \to \mathbb{R}^k$ such that if $X \sim \mathbb{P}_0$, $\mathcal{T}^\star(\boldsymbol{X}_0) \sim \mathbb{P}_1$, and therefore $(\boldsymbol{X}_0, \mathcal{T}^\star(\boldsymbol{X}_0)) \sim \nu^\star$.

Formally, in general settings, such a deterministic correspondence via $\mathcal{T}$ between probability distributions may not exist, in particular if $\mathbb{P}_0$ and $\mathbb{P}_1$ are not Lebesgue absolutely continuous. Solving problem with function Φ is called Kantorovich relaxation of Monge's formulation of optimal transport (from Kantorovich and Rubinshtein (1958)).

In dimension 1, as discussed previously, there is a simple solution to this problem, under mild conditions. Let F_p and F_q denote the cumulative distribution functions associated with measures p and q,

Propoition 9.3.1 *Assume that Φ is strictly submodular, and p has no mass point. Then the Monge–Kantorovich problem has a unique optimal assignment, and this assignment is characterized by $X_1 = \mathcal{T}^\star(X_0)$, where $\mathcal{T}^\star$ is given by $\mathcal{T}^\star = F_q^{-1} \circ F_p$.*

See Villani (2003) or Galichon (2016).

Observe that the optimal assignment is a comonotone solution ($\mathcal{T}^\star$ is an increasing mapping), which could be seen as an extension of Hardy–Littlewood–Pólya inequality (Lemma 9.3.3). In higher dimension, $\mathcal{T}^\star$ is an "increasing" mapping in the (unique) sense that it is the gradient of a convex function.

This idea of transport in the context of quantifying fairness was originally mentioned in Dwork et al. (2012), where the used of "*earthmover distances*" with the Lipschitz condition is mentioned. As mentioned in Section 3.3, the "*earthmover distance*" is actually simply Wasserstein distance with index with 1 (denoted W_1).

9.3.4 Optimal Transport Between Gaussian Distribution

After defining the optimal mapping in the univariate case using quantiles, we mentioned that the optimal transport between two univariate Gaussian distributions (respectively $\mathcal{N}(\mu_0, \sigma_0^2)$ and $\mathcal{N}(\mu_1, \sigma_1^2)$) is

$$x_1 = \mathcal{T}_{\mathcal{N}}^\star(x_0) = \mu_1 + \frac{\sigma_0}{\sigma_1}\big(x_0 - \mu_0\big),$$

and actually, a similar transformation can be obtained in higher dimension,

Propoition 9.3.2 *Suppose that $\mathbb{P}_0$ and $\mathbb{P}_1$ are two Gaussian measures (respectively $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$) then the optimal transport is*

$$\boldsymbol{x}_1 = \mathcal{T}_{\mathcal{N}}^\star(\boldsymbol{x}_0) = \boldsymbol{\mu}_1 + \boldsymbol{A}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0),$$

where $\boldsymbol{A}$ is a symmetric positive matrix that satisfies $\boldsymbol{A}\boldsymbol{\Sigma}_0\boldsymbol{A} = \boldsymbol{\Sigma}_1$, which has a unique solution given by $\boldsymbol{A} = \boldsymbol{\Sigma}_0^{-1/2}\big(\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{1/2}\big)^{1/2}\boldsymbol{\Sigma}_0^{-1/2}$.

See Villani (2003) or Galichon (2016).

Here $\boldsymbol{M}^{1/2}$ is the square root of the square (symmetric) positive matrix $\boldsymbol{M}$ based on the Schur decomposition ($\boldsymbol{M}^{1/2}$ is a positive symmetric matrix), as described in Higham (2008). In R, such a function is obtained using `sqrtm` in the `expm` package.

9.3.5 Transport and Causal Graphs

Plečko et al. (2021) considered a more general approach, to solve problems as the one on Figure 9.1 (d): as previously, s will cause x_1 and x_2, but x_2 will also be influenced by x_1. Therefore, when transporting on x_2, we should consider not the quantile function of x_2 conditional on s, but a quantile regression function, of x_2 conditional on s and x_1. A quantile random forest, as in Meinshausen and Ridgeway (2006) can be considered (or any machine learning algorithm based on a quantile loss function $\ell_{q,\alpha}$) or more generally, a quantile regression based on optimal transport, as in Carlier et al. (2016).

Numerical applications on `toydata2` and `germancredit` dataset, based on causal graphs of Figure 9.1 and 9.8 respectively will be considered in the next section and at the end of this chapter.

9.4 Mutatis Mutandis Counterfactual Fairness

A model satisfies the "counterfactual fairness" (Definition 9.2.3) property if "*had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same*". This is a *ceteris paribus* definition of counterfactual fairness. But a *mutatis mutandis* definition of the conditional average treatment effect can be considered (as in Berk et al. (2021b) and Charpentier et al. (2023)).

Definition 9.4.1 (Mutatis Mutandis Counterfactual fairness) *Kusner et al. (2017) If the prediction in the real world is the same as the prediction in the counterfactual world,* mutatis mutandis *where the individual would have belonged to a different demographic group, we have counterfactual fairness, i.e.*

$$\mathbb{E}[Y^\star_{S\leftarrow \mathtt{A}}|\boldsymbol{X}=\mathcal{T}(\boldsymbol{x})]=\mathbb{E}[Y^\star_{S\leftarrow \mathtt{B}}|\boldsymbol{X}=\boldsymbol{x}],\ \forall \boldsymbol{x},$$

where $\mathcal{T}$ is the optimal transport from the distribution of $\boldsymbol{X}$ conditional on $S=\mathtt{B}$ to the distribution of $\boldsymbol{X}$ condition on $S=\mathtt{A}$.

Similarity Fairness (*Lipschitz*)	Dwork et al. (2012)	9.1.2	$D_y(\widehat{y}_i,\widehat{y}_j)\leq D_x(\boldsymbol{x}_i,\boldsymbol{x}_j),\ \forall i,j$
Proxy Based Fairness,	Kilbertus et al. (2017)	9.2.1	$\mathbb{E}\left[Y\middle\vert \text{do}(S=\mathtt{A})\right]=\mathbb{E}\left[Y\middle\vert \text{do}(S=\mathtt{B})\right]$
Fairness on Average Treatment Effect	Kusner et al. (2017)	9.2.2	$\mathbb{E}\left[Y^\star_{S\leftarrow \mathtt{A}}\right]=\mathbb{E}\left[Y^\star_{S\leftarrow \mathtt{B}}\right]$
Counterfactual Fairness,	Kusner et al. (2017)	9.2.3	$\mathbb{E}\left[Y^\star_{S\leftarrow \mathtt{A}}\middle\vert \boldsymbol{X}=\boldsymbol{x}\right]=\mathbb{E}\left[Y^\star_{S\leftarrow \mathtt{A}}\middle\vert \boldsymbol{X}=\boldsymbol{x}\right]$
Path-Specific Effect	Avin et al. (2005)	9.2.4	$\mathbb{E}[Y\vert \text{do}_\pi(S=\mathtt{A})]=\mathbb{E}[Y\vert \text{do}_\pi(S=\mathtt{B})]$
Path-Specific Counterfactual Effect	Wu et al. (2019)	9.2.5	$\mathbb{E}[Y\vert \text{do}_\pi(S=\mathtt{A}),\mathcal{F}]=\mathbb{E}[Y\vert \text{do}_\pi(S=\mathtt{B}),\mathcal{F}]$
Mutatis Mutandis Counterfactual	Kusner et al. (2017)	9.4.1	$\mathbb{E}[Y^\star_{S\leftarrow \mathtt{A}}\vert \boldsymbol{X}=\mathcal{T}(\boldsymbol{x})]=\mathbb{E}[Y^\star_{S\leftarrow \mathtt{B}}\vert \boldsymbol{X}=\boldsymbol{x}]$

Table 9.3: Definitions of individual fairness, see Carey and Wu (2022) for a complete list.

9.5 Application on the `toydataset2` dataset

Throughout chapter 8, we have studied various group related notions of fairness, on some simulated data, `toydataset2`. Recall that empirical proportions of 1's are respectively 24% and 65%, in groups `A` and `B` respectively. From a demographic parity perspective, we might claim that there is discrimination, since the two proportions are significantly different. Here, instead of that global perspective, we will consider specific individuals (the same as the ones considered in Section 4.1 to illustrate local interpretability concepts).

On top of Table 9.4, in the first block, we compare $\widehat{m}(\boldsymbol{x},\mathtt{A})$ and $\widehat{m}(\boldsymbol{x},\mathtt{B})$ for different models, non-aware and aware, with a logistic regression, a logistic smooth version (GAM) and a random forest (those models were described previously). On top, we have Betty, Brienne and Beatrix, and for the those three individuals, we want to quantify some possible individual discrimination. Below, we have Alex, Ahmad and Anthony, who are somehow "male versions" of Betty, Brienne and Beatrix, in the sense that they share the same legitimate characteristics $\boldsymbol{x}$. Even if the distance between $\boldsymbol{x}$'s is null (as in the Lipschitz property) within each pair, a "proper counterfactual" of Betty is not Alex (neither is Ahmad the counterfactual of Brienne and neither is Anthony the counterfactual of Beatrix). We will use the techniques mentioned previously to construct proper ones on that small dataset.

For the first block, we simply use marginal transformations, as in Definition 9.3.1. For example, for Brienne, $x_1=1$, which is the median value among women (group `B`), if Brienne would have been a man (group `A`), and if we assume that she would have kept the same relative positive (the median), that corresponds to $x_1=-1$. Similarly for x_3. But for x_2, 5 is also the median in group `B`, and since the median is almost the same in group `A`, x_2 will remain the same. Thus, if Brienne would have been a man, we would be discrimination with model m, and individual $(x_1,x_2,x_3,\mathtt{B})$ if $m(\mathcal{T}_1(x_1),\mathcal{T}_2(x_2),\mathcal{T}_3(x_3),\mathtt{A})<m(x_1,x_2,x_3,\mathtt{B})$.

Instead of marginal transformations, it is possible to consider the two other techniques mentioned previously, optimal transport (using `transport` in `R`) and a multivariate Gaussian transport. Formally, we consider here causal graphs of Figure 9.1, in the sense that s will have a causal impact on both x_1 and

Original data

	s	x_1	x_2	x_3	$\widehat{m}_{\text{glm}}(\boldsymbol{x})$	$\widehat{m}_{\text{glm}}(\boldsymbol{x}, s)$	$\widehat{m}_{\text{gam}}(\boldsymbol{x})$	$\widehat{m}_{\text{gam}}(\boldsymbol{x}, s)$	$\widehat{m}_{\text{rf}}(\boldsymbol{x})$	$\widehat{m}_{\text{rf}}(\boldsymbol{x}, s)$
Betty	B	0	2	0	18.22%	24.06%	13.23%	17.63%	17.4%	29.6%
Brienne	B	1	5	1	67.19%	70.47%	66.18%	67.09%	63.60%	61.80%
Beatrix	B	2	8	2	94.95%	94.73%	97.53%	97.58%	96.60%	98.40%
Alex	A	0	2	0	18.22%	13.71%	13.23%	10.05%	17.40%	9.20%
Ahmad	A	1	5	1	67.19%	54.48%	66.18%	50.49%	63.60%	64.40%
Anthony	A	2	8	2	94.95%	90.02%	97.53%	90.51%	96.60%	68.20%

Counterfactual

adjusted data, using marginal quantiles

Betty	A	-1.68	2.1	-1.68	3.51%	3.58%	4.78%	4.85%	10.40%	10.80%
Brienne	A	-0.98	5.1	-0.96	19.39%	17.65%	16.64%	16.13%	29.00%	41.00%
Beatrix	A	-0.27	7.9	-0.26	59.83%	53.65%	51.89%	46.37%	53.60%	49.00%

adjusted data, using optimal transport, Figure 9.1 (c)

Betty	A	-1.96	2.1	-1.9	2.62%	2.82%	4.65%	4.81%	0.00%	0.00%
Brienne	A	0.29	5	0.25	48.24%	38.92%	40.04%	32.14%	21.40%	12.20%
Beatrix	A	0.31	7.8	0.21	72.83%	65.1%	67.5%	58.83%	20.80%	15%

adjusted data, using Gaussian transport, Figure 9.1 (c)

Betty	A	-1.58	2.15	-1.59	3.95%	3.96%	4.96%	4.99%	0.40%	0.40%
Brienne	A	-0.98	4.96	-0.99	18.47%	16.84%	15.84%	15.40%	19.80%	27.20%
Beatrix	A	-0.38	7.79	-0.38	55.71%	50.05%	47.86%	43.16%	51.80%	63.60%

adjusted data, with `fairAdapt`, Figure 9.2 (e)

Betty	A	-1.65	2	-1.32	3.63%	3.54%	4.72%	4.60%	14.60%	8.00%
Brienne	A	-0.97	4.55	-0.94	16.57%	14.96%	13.96%	13.51%	2.20%	5.20%
Beatrix	A	-0.33	7.72	-0.44	56.3%	50.71%	48.49%	43.74%	70.60%	74.80%

adjusted data, with `fairAdapt`, Figure 9.2 (f)

Betty	A	-1.75	2.28	-1.68	3.5%	3.6%	5.03%	5.13%	7.20%	7.00%
Brienne	A	-0.96	5.3	-0.96	20.9%	19.05%	17.91%	17.34%	5.80%	8.40%
Beatrix	A	-0.24	8.12	-0.34	62.31%	56.43%	54.8%	49.3%	45.60%	39.20%

Table 9.4: Creating counterfactuals for `Betty`, `Brienne` and `Beatrix` VERIFIER LES REFERENCES !!!

x_2, and not on x_3. In Table 9.4, counterfactuals are create for Betty, Brienne and Beatrix, to compare the prediction if those individuals would have been in group `A` instead of `B`. On Figure 9.3, we have the optimal transport on a regular grid (x_1, x_3), where the arrow starts at $\boldsymbol{x} = (x_1, x_3)$ and ends at $\mathcal{T}(\boldsymbol{x})$, respectively with marginal quantile transport (as in Definition 9.3.1) and with multivariate Gaussian transport (based on Proposition 9.3.2). For example, Beatrix, corresponding to $(2, 8)$ is mapped with $(-0.27, 7.9)$ in the first case and $(-0.38, 7.82)$ in the second case.

With the `fairadapt` function, in the `fairAdapt` package based on Plečko et al. (2021), it is also possible to create some counterfactuals. To do so, we will consider the causal networks of Figure 9.2 (e) and (f), the first one being the one used to generate data. Compared with the causal networks of Figure 9.1, here, we take into account the existing correlation between x_1 and x_3, in the sense that an intervention on s will change x_1, and even if s has no direct impact on x_3, it will go through the path $\pi = \{s, x_1, x_3\}$. Then all variables

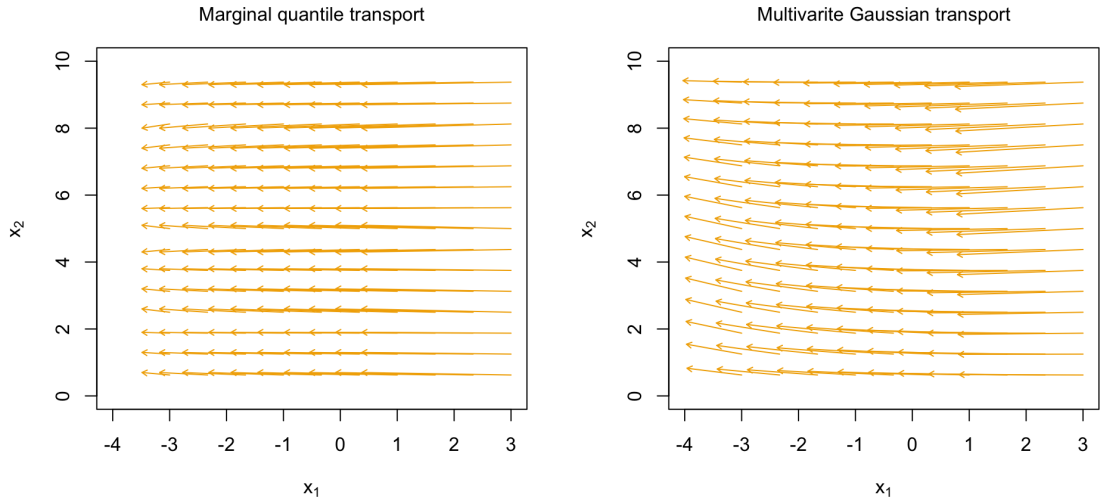


Figure 9.3: Optimal transport on a (x_1, x_2) grid, from `B` to `A`, using `fairadapt` on the left, and, on the right using a parametric estimate of $\mathcal{T}^\star$ (based on Gaussian quantiles). Only individuals in group `B` are transported here.

might have an impact on y. To model properly that complex graph, we use `fairadapt` function to create counterfactuals.

On Figure 9.4, we have scatterplots $(m(\boldsymbol{x}_i), m(\mathcal{T}(\boldsymbol{x}_i)))$ for individuals in groups `A` and `B`. Light points are the ones from the originl dataset, while dark ones are the six individuals described in Table 9.4. Points in group `A` are on the diagonal, since points are not adapted. Points in group `B` are here, for all models, below the diagonal, in the sense that adapted predictions are below initial predictions. Therefore, this could legitimate the idea that people in group `B` are discriminated, compared with people in group `A` since, if treated as someone in group `A`, people in group `B`, predictions would be smaller.

On Figure 9.5, we can visualize on the left the distribution of $m(\boldsymbol{x}_i)$'s in group A and B. On the right, the distribution of $m(\mathcal{T}(\boldsymbol{x}_i))$'s in group B.

On the left of Figure 9.6, we have the scatterplot on (x_1, x_2), with points in the A group mainly on the left and in the group B on the right. On the right of Figure 9.6 gray segments mapping individuals in group B and in group A. The three points on the rights, · are Betty, Brienne and Beatrix, and on the left · the three matches individuals in group A.

Figure 9.7, we can visualize the optimal transport on a (x_1, x_2) grid, from `B` to `A`, using nonparametric estimate of $\mathcal{T}^\star$ (based on empirical quantiles) on the left, and a Gaussian distribution on the right.

9.6 Application on the `germancredit` dataset

The `germancredit` dataset was considered in Section 8.8, where group fairness metrics were computed. Recall that the proportion of empirical defaults in the training dataset was 35% for males (group `M`) while it was 28% for female (group `F`). And here, we will try to quantify potential individual discrimination against women.

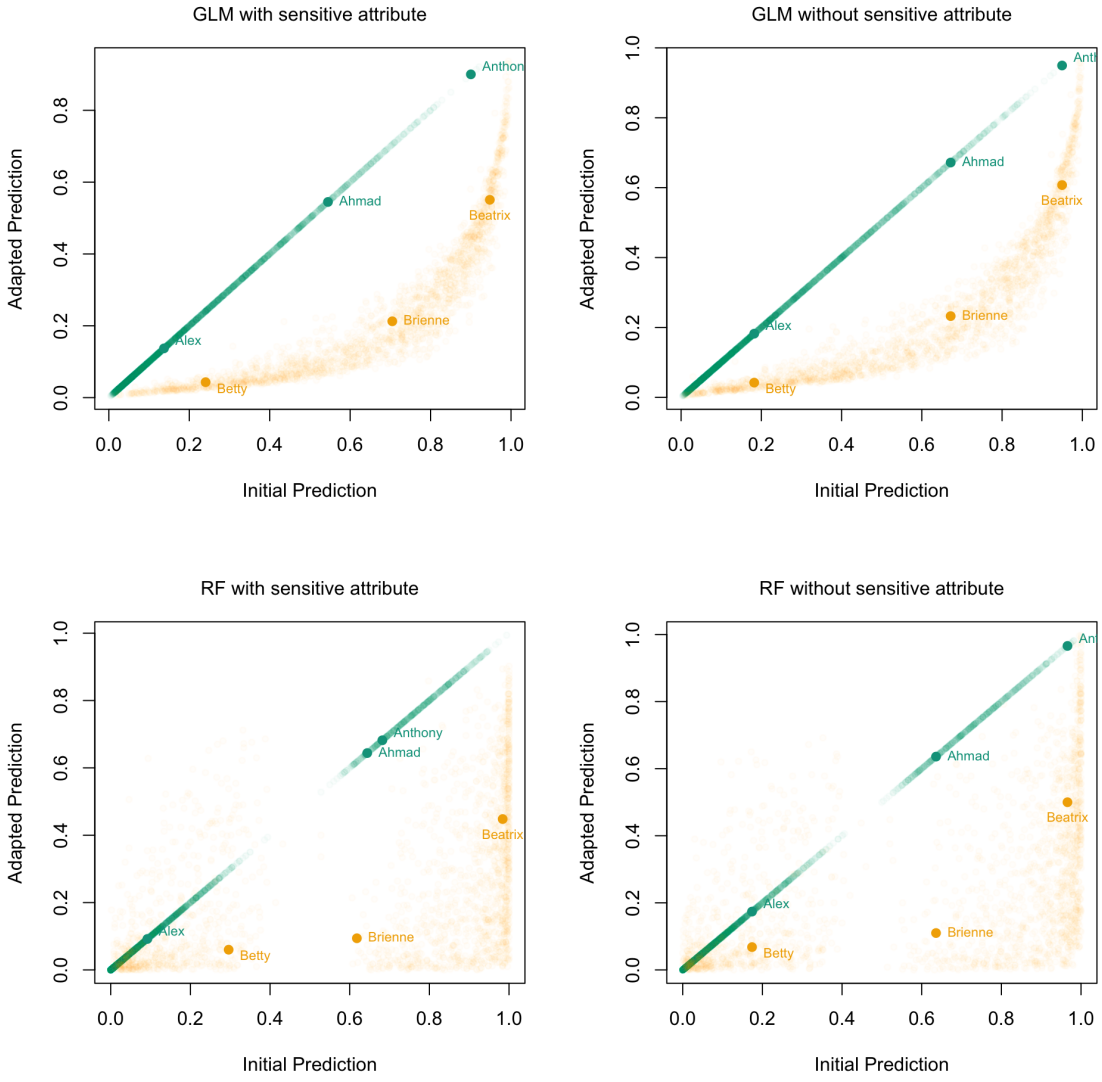


Figure 9.4: Scatterplot $(m(\boldsymbol{x}_i), m(\mathcal{T}^\star(\boldsymbol{x}_i)))$, adapted prediction against the original prediction, for individuals in groups A and B, on the `toydataset2` dataset. Transformation is from group B to group A (therefore predictions in group A remain unchanged). Model m is, on top a plain logistic regression (GLM) and below, a random forest.

On Figure 9.8, we have a simple causal graph on the `germancredit` dataset, with $s \rightarrow \{\texttt{duration}, \texttt{credit}\}$, and $\{\texttt{duration}, \texttt{credit}\} \rightarrow y$ as well as all other variables (that could cause y). The causal graph of Figure 9.10 is the one used in Watson et al. (2021). On that second causal graph, $s \rightarrow \{\texttt{savings}, \texttt{job}\}$ and then multiple causal relationships. Finally, on the causal graph of Figure 9.11, four causal links are added, to the

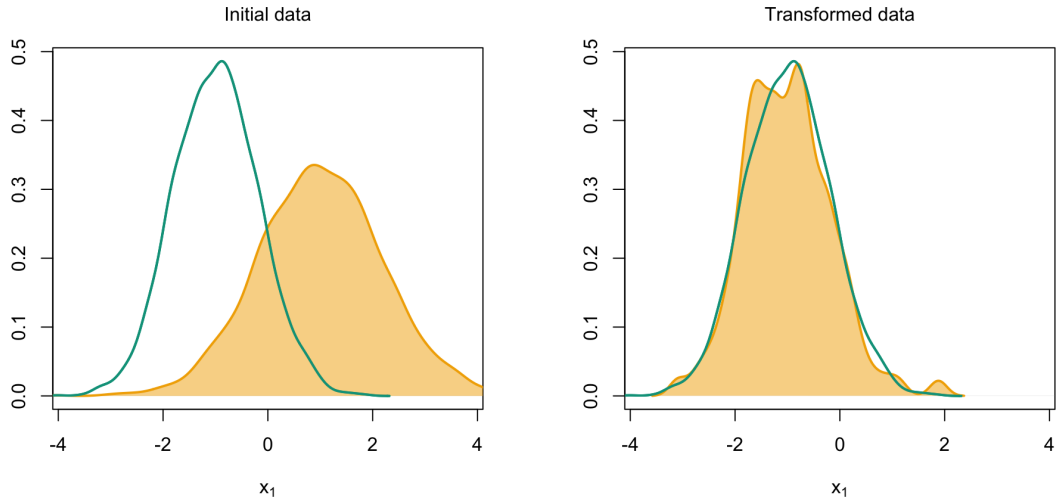


Figure 9.5: The plain line is the density of $x_1$for group A, while the plain area corresponds to group B, on the `toydataset2` dataset. On the left, we have the distribution on the training dataset, and on the right, the density of the adapted variables x_1 (with a transformation from group B to group A).

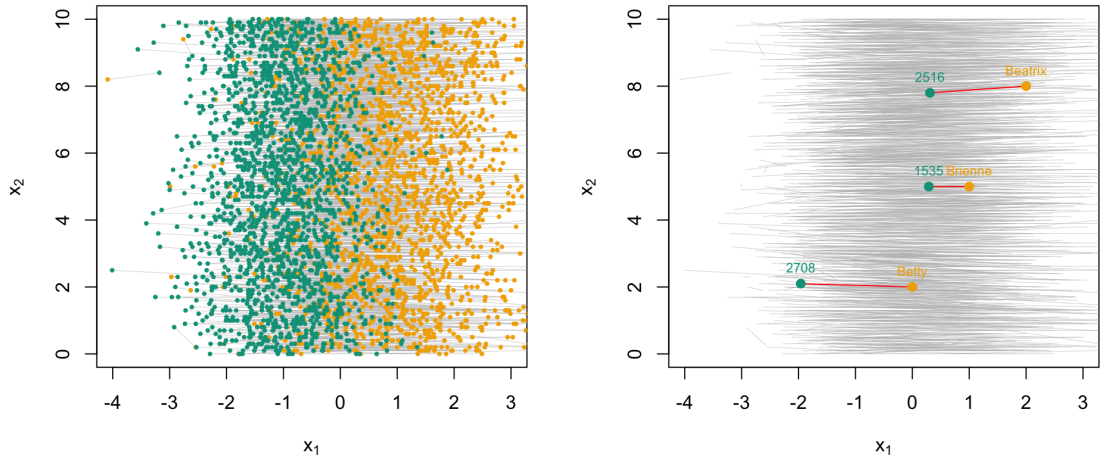


Figure 9.6: Optimal matching, of individuals in group B to individuals in group A, on right, where points • are Betty, Brienne and Beatrix, and • their counterfactual version in group A.

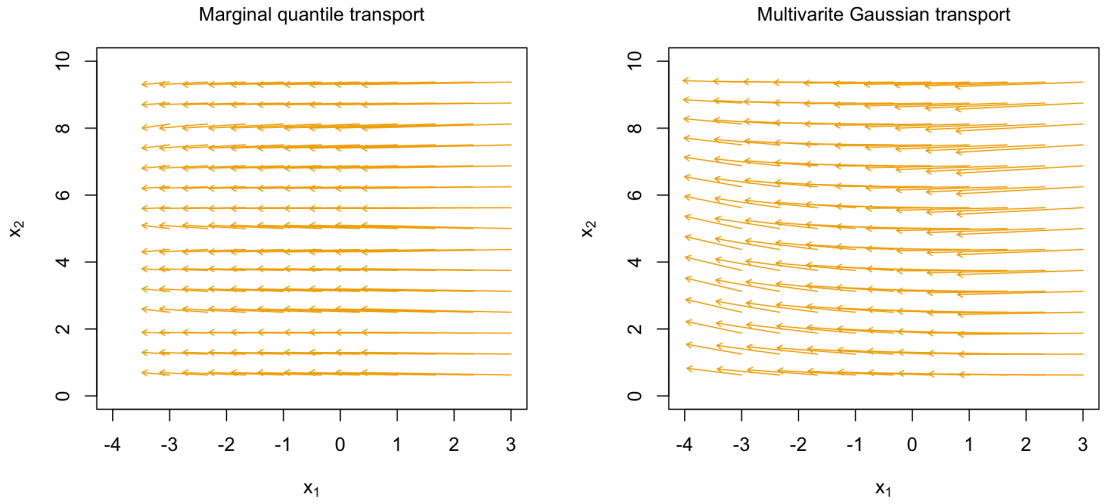


Figure 9.7: Optimal transport on a (x_1, x_2) grid, from `B` to `A`, using nonparametric estimate of $\mathcal{T}^\star$ (based on empirical quantiles) on the left, and a Gaussian distribution on the right.

previous one, $s \rightarrow \{\texttt{duration}, \texttt{credit\_amount}\}$ and $\{\texttt{duration}, \texttt{credit\_amount}\} \rightarrow y$.

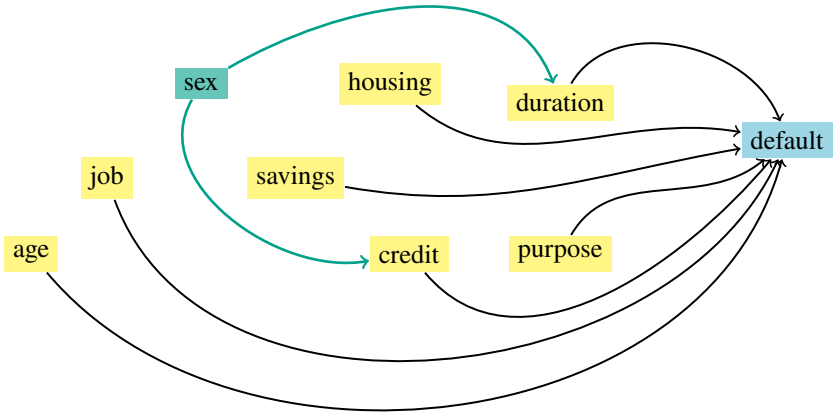


Figure 9.8: Simple causal graph on the `germancredit` dataset, where all variables could have a direct impact on y (or `default`), except the sensitive attribute (`s`), that will have an indirect impact through two possible paths, via either the duration of the credit (`duration`) or the amount (`credit_amount`).

Based on the causal graph of Figure 9.8, when computing the counterfactual, most variables that are

children of the sensitive attribute s are adjusted (here x_1 is `Duration` and x_2 is `Credit_amount`). As in the previous section, we consider three pairs of individuals, within each pairs, individuals share the same $\boldsymbol{x}$'s, only s changes (more details on the other features are also given). And again, the counterfactual version of Betty (F) is not Alex (M). For instance, the amount of credit that was initially 1262 for Betty should be different if Betty would have been a man. If we were using some quantile-based transport, as on Table 9.5, 1262 corresponds to the 25% quantile for men, and 17% quantile for women. So, if Betty was consider as a man, with a credit amount corresponding to the 17% quantile within men, it should be a smaller amount, about 1074. Again, after that transformation, when everyone is considered as a man, distributions of credit and duration are identical.

	Alex	Ahmad	Anthony	Betty	Brienne	Beatrix
s (gender)	M	M	M	F	F	F
x_1 `Duration`	12	18	30	12	18	30
$u = F_{1\|s}(x_1)$	36%	57%	86%	34%	50%	79%
$\mathcal{T}(x_1) = F^{-1}_{1\|s=\texttt{M}}(u)$	12	18	30	12	18	24
x_2 `Credit`	1262	2319	4720	1262	2319	4720
$u = F_{2\|s}(x_2)$	25%	55%	82%	17%	45%	76%
$\mathcal{T}(x_2) = F^{-1}_{2\|s=\texttt{M}}(u)$	1262	2319	4720	1074	1855	3854

Table 9.5: Counterfactuals based on the causal graph of Figure 9.8, based on marginal quantile transformations. We consider here only counterfactual versions of women, considered as the "disfavored".

Here, eight models are considered: a logistic regression (GLM), a boosting (with Adaboost algorithm, as presented in Section 3.3.6), a classification tree (as on Figure 3.21) and a random forest (RF, with 500 trees), each time the unaware version (based on $\boldsymbol{x}$ only, not s) and the aware version (including s). For all unaware models, Alex (M) and Beatrix (F) get the same prediction. But for most models, aware models yield to different prediction, when individuals have different gender. Predictions of those six individuals are given in Table 9.7. Those individual predictions can be visualized on Figure 9.9, when the causal graph is the one of Figure 9.8. Lighter points correspond to the entiere dataset, and darker points are the six individuals. Predictions with the random forest are not very robust here.

For the logistic regression and the boosting algorithm (see Table 9.7), counterfactual predictions are rather close to the original ones. For example, if Brienne had been a man, mutatis mutandis, the default probability would have been 23.95% instead of 24.30% (as Ahamad), when considering the unaware logistic regression. The impact is larger for Beatrix, who had an initial prediction of default of 30.88%, as a women (even if the logistic regression is gender blind), and the same model, on the counterfactual version of Beatrix, would have predicted 24.91%. It could be seen as sufficiently different to consider that Beatrix could legitimately fell discrimated because of her gender But as shown on Table 9.6, we can compute easily confidence intervals for predictions on GLM (it would be more complicated for boosting algorithms and random forests) and the difference is not significantly different.

But more complex models can be considered. Using function `fairTwins`, it is possible to get a counterfactual for each individual in the dataset, as suggested in Szepannek and Lübke (2021), even when we consider categorical covariates. The first "realistic" causal graph we will consider is the one used in Watson et al. (2021), on that same dataset, that can be visualized on Figure 9.10.

In the `germancredit` dataset, several variables in $\boldsymbol{x}$ are categorical. For ordered categorical variables (such as `Savings_bonds`, taking values `< 100 DM`, `100 <= ...< 500 DM`, etc.) it is possible to adapt optimal transport techniques, as suggested in Plečko et al. (2021), assuming that ordered categories are

	Betty		Brienne		Beatrix	
Unaware logistic $m(\boldsymbol{x})$	39.7%	[23.9% ; 57.9%]	24.3%	[13.8% ; 39.1%]	30.9%	[15.7% ; 51.7%]
Unaware logistic $m(\boldsymbol{x})$	39.7%	[23.9% ; 57.9%]	24.3%	[13.8% ; 39.1%]	30.9%	[15.7% ; 51.7%]
Unaware logistic $m(\mathcal{T}(\boldsymbol{x}))$	39.5%	[23.8% ; 57.8%]	24.0%	[13.6% ; 38.7%]	24.9%	[12.2% ; 44.1%]
Aware logistic $m(\boldsymbol{x}, s = \texttt{F})$	36.7%	[21.1% ; 55.6%]	22.6%	[12.5% ; 37.4%]	30.1%	[15.2% ; 50.8%]
Aware logistic $m(\boldsymbol{x}, s = \texttt{M})$	42.1%	[25.4% ; 60.8%]	26.8%	[15.0% ; 43.2%]	35.1%	[17.6% ; 57.8%]
Aware logistic $m(\mathcal{T}(\boldsymbol{x}), s = \texttt{M})$	41.9%	[25.2% ; 60.7%]	26.5%	[14.8% ; 42.8%]	28.6%	[13.7% ; 50.2%]

Table 9.6: Predictions, with 95% confidence intervals, based on the causal graph of Figure 9.8, using marginal quantile transformations. We consider here only counterfactual versions of women only, considered as the "disfavored".

$\{1, 2, \cdots, m\}$ and using a cost function $\gamma(i, j) = |i - j|^k$. For non-ordered categorical variables (such as `Job`, or `Housing`, the later taking values such as '`rent`', '`own`' or '`for free`'), a cost function $\gamma(i, i) = \mathbf{1}(i \neq j)$ is used. And for continuous variables (`Age`, `Credit_amount` or `Duration`), previous techniques can be used. Since `Age` is on no-path from s to y in all causal graphs, it will not change. On Figure 9.12 we can visualize the distributions of x conditional on $s = \texttt{M}$ and $s = \texttt{F}$, respectively when x is `Credit_amount` and `Duration`.

Predictions of the six individuals from Table 9.7 can be visualized on Figure 9.13, when the causal graph is the one of Figure 9.10. As in the previous section, there are three pairs of individuals, within each pairs, individuals share the same $\boldsymbol{x}$'s, only s changes. Then 8 models are considered, a logistic regression (GLM), a boosting (with Adaboost algorithm, as presented in Section 3.3.6), a classification tree (as on Figure 3.21) and a random forest (RF, with 500 trees), each time the unaware version (based on $\boldsymbol{x}$ only, not s) and the aware version (including s). For all unaware models, Alex (M) and Beatrix (F) get the same prediction. But for most models, aware models yield to different prediction, when individuals have different gender.

Firstname	s	Firstname	s	Job	Savings	Housing	Purpose
Alex	M	Betty	F	highly qualified employee	100 DM	rent	radio / television
Ahmad	M	Brienne	F	skilled employee	100<=...<500 DM	own	furniture
Anthony	M	Beatrix	F	unskilled - resident	no savings	for free	car (new)

Original data

	s	Age	Duration	Credit	$\widehat{m}_{\text{glm}}(\boldsymbol{x})$	$\widehat{m}_{\text{glm}}(\boldsymbol{x}, s)$	$\widehat{m}_{\text{gbm}}(\boldsymbol{x})$	$\widehat{m}_{\text{gbm}}(\boldsymbol{x}, s)$	$\widehat{m}_{\text{cart}}(\boldsymbol{x})$	$\widehat{m}_{\text{cart}}(\boldsymbol{x}, s)$	$\widehat{m}_{\text{rf}}(\boldsymbol{x})$
Betty	F	26	12	1262	39.69%	36.66%	42.30%	43.26%	31.75%	31.75%	25.40%
Brienne	F	33	18	2320	24.30%	22.61%	23.88%	21.08%	21.31%	21.31%	43.60%
Beatrix	F	45	30	4720	30.88%	30.08%	28.49%	30.42%	15.38%	15.38%	23.40%
Alex	M	26	12	1262	39.69%	42.10%	42.30%	44.86%	31.75%	31.75%	25.40%
Ahmad	M	33	18	2320	24.30%	26.84%	23.88%	22.18%	21.31%	21.31%	43.60%
Anthony	M	45	30	4720	30.88%	35.08%	28.49%	31.82%	15.38%	15.38%	23.40%

Counterfactual

adjusted data, with marginal quantile transformations, causal graph from Figure 9.8

Betty	M	26	12	1074	39.51%	41.90%	40.69%	44.86%	31.75%	31.75%	23.80%
Brienne	M	33	18	1855	23.95%	26.46%	23.88%	22.18%	21.31%	21.31%	43.00%
Beatrix	M	45	24	3854	24.91%	28.58%	20.55%	20.31%	21.31%	21.31%	17.60%

adjusted data, with `fairAdapt`, causal graph from Figure 9.8

Betty	M	26	12	1110	42.73%	45.18%	44.24%	46.64%	31.75%	31.75%	22.2%
Brienne	M	33	18	1787	23.90%	26.40%	23.88%	22.18%	21.31%	21.31%	43.2%
Beatrix	M	45	24	3990	25.01%	28.70%	22.17%	23.60%	21.31%	21.31%	19.6%

adjusted data, with `fairAdapt`, causal graph from Figure 9.10

Betty	M	26	18	1778	52.23%	54.03%	40.05%	46.81%	21.31%	21.31%	34.80%
Brienne	M	33	15	1864	32.25%	35.85%	31.60%	25.97%	21.31%	21.31%	23.00%
Beatrix	M	45	21	3599	39.70%	43.16%	28.36%	28.90%	21.31%	21.31%	10.60%

adjusted data, with `fairAdapt`, causal graph from Figure 9.11

Betty	M	26	15	1882	49.05%	50.86%	35.32%	40.12%	21.31%	21.31%	27.8%
Brienne	M	33	18	1881	50.76%	53.49%	43.00%	38.77%	21.31%	21.31%	10.8%
Beatrix	M	45	24	3234	24.20%	26.23%	14.63%	16.84%	21.31%	21.31%	22.4%

Table 9.7: Creating counterfactuals for `Betty`, `Brienne` and `Beatrix` in the `germancredit` dataset.

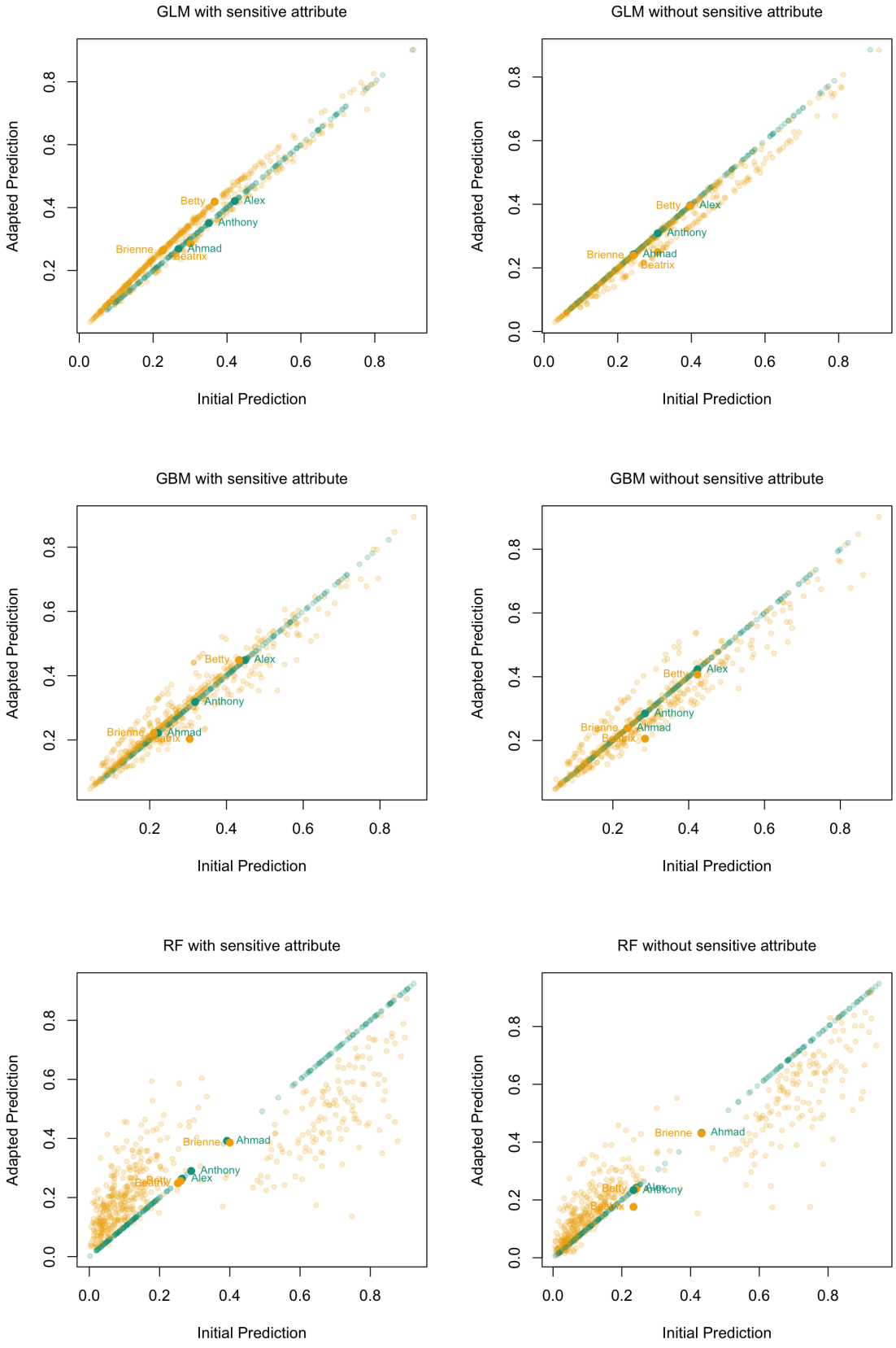


Figure 9.9: Scatterplot $(m(\boldsymbol{x}_i), m(\mathcal{T}^\star(\boldsymbol{x}_i)))$ for individuals in groups `M` and `F`, on the `germancredit` dataset. Transformation is only from group `F` to group `M`, so that all individuals are seen as men, on the y-axis. Model m is, from top to bottom, a plain logistic regression (GLM), a boosting model (GBM) and a random forest (RF). We used `fairadapt` codes and the causal graph of Figure 9.8.

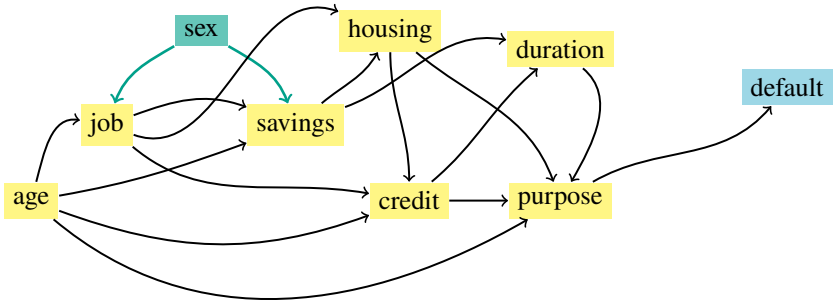


Figure 9.10: Causal graph on the `germancredit` dataset, from Watson et al. (2021).

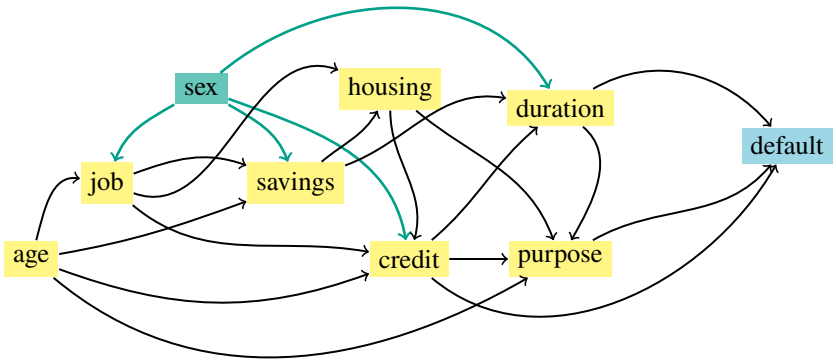


Figure 9.11: Causal graph on the `germancredit` dataset.

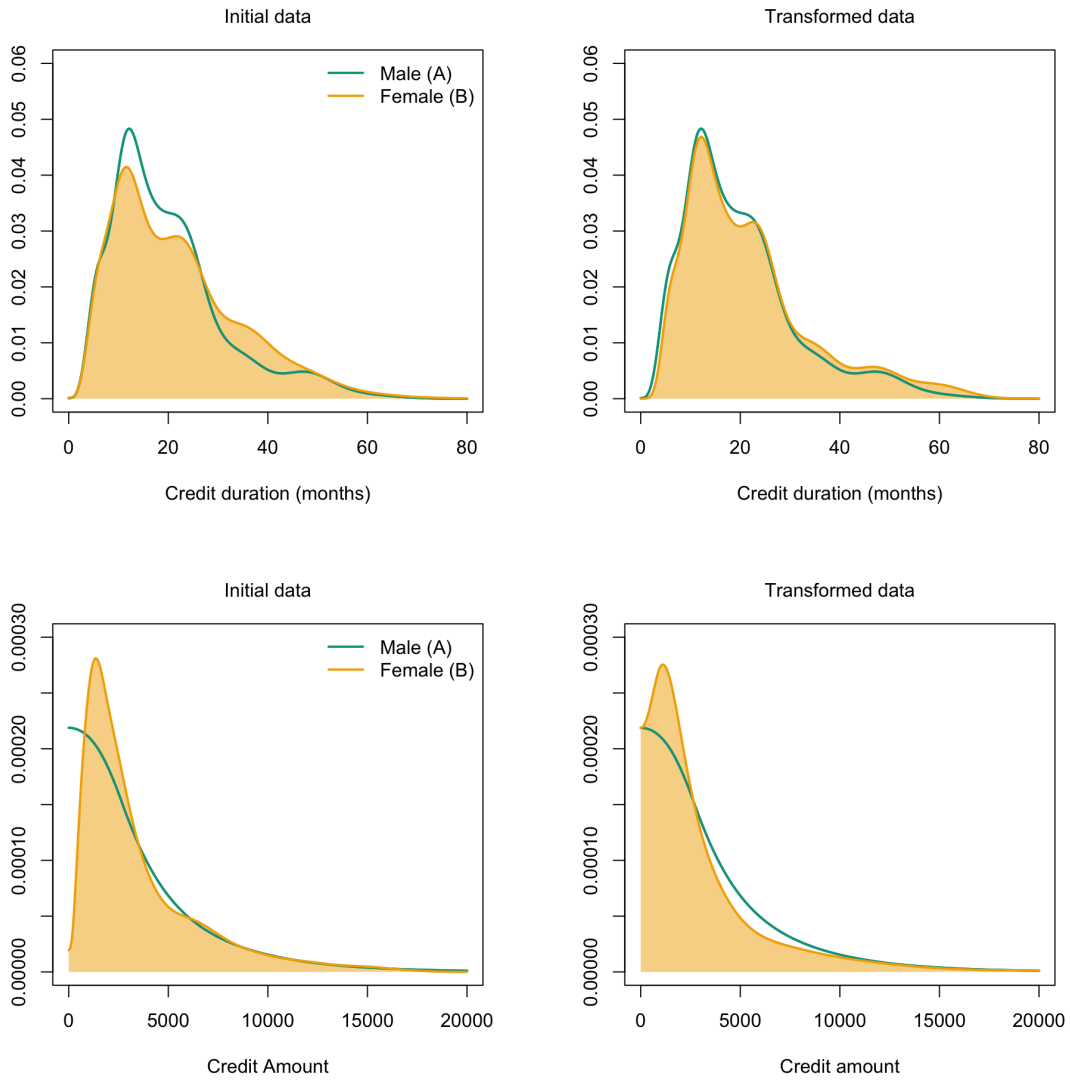


Figure 9.12: Distributions of x (`Credit_amount` on top and `Duration` below) conditional on $s =$ `A` and $s =$ `B`.

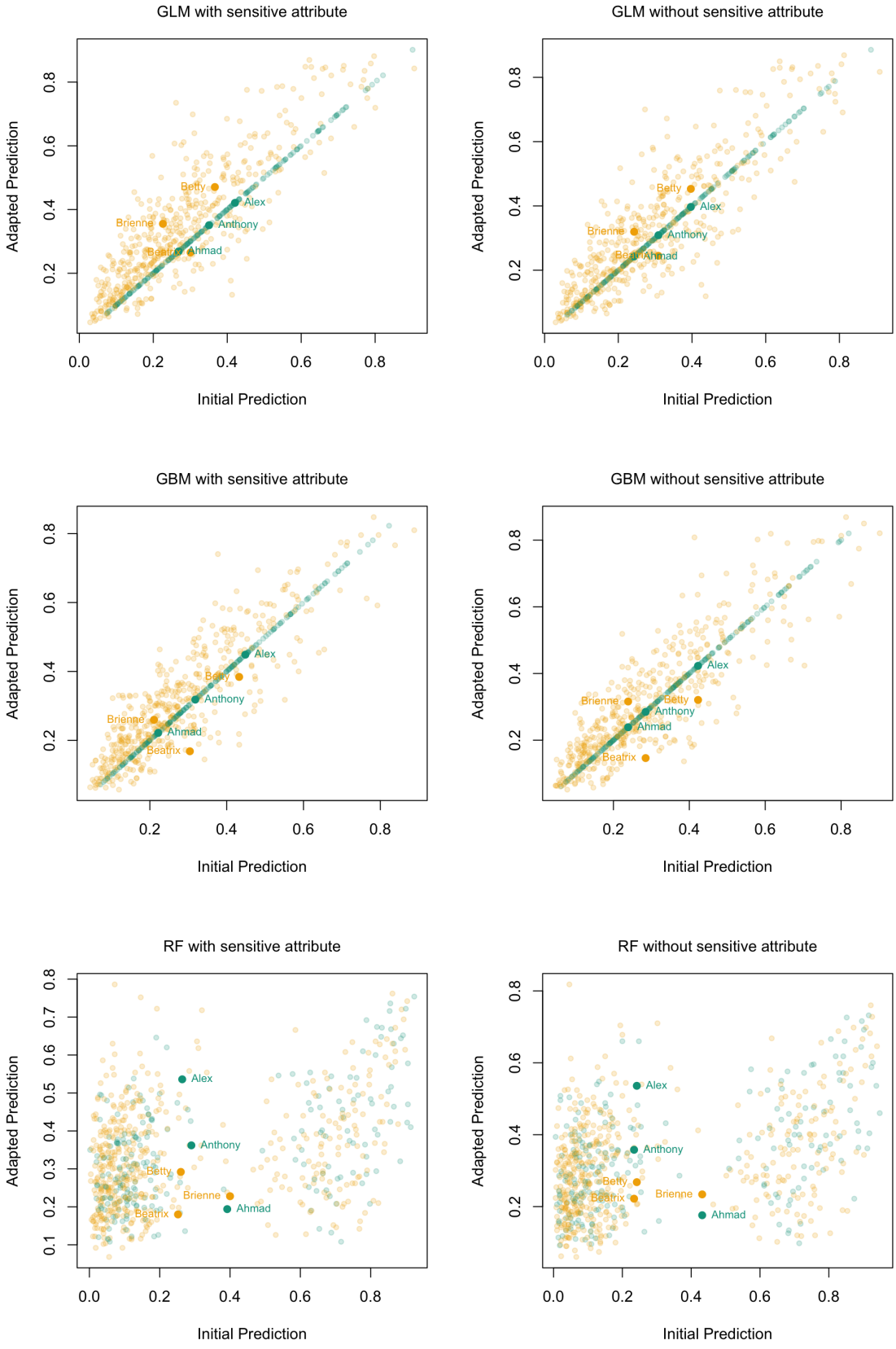


Figure 9.13: Scatterplot $(m(\boldsymbol{x}_i), m(\mathcal{T}(\boldsymbol{x}_i)))$ for individuals in groups `M` and `F`, on the `germancredit` dataset. Transformation is only from group `F` to group `M`, so that all individuals are seen as men, on the y-axis. Model m is, from top to bottom, a plain logistic regression (GLM), a boosting model (GBM) and a random forest (RF). We used `fairadapt` codes and the causal graph of Figure 9.10.

Part IV

Mitigation

"*Technology is neither good nor bad; nor is it neutral,*" Kranzberg (1986)[1]

"*Machine learning won't give you anything like gender neutrality 'for free' that you didn't explicitly ask for,*" Kearns and Roth (2019)

[1]Also called "Kranzberg's First Law". "*By that I mean that technology's interaction with the social ecology is such that technical developments frequently have environmental, social, and human consequences that go far beyond the immediate purposes of the technical devices and practices themselves, and the same technology can have quite different results when introduced into different contexts or under different circumstances,*" Kranzberg (1995).

Chapter 10

Pre-processing

"Pre-processing" is about distorting the training sample to insurance that the model we obtain is "fair", with respect to some criteria (defined in the previous chapters). The two standard techniques are either to modifies the original dataset (and to distort features to make them "fair", or independent to the sensitive attribute), or to use weights (as used in surveys to correct for biases). If there are poor theoretical guarantees, there are also legal issues with those techniques.

As we have seen previously, given a dataset $\mathcal{D}_n$ which is a collection of observations $(\boldsymbol{x}_i, s_i, y_i)$, it is possible to estimate a model $\widehat{m}$ (as discussed in Part II) and to quantify the fairness of the model (as discussed in Part III) based on appropriate metrics. Therefore, there are different ways to mitigate a possible discrimination, if any. The first one is to modify $\mathcal{D}_n$ (namely "pre-processing", described in this Chapter), to modify the training algorithm ("in-processing", possibly in adding a fairness constraint in the objective function, as described in Chapter 8), or by distorting the trained model $\widehat{m}$ ("post-processing", as described in Chapter 9).

Pre-processing techniques can be divided into two categories. The first one modifies the original data, as suggested in Calmon et al. (2017) and Feldman et al. (2015), but it does not have much statistical guarantees (see Propositions 7.2.6 and 7.2.8, where we have seen that if we were able to insure that $\boldsymbol{X}^{\perp} \perp\!\!\!\perp S$, we can still have $\widehat{Y} = m(\boldsymbol{X}^{\perp}) \not\perp\!\!\!\perp S$ – even $\widehat{Y} = m(\boldsymbol{X}^{\perp}) \not\perp S$). Barocas and Selbst (2016) and Krasanakis et al. (2018) also question the legality of that approach where training data are, somehow, falsified. An another approach is based on reweighing, where instead of having observations with equal weights in the training sample, we adjust weights in the training sample (in the training function to be more specific, so it could actually be seen as an "in-processing" approach). Kamiran and Calders (2012) suggests simply to have two weights, depending if s_i is either `A` or `B`, as well as Jiang and Nachum (2020). Heuristically, the idea is to amplify the error from an "underrepresented group" in the training sample, so that the optimization procedure can equally update a model for different group.

10.1 Removing Sensitive Attributes

The most simple "pre-processing" approach is simply to remove the sensitive attribute s. As discussed in Chapter 8, this corresponds to the concept of "fairness by unawarness". That is what is required by the

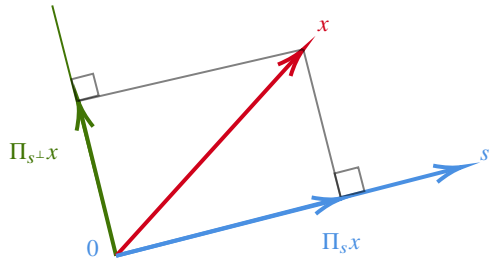


Figure 10.1: Orthogonal projection of a vector $\vec{x}$ on the line generated by $\vec{s}$, $\Pi_s x$, and its orthogonal component, $\Pi_{s^\perp} x$

"*gender directive*" (C-236 / 09, of March 2011), in Europe, where insurers are no longer allowed to use the gender to price insurance products.

As mentioned in Section 1.1.7 , mitigating discrimination is usually seen as paradoxical, because in order to avoid discrimination, we must create another discrimination. The point of the "*reverse discrimination objection*," as defined in Goldman (1979), is that there is an absolute ethical constraint against unfair discrimination. The motto of that approach could be the quote of John G. Roberts of the US Supreme Court: "*the way to stop discrimination on the basis of race is to stop discriminating on the basis of race*", as mentioned in Sabbagh (2007). Under that principle, this approach is actually the only one that should be used, a ban on the use of any sensitive attribute.

In Italy, Porrini and Fusco (2020) used data from the National Institute for the Supervision of Insurance (IVASS) to understand what was the effect of such a ban on the market prices. They measure the influence on the premiums of the gender variable (that was still in the databases, even if insurers were not allowed to use it to discriminate) and others variables such as the age of the driver, the type of vehicle, and the geolocation, for the period 2011-2014. The price paid as an insurance premium by male and female measured was collected in that dataset. They observed that the legal limitation in the use of a rating factor such as gender may have effects on the market with increases in premiums. More precisely, the effect of the gender discrimination ban is that women are not directly discriminated by gender, since after the ban, the premium is the same for male and female with identical other features, but overall, "*there is less gender equality because in the same conditions a woman pays more that a man, given the effects of other risky variables*".

10.2 Orthogonalization

10.2.1 General Case

This approach is a classical one in econometric literature, when linear models are considered. Interestingly, it is possible to use it even if there are multiple sensitive attributes. It is also the one discussed in Frees and Huang (2023). Write the $n \times k$ matrix $\boldsymbol{S}$ as a collection of k vectors in $\mathbb{R}^n$, $\boldsymbol{S} = \begin{pmatrix} \boldsymbol{s}_1 & \cdots & \boldsymbol{s}_k \end{pmatrix}$, that will correspond to k sensitive attributes. The orthogonal projection on variables $\{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_k\}$ is associate to matrix $\Pi_S = \boldsymbol{S}(\boldsymbol{S}^\top \boldsymbol{S})^{-1}\boldsymbol{S}^\top$, while the projection on the orthogonal of $\boldsymbol{S}$ is $\Pi_{S^\perp} = \mathbb{I} - \Pi_S$ (see Gram-Schmidt orthogonalization, and Figure 10.1). Let $\widetilde{\boldsymbol{S}}$ denote the collection of centered vectors (using matrix notations, $\widetilde{\boldsymbol{S}} = \boldsymbol{H}\boldsymbol{S}$ where $\boldsymbol{H} = \mathbb{I} - (\mathbf{1}\mathbf{1}^\top)/n$).

Write the $n \times p$ matrix $\boldsymbol{X}$ as a collection of p vectors in $\mathbb{R}^n$, $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_p \end{pmatrix}$. For any $\boldsymbol{x}_j$, define

$$\boldsymbol{x}_j^\perp = \Pi_{\widetilde{S}^\perp}\boldsymbol{x}_j = \boldsymbol{x}_j - \widetilde{\boldsymbol{S}}(\widetilde{\boldsymbol{S}}^\top\widetilde{\boldsymbol{S}})^{-1}\widetilde{\boldsymbol{S}}^\top\boldsymbol{x}_j$$

One can easily prove that $\boldsymbol{x}_j^\perp$ is then orthogonal to any $\boldsymbol{s}$, since

$$\text{Cov}(\boldsymbol{s}, \boldsymbol{x}_j^\perp) = \frac{1}{n}\boldsymbol{s}^\top \boldsymbol{H}\boldsymbol{x}_j^\perp = \frac{1}{n}\widetilde{\boldsymbol{s}}^\top \Pi_{\widetilde{S}^\perp}\boldsymbol{x}_j = 0$$

And similarly the centred version of $\boldsymbol{x}_j^\perp$ is then also orthogonal to any $\boldsymbol{s}$. From an econometric perspective, $\boldsymbol{x}_j^\perp$ can be seen as the residual of the regression of $\boldsymbol{x}_j$ against $\boldsymbol{s}$'s, obtained from least square estimation

$$\boldsymbol{x}_j = \widetilde{\boldsymbol{s}}^\top \widehat{\boldsymbol{\beta}}_j + \boldsymbol{x}_j^\perp.$$

Quite naturally, instead of training a model on $\mathcal{D}_n = (\boldsymbol{x}_i, s_i, y_i)$, we could use $\mathcal{D}_n = (\boldsymbol{x}_i^\perp, y_i)$, where, by construction all variables $\boldsymbol{x}_j^\perp$ are now orthogonal to any $\boldsymbol{s}$. Unfortunately, as mentioned in Proposition 7.2.6, if $\boldsymbol{X}^\perp \perp\!\!\!\perp S$, we could still have $\widehat{Y} = m(\boldsymbol{X}^\perp) \not\perp\!\!\!\perp S$ if we consider a non-linear class of model for m.

10.2.2 Binary Sensitive Attribute

Asking for orthogonality, $X_j \perp S$ is related to a null correlation between the two variables. But if S is binary, this can be simplified. Recall that if $\boldsymbol{x} = (x_1, \cdots, x_n)$, then

$$\overline{x} = \frac{1}{n}(\mathbf{1}^\top \boldsymbol{x}) \text{ and } \text{Var}(\boldsymbol{x}) = \frac{1}{n}\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} \text{ where } \boldsymbol{H} = \mathbb{I} - \frac{1}{n}(\mathbf{1}\mathbf{1}^\top),$$

$\boldsymbol{H}$ being an idempotent (projection) matrix (since $\boldsymbol{H}^2 = \boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$). The empirical covariance is defined, using matrix notations

$$\text{Cov}(\boldsymbol{s}, \boldsymbol{x}) = \frac{1}{n}\boldsymbol{s}^\top \boldsymbol{H}\boldsymbol{x}$$

Observe that Pearson's linear correlation is

$$\text{Cor}(\boldsymbol{s}, \boldsymbol{x}) = \frac{\boldsymbol{H}\boldsymbol{s}}{\|\boldsymbol{H}\boldsymbol{s}\|} \cdot \frac{\boldsymbol{H}\boldsymbol{x}}{\|\boldsymbol{H}\boldsymbol{x}\|} = \frac{\boldsymbol{s}^\top \boldsymbol{H}\boldsymbol{x}}{\|\boldsymbol{H}\boldsymbol{s}\|\|\boldsymbol{H}\boldsymbol{x}\|},$$

since $\boldsymbol{H}$ is an idempotent (projection) matrix. If $\boldsymbol{s} \in \{0, 1\}^n$, the covariance can be written

$$\text{Cov}(\boldsymbol{s}, \boldsymbol{x}) = \overline{s}_0 \overline{s}_1 \big(\Delta\overline{x}_1 - \Delta\overline{x}_0\big),$$

where

$$\overline{s}_j = \frac{n_j}{n} \text{ and } \Delta\overline{x}_j = \frac{1}{n_j}\sum_{i:s_i=j}(x_i - \overline{x}).$$

Observe that to insure that our predictor is fair, we need to compute the correlation between y and s. If y is also binary, $\boldsymbol{y} \in \{0, 1\}^n$, the covariance between s and y can be written

$$\text{Cov}(\boldsymbol{s}, \boldsymbol{y}) = \frac{n_{11}}{n} - \frac{n_{1\bullet}}{n}\frac{n_{\bullet 1}}{n} \text{ where } n_{jk} = \sum_{i=1}^n \mathbf{1}_{s_i=j}\mathbf{1}_{S_i=k}.$$

10.3 Weights

The propensity score was introduced in Rosenbaum and Rubin (1983), as the probability of treatment assignment conditional on observed baseline covariates.. The propensity score exists in both randomized experiments and in observational studies, the difference is that in randomized experiments, the true propensity score is known (and is related to the design of the experiment) while in observational studies, the propensity score is unknown, and should be estimated. If the logistic regression is the classical technique used to estimate that conditional probability, McCaffrey et al. (2004) suggested some boosted regression approach, while Lee et al. (2010) suggested Bagging techniques.

On Figure 10.2, we can visualise $\omega \mapsto \text{Cor}[\widehat{m}_\omega(\boldsymbol{x}), \mathbf{1}_{\texttt{B}}(s)]$, on two datasets, `toydata2` (on the left) and `germancredit` (on the right), where $\widehat{m}_\omega$ is a logistic regression with weights proportional to 1 in class `A` and ω in class `B`. The large dot is the plain logistic regression (with identical weights).

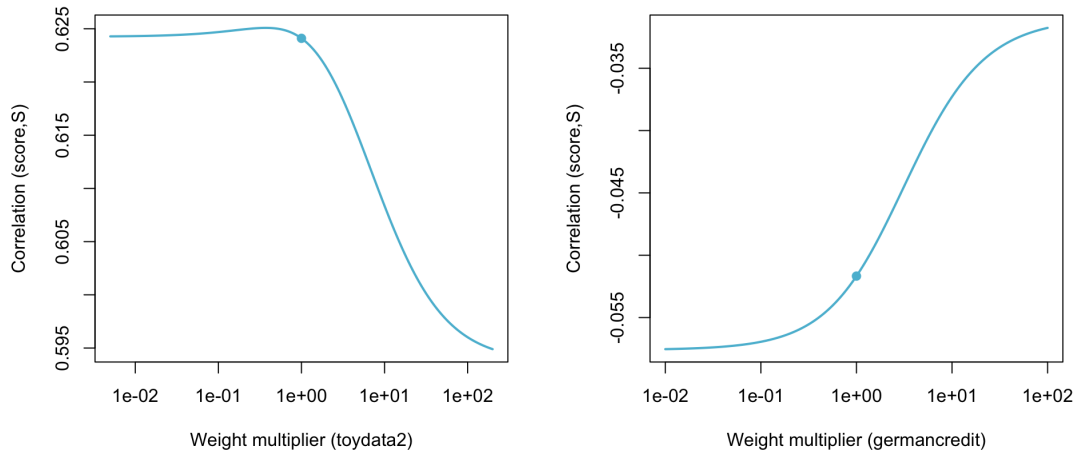


Figure 10.2: $\omega \mapsto \text{Cor}[\widehat{m}_\omega(\boldsymbol{x}, s), \mathbf{1}_{\texttt{B}}(s)]$, on two datasets, `toydata2` (on the left) and `germancredit` (on the right), where $\widehat{m}_\omega$ is a logistic regression with weights proportional to 1 in class `A` and ω in class `B`.

10.4 Application on `toydata2`

In the first block, on top, prediction for six individuals, using a logistic regression trained on `toydata2`, including the sensitive attribute. On the left, original values of the features $(\boldsymbol{x}_i, s_i)$. The second part corresponds to the transformation of the features, $\boldsymbol{x}_i^\perp$, so that each feature is orthogonal to the sensitive attribute ($\boldsymbol{x}_j^\perp$ is then orthogonal to any $\boldsymbol{s}$). Observe that features have on average the same values when conditioning with respect to the sensitive attribute. Base on orthogonalized features $\widehat{m}^\perp(\boldsymbol{x}^\perp)$ is the fitted unaware logistic regression. This model is fair with respect to demographic parity since the demographic parity ratio, $\mathbb{E}[\widehat{m}(\boldsymbol{Z})|S = \texttt{B}]/\mathbb{E}[\widehat{m}(\boldsymbol{Z})|S = \texttt{A}]$, is close to one. Observe further that the empirical correlation between $\widehat{m}(\boldsymbol{x}_i, s_i)'s$ and $\mathbf{1}_{\texttt{B}}(s_i)$'s was initially 0.72, and after orthogonalization, the empirical correlation between $\widehat{m}^\perp(\boldsymbol{x}_i^\perp)$'s and $\mathbf{1}_{\texttt{B}}(s_i)$'s is 0.01. At the bottom (third block), we can also visualize statistics about

equalized odds, with ratios$\mathbb{E}[\widehat{m}(\boldsymbol{Z})|S = \texttt{B}, Y = y]/\mathbb{E}[\widehat{m}(\boldsymbol{Z})|S = \texttt{A}, Y = y]$ for $y = 0$ (on top) and $y = 1$ (below). Values should be close to 1 to have equalized odds, which is not the case here. On the right, at the very last column, we consider a model based on weights in the training sample $\widehat{m}_\omega(\boldsymbol{x})$. For the choice of the weight, based, Figure 10.2 suggested to use very large weights for individuals in class B. This model is slightly more fair than the original plain logistic regression.

	x_1	x_2	x_3	s	y	$\widehat{m}(\boldsymbol{x}, s)$	$x_1^\perp$	$x_2^\perp$	$x_3^\perp$	$\widehat{m}^\perp(\boldsymbol{x}^\perp)$	$\widehat{m}_\omega(\boldsymbol{x})$
Alex	0	2	0	A	GOOD	0.138	0.957	-2.924	0.968	0.355	0.221
Betty	0	2	0	B	BAD	0.274	-0.958	-3.128	-0.938	0.123	0.221
Ahmad	1	5	1	A	BAD	0.546	1.957	0.076	1.968	0.709	0.761
Brienne	1	5	1	B	OOD	0.739	0.042	-0.128	0.062	0.383	0.761
Anthony	2	8	2	A	GOOD	0.900	2.957	3.076	2.968	0.915	0.973
Beatrix	2	8	2	B	BAD	0.955	1.042	2.872	1.062	0.733	0.973
average $\|S = \texttt{A}$	-0.967	4.919	-0.976		0.223	0.223	0.223	-0.010	-0.006	0.402	0.288
average $\|S = \texttt{B}$	0.958	5.132	0.935		0.675	0.675	0.223	-0.010	-0.006	0.408	0.674
(difference)	+2	~ 0	+2		$\times 3.027$	$\times 3.027$	~ 0	~ 0	~ 0	$\times 1.015$	$\times 2.340$
average $\|S = \texttt{A}, Y = 0$	-1.018	4.357	-0.993		0.000	0.184	0.000	-0.061	-0.567	0.362	0.239
average $\|S = \texttt{B}, Y = 0$	0.194	3.569	0.437		0.000	0.463	0.000	-0.061	-0.567	0.228	0.435
(difference)						$\times 2.516$	~ 0	~ 0	~ 0	$\times 0.630$	$\times 1.820$
average $\|S = \texttt{A}, Y = 1$	-0.788	6.873	-0.917		1.000	0.362	1.000	0.169	1.949	0.540	0.458
average $\|S = \texttt{B}, Y = 1$	1.325	5.884	1.175		1.000	0.777	1.000	0.169	1.949	0.494	0.789
(difference)						$\times 2.146$	~ 0	~ 0	~ 0	$\times 0.915$	$\times 1.723$

Table 10.1: Predictions using logistic regressions on `toydata2`, with two "pre-processing" approaches, with orthogonalized features $\widehat{m}^\perp(\boldsymbol{x}^\perp)$ and using weights $\widehat{m}_\omega(\boldsymbol{x})$. The first block describes six individuals. The second block is a check for demographic parity, with averages conditional on s, while the third block is a check for equalized odds, with averages conditional on s and y. At the bottom, statistics related to demographic parity (averages conditional on S), and equalized odds (averages conditional on S and Y).

On Figure 10.3 and 10.4, we can visualize the orthogonalization method, with, on Figure 10.3, the optimal transport plot, between distributions of $\widehat{m}(\boldsymbol{x}_i, s_i)$'s (on the x-axis) to $\widehat{m}^\perp(\boldsymbol{x}_i^\perp)$'s (on the y-axis), for individuals in group A on the left, and in group B on the right. On Figure 10.4, on the left, we can visualize densities of $\widehat{m}^\perp(\boldsymbol{x}_i^\perp)$'s from individuals in group A and in group B (thin lines are densities of scores from the original plain logistic regression $\widehat{m}(\boldsymbol{x}_i, s_i)$'s). On the right, we have the scatterplot of points $(\widehat{m}(\boldsymbol{x}_i, s_i = \texttt{A}), \widehat{m}^\perp(\boldsymbol{x}_i^\perp))$ and $(\widehat{m}(\boldsymbol{x}_i, s_i = \texttt{B}), \widehat{m}^\perp(\boldsymbol{x}_i^\perp))$. The six individuals on top of Table 10.2 are emphasized, with vertical segments showing the (individual) difference between the initial model, and the fair one.

On Figure 10.5 and 10.6, we can visualize similar plots for the weight method, with, on Figure 10.5, the optimal transport plot, between distributions of $\widehat{m}(\boldsymbol{x}_i, s_i)$'s (on the x-axis) to $\widehat{m}_\omega(\boldsymbol{x}_i)$'s (on the y-axis), for individuals in group A on the left, and in group B on the right. On Figure 10.6, on the left, we can visualize densities of $\widehat{m}_\omega(\boldsymbol{x}_i)$'s from individuals in group A and in group B (again, thin lines are the original densities of scores from the plain logistic regression $\widehat{m}(\boldsymbol{x}_i, s_i)$'s). On the right, we have the scatterplot of points $(\widehat{m}(\boldsymbol{x}_i, s_i = \texttt{A}), \widehat{m}^\perp(\boldsymbol{x}_i^\perp))$ and $(\widehat{m}(\boldsymbol{x}_i, s_i = \texttt{B}), \widehat{m}^\perp(\boldsymbol{x}_i^\perp))$. We emphasize again the six individuals on top of Table 10.2.

On Figure 10.7, visualization of the optimal transport between distributions of $\widehat{m}^\perp(\boldsymbol{x}_i^\perp)$'s for individuals in group A (x-axis) to individuals in group B (y-axis) to on the left, and $\widehat{m}^{\omega(\boldsymbol{x}_i)}$'s on the right. If the "transport

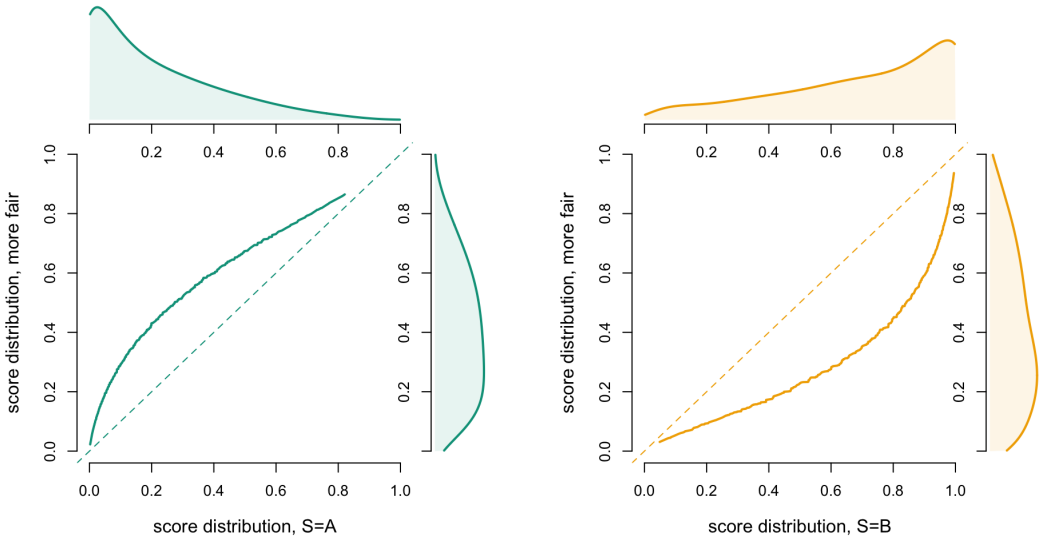


Figure 10.3: Optimal transport between distributions of $\widehat{m}(\boldsymbol{x}_i, s_i)$'s ($x$-axis) to $\widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp})$'s ($y$-axis), for individuals in group A on the left, and in group B on the right.

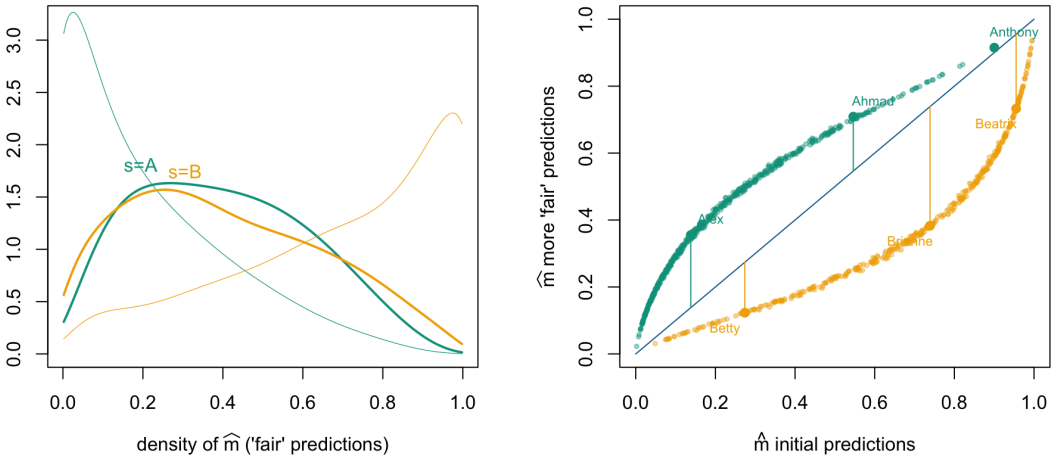


Figure 10.4: On the left, densities of $\widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp})$'s from individuals in group A and in B (thin lines are densities of scores from the plain logistic regression $\widehat{m}(\boldsymbol{x}_i, s_i)$'s). On the right, scatterplot of points $(\widehat{m}(\boldsymbol{x}_i, s_i = \texttt{A}), \widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp}))$ and $(\widehat{m}(\boldsymbol{x}_i, s_i = \texttt{B}), \widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp}))$, from the `toydata2` dataset.

line" is on the first diagonal, Wassertein distance between the two conditional distributions is close to 0, meaning that demographic parity is satisfied.

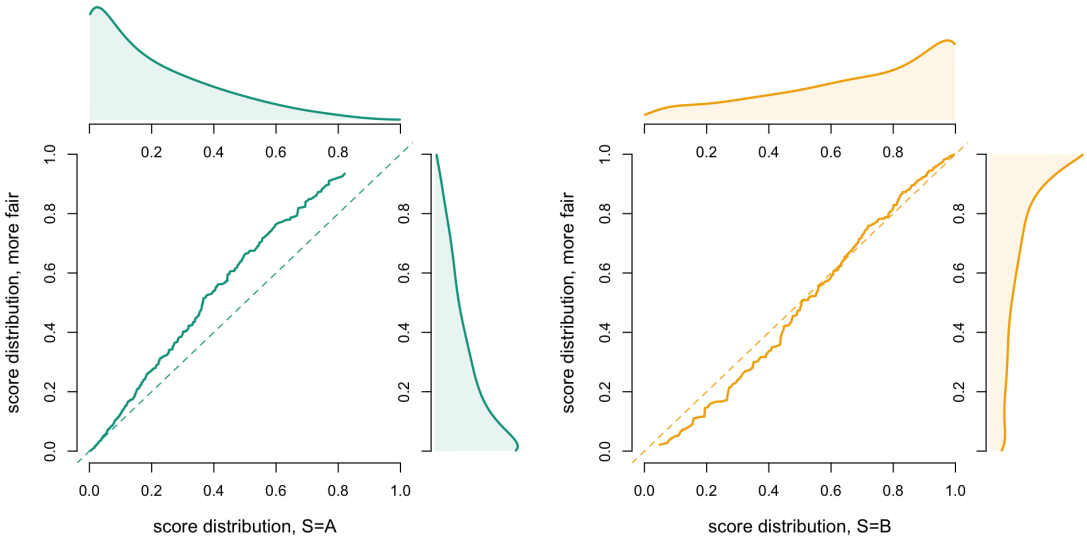


Figure 10.5: Optimal transport between distributions of $\widehat{m}(\boldsymbol{x}_i, s_i)$'s ($x$-axis) to $\widehat{m}_\omega(\boldsymbol{x}_i)$'s ($y$-axis), for individuals in group A on the left, and in group B on the right.

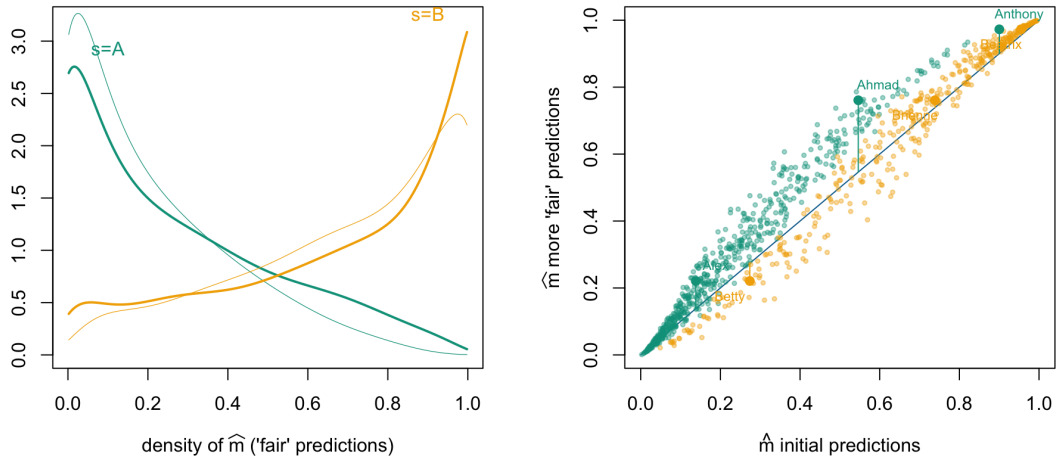


Figure 10.6: On the left, densities of $\widehat{m}_\omega(\boldsymbol{x}_i)$'s from individuals in group A and in B (thin lines are densities of scores from the plain logistic regression $\widehat{m}(\boldsymbol{x}_i, s_i)$'s). On the right, scatterplot of points $(\widehat{m}(\boldsymbol{x}_i, s_i = \text{A}), \widehat{m}_\omega(\boldsymbol{x}_i))$ and $(\widehat{m}(\boldsymbol{x}_i, s_i = \text{B}), \widehat{m}_\omega(\boldsymbol{x}_i))$, from the `toydata2` dataset.

10.5 Application on the `germancredit` Dataset

The same analysis can be performed on the `germancredit` dataset, with here also the plain logistic regression, but other techniques described in Chapter 3 could be considered. For the orthogonalization, it is performed on

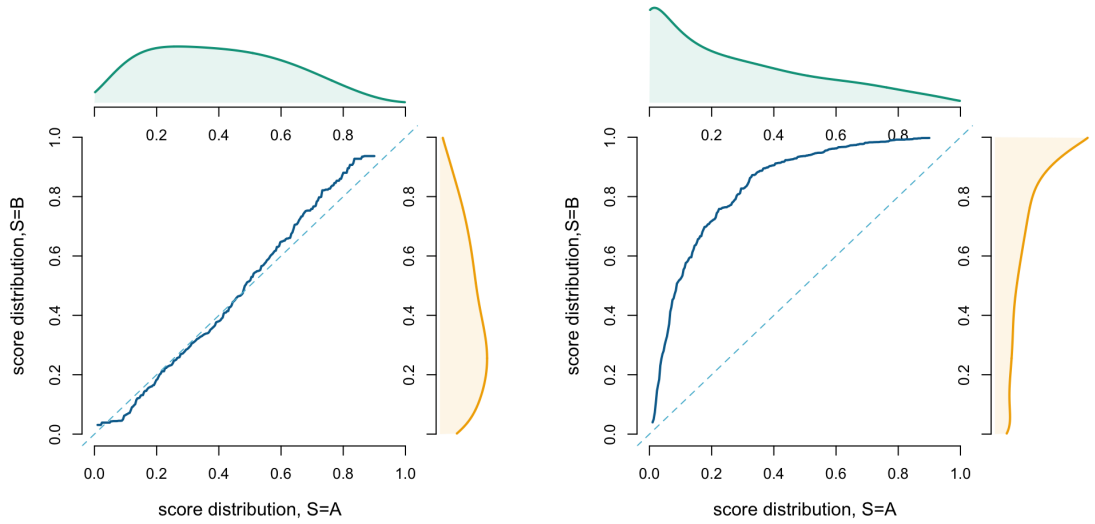


Figure 10.7: Optimal transport between distributions of $\widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp})$'s for individuals in group A (x-axis) to individuals in group B (y-axis) to on the left, and $\widehat{m}_{\omega}(\boldsymbol{x}_i)$'s on the right.

the $\boldsymbol{X}$ design matrix, that contains indicators for all factors (but the reference) for categorical variable. Observe that the empirical correlation between $\widehat{m}(\boldsymbol{x}_i, s_i)'s$ and $\mathbf{1}_{\mathtt{B}}(s_i)$'s was initially -0.195. After orthogonalization, the empirical correlation between $\widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp})$'s and $\mathbf{1}_{\mathtt{B}}(s_i)$'s is now 0.009. In Table 10.2, we have averages of score predictions using the plain logistic regressions on `germandata` on top. Since averages of $\widehat{m}(\boldsymbol{x}_i, s_i = \mathtt{A})$ and $\widehat{m}(\boldsymbol{x}_i, s_i = \mathtt{B})$ are respectively 35.2% and 27.7%, the demographic parity ratio is 78.7% which is far away from 1, it does not even satisfy the "*four-fifths rule*" for disparate impact (corresponding to a 80% threshold in Definition 8.6.1). The second row is the first "pre-processing" approacs, with orthogonalized features $\widehat{m}^{\perp}(\boldsymbol{x}^{\perp})$. The demographic parity ratio is not 1.010, which means that model $\widehat{m}^{\perp}$ could be considered as fair, with respect to this criteria. It could even be considered as fair with respect to the equalized odds criteria, since averages $\widehat{m}(\boldsymbol{x}_i)$ between individuals such that $s_i = \mathtt{A}$ and $s_i = \mathtt{B}$, in the group $y_i = 0$ (or `GOOD` risk), have a ratio of 1.045, while it is 1.054 in the group $y_i = 1$ (or `BAD` risk). The third row is obtained when weights are considered $\widehat{m}_{\omega}(\boldsymbol{x})$. From Figure 10.2, here also a large weight for individuals in class B is considered (so that the correlation between the scores $\widehat{m}_{\omega}(\boldsymbol{x}_i)$ and $\mathbf{1}_{\mathtt{B}}(s_i)$ gets closer to 0). We can observe that this model can also be considered as fair, since averages $\widehat{m}(\boldsymbol{x}_i)$ between individuals such that $s_i = \mathtt{A}$ and $s_i = \mathtt{B}$ could be considered as similar.

On Figure 10.8 and 10.9, we can visualize the orthogonalization method, with, on Figure 10.8, the optimal transport plot, between distributions of $\widehat{m}(\boldsymbol{x}_i, s_i)$'s (on the x-axis) to $\widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp})$'s (on the y-axis), for individuals in group A on the left, and in group B on the right. On Figure 10.9, on the left, we can visualize densities of $\widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp})$'s from individuals in group A and in group B (thin lines are densities of scores from the original plain logistic regression $\widehat{m}(\boldsymbol{x}_i, s_i)$'s). On the right, we have the scatterplot of points $(\widehat{m}(\boldsymbol{x}_i, s_i = \mathtt{A}), \widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp}))$ and $(\widehat{m}(\boldsymbol{x}_i, s_i = \mathtt{B}), \widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp}))$. The six individuals mentioned (in Table 9.7) are again emphasised.

On Figure 10.10 and 10.11, we can visualize similar plots for the weight method, with, on Figure 10.10, the optimal transport plot, between distributions of $\widehat{m}(\boldsymbol{x}_i, s_i)$'s (on the x-axis) to $\widehat{m}_{\omega}(\boldsymbol{x}_i)$'s (on the y-axis), for individuals in group A on the left, and in group B on the right. On Figure 10.11, on the left, we can

	demographic parity			equalized odds ($y = 0$)			equalized odds ($y = 1$)		
	A	B	(ratio)	A	B	(ratio)	A	B	(ratio)
$\widehat{m}(\boldsymbol{x}, s)$	0.352	0.277	×0.787	0.299	0.237	×0.792	0.448	0.381	×0.850
$\widehat{m}^{\perp}(\boldsymbol{x}^{\perp})$	0.298	0.301	×1.010	0.249	0.260	×1.045	0.387	0.408	×1.054
$\widehat{m}_{\omega}(\boldsymbol{x})$	0.289	0.277	×0.958	0.249	0.235	×0.946	0.363	0.386	×1.065

Table 10.2: Averages of score predictions using the plain logistic regressions on `germandata` on top, with two "pre-processing" approaches, with orthogonalized features $\widehat{m}^{\perp}(\boldsymbol{x}^{\perp})$ and using weights $\widehat{m}_{\omega}(\boldsymbol{x})$, below. The second block is a check for demographic parity, with averages conditional on s, while the second block is a check for equalized odds, with averages conditional on s and y.

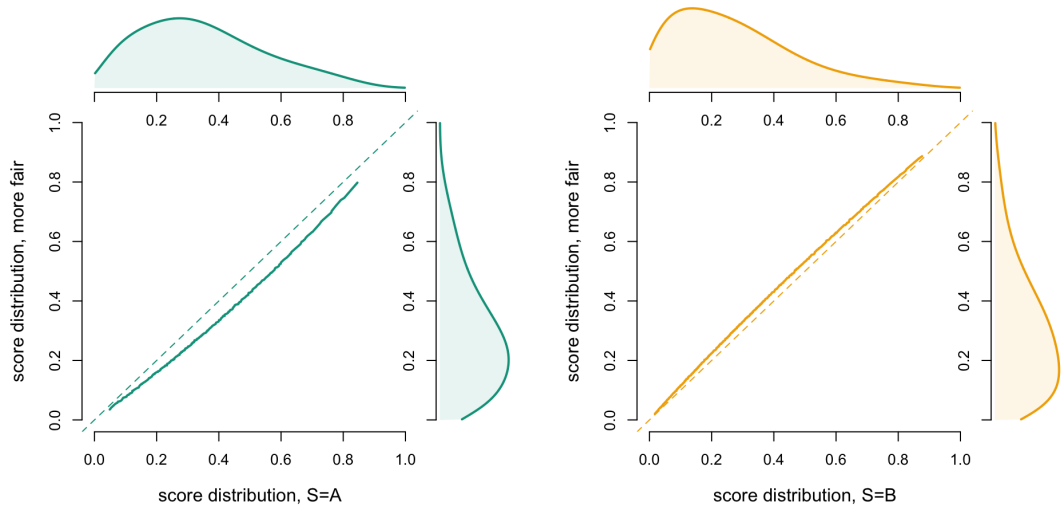


Figure 10.8: Optimal transport between distributions of $\widehat{m}(\boldsymbol{x}_i, s_i)$'s ($x$-axis) to $\widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp})$'s ($y$-axis), for individuals in group A on the left, and in group B on the right.

visualize densities of $\widehat{m}_{\omega}(\boldsymbol{x}_i)$'s from individuals in group A and in group B (again, thin lines are the original densities of scores from the plain logistic regression $\widehat{m}(\boldsymbol{x}_i, s_i)$'s). On the right, we have the scatterplot of points $(\widehat{m}(\boldsymbol{x}_i, s_i = \mathtt{A}), \widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp}))$ and $(\widehat{m}(\boldsymbol{x}_i, s_i = \mathtt{B}), \widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp}))$.

On Figure 10.12, visualization of the optimal transport between distributions of $\widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp})$'s for individuals in group A (x-axis) to individuals in group B (y-axis) to on the left, and $\widehat{m}_{\omega}(\boldsymbol{x}_i)$'s on the right. If the "transport line" is on the first diagonal, Wassertein distance between the two conditional distributions is close to 0, meaning that demographic parity is satisfied.

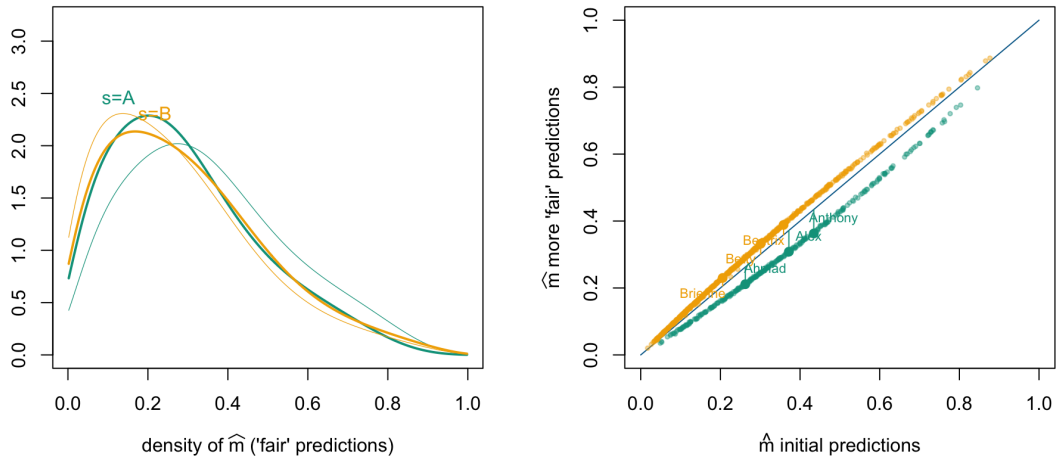


Figure 10.9: On the left, densities of $\widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp})$'s from individuals in group A and in B (thin lines are densities of scores from the plain logistic regression $\widehat{m}(\boldsymbol{x}_i, s_i)$'s). On the right, scatterplot of points $(\widehat{m}(\boldsymbol{x}_i, s_i = \texttt{A}), \widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp}))$ and $(\widehat{m}(\boldsymbol{x}_i, s_i = \texttt{B}), \widehat{m}^{\perp}(\boldsymbol{x}_i^{\perp}))$, from the `germancredit` dataset.

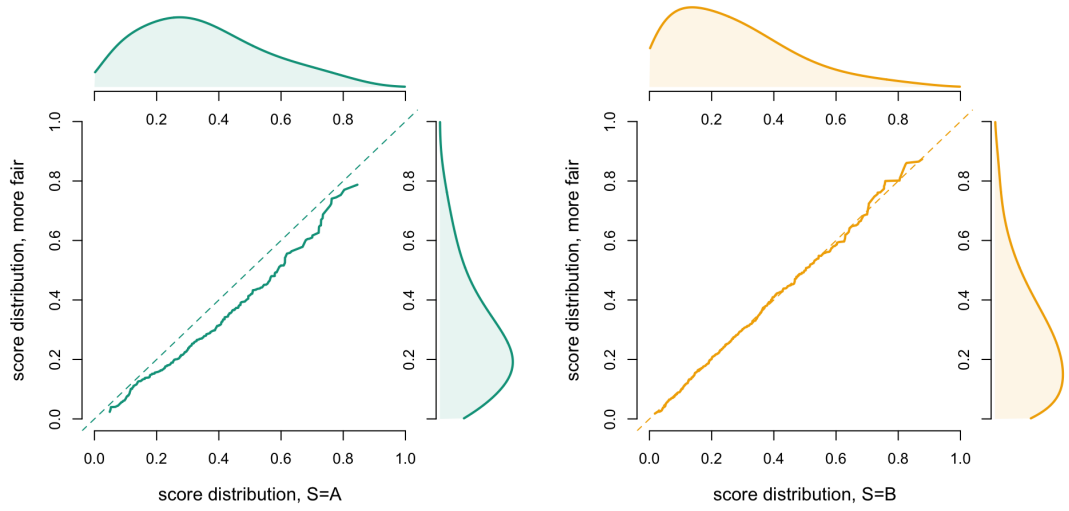


Figure 10.10: Optimal transport between distributions of $\widehat{m}(\boldsymbol{x}_i, s_i)$'s ($x$-axis) to $\widehat{m}_{\omega}(\boldsymbol{x}_i)$'s ($y$-axis), for individuals in group A on the left, and in group B on the right.

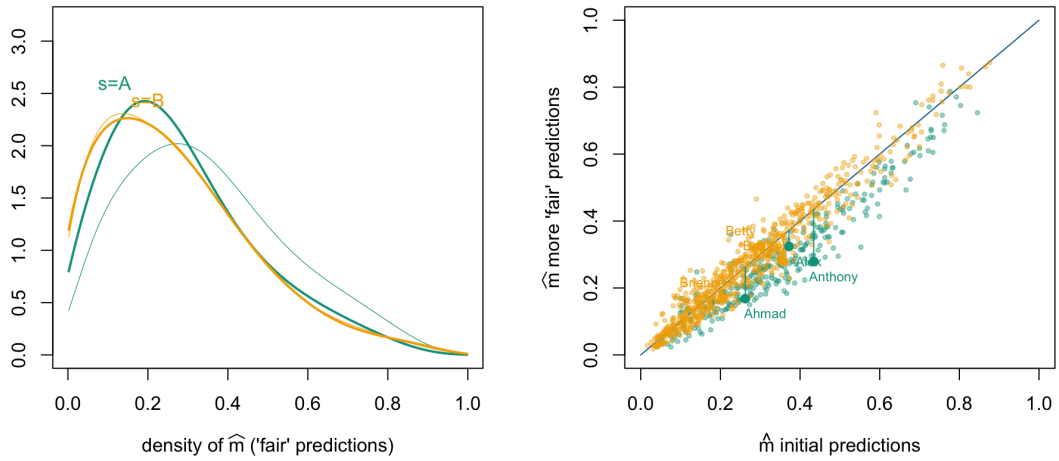


Figure 10.11: On the left, densities of $\widehat{m}_\omega(\boldsymbol{x}_i)$'s from individuals in group A and in B (thin lines are densities of scores from the plain logistic regression $\widehat{m}(\boldsymbol{x}_i, s_i)$'s). On the right, scatterplot of points $(\widehat{m}(\boldsymbol{x}_i, s_i = \texttt{A}), \widehat{m}_\omega(\boldsymbol{x}_i))$ and $(\widehat{m}(\boldsymbol{x}_i, s_i = \texttt{B}), \widehat{m}_\omega(\boldsymbol{x}_i))$, from the `germancredit` dataset.

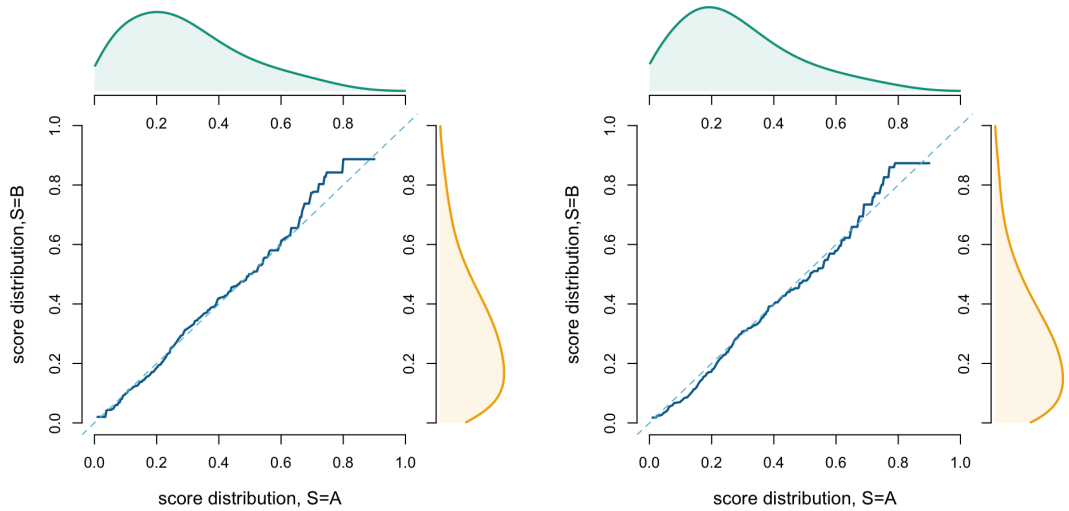


Figure 10.12: Optimal transport between distributions of $\widehat{m}^\perp(\boldsymbol{x}_i^\perp)$'s for individuals in group A (x-axis) to individuals in group B (y-axis) to on the left, and $\widehat{m}_\omega(\boldsymbol{x}_i)$'s on the right, on the `germancredit` dataset.

Chapter 11

In-processing

Classically, to estimate a model, we look for a model (in a pre-defined class) which minimizes a prediction error, or which maximizes the accuracy. If the model is required to satisfy constraints, a natural idea is to add a penalty term in the objective function. The idea of "in-processing" is to get a tradeoff between accuracy and fairness. As previously, we will present that approach on some datasets.

Classically, we have seen (see Definition 3.3.3) that models were obtained by minimizing the empirical risk, the sample based version of $\mathcal{R}(\widehat{m})$

$$\widehat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \Big\{ \mathcal{R}(m) \Big\} \text{ where } \mathcal{R}(m) = \mathbb{E}\Big[\ell\left(Y, m(\boldsymbol{X}, S)\right) \Big],$$

for some set of models, $\mathcal{M}$. Quite naturally, we could consider a constrained optimization problem

$$\widehat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \Big\{ \mathcal{R}(m) \Big\} \text{ s.t. } m \text{ fair},$$

for some fairness criterion. Using standard results in optimization, one could consider a penalized version of that problem

$$\widehat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \Big\{ \mathcal{R}(m) \Big\} + \lambda R(m),$$

where "$R(m)$" denotes a positive regularizer, indicating the extent to which the fairness criterion is violated, as suggested in Zafar et al. (2017) Zhang and Bareinboim (2018), Agarwal et al. (2018), Kearns et al. (2018), Li and Fan (2020) and Li and Liu (2022). As a (technical) drawback, adding a regularizer that is non-convex could rise optimization complexity, as mentioned in Roth et al. (2017) and Cotter et al. (2019).

11.1 Adding a Group Discrimination Penalty

Recall that weak demographic parity (Section 8.2) is achieved if

$$\mathbb{E}\Big[\widehat{m}(X, S)\Big|S = s\Big] = \mathbb{E}\Big[\widehat{m}(X, S)\Big], \ \forall s,$$

while weak equal opportunity (Section 8.3) is achieved if

$$\mathbb{E}\big[\widehat{m}(X,S)\big|S=s,Y=1\big]=\mathbb{E}\big[\widehat{m}(X,S)\big|Y=1\big],\ \forall s.$$

And since strong demographic parity is achieved if

$$\mathbb{P}\big[\widehat{m}(X,S)\in\mathcal{A}\big|S=s\big]=\mathbb{P}\big[\widehat{m}(X,S)\in\mathcal{A}\big],\ \forall s,\forall\mathcal{A},$$

or, with synthetic notations, $\mathbb{P}_{\mathtt{A}}[\mathcal{A}]=\mathbb{P}_{\mathtt{B}}[\mathcal{A}]=\mathbb{P}[\mathcal{A}]$. Thus, a classical metric for strong demographic parity would be

$$w_{\mathtt{A}}D_{\mathrm{KL}}(\mathbb{P}_{\mathtt{A}}\|\mathbb{P})+w_{\mathtt{B}}D_{\mathrm{KL}}(\mathbb{P}_{\mathtt{B}}\|\mathbb{P})$$

for some appropriate weights $w_{\mathtt{A}}$ and $w_{\mathtt{B}}$. Strong equal opportunity is achieved if

$$\mathbb{P}\big[\widehat{m}(X,S)\in\mathcal{A}\big|S=s,Y=1\big]=\mathbb{P}\big[\widehat{m}(X,S)\in\mathcal{A}\big|Y=1\big],\ \forall s,\forall\mathcal{A},$$

and therefore, a metric for strong equal opportunity would be

$$w_{\mathtt{A}}D_{\mathrm{KL}}(\mathbb{P}_{\mathtt{A},Y=1}\|\mathbb{P}_{Y=1})+w_{\mathtt{B}}D_{\mathrm{KL}}(\mathbb{P}_{\mathtt{B},Y=1}\|\mathbb{P}_{Y=1}).$$

In the classical logistic regression, we solve

$$\widehat{\boldsymbol{\beta}}=\underset{\boldsymbol{\beta}}{\operatorname{argmax}}\Big\{\log\mathcal{L}(\boldsymbol{\beta})\Big\},$$

$$\log\mathcal{L}(\boldsymbol{\beta})=\sum_{i=1}^{n}y_i\log[m(\boldsymbol{x}_i)]+(1-y_i)\log[1-m(\boldsymbol{x}_i)]$$

and $m(\boldsymbol{x}_i)=\dfrac{\exp[\boldsymbol{x}_i^\top\boldsymbol{\beta}]}{1+\exp[\boldsymbol{x}_i^\top\boldsymbol{\beta}]}$. Inspired by Zafar et al. (2019), fairness constraints related to disparate impact and disparate mistreatment criteria discussed previously, that should be strictly satisfied, could be introduced. For example, weak demographic parity is achieved if $\mathbb{E}[\widehat{m}(\boldsymbol{X})|S=\mathtt{A}]=\mathbb{E}[\widehat{m}(\boldsymbol{X})|S=\mathtt{B}]$ and therefore

$$\widehat{\boldsymbol{\beta}}_*=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\Big\{-\log\mathcal{L}(\boldsymbol{\beta})\Big\},\ \text{s.t. }\mathbb{E}[\widehat{m}(\boldsymbol{X})|S=\mathtt{A}]=\mathbb{E}[\widehat{m}(\boldsymbol{X})|S=\mathtt{B}]$$

One could consider a more flexible version, where mistreatment is bounded,

$$\widehat{\boldsymbol{\beta}}_\epsilon=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\Big\{-\log\mathcal{L}(\boldsymbol{\beta})\Big\},\ \text{s.t. }\Big|\mathbb{E}[\widehat{m}(\boldsymbol{X})|S=\mathtt{A}]-\mathbb{E}[\widehat{m}(\boldsymbol{X})|S=\mathtt{B}]\Big|\leq\epsilon.$$

Quite naturally, one can consider a penalized version of that constrained optimization problem

$$\widehat{\boldsymbol{\beta}}_\lambda=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\Big\{-\log\mathcal{L}(\boldsymbol{\beta})+\lambda\Big|\mathbb{E}[\widehat{m}(\boldsymbol{X})|S=\mathtt{A}]-\mathbb{E}[\widehat{m}(\boldsymbol{X})|S=\mathtt{B}]\Big|\Big\}.$$

Following Scutari et al. (2022), observe that one could write also

$$\widehat{\boldsymbol{\beta}}_\epsilon=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\Big\{-\log\mathcal{L}(\boldsymbol{\beta})\Big\},\ \text{s.t. }|\operatorname{cov}[m(\boldsymbol{x}),\mathbf{1}_{\mathtt{B}}(s)]|\leq\epsilon,$$

for some $\epsilon \geq 0$, or the penalized version

$$\widehat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\Big\{ -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda |\operatorname{cov}[m(\boldsymbol{x}), \mathbf{1}_{\mathtt{B}}(s)]| \Big\},$$

where $\widehat{m}(\boldsymbol{x}_i) = \dfrac{\exp[\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}]}{1+\exp[\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}]}$.

The sample based version of the penalty would be

$$\widehat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\Big\{ -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \Big| \frac{1}{n_{\mathtt{A}}} \sum_{i:s_i=\mathtt{A}} \widehat{m}(\boldsymbol{x}_i) - \frac{1}{n_{\mathtt{B}}} \sum_{i:s_i=\mathtt{B}} \widehat{m}(\boldsymbol{x}_i) \Big| \Big\},$$

or, since

$$\operatorname{cov}[\widehat{m}(\boldsymbol{x}), \mathbf{1}_{\mathtt{B}}(s)] = \mathbb{E}\big[\widehat{m}(\boldsymbol{x})[\mathbf{1}_{\mathtt{B}}(s) - \mathbb{E}(\mathbf{1}_{\mathtt{B}}(s))]\big] - \underbrace{\mathbb{E}\big[\mathbf{1}_{\mathtt{B}}(s) - \mathbb{E}[\mathbf{1}_{\mathtt{B}}(s)]\big]}_{=0} \widehat{m}(\boldsymbol{x})$$

the sample-based version of the penalty can also be written

$$\operatorname{cov}[\widehat{m}(\boldsymbol{x}), \mathbf{1}_{\mathtt{B}}(s)] = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{1}_{\mathtt{B}}(s_i) - \overline{\mathbf{1}_{\mathtt{B}}}) \cdot \widehat{m}(\boldsymbol{x}_i), \text{ where } \overline{\mathbf{1}_{\mathtt{B}}} = \frac{n_{\mathtt{B}}}{n}.$$

Komiyama et al. (2018) actually initiated that idea: it regresses the fitted values against the sensitive attributes and the response, and it bounds the proportion of the variance explained by the sensitive attributes over the total explained variance in that model. In R, functions `frrm` and `nclm` in the package `fairml` performs such regression, where in the first one, the constraint is actually written as a Ridge penalty (see Definition 3.3.15). Demographic parity is achieved with option `"sp-komiyama"` while equalized odds is obtained with option `"eo-komiyama"`.

11.2 Adding an Individual Discrimination Penalty

Observe that instead of a regularization based on a group fairness criteria, one could consider some individual fairness criteria. With notations of Section 11.1, the penalized risk can be defined, for a binary sensitive attribute (if S is not equal to s, it is equal to s') as

$$\mathcal{R}_\lambda(\widehat{m}) = \mathbb{E}\Big[\ell\left(Y, \widehat{m}(\boldsymbol{X}, S)\right)\Big] + \lambda \sum_{s\in\mathcal{S}} \mathbb{P}[S=s] \cdot \mathbb{E}\Big[r\left(\boldsymbol{X}_s, s, \boldsymbol{X}_{s'}, s'\right)\Big],$$

s' being the other category, and where $r(\boldsymbol{x}, s, \boldsymbol{x}', s')$ is penalty that enforces the difference between the outcomes for an individual and its counterfactual version. Russell et al. (2017) considered

$$r(\boldsymbol{x}, s, \boldsymbol{x}', s') = \big|\widehat{m}(\boldsymbol{x}, s) - \widehat{m}(\boldsymbol{x}', s')\big|,$$

and more generally

$$r_\varepsilon(\boldsymbol{x}, s, \boldsymbol{x}', s') = \max\big\{0, \big|\widehat{m}(\boldsymbol{x}, s) - \widehat{m}(\boldsymbol{x}', s')\big| - \varepsilon\big\},$$

for ε-fairness, as introduced in Section 8.6, that corresponds to a convex relaxation of the previous condition. For instance, de Lara (2023) considered

$$r(\boldsymbol{x}, s, \boldsymbol{x}', s') = \big(\widehat{m}(\boldsymbol{x}, s) - \widehat{m}(\boldsymbol{x}', s')\big)^2.$$

11.3 Application on `toydata2`

11.3.1 Demographic parity

On Figure 11.1 and 11.2, we can visualize AUC($\widehat{m}_\lambda$) against the demographic parity ratio $\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \mathtt{B}]/\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \mathtt{A}]$, on Figure 11.1 when $\boldsymbol{Z} = (\boldsymbol{X}, S)$ and on Figure 11.2 when $\boldsymbol{Z} = \boldsymbol{X}$ (unaware model), where, with general notations, the prediction is obtained from a logistic regression

$$\widehat{m}_{\widehat{\beta}_\lambda}(z) = \frac{\exp[z^\top \widehat{\boldsymbol{\beta}}_\lambda]}{1 + \exp[z^\top \widehat{\boldsymbol{\beta}}_\lambda]}$$

where $\widehat{\boldsymbol{\beta}}_\lambda$ is a solution of a penalized maximum likelihood problem,

$$\widehat{\beta}_\lambda \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\big\{ -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \cdot |\operatorname{cov}[m_{\boldsymbol{\beta}}(\boldsymbol{x}), \mathbf{1}_{\mathtt{B}}(s)]|\big\}.$$

In Table 11.1, we have the outcome of penalized logistic regressions, for different values of λ, including s on the left and excluding s on the right. On top, values of $\widehat{\boldsymbol{\beta}}_\lambda$ in the first block, and predictions $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$ and $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i)$ for a series of individuals in the second block. Below, the demographic parity ratio, $\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \mathtt{B}]/\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \mathtt{A}]$ (fairness is achieved when the ratio is 1), and AUC (the higher the value, the higher the accuracy of the model)

	$\widehat{m}(\boldsymbol{x}, s)$, aware					$\widehat{m}(\boldsymbol{x})$, unaware			
	← less fair			more fair →		← less fair		more fair →	
$\widehat{\boldsymbol{\beta}}_0$ (Intercept)	-2.55	-2.29	-1.97	-1.51	-1.03	-2.14	-1.98	-1.78	-1.63
$\widehat{\boldsymbol{\beta}}_1$ (x_1)	0.88	0.88	0.85	0.77	0.62	1.01	0.84	0.57	0.26
$\widehat{\boldsymbol{\beta}}_2$ (x_2)	0.37	0.37	0.35	0.32	0.25	0.37	0.35	0.31	0.24
$\widehat{\boldsymbol{\beta}}_3$ (x_3)	0.02	0.02	0.02	0.02	0.03	0.15	0.02	-0.15	-0.29
$\widehat{\boldsymbol{\beta}}_{\mathtt{B}}$ ($\mathbf{1}_{\mathtt{B}}$)	0.82	0.44	-0.03	-0.70	-1.31	-	-	-	-
Betty	0.27	0.25	0.22	0.17	0.14	0.20	0.22	0.24	0.24
Brienne	0.74	0.71	0.66	0.54	0.40	0.70	0.66	0.55	0.38
Beatrix	0.95	0.95	0.93	0.87	0.73	0.96	0.93	0.82	0.55
Alex	0.14	0.17	0.22	0.29	0.37	0.20	0.22	0.24	0.24
Ahmad	0.55	0.61	0.66	0.70	0.71	0.70	0.66	0.55	0.38
Anthony	0.90	0.92	0.93	0.93	0.91	0.96	0.93	0.82	0.55
$\mathbb{E}[\widehat{m}(\boldsymbol{x}_i, s_i)\|S = \mathtt{A}]$	0.23	0.26	0.31	0.36	0.42	0.25	0.30	0.37	0.41
$\mathbb{E}[\widehat{m}(\boldsymbol{x}_i, s_i)\|S = \mathtt{B}]$	0.67	0.65	0.61	0.53	0.42	0.64	0.61	0.54	0.41
(ratio)	×2.97	×2.49	×2.01	×1.46	×1.00	×2.53	×2.02	×1.48	×1.00
AUC	0.86	0.86	0.85	0.82	0.74	0.86	0.85	0.82	0.70

Table 11.1: Penalized logistic regression, for different values of λ, including s on the left and excluding s on the right. On top, values of $\widehat{\boldsymbol{\beta}}_\lambda$ in the first block, and predictions $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$ and $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i)$ for a series of individuals in the second block. Below, the demographic parity ratio, $\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \mathtt{B}]/\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \mathtt{A}]$ (fairness is achieved when the ratio is 1), and AUC (the higher the value, the higher the accuracy of the model).

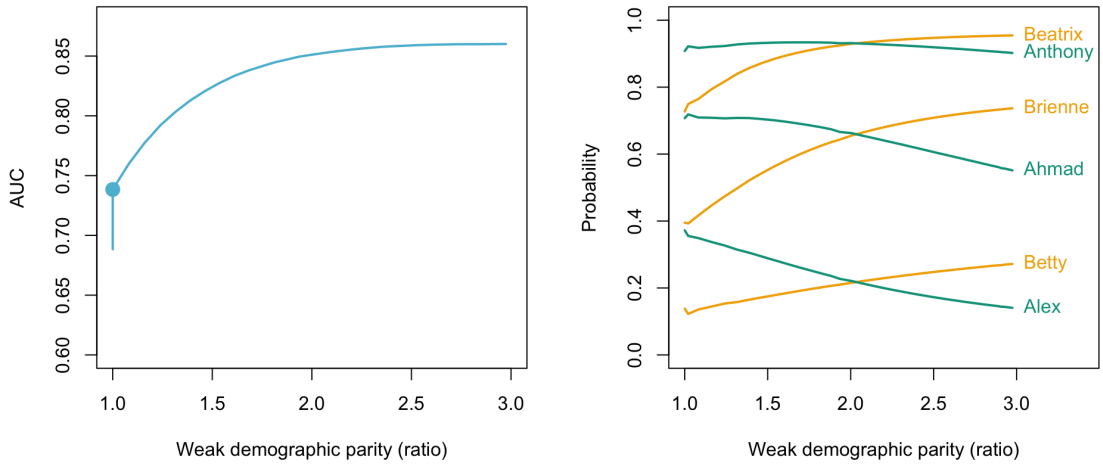


Figure 11.1: On the left, accuracy-fairness tradeoff plot (base on Table 11.1), with the AUC of $\widehat{m}_{\widehat{\beta}_\lambda}$ on the y-axis and the fairness ratio (demographic parity) on the x-axis. Top left corresponds to accurate and fair. On the right, evolution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$ (with a logistic regression) for three individuals in group A and three in B, on `toydata2` dataset.

In Figure 11.11, we have the optimal transport plot, between distributions of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s from individuals in group A and in B, for different values of λ (low value on the left and high value on the right), associated with a demographic parity penalty criteria.

On Figure 11.4, we have, on the left, densities of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s from individuals in group A and in B (thin lines are densities of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s). On the right, scatterplot of points $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \texttt{A}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s = \texttt{A}))$ and $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \texttt{B}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s = \texttt{B}))$, where $\widehat{m}_{\widehat{\beta}}$ is the plain logistic regression, and $\widehat{m}_{\widehat{\beta}_\lambda}$ is the penalized logistic regression, from `toydata2` dataset, associated with a demographic parity penalty criteria.

On Figure 11.5, optimal transport plot, from distribution of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s to the distribution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s, for individuals in group A on the left, and in group B on the right, for different values of λ, with a low value on the top and a high value on the bottom (fair model, associated with a demographic parity penalty criteria).

11.3.2 Equalized odds and class balance

On Figure 11.6 and 11.7, we can visualize $\text{AUC}(\widehat{m}_\lambda)$ against the demographic parity ratio $\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \texttt{B}]/\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \texttt{A}]$, on Figure 11.6 when $\boldsymbol{Z} = (\boldsymbol{X}, S)$ and on Figure **??** when $\boldsymbol{Z} = \boldsymbol{X}$ (unaware model), where, with general notations, the prediction is obtained from a logistic regression

$$\widehat{m}_{\widehat{\beta}_\lambda}(z) = \frac{\exp[z^\top \widehat{\boldsymbol{\beta}}_\lambda]}{1 + \exp[z^\top \widehat{\boldsymbol{\beta}}_\lambda]}$$

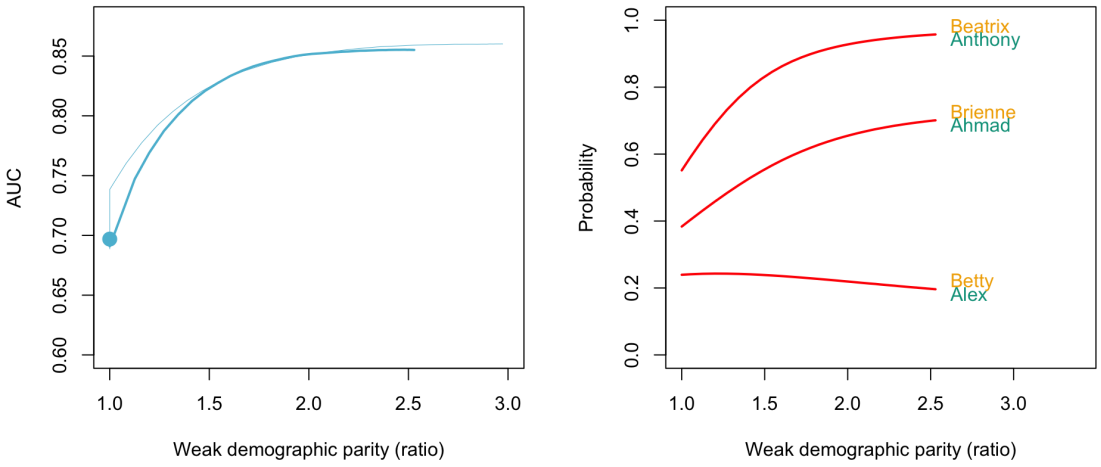


Figure 11.2: On the left, accuracy-fairness (demographic parity) tradeoff plot (base on Table 11.1), with the AUC of $\widehat{m}_{\widehat{\beta}_\lambda}$ on the y-axis and the fairness ratio on the x-axis. The thin line in the one with the model taking into account s (from Figure 11.1). On the right, evolution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i)$ (with a logistic regression) for three individuals in group A and three in B, on the `toydata2` dataset.

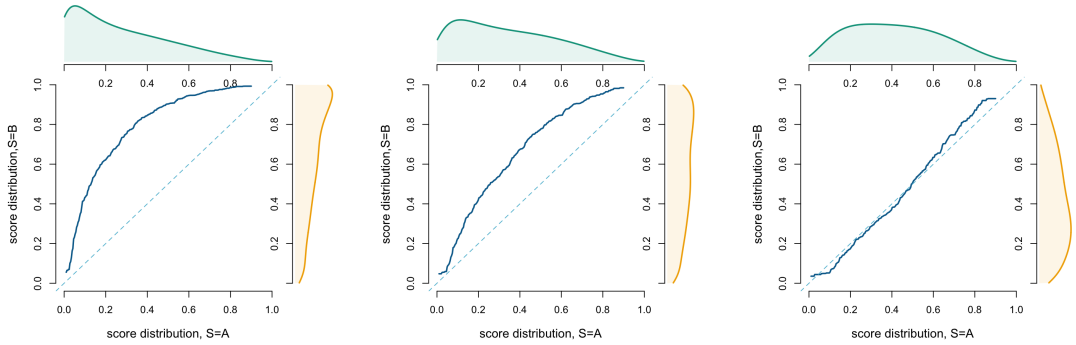


Figure 11.3: Optimal transport between distributions of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s from individuals in group A and in B, for different values of λ (low value on the left and high value on the right), associated with a demographic parity penalty criteria.

where $\widehat{\boldsymbol{\beta}}_\lambda$ is a solution of a penalized maximum likelihood problem,

$$\widehat{\beta}_\lambda \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\Big\{ -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \cdot |\operatorname{cov}[m_{\boldsymbol{\beta}}(\boldsymbol{x}), \mathbf{1}_{\mathrm{B}}(s)]|\Big\}.$$

In Table 11.2, we can visualize the outcome of some penalized logistic regressions, for different values of λ, including s on the left and excluding s on the right. On top, values of $\widehat{\boldsymbol{\beta}}_\lambda$ in the first block, and predictions

$\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$ and $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i)$ for a series of individuals in the second block. Below, the class balance ratios, $\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \mathtt{B}, Y = y]/\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \mathtt{A}, Y = y]$ (fairness is achieved when the ratio is 1), and AUC (the higher the value, the higher the accuracy of the model).

	$\widehat{m}(\boldsymbol{x}, s)$, aware							
	← less fair						more fair →	
$\widehat{\boldsymbol{\beta}}_0$ (Intercept)	-2.55	-2.45	-2.34	-2.21	-2.27	-2.08	-1.44	-2.61
$\widehat{\boldsymbol{\beta}}_1$ (x_1)	0.89	0.90	0.92	0.94	0.88	0.82	0.52	0.39
$\widehat{\boldsymbol{\beta}}_2$ (x_2)	0.37	0.37	0.37	0.37	0.41	0.40	0.30	0.39
$\widehat{\boldsymbol{\beta}}_3$ (x_3)	0.02	0.03	0.03	0.03	0.01	0.00	0.04	-0.42
$\widehat{\boldsymbol{\beta}}_\mathtt{B}$ ($\mathbf{1}_\mathtt{B}$)	0.81	0.63	0.39	0.11	-0.03	-0.48	-0.60	0.40
Betty	0.27	0.25	0.23	0.20	0.19	0.15	0.19	0.20
Brienne	0.74	0.72	0.70	0.67	0.66	0.57	0.51	0.44
Beatrix	0.95	0.95	0.95	0.94	0.94	0.91	0.81	71
Alex	0.14	0.15	0.17	0.19	0.19	0.22	0.30	0.14
Ahmad	0.55	0.58	0.61	0.65	0.66	0.68	0.65	0.33
Anthony	0.90	0.91	0.93	0.94	0.94	0.94	0.89	0.60
$\mathbb{E}[\widehat{m}(\boldsymbol{x}_i, s_i)\|S = \mathtt{A}, Y = 0]$	0.19	0.20	0.21	0.23	0.26	0.29	0.36	0.33
$\mathbb{E}[\widehat{m}(\boldsymbol{x}_i, s_i)\|S = \mathtt{B}, Y = 0]$	0.46	0.44	0.42	0.39	0.38	0.32	0.33	0.33
(ratio)	2.47	2.25	2.00	1.74	1.47	1.10	0.91	1.00
$\mathbb{E}[\widehat{m}(\boldsymbol{x}_i, s_i)\|S = \mathtt{A}, Y = 1]$	0.37	0.38	0.40	0.43	0.48	0.52	0.55	0.54
$\mathbb{E}[\widehat{m}(\boldsymbol{x}_i, s_i)\|S = \mathtt{B}, Y = 1]$	0.78	0.76	0.75	0.73	0.72	0.66	0.59	0.54
(ratio)	2.12	2.00	1.86	1.72	1.50	1.26	1.09	1.00
(global ratio)	2.47	2.25	2.00	1.74	1.50	1.26	1.09	1.00
AUC	0.86	0.86	0.86	0.86	0.85	0.83	0.80	0.75

Table 11.2: Penalized logistic regression, for different values of λ, including s on the left and excluding s on the right. On top, values of $\widehat{\boldsymbol{\beta}}_\lambda$ in the first block, and predictions $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$ and $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i)$ for a series of individuals in the second block. Below, the class balance ratios, $\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \mathtt{B}, Y = y]/\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \mathtt{A}, Y = y]$ (fairness is achieved when the ratio is 1), and AUC (the higher the value, the higher the accuracy of the model).

In Figure 11.8, we can visualize, on the left, densities of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s from individuals in group A and in B (thin lines are densities of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s). On the right, scatterplot of points $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \mathtt{A}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i), s = \mathtt{A})$ and $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \mathtt{B}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i), s = \mathtt{B})$, where $\widehat{m}_{\widehat{\beta}}$ is the plain logistic regression, and $\widehat{m}_{\widehat{\beta}_\lambda}$ is the penalized logistic regression, from `toydata2` dataset, associated with a class balance penalty criteria

On Figure 11.9, we have the optimal transport plot, from distribution of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s to the distribution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s, for individuals in group A on the left, and in group B on the right, for different values of λ, with a low value on the top and a high value on the bottom (fair model, associated with a class balance (equalized odds) penalty criteria).

11.4 Application on the `germancredit` Dataset

11.4.1 Demographic parity

On Figure 11.10 and 11.11, we can visualize AUC($\widehat{m}_\lambda$) against the demographic parity ratio $\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \texttt{B}]/\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \texttt{A}]$, on Figure 11.1 when $\boldsymbol{Z} = (\boldsymbol{X}, S)$ and on Figure 11.2 when $\boldsymbol{Z} = \boldsymbol{X}$ (unaware model), where, with general notations, the prediction is obtained from a logistic regression

$$\widehat{m}_{\widehat{\beta}_\lambda}(z) = \frac{\exp[z^\top \widehat{\boldsymbol{\beta}}_\lambda]}{1 + \exp[z^\top \widehat{\boldsymbol{\beta}}_\lambda]}$$

where $\widehat{\boldsymbol{\beta}}_\lambda$ is a solution of a penalized maximum likelihood problem,

$$\widehat{\beta}_\lambda \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\big\{ -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \cdot |\operatorname{cov}[m_{\boldsymbol{\beta}}(\boldsymbol{x}), \mathbf{1}_{\texttt{B}}(s)]|\big\}.$$

On Figure 11.12, we can visualize densities of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s from individuals in group A and in B (thin lines are densities of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s), on the left. On the right, scatterplot of points $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \texttt{A}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i), s = \texttt{A})$ and $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \texttt{B}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i), s = \texttt{B})$, where $\widehat{m}_{\widehat{\beta}}$ is the plain logistic regression, and $\widehat{m}_{\widehat{\beta}_\lambda}$ is the penalized logistic regression, from `toydata2` dataset, associated with a demographic parity penalty criteria

On Figure 11.13, we have the optimal transport plot, from distribution of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s to the distribution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s, for individuals in group A on the left, and in group B on the right, for different values of λ, with a low value on the top and a high value on the bottom (fair model, associated with a demographic parity penalty criteria)

11.4.2 Equalized odds and class balance

On Figure 11.14 and 11.15, we can visualize AUC($\widehat{m}_\lambda$) against the demographic parity ratio $\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \texttt{B}]/\mathbb{E}[\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{Z})|S = \texttt{A}]$, on Figure 11.1 when $\boldsymbol{Z} = (\boldsymbol{X}, S)$ and on Figure 11.2 when $\boldsymbol{Z} = \boldsymbol{X}$ (unaware model), where, with general notations, the prediction is obtained from a logistic regression

$$\widehat{m}_{\widehat{\beta}_\lambda}(z) = \frac{\exp[z^\top \widehat{\boldsymbol{\beta}}_\lambda]}{1 + \exp[z^\top \widehat{\boldsymbol{\beta}}_\lambda]}$$

where $\widehat{\boldsymbol{\beta}}_\lambda$ is a solution of a penalized maximum likelihood problem,

$$\widehat{\beta}_\lambda \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\big\{ -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \cdot |\operatorname{cov}[m_{\boldsymbol{\beta}}(\boldsymbol{x}), \mathbf{1}_{\texttt{B}}(s)]|\big\}.$$

On Figure 11.16, we can visualize densities of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s from individuals in group A and in B (thin lines are densities of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s), on the left. On the right, scatterplot of points $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \texttt{A}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i), s = \texttt{A})$ and $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \texttt{B}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i), s = \texttt{B})$, where $\widehat{m}_{\widehat{\beta}}$ is the plain logistic regression, and $\widehat{m}_{\widehat{\beta}_\lambda}$ is the penalized logistic regression, from `toydata2` dataset, associated with a class balance penalty criteria.

On Figure 11.17, we can visualize the optimal transport plot, from distribution of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s to the distribution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s, for individuals in group A on the left, and in group B on the right, for different values of λ, with a low value on the top and a high value on the bottom (fair model, associated with a class balance (equalized odds) penalty criteria).

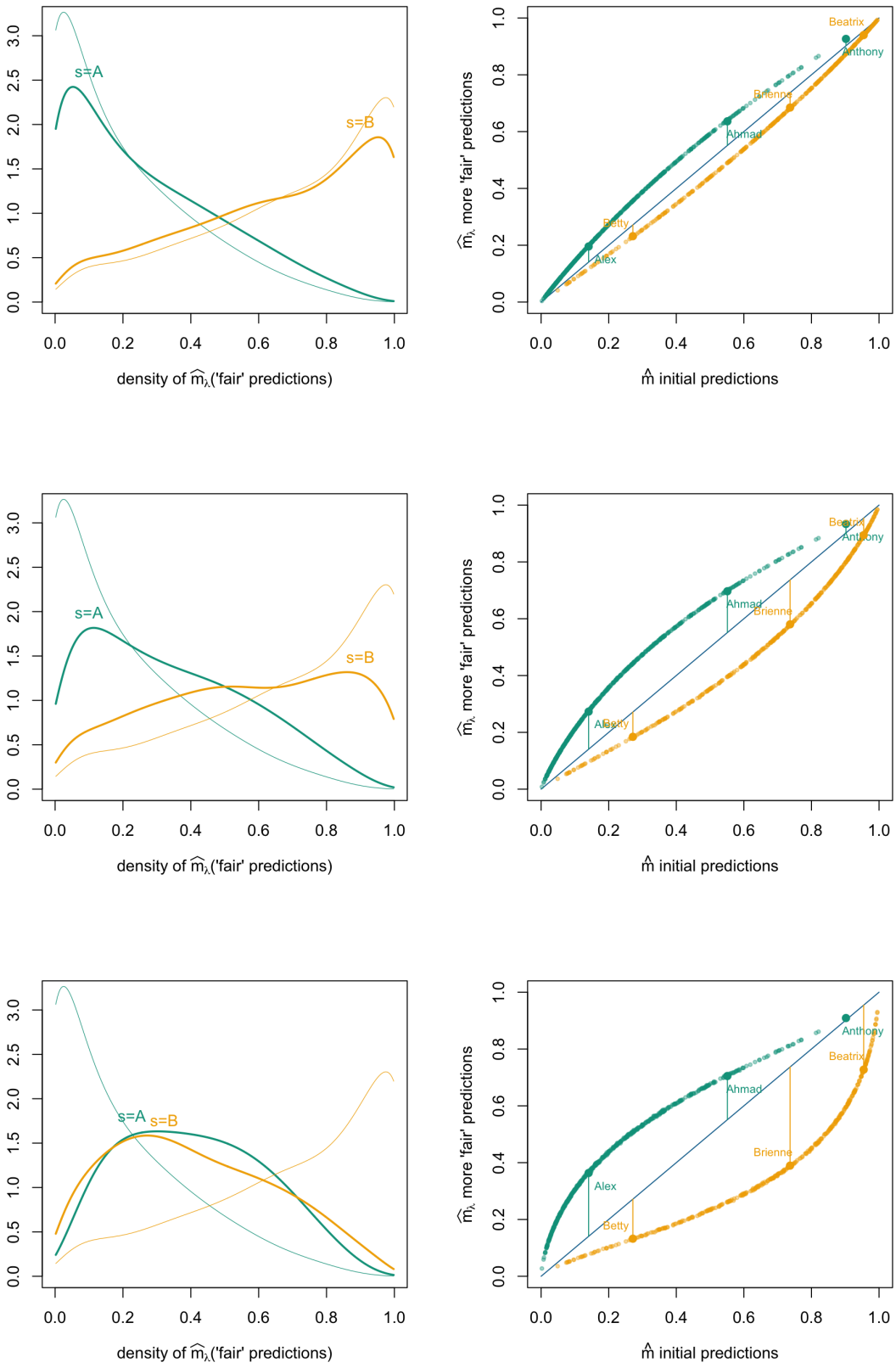


Figure 11.4: On the left, densities of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s from individuals in group A and in B (thin lines are densities of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s). On the right, scatterplot of points $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \texttt{A}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i), s = \texttt{A})$ and $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \texttt{B}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i), s = \texttt{B})$, where $\widehat{m}_{\widehat{\beta}}$ is the plain logistic regression, and $\widehat{m}_{\widehat{\beta}_\lambda}$ is the penalized logistic regression, from `toydata2` dataset, associated with a demographic parity penalty criteria.

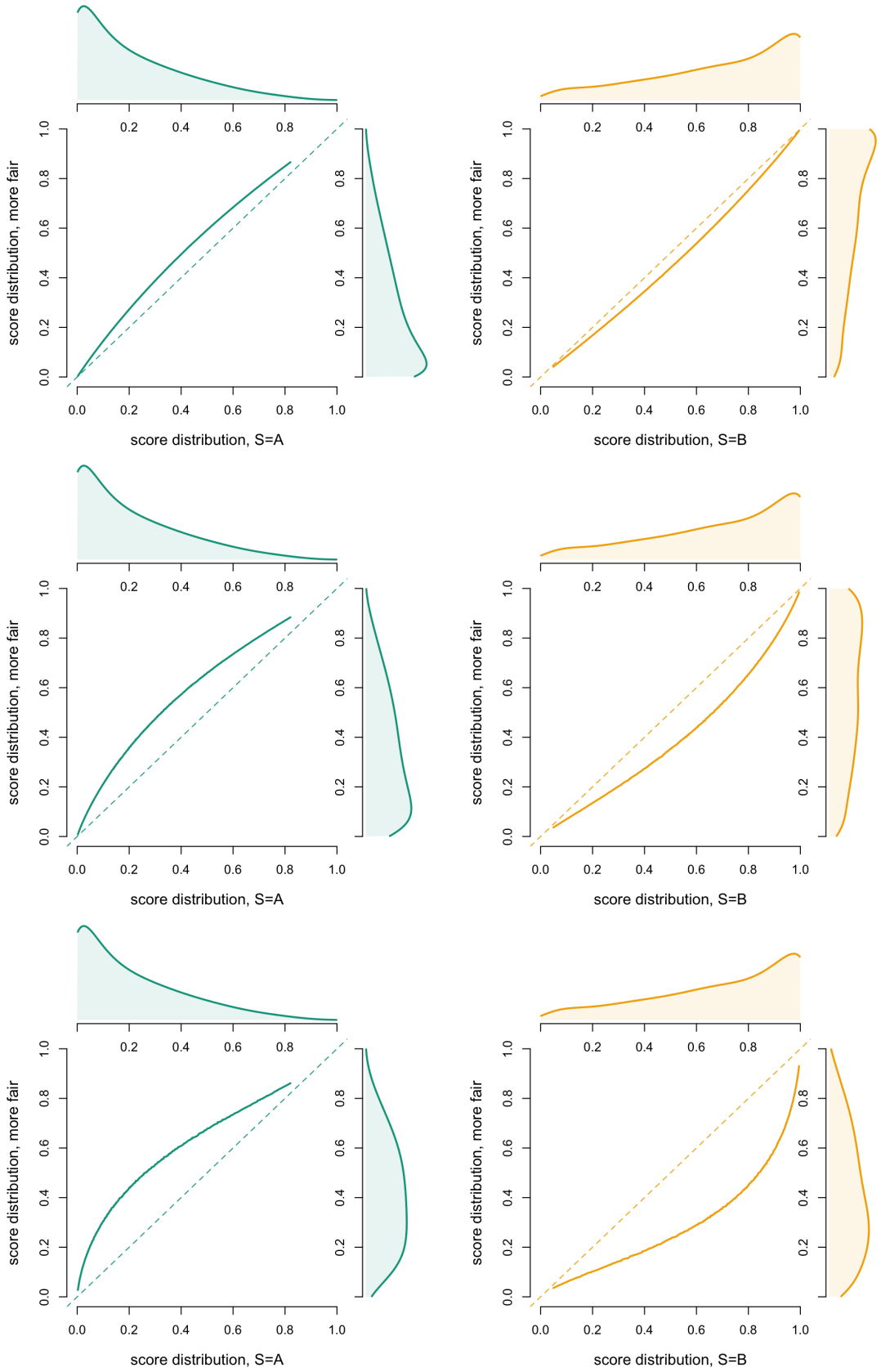


Figure 11.5: Optimal transport from distribution of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s to the distribution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s, for individuals in group A on the left, and in group B on the right, for different values of λ, with a low value on the top and a high value on the bottom (fair model, associated with a demographic parity penalty criteria).

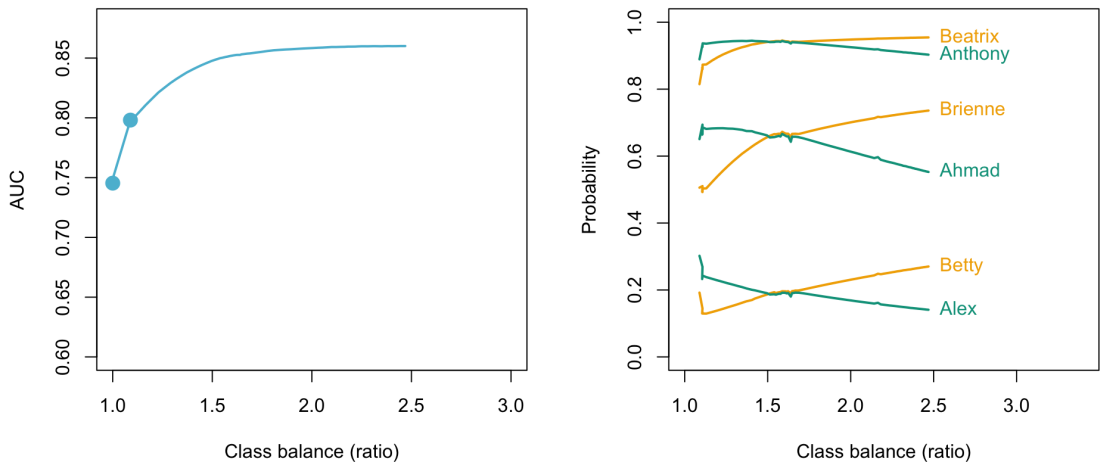


Figure 11.6: On the left, accuracy-fairness trade-off plot (base on Table 11.1), with the AUC of $\widehat{m}_{\widehat{\beta}_\lambda}$ on the y-axis and the fairness ratio (class balance) on the x-axis. Top left corresponds to accurate and fair. On the right, evolution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$ (with a logistic regression) for three individuals in group A and three in B, on `toydata2` dataset.

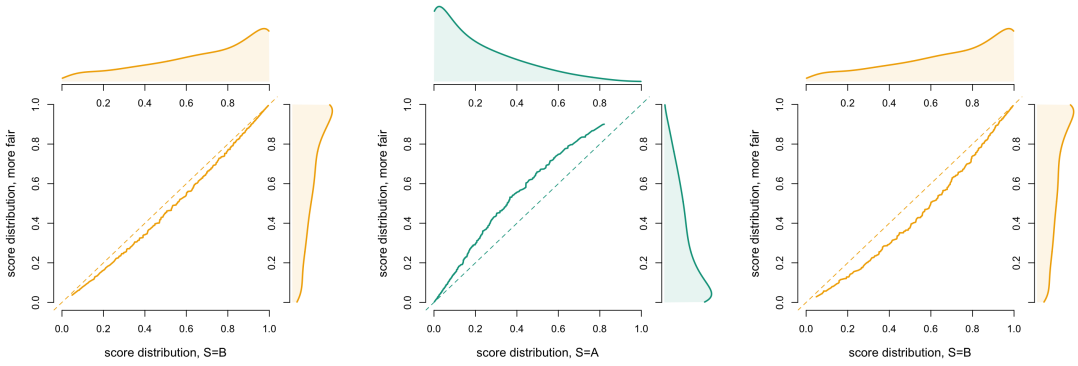


Figure 11.7: Optimal transport between distributions of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s from individuals in group A and in B, for different values of λ (low value on the left and high value on the right), associated with a class balance penalty criteria.

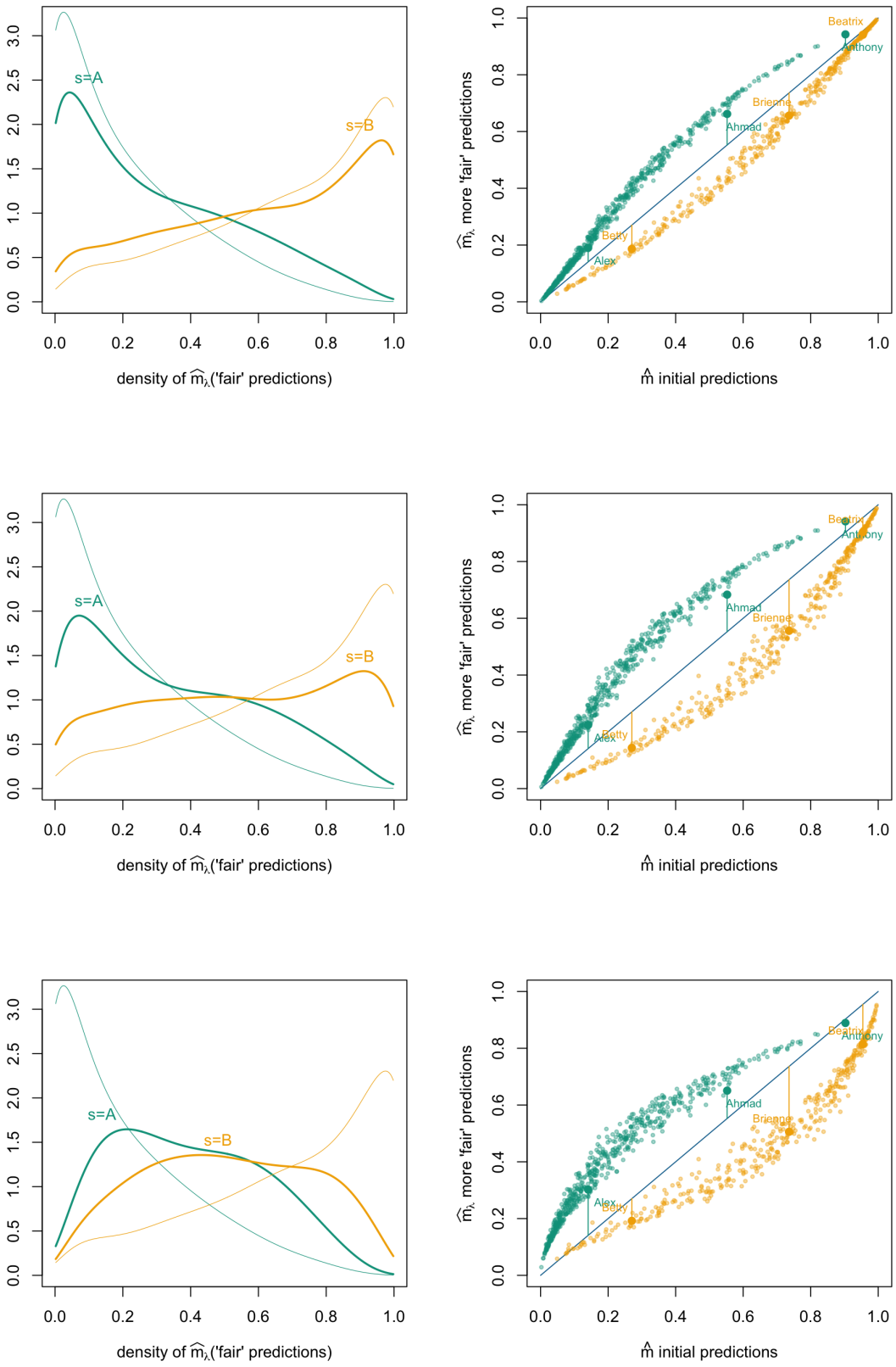


Figure 11.8: On the left, densities of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s from individuals in group A and in B (thin lines are densities of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s). On the right, scatterplot of points $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \texttt{A}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i), s = \texttt{A})$ and $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \texttt{B}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i), s = \texttt{B})$, where $\widehat{m}_{\widehat{\beta}}$ is the plain logistic regression, and $\widehat{m}_{\widehat{\beta}_\lambda}$ is the penalized logistic regression, from `toydata2` dataset, associated with a class balance penalty criteria.

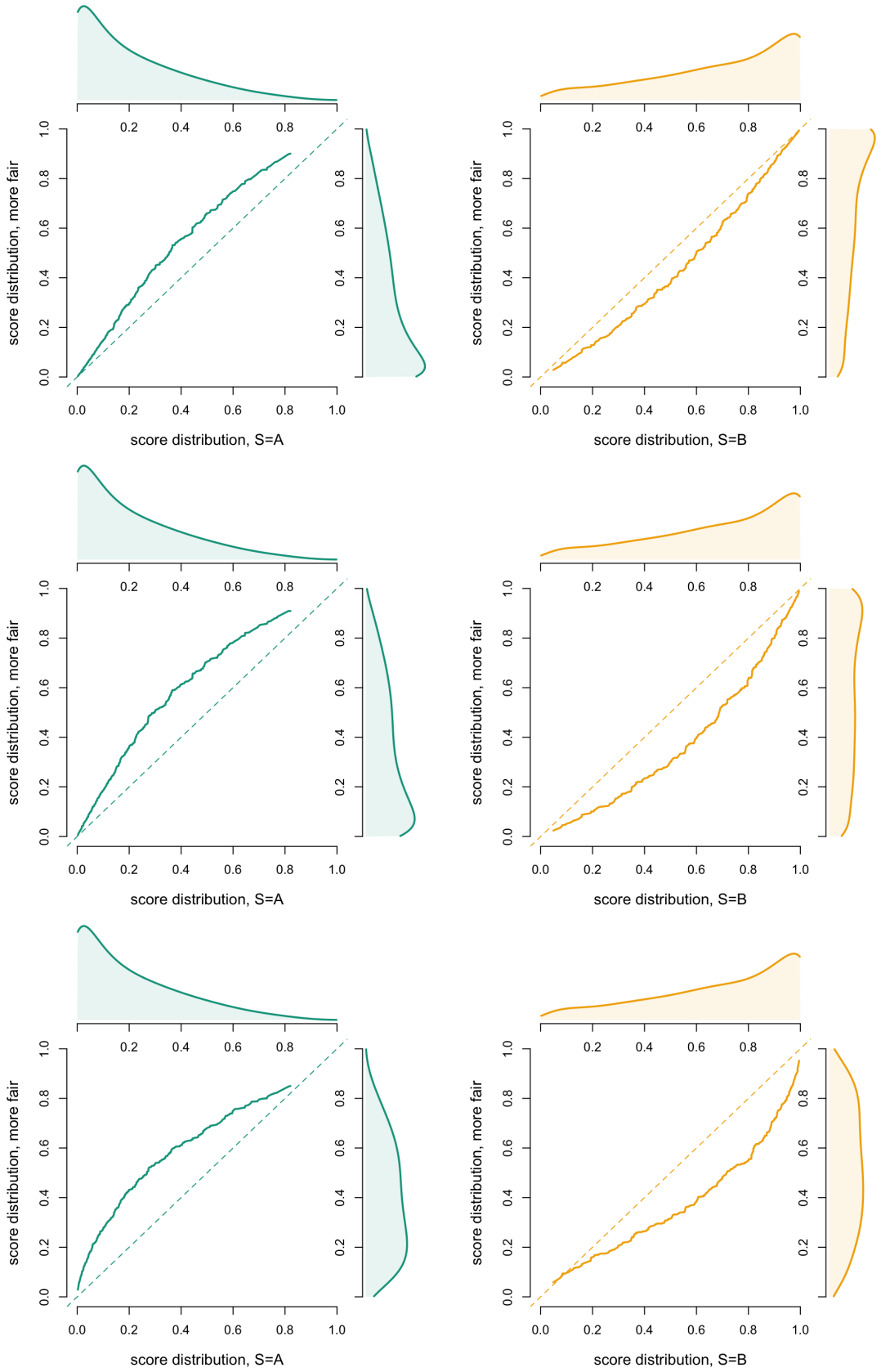


Figure 11.9: Optimal transport from distribution of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s to the distribution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s, for individuals in group A on the left, and in group B on the right, for different values of λ, with a low value on the top and a high value on the bottom (fair model, associated with a class balance (equalized odds) penalty criteria).

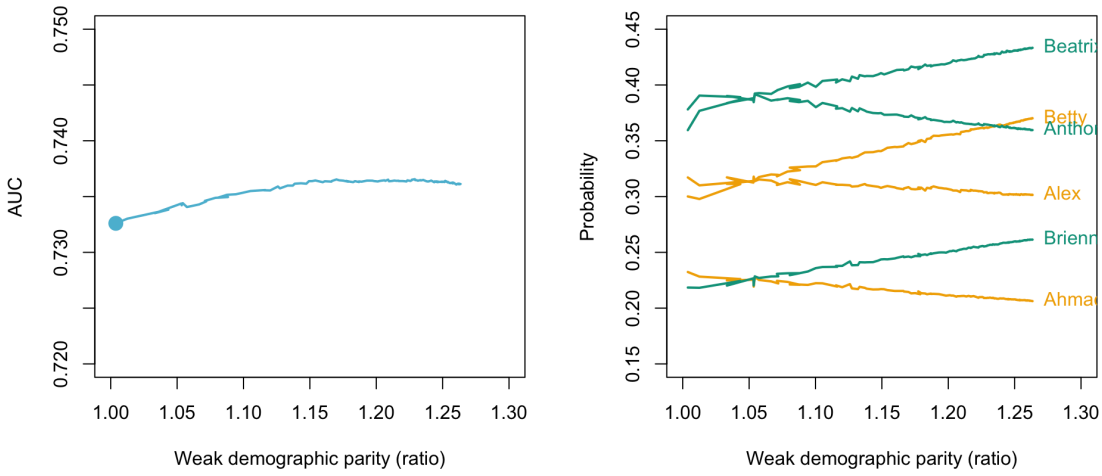


Figure 11.10: On the left, accuracy-fairness tradeoff plot (base on Table 11.1), with the AUC of $\widehat{m}_{\widehat{\beta}_\lambda}$ on the y-axis and the fairness ratio (demographic parity) on the x-axis. Top left corresponds to accurate and fair. On the right, evolution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$ (with a logistic regression) for three individuals in group A and three in B, on `toydata2` dataset.

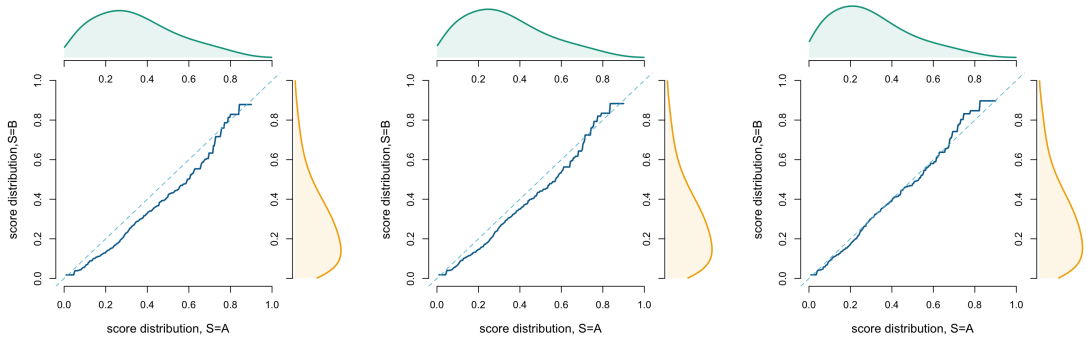


Figure 11.11: Optimal transport between distributions of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s from individuals in group A and in B, for different values of λ (low value on the left and high value on the right), associated with a demographic parity penalty criteria.

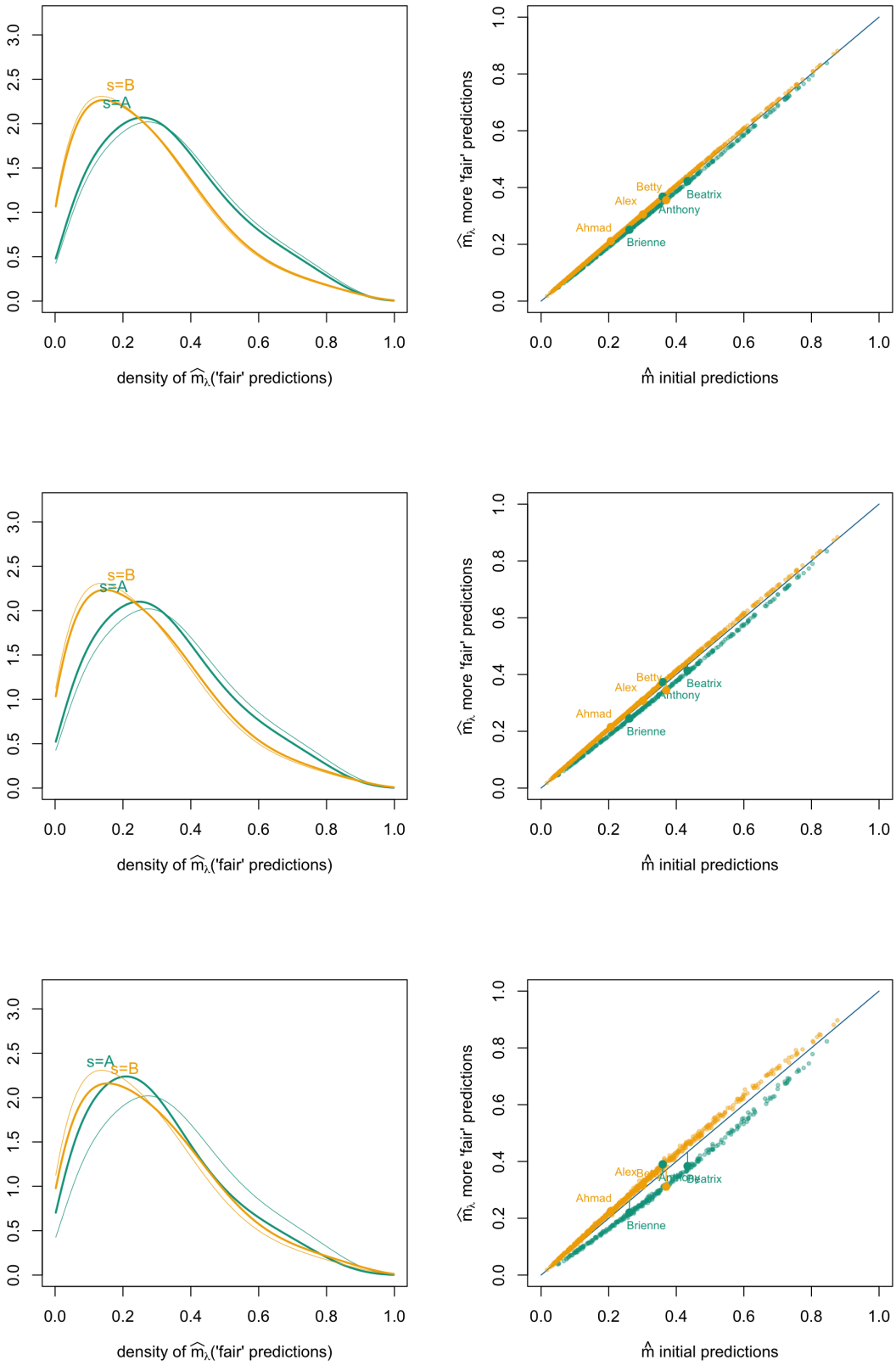


Figure 11.12: On the left, densities of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s from individuals in group A and in B (thin lines are densities of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s). On the right, scatterplot of points $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \mathtt{A}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i), s = \mathtt{A})$ and $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \mathtt{B}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i), s = \mathtt{B})$, where $\widehat{m}_{\widehat{\beta}}$ is the plain logistic regression, and $\widehat{m}_{\widehat{\beta}_\lambda}$ is the penalized logistic

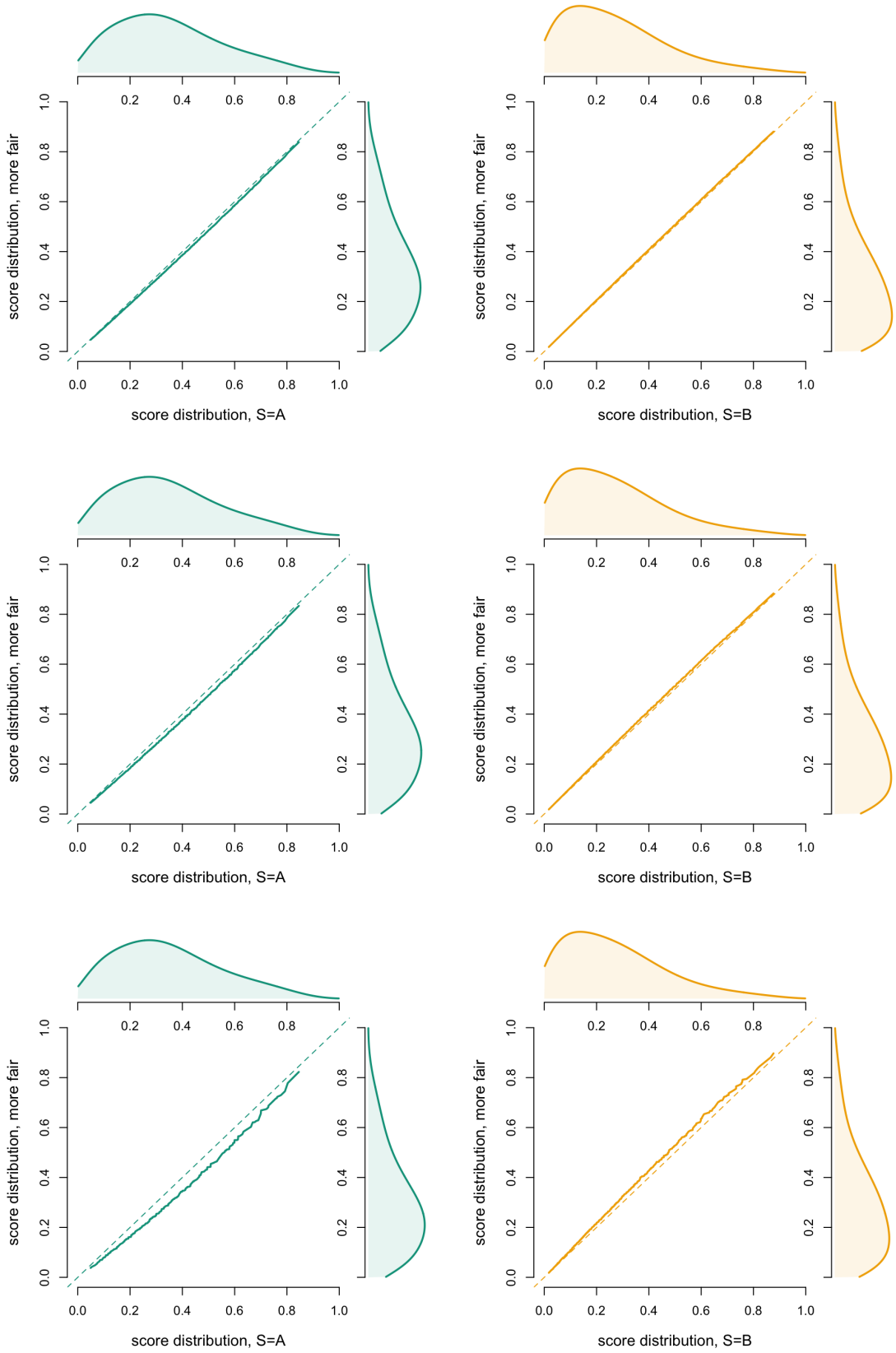


Figure 11.13: Optimal transport from distribution of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s to the distribution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s, for individuals in group A on the left, and in group B on the right, for different values of λ, with a low value on the top and a high value on the bottom (fair model, associated with a demographic parity penalty criteria).

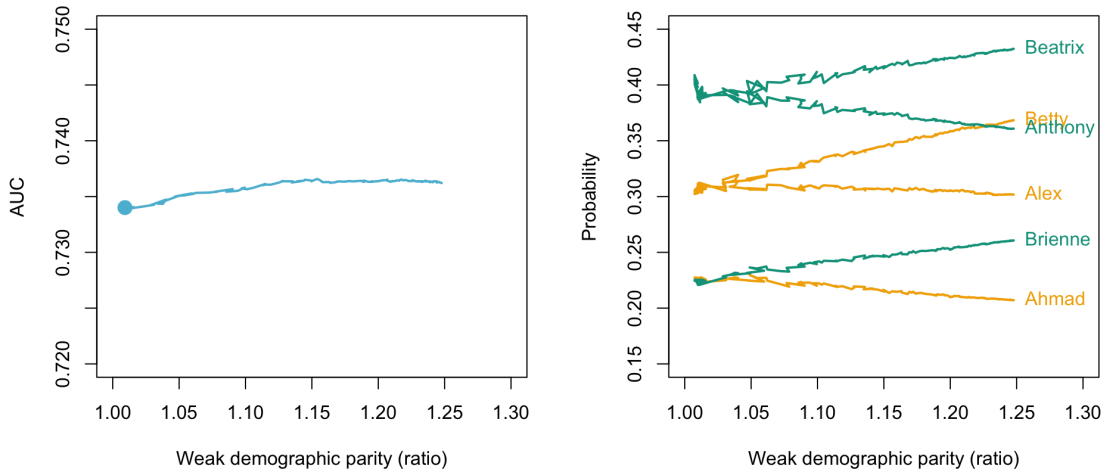


Figure 11.14: On the left, accuracy-fairness trade-off plot (base on Table 11.1), with the AUC of $\widehat{m}_{\widehat{\beta}_\lambda}$ on the y-axis and the fairness ratio (class balance) on the x-axis. Top left corresponds to accurate and fair. On the right, evolution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$ (with a logistic regression) for three individuals in group A and three in B, on `toydata2` dataset.

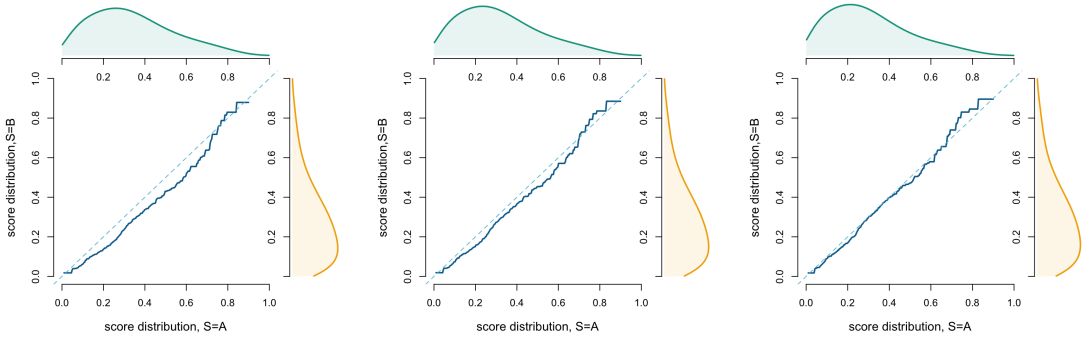


Figure 11.15: Optimal transport between distributions of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s from individuals in group A and in B, for different values of λ (low value on the left and high value on the right), associated with a class balance penalty criteria.

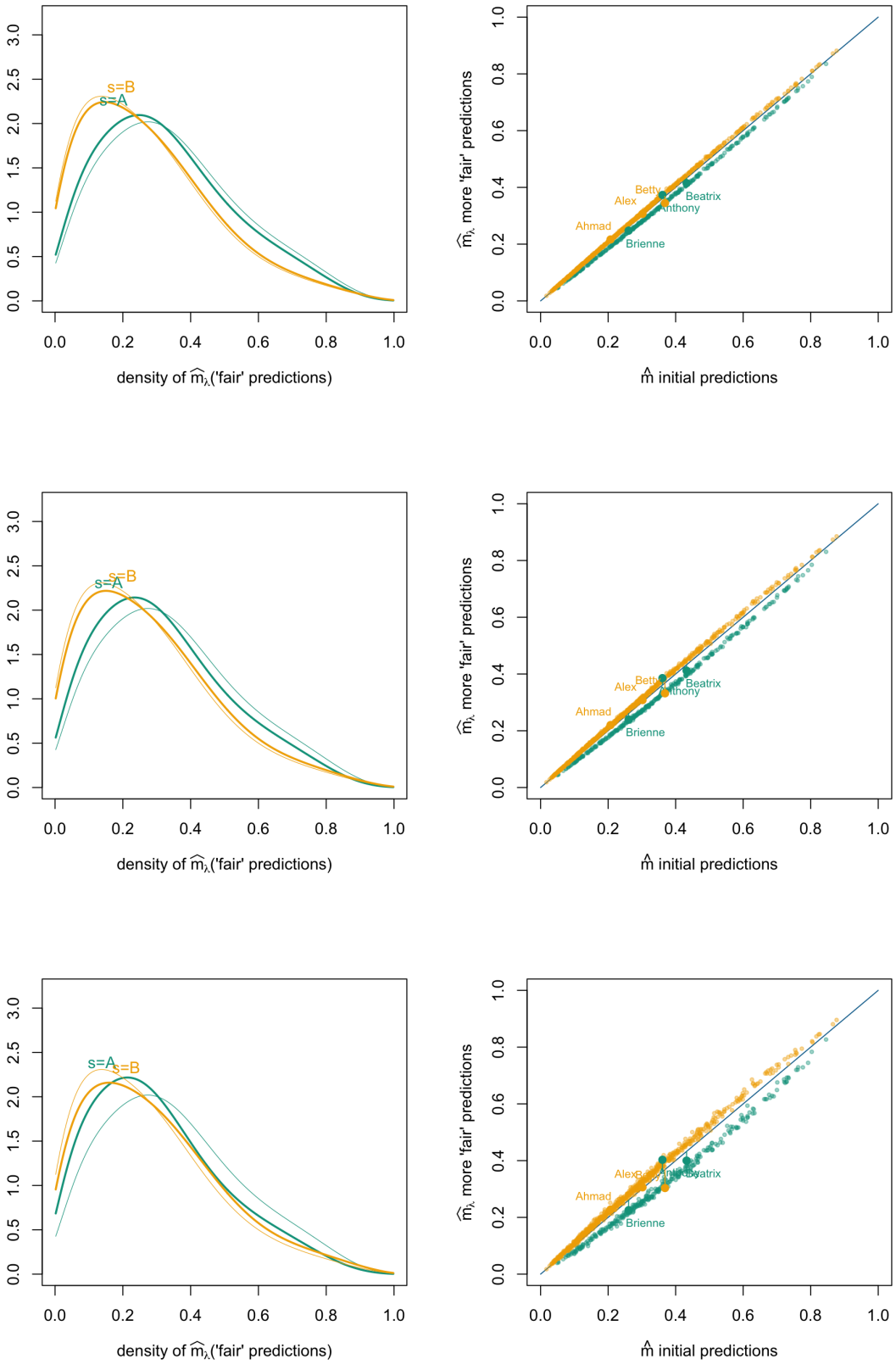


Figure 11.16: On the left, densities of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s from individuals in group A and in B (thin lines are densities of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s). On the right, scatterplot of points $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \texttt{A}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i), s = \texttt{A})$ and $(\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i = \texttt{B}), \widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i), s = \texttt{B})$, where $\widehat{m}_{\widehat{\beta}}$ is the plain logistic regression, and $\widehat{m}_{\widehat{\beta}_\lambda}$ is the penalized logistic

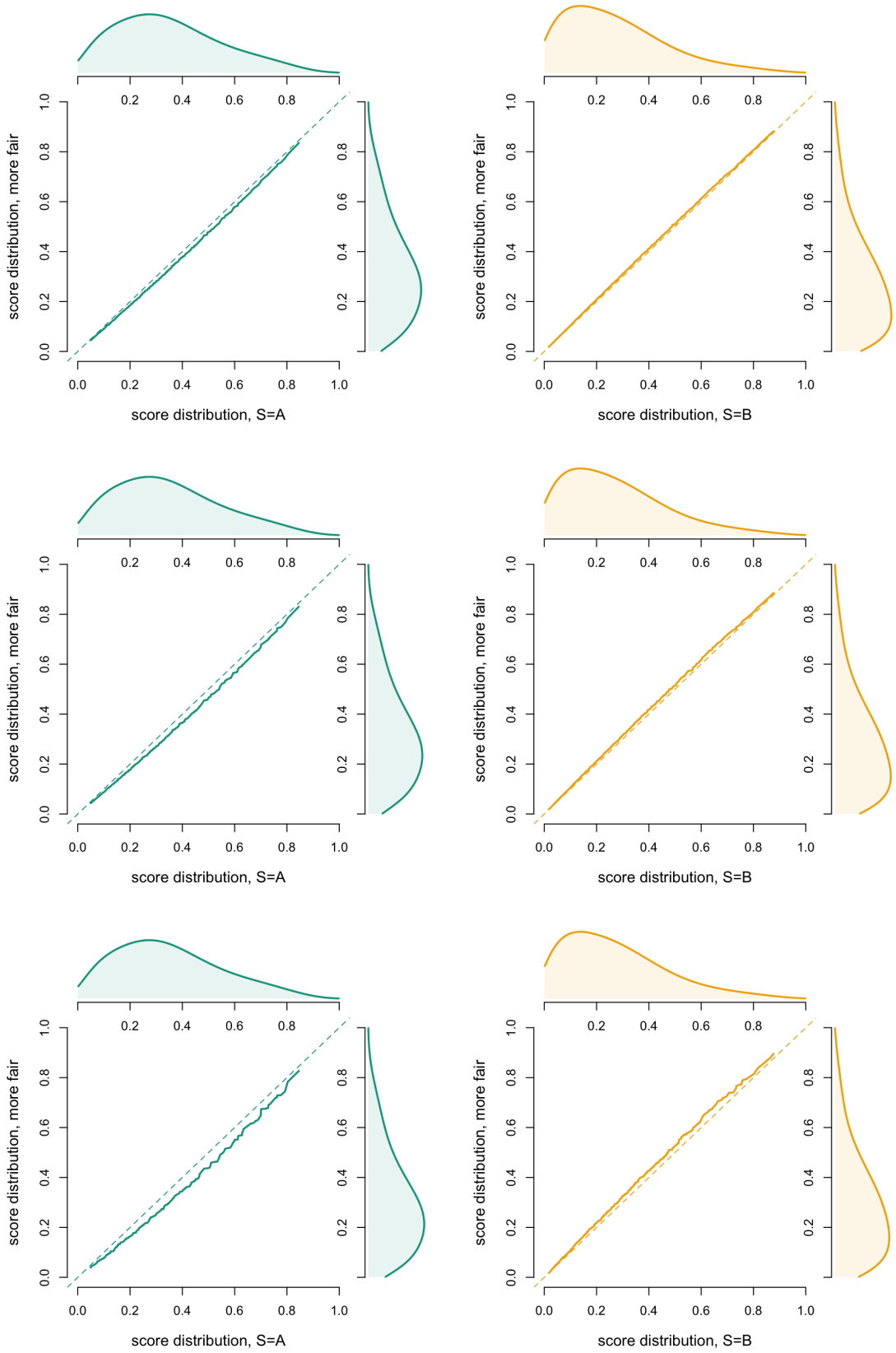


Figure 11.17: Optimal transport from distribution of $\widehat{m}_{\widehat{\beta}}(\boldsymbol{x}_i, s_i)$'s to the distribution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$'s, for individuals in group A on the left, and in group B on the right, for different values of λ, with a low value on the top and a high value on the bottom (fair model, associated with a class balance (equalized odds) penalty criteria).

Chapter 12

Post-processing

The idea of “post-processing" is relatively simple, since we change neither the training data, nor the model that has been estimated, we will simply transform the predictions obtained, to make them “fair" (according to some specific criteria). Since actuaries care about calibration, and the associated concept of “well-balanced" model, quite naturally, we will use averages and barycenters. Using optimal transport, we will describe techniques, with strong mathematical guarantess, that could be used to get a “fair" pricing model.

In some applications, training a model is a long and painful process, so we do not want to re-train it, and “in-processing" (discussed in Chapter 11) is not an option. As we will see in this chapter, an alternatative is to re-calibrate the outcomes independently from a model, as suggested in Hardt et al. (2016) or Pleiss et al. (2017).

12.1 Post-Processing for Binary Classifiers

If weak demographic parity is not satisfied, in the sense that $\mathbb{E}_{\boldsymbol{X}|S=\texttt{A}}[m(\boldsymbol{X})] \neq \mathbb{E}_{\boldsymbol{X}|S=\texttt{B}}[m(\boldsymbol{X})]$, a simple technique to get a fair model is to consider

$$m^\star(\boldsymbol{x}, s) = \frac{\mathbb{E}[m(\boldsymbol{X}, S)]}{\mathbb{E}[m(\boldsymbol{X}, s)]} \cdot m(\boldsymbol{x}, s) \text{ for a policyholder in group } s.$$

For example, in the `frenchmotor` dataset, overall, a single policyholder has 8.67% chance to claim a loss, 8.94% for a man (group `A`) and 8.20% for a woman (group `B`). Because of this difference, in order to get a fair model, “gender-neutral", the premium for a woman should be $8.67/8.20 = 1.058$ (or 5.8%) higher, $m^\star(\boldsymbol{x}, s) = 1.058 \cdot m(\boldsymbol{x}, s)$, and 3% lower than the predicted one, for men. Of course, this is perhaps simplistic, so let us consider the use of barycenters of distributions.

12.2 Weighted Averages of Outputs

Another natural approach, inspired from techniques used in sampling theory, is to use poststratification techniques, which is standard when dealing with a “biased sample". This is the idea discussed in Lindholm et al. (2022a). Let us extend here some concepts introduced in Chapter 7, and more specifically Section 7.4. Recall that the regression function (see Definition 3.3.1) is defined a

$$\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}\big[\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, S]\big] = \int_{\mathcal{S}} \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, S = s]\mathrm{d}\mathbb{P}[S = s].$$

Following Moodie and Stephens (2022), the later can be written

$$\mu(\boldsymbol{x}) = \int_{\mathcal{S}} \mathbb{E}[YW|\boldsymbol{X} = \boldsymbol{x}, S = s]\mathrm{d}\mathbb{P}[S = s|\boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[YW|\boldsymbol{X} = \boldsymbol{x}],$$

where W is a version of the Radon-Nikodym derivative

$$W = \frac{\mathrm{d}\mathbb{P}[S = s]}{\mathrm{d}\mathbb{P}[S = s|\boldsymbol{X} = \boldsymbol{x}]},$$

corresponding to the change of measure that will give independence between $\boldsymbol{X}$ and the sensitive attribute S. Following Côté (2023), we have the following interesting property,

Propoition 12.2.1 *Let W be a version of the Radon-Nikodym derivative*

$$W = \frac{\mathrm{d}\mathbb{P}[S = s]}{\mathrm{d}\mathbb{P}[S = s|\boldsymbol{X} = \boldsymbol{x}]},$$

then $\mathbb{E}[W] = 1$, $\mathbb{E}[SW] = \mathbb{E}[S]$, $\mathbb{E}[\boldsymbol{X}W] = \mathbb{E}[\boldsymbol{X}]$ *and* $\mathbb{E}[XSW] = \mathbb{E}[X]\mathbb{E}[S]$.

As proved in Fong et al. (2018),

$$\mathbb{E}[W] = \iint w\mathrm{d}\mathbb{P}[S = s, \boldsymbol{X} = \boldsymbol{x}] = \iint w\mathrm{d}\mathbb{P}[S = s|\boldsymbol{X} = \boldsymbol{x}]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}]$$

that can be written

$$\mathbb{E}[W] = \iint \frac{\mathrm{d}\mathbb{P}[S = s]}{\mathrm{d}\mathbb{P}[S = s|\boldsymbol{X} = \boldsymbol{x}]}\mathrm{d}\mathbb{P}[S = s|\boldsymbol{X} = \boldsymbol{x}]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}],$$

and therefore

$$\mathbb{E}[W] = \iint \mathrm{d}\mathbb{P}[S = s]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}] = 1.$$

Similarly

$$\mathbb{E}[SW] = \iint sw\mathrm{d}\mathbb{P}[S = s, \boldsymbol{X} = \boldsymbol{x}] = \iint sw\mathrm{d}\mathbb{P}[S = s|\boldsymbol{X} = \boldsymbol{x}]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}],$$

and

$$\mathbb{E}[SW] = \iint s\mathrm{d}\mathbb{P}[S = s]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}] = \int \mathbb{E}[S]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[S].$$

The proof of $\mathbb{E}[\boldsymbol{X}W] = \mathbb{E}[\boldsymbol{X}]$ is similar, and finally

$$\mathbb{E}[XSW] = \iint xswd\mathbb{P}[S = s|\boldsymbol{X} = \boldsymbol{x}]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}] = \iint xs\mathrm{d}\mathbb{P}[S = s]d\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}]$$

$$\mathbb{E}[XSW] = \int x\mathbb{E}[S]d\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[X]\mathbb{E}[S].$$

As a consequence, observe that $\text{Cov}[\boldsymbol{X}W, S] = \boldsymbol{0}$.In statistics, this Radon-Nikodym derivative is related to the propensity score, see Definition 7.4.7, as discussed in Freedman and Berk (2008), Li and Li (2019) and Karimi et al. (2022).

12.3 Average and Barycenters

Recall that "weighted averages" are solutions of

$$y^\star = \underset{y}{\text{argmin}} \left\{ \sum_{i=1}^n \omega_i (y - y_i)^2 \right\},$$

for some collection of observations $\{y_1, \cdots, y_n\}$ and some weights $\omega_1, \cdots, \omega_n \geq 0$. The extension in standard Euclidean spaces, named "barycenters", or "centroids" are defined as the solution of

$$z^\star = \underset{z}{\text{argmin}} \left\{ \sum_{i=1}^n \omega_i d(z, z_i)^2 \right\},$$

for some collection of points $\{z_1, \cdots, z_n\}$, some weights $\omega_1, \cdots, \omega_n \geq 0$, and where d is the Euclidean distance. This can be extended to more general spaces, where points are measures. We can therefore define some sort of average measure, solution of

$$\mathbb{P}^\star = \underset{\mathbb{P}}{\text{argmin}} \left\{ \sum_{i=1}^n \omega_i d(\mathbb{P}, \mathbb{P}_i)^2 \right\},$$

for some distance (or divergence) d, as in Nielsen and Boltz (2011). Those are also called "centroïds" associated with measures $\mathcal{P} = \{\mathbb{P}_1, \cdots, \mathbb{P}_n\}$, and weights $\boldsymbol{\omega}$. Instead of theoretical measures $\mathbb{P}_i$, the idea of "*averaging histograms*" (or empirical measures) as been considered in Nielsen and Nock (2009) using "*generalized Bregman centroïd*," and in Nielsen (2013) that introduced the "*generalized Kullback–Leibler centroïd*," based on Jeffreys'divergence, introduced in Jeffreys (1946), that corresponds to a symmetric divergence, extending Kullback–Leibler divergence (see Definition 3.3.7).

An alternative (see Agueh and Carlier (2011) and Definition 3.3.11) is to use the Wassertein distance W_2. As shown in Santambrogio (2015), if one of the measures $\mathbb{P}_i$ is absolutely continuous, the minimization problem has a unique solution. As discussed in Section 5.5.5 in Santambrogio (2015), it is possible to simple a simple version for univariate measures. Given a reference measure, say $\mathbb{P}_1$, it is possible to write the barycenter as the "*average push-forward*" transformation of $\mathbb{P}_1$: if $\mathbb{P}_i = \mathcal{T}_\#^{1\to i}\mathbb{P}_1$ (with the convention that $\mathcal{T}_\#^{1\to 1}$ is the identity),

$$\mathbb{P}^\star = \left(\sum_{i=1}^n \omega_i \mathcal{T}^{1\to i} \right)_\# \mathbb{P}_1.$$

And in the univariate case, $\mathcal{T}^{1\to i}$ is simply a rearrangement, defined as $\mathcal{T}^{1\to i} = F_i^{-1} \circ F_1$, where $F_i(t) = \mathbb{P}_i((-\infty, t])$ and F_i^{-1} is its generalized inverse. Note that Wasserstein Barycenter is also "*Fréchet mean of distributions*" in Petersen and Müller (2019), with the associate R package `frechet`. See Le Gouic and Loubes (2017) for technical results. As discussed in Alvarez-Esteban et al. (2018), moments and risk measures associated with $\mathbb{P}^\star$ can be expressed simply from associated measures on $\mathbb{P}_i$'s and $\boldsymbol{\omega}$.

To illustrate, consider as Mallasto and Feragen (2017) the case of Gaussian distributions, $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, that works in any dimension actually. Jeffrey-Kullback–Leibler-centroid of those distribution would be

$$\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \text{ where } \boldsymbol{\mu}^* = \sum_{i=1}^{n} \omega_i \boldsymbol{\mu}_i \text{ and } \boldsymbol{\Sigma}^* = \sum_{i=1}^{n} \omega_i \boldsymbol{\Sigma}_i.$$

But this is not the Wassertein barycenter, which is actually

$$\mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star), \text{ where } \boldsymbol{\mu}^\star = \sum_{i=1}^{n} \omega_i \boldsymbol{\mu}_i,$$

and where $\boldsymbol{\Sigma}^\star$ is the unique positive definite matrix such that

$$\boldsymbol{\Sigma}^\star = \sum_{i=1}^{n} \omega_i \left(\boldsymbol{\Sigma}^{\star 1/2} \boldsymbol{\Sigma}_i \boldsymbol{\Sigma}^{\star 1/2}\right)^{1/2}.$$

In the univariate case, with two Gaussian measures, the difference is that in the first case, the variance is the average of variances, while in the second case, the standard deviation is the average of standard deviations,

$$\begin{cases} \sigma^* = \sqrt{\omega_1 \sigma_1^2 + \omega_2 \sigma_2^2} : \text{ Jeffrey-Kullback–Leibler centroid} \\ \sigma^\star = \omega_1 \sigma_1 + \omega_2 \sigma_2 : \text{ Wassertein barycenter.} \end{cases}$$

On Figure 12.1 we can visualize the barycenters of two Gaussian distributions, on the left, and the empirical version with kernel density estimators on the right (based on two Gaussian samples of sizes $n = 1,000$). More specifically, if $\widehat{f_1}$ is the kernel density estimator of sample x_1, if $\widehat{F_1}$ is the integral of $\widehat{f_1}$ and if

$$\widehat{\mathcal{T}}^\star(x) = \omega_1 x + \omega_2 \widehat{F}_2^{-1}\left(\widehat{F}_1(x)\right),$$

the density of the barycenter is

$$\widehat{f}^\star(x) = \widehat{f}_1(\widehat{T}^{\star -1}(x)) \cdot |d\widehat{T}^{\star -1}(x)|.$$

Fairness will be achieved by considering barycenters of regression predictions. For unbounded predictions, classical kernel density estimators can be considered (using `density` in R), but in classification, scores take values in $[0, 1]$. One could use transformed kernel techniques (as discussed in Geenens (2014)), or Beta kernels (as in Chen (1999)) using `kdensity` from the `kdensity` R package (and the option `kernel="beta"`).

On Figure 12.2 we can visualize the barycenters of two samples generated from some Beta distributions. On the left, histograms are used, and Jeffrey centroïd is plotted. In the middle and on the right, smooth density estimators of f_1 and f_2 are considered, using Beta kernels. In the middle, we consider Jeffrey centroïd and on the right, Wassertein barycenter. The transport, from f_1 or f_2 to $f^\star$ (all three on the right of Figure 12.2) can be visualized on Figure 12.3, on the left and on the right respectively

Given two scoring function $m(\boldsymbol{x}, s = \texttt{A})$ and $m(\boldsymbol{x}, s = \texttt{B})$, it is possible, post-processing, to construct a fair score $m^\star$ using the approach we just described,

Definition 12.3.1 (Fair barycenter score) *Given two scores $m(\boldsymbol{x}, s = \texttt{A})$ and $m(\boldsymbol{x}, s = \texttt{B})$, the "fair barycenter score" is*

$$\begin{cases} m^\star(\boldsymbol{x}, s = \texttt{A}) = \mathbb{P}[S = \texttt{A}] \cdot m(\boldsymbol{x}, s = \texttt{A}) + \mathbb{P}[S = \texttt{B}] \cdot F_{\texttt{B}}^{-1} \circ F_{\texttt{A}}\big(m(\boldsymbol{x}, s = \texttt{A})\big) \\ m^\star(\boldsymbol{x}, s = \texttt{B}) = \mathbb{P}[S = \texttt{A}] \cdot F_{\texttt{A}}^{-1} \circ F_{\texttt{B}}\big(m(\boldsymbol{x}, s = \texttt{B})\big) + \mathbb{P}[S = \texttt{B}] \cdot m(\boldsymbol{x}, s = \texttt{B}) \end{cases}$$

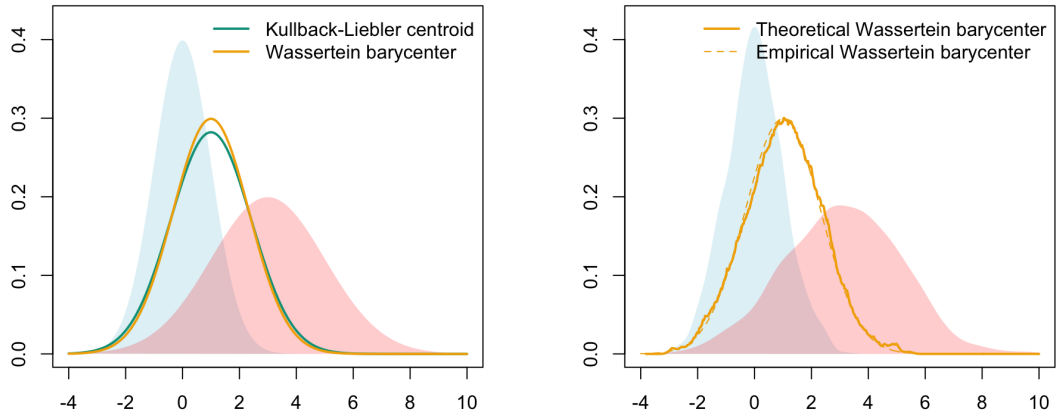


Figure 12.1: Wasserstein barycenter and Jeffrey-Kullback-Leibler centroid of two Gaussian distributions on the left, and empirical estimate of the density of Wasserstein barycenter and Jeffrey-Kullback-Leibler centroid of two samples x_1 and x_2 (drawn from normal distributions).

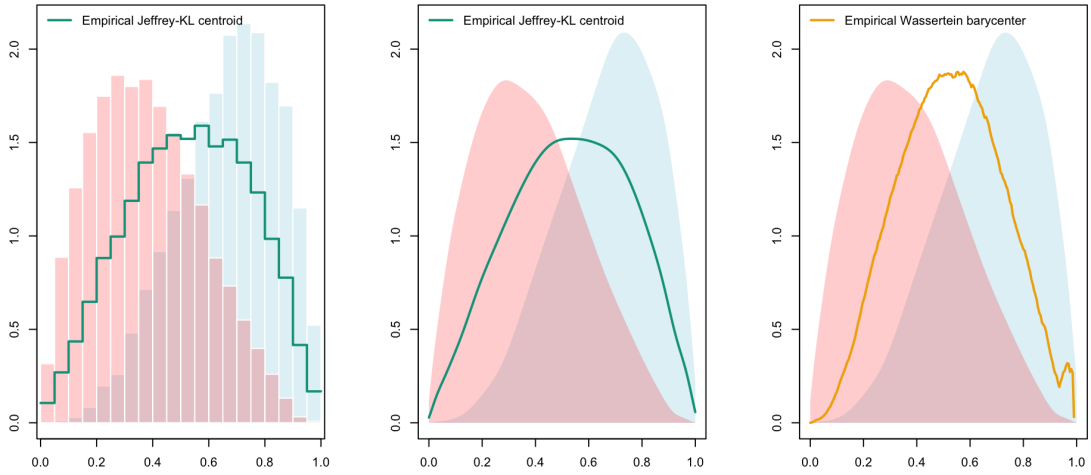


Figure 12.2: Empirical Jeffrey-Kullback-Leibler centroid of two samples generated from Beta distributions on the left (based on histograms) and in the middle (based on Beta kernel estimated of the density), Wasserstein barycenter on the right (based on Beta kernel estimated of the density).

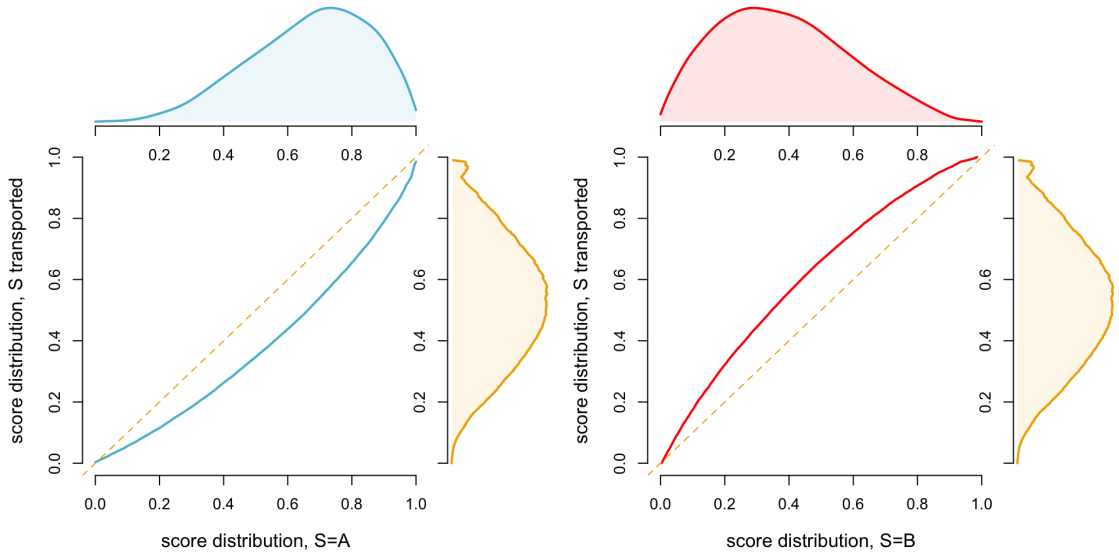


Figure 12.3: Optimal transport for two samples drawn from two Beta distributions (one skewed to the left (on the left) and one to the right (on the right), on the x-axis) to the barycenter (on the y-axis).

In Definition 2.5.1, we defined “balanced" scores,

Propoition 12.3.1 *The score $m^\star$ is balanced.*

On Figure 12.4, inspired from Figure 8.8, we can visualize the matching between $m(\boldsymbol{x}, s = \texttt{A})$ and $m^\star(\boldsymbol{x}, s = \texttt{A})$ on top, and between $m(\boldsymbol{x}, s = \texttt{B})$ and $m^\star(\boldsymbol{x}, s = \texttt{B})$ below.

On Figure 12.5, we have scatterplots of points $(m(\boldsymbol{x}_i, s_i = \texttt{A}), m^\star(\boldsymbol{x}_i, s = \texttt{A})$ and $(m(\boldsymbol{x}_i, s_i = \texttt{B}), m^\star(\boldsymbol{x}_i, s = \texttt{B})$, with three models (GLM, GBM, RF), on the probability to claim a loss in motor insurance when s is the gender of the driver, from the left to the right.

12.4 Application on `toydata1`

On Figure 12.6, computations of the distribution of the “fair score" defined as the barycenter of the distributions of scores $(m(\boldsymbol{x}_i, s_i = \texttt{A})$ and $(m(\boldsymbol{x}_i, s_i = \texttt{B})$, in the two groups. On the left, Wasserstein barycenter, and then two computations of effrey-Kullback-Leibler centroid.

On Figure 12.7 optimal transport plot, with the optimal matching between $m(\boldsymbol{x})$ and $m^\star(\boldsymbol{x})$ for individuals in group $s = \texttt{A}$, on the left, and between $m(\boldsymbol{x})$ and $m^\star(\boldsymbol{x})$ for individuals in group $s = \texttt{B}$ on the right, with unaware scores on the `toydata1` dataset.

On Figure 12.8, optimal transport plot, with matching between $m(\boldsymbol{x}, s = \texttt{A})$ and $m^\star(\boldsymbol{x})$ for individuals in group $s = \texttt{A}$, on the left, and between $m(\boldsymbol{x}, s = \texttt{B})$ and $m^\star(\boldsymbol{x})$ for individuals in group $s = \texttt{B}$ on the right, with aware scores on the `toydata1` dataset (**plain** lines, thin lines are unaware score from Figure 12.7).

On Table 12.1, we can visualize the predictions for six individuals, with an unaware model $\widehat{m}(x, s)$, an aware model $\widehat{m}(x)$, and two barycenters, Wasserstein barycenter $\widehat{m}^*_{\text{w}}(x)$ and Jeffrey-Kullback-Leibler centroid, $\widehat{m}^*_{\text{jkl}}(x)$.

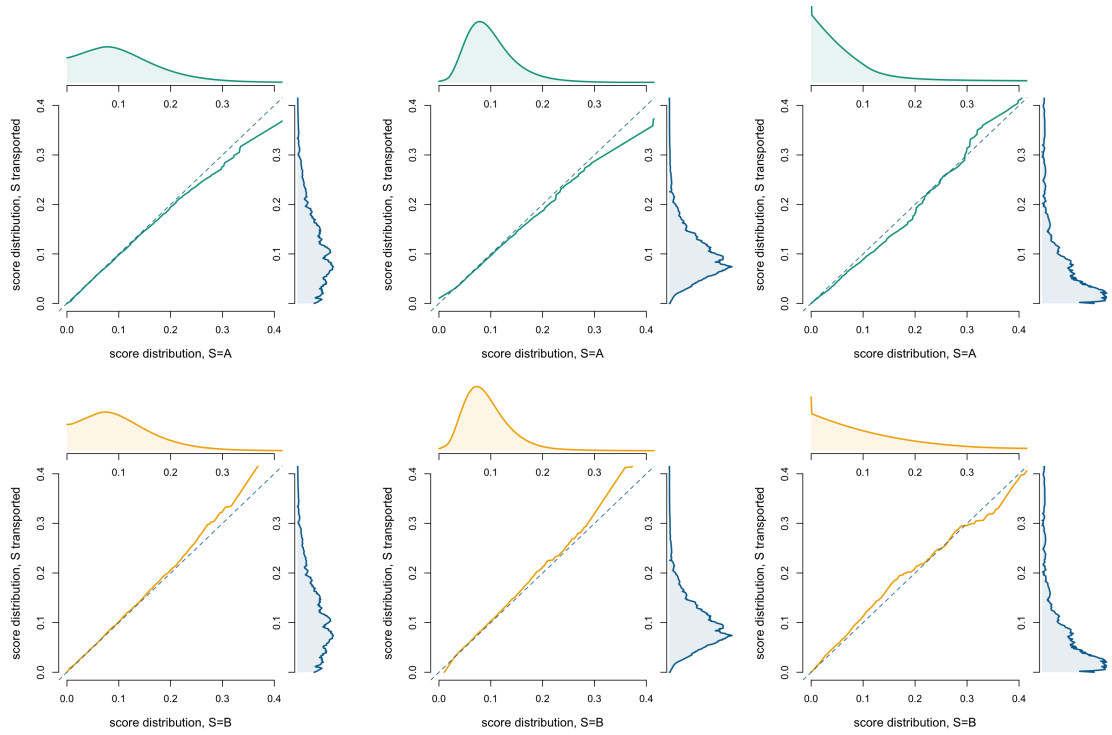


Figure 12.4: Matching between $m(\boldsymbol{x}, s = \texttt{A})$ and $m^\star(\boldsymbol{x})$, on top, and between $m(\boldsymbol{x}, s = \texttt{B})$ and $m^\star(\boldsymbol{x})$, below, on the probability to claim a loss in motor insurance when s is the gender of the driver, from `frenchmotor`.

	x	s	$\overline{y}$	$\widehat{m}(x,s)$	$\widehat{m}(x)$	$\widehat{m}^*_{\text{w}}(x)$	$\widehat{m}^*_{\text{jkl}}(x)$
Alex	-1	A	0.475	0.250	0.219	0.154	0.094
Betty	-1	B	0.475	0.205	0.219	0.459	0.357
Ahmad	0	A	0.475	0.490	0.465	0.341	0.279
Brienne	0	B	0.475	0.426	0.465	0.719	0.692
Anthony	+1	A	0.475	0.734	0.730	0.571	0.521
Beatrix	+1	B	0.475	0.681	0.730	0.842	0.932

Table 12.1: Individual predictions for six fictitious individuals. With two models (aware and unaware) and two barycenters, on `toydata1`.

On Figure 12.9, we have distributions of the scores in the two groups, A and B, after optimal transport to the barycenyter, with Jeffrey-Kullback-Leibler centroid on top and Wassertein barycenter below, with Given two scores $m(\boldsymbol{x}, s = \texttt{A})$ and $m(\boldsymbol{x}, s = \texttt{B})$, the "fair barycenter score" being the one of Definition 12.4,

$$\begin{cases} m^\star(\boldsymbol{x}, s = \texttt{A}) = \mathbb{P}[S = \texttt{A}] \cdot m(\boldsymbol{x}, s = \texttt{A}) + \mathbb{P}[S = \texttt{B}] \cdot F_{\texttt{B}}^{-1} \circ F_{\texttt{A}}\big(m(\boldsymbol{x}, s = \texttt{A})\big) \\ m^\star(\boldsymbol{x}, s = \texttt{B}) = \mathbb{P}[S = \texttt{A}] \cdot F_{\texttt{A}}^{-1} \circ F_{\texttt{B}}\big(m(\boldsymbol{x}, s = \texttt{B})\big) + \mathbb{P}[S = \texttt{B}] \cdot m(\boldsymbol{x}, s = \texttt{B}). \end{cases}$$

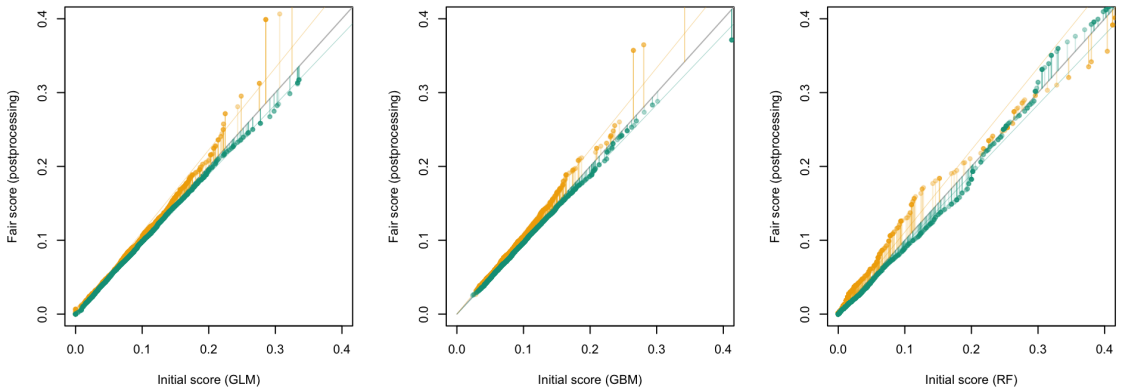


Figure 12.5: Scatterplot of points $(m(\boldsymbol{x}_i, s_i = \texttt{A}), m^\star(\boldsymbol{x}_i, s = \texttt{A})$ and $(m(\boldsymbol{x}_i, s_i = \texttt{B}), m^\star(\boldsymbol{x}_i, s = \texttt{B})$, with three models (GLM, GBM, RF), on the probability to claim a loss in motor insurance when s is the gender of the driver.

Figure 12.6: Barycenter of the two distributions of scores, of the distributions of scores $(m(\boldsymbol{x}_i, s_i = \texttt{A})$ and $(m(\boldsymbol{x}_i, s_i = \texttt{B})$. On the left, Wasserstein barycenter, and then two computations of effrey-Kullback-Leibler centroid.

Figure 12.7: Matching between $m(\boldsymbol{x})$ and $m^\star(\boldsymbol{x})$ for individuals in group $s = \texttt{A}$, on the left, and between $m(\boldsymbol{x})$ and $m^\star(\boldsymbol{x})$ for individuals in group $s = \texttt{B}$ on the right, with unaware scores on the `toydata1` dataset.

12.5 Application on `frenchmotor`

In the entire dataset, we have 64% men (7973) and 36% women (4464) registered as “main driver". Overall, if we consider “weak demographic parity", 8.2% women claim a loss, against 8.9% women. In Table 12.2, we can visualize “gender-neutral" predictions, derived from the logistic regression (GLM), a boosting algorithm (GBM) and a random forest (RF). The first column corresponds to the proportional approach discussed in Section 12.1.

In Table 12.2, we have, for the two groups, the global correction discussed in Section 12.2, with −6% for the men (×0.94, group

On Figures 12.4 and 12.5, we have seen how to get a “fair prediction", with the matching between $m(\boldsymbol{x}, s = \texttt{A})$ and $m^\star(\boldsymbol{x}, s = \texttt{A})$, on top, and between $m(\boldsymbol{x}, s = \texttt{B})$ and $m^\star(\boldsymbol{x}, s = \texttt{B})$, on Figure 12.4, and with scatterplot of points $(m(\boldsymbol{x}_i, s_i = \texttt{A}), m^\star(\boldsymbol{x}_i, s = \texttt{A}))$ and $(m(\boldsymbol{x}_i, s_i = \texttt{B}), m^\star(\boldsymbol{x}_i, s = \texttt{B}))$ on Figure 12.5.

We can also consider a binary sensitive attribute, related to the age, with $s = \mathbf{1}(\texttt{age} > 65)$ (discrimination against old people), in Table 12.3 and $s = \mathbf{1}(\texttt{age} < 30)$ (discrimination against young people), in Table 12.4.

Tables 12.4 and 12.3 extend Table , from the case where the sensitive attribute was the gender to the case where the sensitive attribute is the age, with young / non-young in Table 12.4, old / non-old in Table 12.3.

On Figures 12.10 and 12.12 we can visualize matchings between $m(\boldsymbol{x}, s = \texttt{A})$ and $m^\star(\boldsymbol{x}, s = \texttt{A})$, on top, and between $m(\boldsymbol{x}, s = \texttt{B})$ and $m^\star(\boldsymbol{x}, s = \texttt{B})$ below, respectively with $s = \mathbf{1}(\texttt{age} > 65)$ (discrimination against

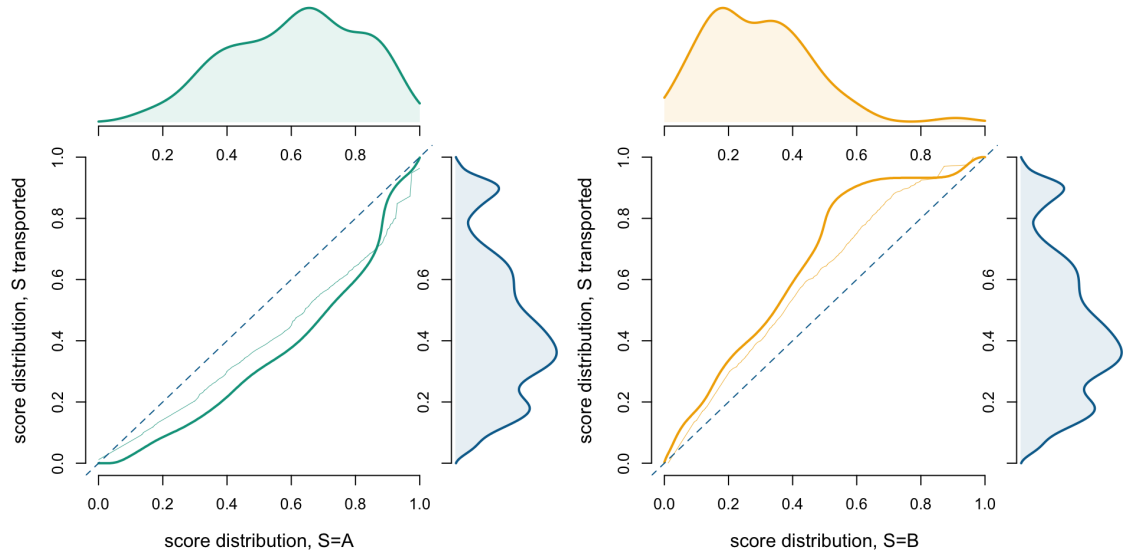


Figure 12.8: Matching between $m(\boldsymbol{x}, s = \texttt{A})$ and $m^\star(\boldsymbol{x})$ for individuals in group $s = \texttt{A}$, on the left, and between $m(\boldsymbol{x}, s = \texttt{B})$ and $m^\star(\boldsymbol{x})$ for individuals in group $s = \texttt{B}$ on the right, with aware scores on the `toydata1` dataset (**plain** lines, thin lines are unaware score from Figure 12.7).

Figure 12.9: Distributions of the scores in the two groups, A and B, after optimal transport to the barycenyter, with Jeffrey-Kullback-Leibler centroid on top and Wassertein barycenter below.

	A (men)				B (women)			
	×0.94	GLM	GBM	RF	×1.11	GLM	GBM	RF
$m(\boldsymbol{x}) = 5\%$	4.73%	4.94%	4.80%	4.42%	5.56%	5.16%	5.25%	6.15%
$m(\boldsymbol{x}) = 10\%$	9.46%	9.83%	9.66%	8.92%	11.12%	10.38%	10.49%	12.80%
$m(\boldsymbol{x}) = 20\%$	18.91%	19.50%	18.68%	18.26%	22.25%	20.77%	21.63%	21.12%

Table 12.2: “Gender-free" prediction if the initial prediction was 5% (on top), 10% (in the middle) and 20% (below). The first approach is the simple “benchmark" based on $\mathbb{P}[Y = 1]/\mathbb{P}[Y = 1|S = s]$, and then three models are considered, GLM, GBM and RF.

	A (younger < 65)				B (old > 65)			
	×1.01	GLM	GBM	RF	×0.94	GLM	GBM	RF
$m(\boldsymbol{x}) = 5\%$	5.05%	5.17%	5.10%	5.27%	4.71%	3.84%	3.84%	3.96%
$m(\boldsymbol{x}) = 10\%$	10.09%	10.37%	10.16%	11.00%	9.42%	7.81%	9.10%	6.88%
$m(\boldsymbol{x}) = 20\%$	20.19%	19.98%	19.65%	21.26%	18.85%	19.78%	23.79%	12.54%

Table 12.3: “Age-free" prediction (against old driver) if the initial prediction was 5% (on top), 10% (in the middle) and 20% (below).

	A (young < 25)				B (older > 25)			
	×0.74	GLM	GBM	RF	×1.06	GLM	GBM	RF
$m(\boldsymbol{x}) = 5\%$	3.71%	3.61%	4.45%	2.41%	5.29%	5.29%	5.14%	6.05%
$m(\boldsymbol{x}) = 10\%$	7.42%	7.89%	8.69%	5.17%	10.59%	10.29%	10.19%	11.95%
$m(\boldsymbol{x}) = 20\%$	14.84%	21.82%	18.09%	9.93%	21.17%	19.87%	20.33%	21.29%

Table 12.4: “Age-free" (against young drivers) prediction if the initial prediction was 5% (on top), 10% (in the middle) and 20% (below).

old people) and $s = \mathbf{1}(\texttt{age} < 30)$ (discrimination against young people).

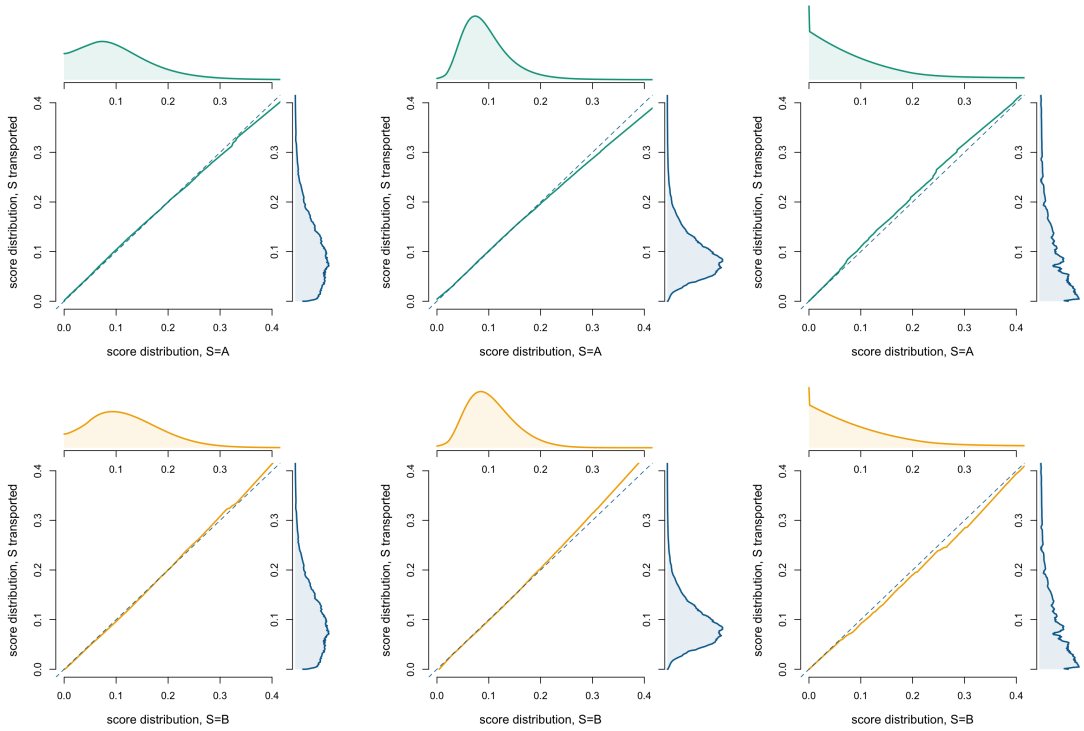


Figure 12.10: Matching between $m(\boldsymbol{x}, s = \texttt{A})$ and $m^\star(\boldsymbol{x}, s = \texttt{A})$, on top, and between $m(\boldsymbol{x}, s = \texttt{B})$ and $m^\star(\boldsymbol{x}, s = \texttt{B})$, below, on the probability to claim a loss in motor insurance when s is the indicator that the driver is “old" $\mathbf{1}(\texttt{age} > 65)$.

On Figure 12.11, we have a scatterplot of points $(m(\boldsymbol{x}_i, s_i = \texttt{A}), m(^\star\boldsymbol{x}_i))$ and $(m(\boldsymbol{x}_i, s_i = \texttt{B}), m(^\star\boldsymbol{x}_i))$, with three models (GLM, GBM, RF), on the probability to claim a loss in motor insurance when s is the indicator that the driver is “old" $\mathbf{1}(\texttt{age} > 65)$ (more than 65 years old).

On Figure 12.12, we can visualize the optimal transport plot, with t matching between $m(\boldsymbol{x}, s = \texttt{A})$ and $m^\star(\boldsymbol{x}, s = \texttt{A})$, on top, and between $m(\boldsymbol{x}, s = \texttt{B})$ and $m^\star(\boldsymbol{x}, s = \texttt{B})$, below, on the probability to claim a loss in motor insurance when s is the indicator that the driver is “young" $\mathbf{1}(\texttt{age} < 30)$ (less than 30 years old).

On Figure 12.13, we have the scatterplot of points $(m(\boldsymbol{x}_i, s_i = \texttt{A}), m(^\star\boldsymbol{x}_i))$ and $(m(\boldsymbol{x}_i, s_i = \texttt{B}), m(^\star\boldsymbol{x}_i))$,

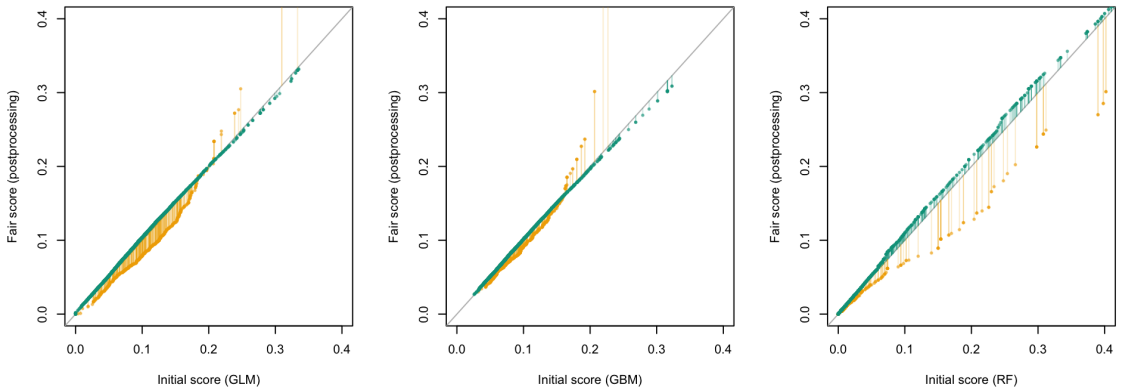


Figure 12.11: Scatterplot of points $(m(\boldsymbol{x}_i, s_i = \texttt{A}), m(^\star\boldsymbol{x}_i))$ and $(m(\boldsymbol{x}_i, s_i = \texttt{B}), m(^\star\boldsymbol{x}_i))$, with three models (GLM, GBM, RF), on the probability to claim a loss in motor insurance when s is the indicator that the driver is "old" $\mathbf{1}(\texttt{age} > 65)$.

with three models (GLM, GBM, RF), on the probability to claim a loss in motor insurance when s is the indicator that the driver is "young" $\mathbf{1}(\texttt{age} < 30)$.

12.6 Penalized Bagging

If m is a random forest, instead of using equal weights in the bagging procedure, one can consider to seek for weights to that the outcome will be "fair", as suggested in Fermanian and Guegan (2021). Formally, consider k models, $m_1, \cdots, m_k$. In a classification problem, $m_j(\boldsymbol{x}) \approx \mathbb{P}[Y = 1|\boldsymbol{X} = \boldsymbol{x}]$ and in the case of a regression $m_j(\boldsymbol{x}) \approx \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$.

Let $M_{\boldsymbol{\omega}}(\boldsymbol{x}) = \sum_{j=1}^{k} \omega_j m_j(\boldsymbol{x}) = \boldsymbol{\omega}^\top \boldsymbol{m}(\boldsymbol{x})$. For example, with random forests, k is large, and $\omega_j = 1/k$. But we can consider an ensemble approach, on few models.

Recall that "demographic parity" (Section 8.2) is achieved if

$$\mathbb{E}\big[\widehat{Y}\big|S = \texttt{A}\big] = \mathbb{E}\big[\widehat{Y}\big|S = \texttt{B}\big] = \mathbb{E}\big[\widehat{Y}\big],$$

meaning here

$$\mathbb{E}\big[\boldsymbol{\omega}^\top \boldsymbol{m}(\boldsymbol{X})\big|S = \texttt{A}\big] = \mathbb{E}\big[\boldsymbol{\omega}^\top \boldsymbol{m}(\boldsymbol{X})\big|S = \texttt{B}\big].$$

Empirically, we can compute (for some loss ℓ)

$$\ell\left(\frac{\sum_{i=1}^n \mathbf{1}\left(s_i = \texttt{A}\right)\boldsymbol{\omega}^\top \boldsymbol{m}(\boldsymbol{x}_i)}{\sum_{i=1}^n \mathbf{1}\left(s_i = \texttt{A}\right)}, \frac{\sum_{i=1}^n \mathbf{1}\left(s_i = \texttt{B}\right)\boldsymbol{\omega}^\top \boldsymbol{m}(\boldsymbol{x}_i)}{\sum_{i=1}^n \mathbf{1}\left(s_i = \texttt{B}\right)}\right).$$

The problem we should solve could be

$$\underset{\boldsymbol{\omega}}{\text{argmin}}\big\{R(\boldsymbol{\omega}) + \lambda\mathcal{R}(\boldsymbol{\omega})\big\},$$

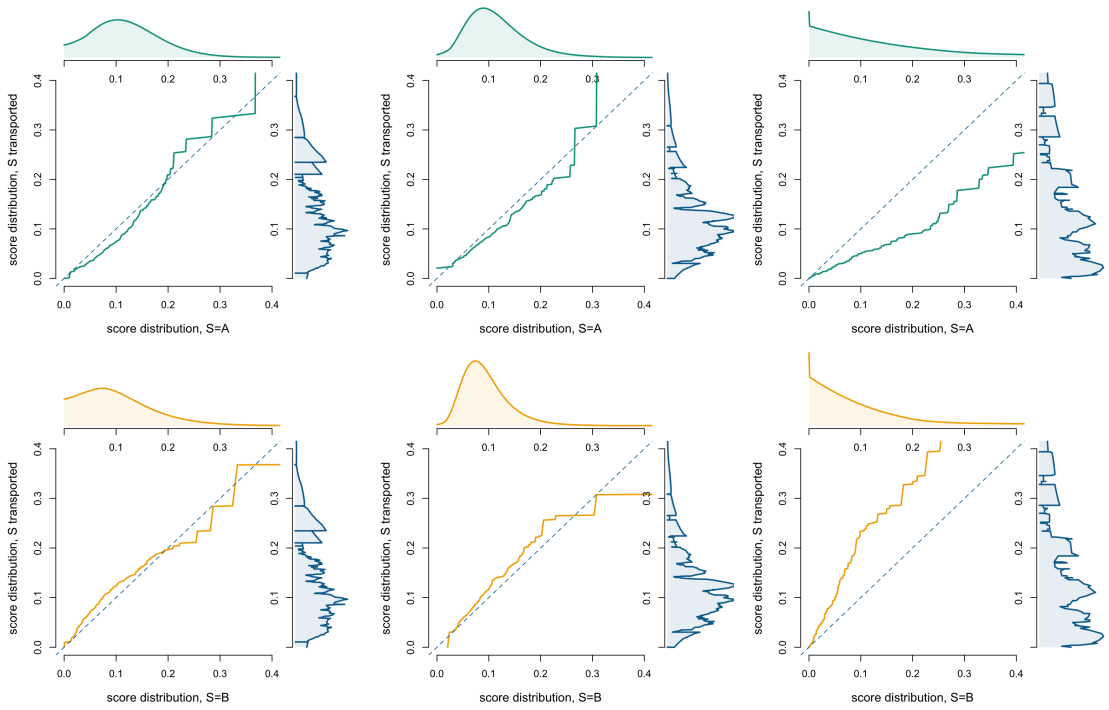


Figure 12.12: Matching between $m(\boldsymbol{x}, s = \texttt{A})$ and $m^\star(\boldsymbol{x}, s = \texttt{A})$, on top, and between $m(\boldsymbol{x}, s = \texttt{B})$ and $m^\star(\boldsymbol{x}, s = \texttt{B})$, below, on the probability to claim a loss in motor insurance when s is the indicator that the driver is "young" $\mathbf{1}(\texttt{age} < 30)$.

where the empirical risk (associated with accuracy) could be, if the risk is associated to loss ℓ,

$$\mathcal{R}(\boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\boldsymbol{\omega}^\top \boldsymbol{m}(\boldsymbol{x}_i), y_i\right),$$

and where $R(\boldsymbol{\omega})$ is some fairness criteria as discussed in Section 11.1. Following Friedler et al. (2019), an alternative could be to consider α-disparate impact

$$R_\alpha = \frac{\mathbb{E}[|\widehat{Y}|^\alpha | S = \texttt{A}]}{\mathbb{E}[|\widehat{Y}|^\alpha | S = \texttt{B}]} \text{ for } \alpha > 0,$$

and β-equalized odds

$$R_\beta = 1 - \mathbb{E}_Y\left[\left|\mathbb{E}[\widehat{Y}|S = \texttt{A}, Y] - \mathbb{E}[\widehat{Y}|S = \texttt{B}, Y]\right|^\beta\right].$$

As in Section 11.1, we want to find some weights that will give a trade-off between fairness and accuracy. The only difference in that in Chapter 11, "in-processing", we were still training the model. Here, we already have a collection of models, we just want to consider a weighted average of the model. Given a sample $\{(y_i, \boldsymbol{x}_i)\}$, consider the following penalized problem, where the fairness criteria is related to α-"demographic

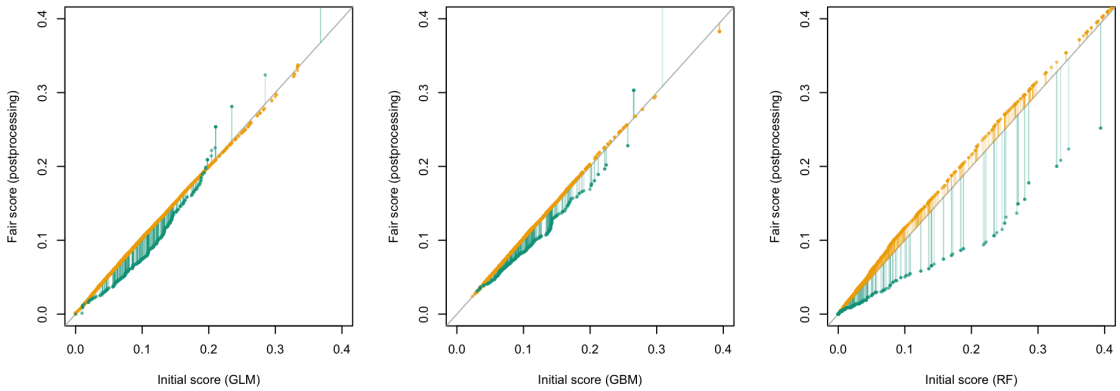


Figure 12.13: Scatterplot of points $(m(\boldsymbol{x}_i, s_i = \texttt{A}), m(^\star\boldsymbol{x}_i))$ and $(m(\boldsymbol{x}_i, s_i = \texttt{B}), m(^\star\boldsymbol{x}_i))$, with three models (GLM, GBM, RF), on the probability to claim a loss in motor insurance when s is the indicator that the driver is "young" $\mathbf{1}(\texttt{age} < 30)$.

parity", and accuracy is characterized by some loss ℓ,

$$\min_{\omega\in\mathcal{S}_k}\left\{\left|\frac{1}{n_0}\sum_{i:S_i=0}\boldsymbol{\omega}^\top\boldsymbol{m}(\boldsymbol{x}_i)-\frac{1}{n}\sum_i\boldsymbol{\omega}^\top\boldsymbol{m}(\boldsymbol{x}_i)\right|^\alpha+\frac{\lambda}{n}\sum_i\ell\big(\boldsymbol{\omega}^\top\boldsymbol{m}(\boldsymbol{x}_i),y_i\big)\right\},$$

for some $\alpha > 0$, where $\mathcal{S}_k$ is the standard probability simplex (as defined in Boyd and Vandenberghe (2004)), $\mathcal{S}_k = \left\{\boldsymbol{w}\in\mathbb{R}_+^k : \boldsymbol{w}^\top\mathbf{1} = 1\right\}$. As proved in Fermanian and Guegan (2021), if $\alpha = 2$ and $\ell = \ell_2$, then

$$\boldsymbol{\omega}^\star_{\text{dp}} = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^\top\Sigma^{-1}\mathbf{1}} \text{ where } \Sigma = \boldsymbol{A}\boldsymbol{A}^\top + \lambda\boldsymbol{B},$$

where

$$\boldsymbol{A} = \frac{1}{n_0}\sum_{i:S_i=0}\boldsymbol{m}(\boldsymbol{x}_i) - \frac{1}{n}\sum_i\boldsymbol{m}(\boldsymbol{x}_i)$$

and

$$\boldsymbol{B} = \frac{1}{n}\sum_{i=1}^n\big(\boldsymbol{m}(\boldsymbol{x}_i)-y_i\mathbf{1}\big)\big(\boldsymbol{m}(\boldsymbol{x}_i)-y_i\mathbf{1}\big)^\top.$$

The tuning parameter λ is positive, and selecting $\lambda > 0$ will yield to a unique solution.

For β-"equalized odds", the optimization problem is

$$\min_{\omega\in\mathcal{S}_k}\left\{\frac{1}{n}\sum_{i=1}^n\left|\widehat{e}^{(\texttt{A})}(y_i)-\widehat{e}(y_i)\right|^\alpha+\frac{\lambda}{n}\sum_i\ell\big(\boldsymbol{\omega}^\top\boldsymbol{m}(\boldsymbol{x}_i),y_i\big)\right\}$$

where $\widehat{e}(y_i)$ is an estimation of $\mathbb{E}[\widehat{Y}|Y = y_i]$ while $\widehat{e}^{(\texttt{A})}(y_i)$ is an estimation of $\mathbb{E}[\widehat{Y}|Y = y_i, S = \texttt{A}]$. For instance, consider some standard nonparametric estimators, kernel based,

$$\widehat{e}(y) \propto \sum_{i=1}^n\boldsymbol{\omega}^\top\boldsymbol{m}(\boldsymbol{x}_i)K_h(y_i - y) = \boldsymbol{\omega}^\top\boldsymbol{v}_h(y),$$

and

$$\widehat{e}^{(\mathtt{A})}(y) \propto \sum_{i:s_i=\mathtt{A}} \boldsymbol{\omega}^\top \boldsymbol{m}(\boldsymbol{x}_i) K_h(y_i - y) = \boldsymbol{\omega}^\top \boldsymbol{v}_h^{(0)}(y)$$

Again, if $\alpha = 2$ and with $\ell = \ell_2$,

$$\boldsymbol{\omega}_{\text{eo}}^\star = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^\top \Sigma^{-1}\mathbf{1}} \text{ where } \Sigma = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^\top + \lambda \boldsymbol{B},$$

where

$$\boldsymbol{\gamma}_i = \boldsymbol{v}_h^{(0)}(y_i) - \boldsymbol{v}_h(y_i)$$

and

$$\boldsymbol{B} = \frac{1}{n}\sum_{i=1}^n \big(\boldsymbol{m}(\boldsymbol{x}_i) - y_i\mathbf{1}\big)\big(\boldsymbol{m}(\boldsymbol{x}_i) - y_i\mathbf{1}\big)^\top.$$

We should keep in mind here that we can always solve numerically this problem.

References

Bibliography

Aas K, Jullum M, Løland A (2021) Explaining individual predictions when features are dependent: More accurate approximations to shapley values. Artificial Intelligence 298:103502

Abraham K (1986) Distributing risk: Insurance, legal theory and public policy. Yale University Press,

Abrams M (2014) The origins of personal data and its implications for governance. SSRN 2510927

Achenwall G (1749) Abriß der neuesten Staatswissenschaft der vornehmsten Europäischen Reiche und Republicken zum Gebrauch in seinen Academischen Vorlesungen. Schmidt

Adams SJ (2004) Age discrimination legislation and the employment of older workers. Labour Economics 11(2):219–241

Agarwal A, Beygelzimer A, Dudík M, Langford J, Wallach H (2018) A reductions approach to fair classification. In: Dy J, Krause A (eds) International Conference on Machine Learning, Proceedings of Machine Learning Research, Stockholmsmässan, Stockholm Sweden, Proceedings of Machine Learning Research, vol 80, pp 60–69

Agrawal T (2013) Are there glass-ceiling and sticky-floor effects in india? an empirical examination. Oxford Development Studies 41(3):322–342

Agresti A (2012) Categorical data analysis. John Wiley & Sons

Agresti A (2015) Foundations of linear and generalized linear models. John Wiley & Sons

Agueh M, Carlier G (2011) Barycenters in the wasserstein space. SIAM Journal on Mathematical Analysis 43(2):904–924

Ahmed AM (2010) What is in a surname? the role of ethnicity in economic decision making. Applied Economics 42(21):2715–2723

Aigner DJ, Cain GG (1977) Statistical theories of discrimination in labor markets. Industrial and Labor Relations Review 30(2):175–187

Ajunwa I (2014) Genetic testing meets big data: Tort and contract law issues. Ohio State Law Journal 75:1225

Ajunwa I (2016) Genetic data and civil rights. Harvard Civil Rights-Civil Liberties Law Review 51:75

Akerlof GA (1970) The market for "lemons": Quality uncertainty and the market mechanism. The Quarterly Journal of Economics 84(3)

Al Ramiah A, Hewstone M, Dovidio JF, Penner LA (2010) The social psychology of discrimination: Theory, measurement and consequences. In: Making Equality Count, The Liffey Press, pp 84–112

Alexander L (1992) What makes wrongful discrimination wrong? biases, preferences, stereotypes, and proxies. University of Pennsylvania Law Review 141(1):149–219

Alexander W (1924) Insurance fables for life underwriters. The Spectator Company

Alipourfard N, Fennell PG, Lerman K (2018) Can you trust the trend? discovering simpson's paradoxes in social data. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp 19–27

Allen CG (1975) Plato on women. Feminist Studies 2(2):131

Allerhand L, Youngmann B, Yom-Tov E, Arkadir D (2018) Detecting parkinson's disease from interactions with a search engine: Is expert knowledge sufficient? In: Proceedings of the 27th ACM international conference on information and knowledge management, pp 1539–1542

Altman A (2011) Discrimination. Stanford Encyclopedia of Philosophy

Altman N, Krzywinski M (2015) Association, correlation and causation. Nature methods 12(10)

Alvarez-Esteban PC, del Barrio E, Cuesta-Albertos JA, Matrán C (2018) Wide consensus aggregation in the wasserstein space. application to location-scatter families. Bernoulli 24:3147–3179

Amadieu JF (2008) Vraies et fausses solutions aux discriminations. Formation emploi Revue française de sciences sociales (101):89–104

Amari SI (1982) Differential geometry of curved exponential families-curvatures and information loss. The Annals of Statistics 10(2):357–385

American Academy of Actuaries (2011) Market consistent embedded value. Life Financial Reporting Committee

Amossé T, De Peretti G (2011) Hommes et femmes en ménage statistique: une valse à trois temps. Travail, genre et sociétés (2):23–46

Anderson TH (2004) The pursuit of fairness: A history of affirmative action. Oxford University Press

de Andrade N (2012) Oblivion: The right to be different from oneself-reproposing the right to be forgotten. In: Cerrillo Martínez A, Peguera Poch M, Peña López I, Vilasau Solana M (eds) VII international conference on internet, law & politics. Net neutrality and other challenges for the future of the Internet, IDP. Revista de Internet, Derecho y Política, 13, pp 122–137

Anguraj K, Padma S (2012) Analysis of facial paralysis disease using image processing technique. International Journal of Computer Applications 54(11)

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. ProPublica May 23

Antoniak M, Mimno D (2021) Bad seeds: Evaluating lexical methods for bias measurement. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp 1889–1904

Antonio K, Beirlant J (2007) Actuarial statistics with generalized linear mixed models. Insurance: Mathematics and Economics 40(1):58–76

Apfelbaum EP, Pauker K, Sommers SR, Ambady N (2010) In blind pursuit of racial equality? Psychological science 21(11):1587–1592

Apley DW, Zhu J (2020) Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82(4):1059–1086

Aran XF, Such JM, Criado N (2019) Attesting biases and discrimination using language semantics. arXiv 1909.04386

Armstrong JS (1985) Long-range Forecasting: from crystal ball to computer

Arneson RJ (1999) Egalitarianism and responsibility. The Journal of Ethics 3:225–247

Arneson RJ (2007) Desert and equality. In: Egalitarianism: New essays on the nature and value of equality, Oxford University Press Oxford, pp 262–293

Arneson RJ (2013) Discrimination, disparate impact, and theories of justice. In: Hellman D, Moreau S (eds) Philosophical foundations of discrimination law, vol 87, Oxford University Press Oxford, p 105

Arrow KJ (1963) Uncertainty and the welfare eco-nomicsof medical care. American Economic Review 53:941–973

Arrow KJ (1973) The theory of discrimination. In: Ashenfelter O, Rees A (eds) Discrimination in labor markets, Princeton University Press

Artıs M, Ayuso M, Guillen M (1999) Modelling different types of automobile insurance fraud behaviour in the spanish market. Insurance: Mathematics and Economics 24(1-2):67–81

Artís M, Ayuso M, Guillén M (2002) Detection of automobile insurance fraud with discrete choice models and misclassified claims. Journal of Risk and Insurance 69(3):325–340

Ashenfelter O, Oaxaca R (1987) The economics of discrimination: Economists enter the courtroom. The American Economic Review 77(2):321–325

Ashley F (2018) Man who changed legal gender to get cheaper insurance exposes the unreliability of gender markers. CBC (Canadian Broadcasting Corporation) - Radio Canada July 28

Atkin A (2012) The philosophy of race. Acumen

Ausloos J (2020) The right to erasure in EU data protection law. Oxford University Press

Austin PC, Steyerberg EW (2012) Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Medical Research Methodology 12:1–8

Austin R (1983) The insurance classification controversy. University of Pennsylvania Law Review 131(3):517–583

Automobile Insurance Rate Board (2022) Technical guidance: Change in rates and rating programs. Albera AIRB

Autor D (2003) Lecture note: the economics of discrimination-theory. Graduate Labor Economics, Massachusetts Institute of Technology pp 1–18

Avery RB, Calem PS, Canner GB (2004) Consumer credit scoring: do situational circumstances matter? Journal of Banking & Finance 28(4):835–856

Avin C, Shpitser I, Pearl J (2005) Identifiability of path-specific effects. IJCAI International Joint Conference on Artificial Intelligence pp 357–363

Avraham R (2017) Discrimination and insurance. In: Lippert-Rasmussen K (ed) Handbook of the Ethics of Discrimination, Routledge, pp 335–347

Avraham R, Logue KD, Schwarcz D (2013) Understanding insurance antidiscrimination law. South California Law Review 87:195

Avraham R, Logue KD, Schwarcz D (2014) Towards a universal framework for insurance anti-discrimination laws. Connecticut Insurance Law Journal 21:1

Ayalon L, Tesch-Römer C (2018) Introduction to the section: Ageism—concept and origins. Contemporary perspectives on ageism pp 1–10

Ayer AJ (1972) Probability and evidence. Columbia University Press

Azen R, Budescu DV (2003) The dominance analysis approach for comparing predictors in multiple regression. Psychological methods 8(2):129

Bachelard G (1927) Essai sur la connaissance approchée. Vrin

Backer DC (2017) Risk profiling in the auto insurance industry. Gracey-Backer, Inc Blog March 14

Baer BR, Gilbert DE, Wells MT (2019) Fairness criteria through the lens of directed acyclic graphical models. arXiv 1906.11333

Bagdasaryan E, Poursaeed O, Shmatikov V (2019) Differential privacy has disparate impact on model accuracy. Advances in Neural Information Processing Systems 32:15479–15488

Bailey RA, Simon LJ (1959) An actuarial note on the credibility of experience of a single private passenger car. Proceedings of the Casualty Actuarial Society XLVI:159

Bailey RA, Simon LJ (1960) Two studies in automobile insurance ratemaking. ASTIN Bulletin: The Journal of the IAA 1(4):192–217

Baird IM (1994) Obesity and insurance risk. Pharmacoeconomics 5(1):62–65

Baker T (2011) Health insurance, risk, and responsibility after the patient protection and affordable care act. University of Pennsylvania Law Review pp 1577–1622

Baker T, McElrath K (1997) Insurance claims discrimination. In: Insurance redlining: Disinvestment, reinvestment, and the evolving role of financial institutions, The Urban Institute Press, pp 141–156

Baker T, Simon J (2002) Embracing Risk: The Changing Culture of Insurance and Responsibility. Chicago: Univ. Chicago Press

Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16(5):412–424

Ban GY, Keskin NB (2021) Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. Management Science 67(9):5549–5568

Banerjee A, Guo X, Wang H (2005) On the optimality of conditional expectation as a bregman predictor. IEEE Transactions on Information Theory 51(7):2664–2669

Banham R (2015) Price optimization or price discrimination? regulators weigh in. Carrier Management May 17

Barbosa JJR (2019) The business opportunities of implementing wearable based products in the health and life insurance industries. PhD thesis, Universidade Católica Portuguesa

Barbour V (1911) Privateers and pirates of the west indies. The American Historical Review 16(3):529–566

Barocas S, Selbst AD (2016) Big data's disparate impact. California Law Review 104:671–732

Barocas S, Hardt M, Narayanan A (2017) Fairness in machine learning. Nips tutorial 1:2017

Barocas S, Hardt M, Narayanan A (2019) Fairness and Machine Learning. fairmlbook.org

Barry L (2020a) Insurance, big data and changing conceptions of fairness. European Journal of Sociology 61:159 – 184

Barry L (2020b) L'invention du risque catastrophes naturelles. Chaire PARI, Document de Travail 18

Barry L, Charpentier A (2020) Personalization as a promise: Can big data change the practice of insurance? Big Data & Society 7(1):2053951720935143

Bartik A, Nelson S (2016) Deleting a signal: Evidence from pre-employment credit checks. SSRN 2759560

Bartlett R, Morse A, Stanton R, Wallace N (2018) Consumer-lending discrimination in the era of fintech. University of California, Berkeley, Working Paper

Bartlett R, Morse A, Stanton R, Wallace N (2021) Consumer-lending discrimination in the fintech era. Journal of Financial Economics

Bath C, Edgar K (2010) Time is money: Financial responsibility after prison. Prison Reform Trust London

Baumann J, Loi M (2023) Fairness and risk: An ethical argument for a group fairness definition insurers can use. Philosophy & Technology 36(3):45

Bayer PB (1986) Mutable characteristics and the definition of discrimination under title vii. UC Davis Law Review 20:769

Bayes T (1763) An essay towards solving a problem in the doctrine of chances. Philosophical transactions of the Royal Society of London (53):370–418

Becker GS (1957) The economics of discrimination. University of Chicago press

Beckett L (2014) Everything we know about what data brokers know about you. ProPublica June 13

Beider P (1987) Sex discrimination in insurance. Journal of Applied Philosophy 4:65–75

Belhadji EB, Dionne G, Tarkhani F (2000) A model for the detection of insurance fraud. Geneva Papers on Risk and Insurance Issues and Practice pp 517–538

Bélisle-Pipon JC, Vayena E, Green RC, Cohen IG (2019) Genetic testing, insurance discrimination and medical research: what the united states can learn from peer countries. Nature medicine 25(8):1198–1204

Bell ET (1945) The development of mathematics. Courier Corporation

Bender M, Dill C, Hurlbert M, Lindberg C, Mott S (2022) Understanding potential influences of racial bias on p&c insurance: Four rating factors explored

Beniger J (2009) The control revolution: Technological and economic origins of the information society. Harvard university press

Benjamin B, Michaelson R (1988) Mortality differences between smokers and non-smokers. Journal of the Institute of Actuaries 115(3):519–525

Bennett M (1978) Models in motor insurance. Journal of the Staple Inn Actuarial Society 22:134–160

Bergstrom CT, West JD (2021) Calling bullshit: the art of skepticism in a data-driven world. Random House Trade Paperbacks

Berk R, Heidari H, Jabbari S, Joseph M, Kearns M, Morgenstern J, Neel S, Roth A (2017) A convex framework for fair regression. arXiv 1706.02409

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2021a) Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research 50(1):3–44

Berk RA, Kuchibhotla AK, Tchetgen ET (2021b) Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets. arXiv 2111.09211

Berkson J (1944) Application of the logistic function to bio-assay. Journal of the American statistical association 39(227):357–365

Bernard DS, Farr SL, Fang Z (2011) National estimates of out-of-pocket health care expenditure burdens among nonelderly adults with cancer: 2001 to 2008. Journal of Clinical Oncology 29(20):2821

Bernoulli J (1713) Ars conjectandi: opus posthumum: accedit Tractatus de seriebus infinitis; et Epistola gallice scripta de ludo pilae reticularis. Impensis Thurnisiorum

Bernstein A (2013) What's wrong with stereotyping. Arizona Law Review 55:655

Bernstein E (2007) Temporarily Yours: Intimacy, Authenticity, and the Commerce of Sex. University of Chicago Press

Bertillon A, Chervin A (1909) Anthropologie métrique: conseils pratiques aux missionnaires scientifiques sur la manière de mesurer, de photographier et de décrire des sujets vivants et des pièces anatomiques. Imprimerie nationale

Bertrand M, Duflo E (2017) Field experiments on discrimination. Handbook of economic field experiments 1:309–393

Bertrand M, Mullainathan S (2004) Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. American economic review 94(4):991–1013

Besnard P, Grange C (1993) La fin de la diffusion verticale des gouts?(prénoms de l'élite et du vulgum). L'Année sociologique pp 269–294

Besse P, del Barrio E, Gordaliza P, Loubes JM (2018) Confidence intervals for testing disparate impact in fair learning. arXiv 1807.06362

Beutel A, Chen J, Doshi T, Qian H, Wei L, Wu Y, Heldt L, Zhao Z, Hong L, Chi EH, et al. (2019) Fairness in recommendation ranking through pairwise comparisons. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 2212–2220

Bhattacharya A (2015) Facebook patent: Your friends could help you get a loan - or not. CNN Business 2015/08/04

Bickel PJ, Hammel EA, O'Connell JW (1975) Sex bias in graduate admissions: Data from berkeley. Science 187(4175):398–404

Bidadanure J (2017) Discrimination and age. In: Lippert-Rasmussen K (ed) Handbook of the Ethics of Discrimination, Routledge, pp 243–253

Biddle D (2017) Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing. Routledge

Biecek P, Burzykowski T (2021) Explanatory model analysis: explore, explain, and examine predictive models. CRC Press

Bielby WT, Baron JN (1986) Men and women at work: Sex segregation and statistical discrimination. American journal of sociology 91(4):759–799

Biemer PP, Christ SL (2012) Weighting survey data. In: International handbook of survey methodology, Routledge, pp 317–341

Bigot R, Cayol A (2020) Le droit des assurances en tableaux. Ellipses

Bigot R, Cocteau-Senn D, Arthur C (2019) La protection des données personnelles en assurance : dialogue du juriste avec l'actuaire. In: Netter E (ed) Regards sur le nouveau droit des données personnelles, CEPRISCA, collection Colloques

Billingsley P (2008) Probability and measure. John Wiley & Sons

Birnbaum B (2020) Insurance consumer protection issues resulting from, or heightened bycovid-19. Center for Economic Justice Report

Blanchet P (2017) Discriminations: combattre la glottophobie. Éditions Textuel

Blanpain N (2018) L'espérance de vie par niveau de vie-méthode et principaux résultats. INSEE Document de Travail F1801

Blier-Wong C, Cossette H, Lamontagne L, Marceau E (2021) Geographic ratemaking with spatial embeddings. ASTIN Bulletin: The Journal of the IAA pp 1–31

Blinder AS (1973) Wage discrimination: Reduced form and structural estimates. The Journal of Human Resources 8(4):436–455

Bloch M (1932) Noms de personne et histoire sociale. Annales d'histoire économique et sociales 4(13):67–69

Blodgett SL, O'Connor B (2017) Racial disparity in natural language processing: A case study of social media african-american english. arXiv 1707.00061

Blumenbach JF (1775) De generis humani varietate nativa. Vandenhoek & Ruprecht

Boczar D, Avila FR, Carter RE, Moore PA, Giardi D, Guliyeva G, Bruce CJ, McLeod CJ, Forte AJ (2021) Using facial recognition tools for health assessment. Plastic Surgical Nursing 41(2):112–116

Bohren JA, Haggag K, Imas A, Pope DG (2019) Inaccurate statistical discrimination: An identification problem. Tech. rep., National Bureau of Economic Research

Bolton LE, Warlop L, Alba JW (2003) Consumer perceptions of price (un) fairness. Journal of consumer research 29(4):474–491

Bongers S, Forré P, Peters J, Mooij JM (2021) Foundations of structural causal models with cycles and latent variables. The Annals of Statistics 49(5):2885–2915

Bonnefon JF (2019) La voiture qui en savait trop. L'intelligence artificielle a-t-elle une morale ? Humensciences Editions

Boonekamp C, Donaldson D (1979) Certain alternatives for price uncertainty. The Canadian Journal of Economics / Revue canadienne d'Economique 12(4):718–728

Boonen TJ, Liu F (2022) Insurance with heterogeneous preferences. Journal of Mathematical Economics 102:102742

Borch K (1962) Application of game theory to some problems in automobile insurance*. ASTIN Bulletin: The Journal of the IAA 2(2):208–221

Borgelt C, Steinbrecher M, Kruse RR (2009) Graphical models: representations for learning, reasoning and data mining. John Wiley & Sons

Borges JL (1946) Del rigor en la ciencia. Los Anales de Buenos Aires

Borkan D, Dixon L, Sorensen J, Thain N, Vasserman L (2019) Nuanced metrics for measuring unintended bias with real data for text classification. In: Companion proceedings of the 2019 world wide web conference, pp 491–500

Bornstein S (2018) Antidiscriminatory algorithms. Alabama Law Review 70:519

Bosmajian HA (1974) The language of oppression, vol 10. Public Affairs Press

Bouk D (2015) How Our Days Became Numbered: Risk and the Rise of the Statistical Individual. The University of Chicago Press

Bouk D (2022) Democracy's Data: The Hidden Stories in the U.S. Census and How to Read Them. MCD

Bourdieu P (2018) Distinction a social critique of the judgement of taste. In: Inequality Classic Readings in Race, Class, and Gender, Routledge, pp 287–318

Bowles S, Gintis H (2004) The evolution of strong reciprocity: cooperation in heterogeneous populations. Theoretical population biology 65(1):17–28

Box GE, Luceño A, del Carmen Paniagua-Quinones M (2011) Statistical control by monitoring and adjustment, vol 700. John Wiley & Sons

Boxill BR (1992) Blacks and social justice. Rowman & Littlefield

Boyd D, Levy K, Marwick A (2014) The networked nature of algorithmic discrimination. Data and Discrimination: Collected Essays Open Technology Institute

Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge university press

Brams SJ, Brams SJ, Taylor AD (1996) Fair Division: From cake-cutting to dispute resolution. Cambridge University Press

Brant S (1494) Das Narrenschiff. von Jakob Locher

Breiman L (1995) Better subset regression using the nonnegative garrote. Technometrics 37(4):373–384

Breiman L (1996a) Bagging predictors. Machine learning 24:123–140

Breiman L (1996b) Bias, variance, and arcing classifiers. Tech. rep., University of California, Berkeley

Breiman L (1996c) Stacked regressions. Machine learning 24:49–64

Breiman L (2001) Random forests. Machine learning 45:5–32

Breiman L, Stone C (1977) Parsimonious binary classification trees. technology service co. rporation, santa monica. Tech. rep., Ca., Technical Report

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Taylor & Francis

Brenier Y (1991) Polar factorization and monotone rearrangement of vector-valued functions. Communications on pure and applied mathematics 44(4):375–417

Brilmayer L, Hekeler RW, Laycock D, Sullivan TA (1979) Sex discrimination in employer-sponsored insurance plans: A legal and demographic analysis. University of Chicago Law Review 47:505

Brilmayer L, Laycock D, Sullivan TA (1983) The efficient use of group averages as nondiscrimination: A rejoinder to professor benston. The University of Chicago Law Review 50(1):222–249

Bröcker J (2009) Reliability, sufficiency, and the decomposition of proper scores. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography 135(643):1512–1519

Brockett PL, Golden LL (2007) Biological and psychobehavioral correlates of credit scores and automobile insurance losses: Toward an explication of why credit scoring works. Journal of Risk and Insurance 74(1):23–63

Brockett PL, Xia X, Derrig RA (1998) Using kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. Journal of Risk and Insurance pp 245–274

Brosnan SF (2006) Nonhuman species' reactions to inequity and their implications for fairness. Social Justice Research 19(2):153–185

Brown RL, Charters D, Gunz S, Haddow N (2007) Colliding interests–age as an automobile insurance rating variable: Equitable rate-making or unfair discrimination? Journal of Business Ethics 72(2):103–114

Brown RS, Moon M, Zoloth BS (1980) Incorporating occupational attainment in studies of male-female earnings differentials. Journal of Human Resources pp 3–28

Browne S (2015) Dark matters: On the surveillance of blackness. Duke University Press

Brownstein M, Saul J (2016a) Implicit bias and philosophy, volume 1: Metaphysics and epistemology. Oxford University Press

Brownstein M, Saul J (2016b) Implicit bias and philosophy, volume 2: Moral responsibility, structural injustice, and ethics. Oxford University Press

Brualdi RA (2006) Combinatorial matrix classes, vol 13. Cambridge University Press

Brubaker R (2015) Grounds for Difference. Harvard University Press

Brudno B (1976) Poverty, Inequality, and the Law. West Publishing Company

Bruner JS (1957) Going beyond the information given. In: Bruner J, Brunswik E, Festinger L, Heider F, Muenzinger K, Osgood C, Rapaport D (eds) Contemporary approaches to cognition, Harvard University Press, pp 119–160

Brunet G, Bideau A (2000) Surnames: history of the family and history of populations. The History of the Family 5(2):153–160

Buchanan R, Priest C (2006) Deductible. Encyclopedia of Actuarial Science

Budd LP, Moorthi RA, Botha H, Wicks AC, Mead J (2021) Automated hiring at amazon. Universiteit van Amsterdam E-0470

Bugbee M, Matthews B, Callanan S, Ewert J, Guven S, Boison L, Liao C (2014) Price optimization overview. Casualty Actuarial Society

Bühlmann H, Gisler A (2005) A course in credibility theory and its applications, vol 317. Springer

Buntine WL, Weigend AS (1991) Bayesian back-propagation. Complex Systems 5

Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency, Proceedings of Machine Learning Research, pp 77–91

Burgdorf MP, Burgdorf Jr R (1974) A history of unequal treatment: The qualifications of handicapped persons as a suspect class under the equal protection clause. Santa Clara Lawyer 15:855

Butler P, Butler T (1989) Driver record: A political red herring that reveals the basic flaw in automobile insurance pricing. Journal of Insurance Regulation 8(2):200–234

Butler RN (1969) Age-ism: Another form of bigotry. The gerontologist 9(4_Part_1):243–246

Cain GG (1986) The economic analysis of labor market discrimination: A survey. Handbook of labor economics 1:693–785

Calders T, Jaroszewicz S (2007) Efficient auc optimization for classification. In: Knowledge Discovery in Databases: PKDD 2007: 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007. Proceedings 11, Springer, pp 42–53

Calders T, Verwer S (2010) Three naive bayes approaches for discrimination-free classification. Data mining and knowledge discovery 21(2):277–292

Calders T, Žliobaite I (2013) Why unbiased computational processes can lead to discriminative decision procedures. In: Discrimination and privacy in the information society, Springer, pp 43–57

Calisher CH (2007) Taxonomy: what's in a name? doesn'ta rose by any other name smell as sweet? Croatian medical journal 48(2):268

Calmon FP, Wei D, Ramamurthy KN, Varshney KR (2017) Optimized data pre-processing for discrimination prevention. arXiv 1704.03354

Cameron J (2004) Calibration - i. Encyclopedia of Statistical Sciences 2

Campbell M (1986) An integrated system for estimating the risk premium of individual car models in motor insurance. ASTIN Bulletin: The Journal of the IAA 16(2):165–183

Candille G, Talagrand O (2005) Evaluation of probabilistic prediction systems for a scalar variable. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography 131(609):2131–2150

Cantwell GT, Kirkley A, Newman MEJ (2021) The friendship paradox in real and model networks. Journal of Complex Networks 9(2)

Cao D, Chen C, Piccirilli M, Adjeroh D, Bourlai T, Ross A (2011) Can facial metrology predict gender? In: 2011 International Joint Conference on Biometrics (IJCB), IEEE, pp 1–8

Cardano G (1564) Liber de ludo aleae. Franco Angeli

Cardon D (2019) Culture numérique. Presses de Sciences Po

Carey AN, Wu X (2022) The causal fairness field guide: Perspectives from social and formal sciences. Frontiers in Big Data 5

Carlier G, Chernozhukov V, Galichon A (2016) Vector quantile regression: an optimal transport approach. The Annals of Statistics 44

Carnis L, Lassarre S (2019) Politique et management de la sécurité routière. In: Laurent C, Catherine G, Marie-Line G (eds) La sécurité routière en France, Quand la recherche fait son bilan et trace des perspectives, L'Harmattan

Carpusor AG, Loges WE (2006) Rental discrimination and ethnicity in names. Journal of applied social psychology 36(4):934–952

Carrasco V (2007) Le pacte civil de solidarité: une forme d'union qui se banalise. Infostat Justice 97(4)

Cartwright N (1983) How the Laws of Physics Lie. Oxford University Press

Casella G, Berger RL (1990) Statistical Inference. Duxbury Advanced Series

Casey B, Pezier J, Spetzler C (1976) The Role of Risk Classification in Property and Casualty Insurance: A Study of the Risk Assessment Process : Final Report. Stanford Research Institute

Cassedy JH (2013) Demography in early America. Harvard University Press

Castelvecchi D (2016) Can we open the black box of ai? Nature News 538(7623):20

Caton S, Haas C (2020) Fairness in machine learning: A survey. arXiv 2010.04053

Cavanagh M (2002) Against equality of opportunity. Clarendon Press

Central Bank of Ireland (2021) Review of differential pricing in the private car and home insurance markets

Chakraborty S, Raghavan KR, Johnson MP, Srivastava MB (2013) A framework for context-aware privacy of sensor data on mobile systems. In: Proceedings of the 14th Workshop on Mobile Computing Systems and Applications, Association for Computing Machinery, HotMobile '13

Chardenon A (2019) Voici maxime, le chatbot juridique d'axa, fruit d'une démarche collaborative

Charles KK, Guryan J (2011) Studying discrimination: Fundamental challenges and recent progress. Annual Review of Economics 3(1):479–511

Charpentier A (2014) Computational actuarial science with R. CRC press

Charpentier A, Denuit M (2004) Mathématiques de l'assurance non-vie - Tarification et provisionnement (Tome 1). Economica

Charpentier A, Flachaire E, Ly A (2018) Econometrics and machine learning. Economie et Statistique 505(1):147–169

Charpentier A, Élie R, Remlinger C (2021) Reinforcement Learning in Economics and Finance. Computational Economics 2003.10014

Charpentier A, Flachaire E, Gallic E (2023) Optimal transport for counterfactual estimation: A method for causal inference. In: Thach NN, Kreinovich V, Ha DT, Trung ND (eds) Optimal Transport Statistics for Economics and Related Topics, Springer Verlag

Chassagnon A (1996) Sélection adverse: modèle générique et applications. PhD thesis, Paris, EHESS

Chassonnery-Zaïgouche C (2020) How economists entered the 'numbers game': Measuring discrimination in the us courtrooms, 1971–1989. Journal of the History of Economic Thought 42(2):229–259

Chatterjee S, Barcun S (1970) A nonparametric approach to credit screening. Journal of the American statistical Association 65(329):150–154

Chaufton A (1886) Les assurances, leur passé, leur présent, leur avenir, au point de vue rationnel, technique et pratique, moral, économique et social, financier et administratif, légal, législatif et contractuel, en France et à l'étranger. Chevalier-Marescq

Chen SX (1999) Beta kernel estimators for density functions. Computational Statistics & Data Analysis 31(2):131–145

Chen Y, Liu Y, Zhang M, Ma S (2017) User satisfaction prediction with mouse movement information in heterogeneous search environment. IEEE Transactions on Knowledge and Data Engineering 29(11):2470–2483

Cheney-Lippold J (2017) We are data. In: We Are Data, New York University Press

Cheng M, De-Arteaga M, Mackey L, Kalai AT (2023) Social norm bias: residual harms of fairness-aware algorithms. Data Mining and Knowledge Discovery pp 1–27

Chetty R, Stepner M, Abraham S, Lin S, Scuderi B, Turner N, Bergeron A, Cutler D (2016) The association between income and life expectancy in the united states, 2001-2014. Jama 315(16):1750–1766

Cheung I, McCartt AT (2011) Declines in fatal crashes of older drivers: Changes in crash risk and survivability. Accident Analysis & Prevention 43(3):666–674

Chiappa S (2019) Path-specific counterfactual fairness. Proceedings of the AAAI Conference on Artificial Intelligence 33(01):7801–7808

Chicco D, Jurman G (2020) The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC genomics 21(1):1–13

Chollet F (2021) Deep learning with Python. Simon and Schuster

Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data 5(2):153–163

Christensen CM, Dillon K, Hall T, Duncan DS (2016) Competing against luck: The story of innovation and customer choice. Harper Business

Churchill G, Nevin JR, Watson RR (1977) The role of credit scoring in the loan decision. Credit World 3(3):6–10

Cinelli C, Hazlett C (2020) Making sense of sensitivity: Extending omitted variable bias. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82(1):39–67

Clark G, Clark GW (1999) Betting on lives: the culture of life insurance in England, 1695-1775. Manchester University Press

Clarke DD, Ward P, Bartle C, Truman W (2010) Older drivers' road traffic crashes in the uk. Accident Analysis & Prevention 42(4):1018–1024

Cohen I, Goldszmidt M (2004) Properties and benefits of calibrated classifiers. In: 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Springer, vol 3202, pp 125–136

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and psychological measurement 20(1):37–46

Cohen JE (1986) An uncertainty principle in demography and the unisex issue. The American Statistician 40(1):32–39

Coldman AJ, Braun T, Gallagher RP (1988) The classification of ethnic status using name information. Journal of Epidemiology & Community Health 42(4):390–395

Collins BW (2007) Tackling unconscious bias in hiring practices: The plight of the rooney rule. New York University Law Review 82:870

Collins E (2018) Punishing risk. Georgetown Law Journal 107:57

de Condorcet N (1785) Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Paris: Imprimerie royale

Constine J (2017) Facebook rolls out ai to detect suicidal posts before they're reported. Techcrunch November 27

Conway DA, Roberts HV (1983) Reverse regression, fairness, and employment discrimination. Journal of Business & Economic Statistics 1(1):75–85

Cook TD, Campbell DT, Shadish W (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin Boston, MA

Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H (2013) Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Human genetics 132:1077–1130

Cooper PJ (1990) Differences in accident characteristics among elderly drivers and between elderly and middle-aged drivers. Accident analysis & prevention 22(5):499–508

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. arXiv 1701.08230

Corlier F (1998) Segmentation : le point de vue de l'assureur. In: Cousy H, Classens H, Van Schoubroeck C (eds) Compétitivité, éthique et assurance, Academia Bruylant

Cornell B, Welch I (1996) Culture, information, and screening discrimination. Journal of Political Economy 104(3):542–571

Correll J, Judd CM, Park B, Wittenbrink B (2010) Measuring prejudice, stereotypes and discrimination. The SAGE handbook of prejudice, stereotyping and discrimination pp 45–62

Correll SJ, Benard S (2006) Biased estimators? comparing status and statistical theories of gender discrimination. In: Advances in group processes, vol 23, Emerald Group Publishing Limited, pp 89–116

Cortina A (2022) Aporophobia: Why We Reject the Poor Instead of Helping Them. Princeton University Press

Côté O (2023) Methodology applied to build a non-discriminatory general insurance rate according to a pre-specified sensitive variable. MSc Thesis, Université Laval

Cotter A, Jiang H, Gupta MR, Wang S, Narayan T, You S, Sridharan K (2019) Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. Journal of Machine Learning Research 20(172):1–59

Cotton J (1988) On the decomposition of wage differentials. The review of economics and statistics pp 236–243

Coulmont B (2011) Sociologie des prénoms. la Découverte

Coulmont B, Simon P (2019) Quels prénoms les immigrés donnent-ils à leurs enfants en france? Population Societes (4):1–4

Council of the European Union (2004) Council directive 2004/113/ec of 13 december 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services. Official Journal of the European Union (373):37–43

Cournot AA (1843) Exposition de la théorie des chances et des probabilités. Hachette

Coutts S (2016) Anti-choice groups use smartphone surveillance to target 'abortion-minded women'during clinic visits. Rewire News Group May 25

Cowell F (2011) Measuring inequality. Oxford University Press

Cragg JG (1971) Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica: Journal of the Econometric Society pp 829–844

Crawford JT, Leynes PA, Mayhorn CB, Bink ML (2004) Champagne, beer, or coffee? a corpus of gender-related and neutral words. Behavior Research Methods, Instruments, & Computers 36:444–458

Cresta J, Laffont J (1982) The value of statistical information in insurance contracts. GREMAQ Working Paper 8212

Crizzle AM, Classen S, Uc EY (2012) Parkinson disease and driving: an evidence-based review. Neurology 79(20):2067–2074

Crocker KJ, Snow A (2013) The theory of risk classification. In: Loubergé H, Dionne G (eds) Handbook of insurance, Springer, pp 281–313

Crossney KB (2016) Redlining. https://philadelphiaencyclopediaorg/essays/redlining/

Cudd AE, Jones LE (2005) Sexism. A companion to applied ethics pp 102–117

Cummins JD, Smith BD, Vance RN, Vanderhel J (2013) Risk classification in life insurance, vol 1. Springer Science & Business Media

Cunha HS, Sclauser BS, Wildemberg PF, Fernandes EAM, Dos Santos JA, Lage MdO, Lorenz C, Barbosa GL, Quintanilha JA, Chiaravalloti-Neto F (2021) Water tank and swimming pool detection based on remote sensing and deep learning: Relationship with socioeconomic level and applications in dengue control. Plos one 16(12):e0258681

Cunningham S (2021) Causal inference. Yale University Press

Cybenko G (1989) Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems 2(4):303–314

Czerniawski AM (2007) From average to ideal: The evolution of the height and weight table in the united states, 1836-1943. Social Science History 31(2):273–296

Da Silva N (2023) La bataille de la Sécu: une histoire du système de santé. La fabrique éditions

Dalenius T (1977) Towards a methodology for statistical disclosure control. statistik Tidskrift 15(429-444):2–1

Dalziel JR, Job RS (1997) Motor vehicle accidents, fatigue and optimism bias in taxi drivers. Accident Analysis & Prevention 29(4):489–494

Dambrum M, Despres G, Guimond S (2003) On the multifaceted nature of prejudice: Psychophysiological responses to ingroup and outgroup ethnic stimuli. Current Research in Social Psychology 8(14):187–206

Dane SM (2006) The potential for racial discrimination by homeowners insurers through the use of geographic rating territories. Journal of Insurance Regulation 24(4):21

Daniel JE, Daniel JL (1998) Preschool children's selection of race-related personal names. Journal of Black Studies 28(4):471–490

Daniels N (1990) Insurability and the hiv epidemic: ethical issues in underwriting. The Milbank Quarterly pp 497–525

Daniels N (1998) Am I my parents' keeper? An Essay On Justice Between The Young And The Old. Oxford University Press

Dar-Nimrod I, Heine SJ (2011) Genetic essentialism: on the deceptive determinism of dna. Psychological bulletin 137(5):800

Darlington RB (1971) Another look at "cultural fairness" 1. Journal of educational measurement 8(2):71–82

Daston L (1992) Objectivity and the escape from perspective. Social studies of science 22(4):597–618

Davenport T (2006) Competing on analytics. harvard Business Review, 84:1–10

David H (2015) Why are there still so many jobs? the history and future of workplace automation. Journal of economic perspectives 29(3):3–30

Davidson R, MacKinnon JG, et al. (2004) Econometric theory and methods, vol 5. Oxford University Press New York

Davis GA (2004) Possible aggregation biases in road safety research and a mechanism approach to accident modeling. Accident Analysis & Prevention 36(6):1119–1127

Dawid AP (1979) Conditional independence in statistical theory. Journal of the Royal Statistical Society: Series B (Methodological) 41(1):1–15

Dawid AP (1982) The well-calibrated bayesian. Journal of the American Statistical Association 77(379):605–610

Dawid AP (2000) Causal inference without counterfactuals. Journal of the American statistical Association 95(450):407–424

Dawid AP (2004) Probability forecasting. Encyclopedia of statistical sciences 10

De Alba E (2004) Bayesian claims reserving. Encyclopedia of Actuarial Science

De Baere G, Goessens E (2011) Gender differentiation in insurance contracts after the judgment in case c-236/09, association belge des consomma teurs test-achats asbl v. conseil des ministres. Colum J Eur L 18:339

De Pril N, Dhaene J (1996) Segmentering in verzekeringen. DTEW Research Report 9648 pp 1–56

De Wit GW, Van Eeghen J (1984) Rate making and society's sense of fairness. ASTIN Bulletin: The Journal of the IAA 14(2):151–163

De Witt J (1671) Value of life annuities in proportion to redeemable annuities. Originally in Dutch Translated in Hendriks (1853) pp 232–49

Dean LT, Nicholas LH (2018) Using credit scores to understand predictors and consequences of disease

Dean LT, Schmitz KH, Frick KD, Nicholas LH, Zhang Y, Subramanian S, Visvanathan K (2018) Consumer credit as a novel marker for economic burden and health after cancer in a diverse population of breast cancer survivors in the usa. Journal of Cancer Survivorship 12(3):306–315

Debet A (2007) Mesure de la diversité et protection des données personnelles. Commission Nationale de l'Informatique et des Libertés 16/05/2007 08:40 DECO / IRC

Défenseur des droits (2020) Algorithmes : prévenir l'automatisation des discriminations

Degroot MH (2004) Well-calibrated forecasts. Encyclopedia of Statistical Sciences

Dehejia RH, Wahba S (1999) Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association 94(448):1053–1062

Delaporte P (1962) Sur l'efficacité des critères de tarification de l'assurance contre les accidents d'automobiles. ASTIN Bulletin: The Journal of the IAA 2(1):84–95

Delaporte PJ (1965) Tarification du risque individuel d'accidents d'automobiles par la prime modelée sur le risque. ASTIN Bulletin: The Journal of the IAA 3(3):251–271

Demakakos P, Biddulph JP, Bobak M, Marmot MG (2016) Wealth and mortality at older ages: a prospective cohort study. Journal of Epidemiology and Community Health 70(4):346–353

Dennis RM (2004) Racism. In: Kuper A, Kuper J (eds) The Social Science Encyclopedia, Routledge

Denuit M, Charpentier A (2004) Mathématiques de l'assurance non-vie: Tome I Principes fondamentaux de théorie du risque. Economica

Denuit M, Charpentier A (2005) Mathématiques de l'assurance non-vie: Tome II Tarification et provisionnement. Economica

Denuit M, Maréchal X, Pitrebois S, Walhin JF (2007) Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems. John Wiley & Sons

Denuit M, Hainault D, Trufin J (2019a) Effective Statistical Learning Methods for Actuaries I (GLMs and Extensions). Springer Verlag

Denuit M, Hainault D, Trufin J (2019b) Effective Statistical Learning Methods for Actuaries III (Neural Networks and Extensions). Springer Verlag

Denuit M, Hainault D, Trufin J (2020) Effective Statistical Learning Methods for Actuaries II (Tree-based methods and Extensions). Springer Verlag

Denuit M, Charpentier A, Trufin J (2021) Autocalibration and tweedie-dominance for insurance pricing with machine learning. Insurance: Mathematics & Economics

Depoid P (1967) Applications de la statistique aux assurances accidents et dommages: cours professé à l'Institut de statistique de l'Université de Paris. 2e édition revue et augmentée... Berger-Levrault

Derrig RA, Ostaszewski KM (1995) Fuzzy techniques of pattern recognition in risk and claim classification. Journal of Risk and Insurance pp 447–482

Derrig RA, Weisberg HI (1998) Aib pip claim screening experiment final report. understanding and improving the claim investigation process. AIB Filing on Fraudulent Claims Payment

Desrosières A (2016) La politique des grands nombres: histoire de la raison statistique. La découverte

Devine PG (1989) Stereotypes and prejudice: Their automatic and controlled components. Journal of personality and social psychology 56(1):5

Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26(3):297–302

Dierckx G (2006) Logistic regression model. Encyclopedia of Actuarial Science

Dieterich W, Mendoza C, Brennan T (2016) Compas risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc 7(7.4):1

Dilley S, Greenwood G (2017) Abandoned 999 calls to police more than double. BBC 19 September 2017

DiNardo J (2016) Natural Experiments and Quasi-Natural Experiments, Palgrave Macmillan UK, London, pp 1–12

DiNardo J, Fortin N, Lemieux T (1995) Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. National bureau of economic research (NBER)

Dingman H (1927) Insurability, prognosis and selection. The spectator company

Dinur R, Beit-Hallahmi B, Hofman JE (1996) First names as identity stereotypes. The Journal of social psychology 136(2):191–200

Dionne G (2000) Handbook of insurance. Springer

Dionne G (2013) Contributions to insurance economics. Springer

Dionne G, Harrington SE (1992) An introduction to insurance economics. Springer

Dobbin F (2001) Do the social sciences shape corporate anti-discrimination practice: The united states and france. Comparative Labor Law & Policy Journal 23:829

Donoghue JD (1957) An eta community in japan: the social persistence of outcaste groups. American Anthropologist 59(6):1000–1017

Dorlin E (2005) Sexe, genre et intersexualité: la crise comme régime théorique. Raisons politiques (2):117–137

Dostie G (1974) Entrevue de michèle lalonde. Le Journal 1er juin 1974

Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Science Advances 4(1):eaao5580

Du Bois W (1896) Review of race traits and tendencies of the american negro. Annals of the American Academy pp 127–33

Duan T, Anand A, Ding DY, Thai KK, Basu S, Ng A, Schuler A (2020) Ngboost: Natural gradient boosting for probabilistic prediction. In: International Conference on Machine Learning, PMLR, pp 2690–2700

Dubet F (2014) La Préférence pour l'inégalité. Comprendre la crise des solidarités: Comprendre la crise des solidarités. Seuil - La République des idées

Dubet F (2016) Ce qui nous unit : Discriminations, égalité et reconnaissance. Seuil - La République des idées

Dublin L (1925) Report of the joint committee on mortality of the association of life insurance medical directors. The Actuarial Society of America

Dudley RM (2010) Distances of probability measures and random variables. In: Selected Works of RM Dudley, Springer, pp 28–37

Duggan JE, Gillingham R, Greenlees JS (2008) Mortality and lifetime income: evidence from us social security records. IMF Staff Papers 55(4):566–594

Duhigg C (2019) How companies learn your secrets. The New York Times 02-16-2019

Duivesteijn W, Feelders A (2008) Nearest neighbour classification with monotonicity constraints. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, pp 301–316

Dulisse B (1997) Older drivers and risk to other road users. Accident Analysis & Prevention 29(5):573–582

Dumas A, Allodji R, Fresneau B, Valteau-Couanet D, El-Fayech C, Pacquement H, Laprie A, Nguyen TD, Bondiau PY, Diallo I, et al. (2017) The right to be forgotten: a change in access to insurance and loans after childhood cancer? Journal of Cancer Survivorship 11:431–437

Duncan A, McPhail M (2013) Price optimization for the us market. techniques and implementation strategies.". In: Ratemaking and Product Management Seminar

Duncan C, Loretto W (2004) Never the right age? gender and age-based discrimination in employment. Gender, Work & Organization 11(1):95–115

Durkheim É (1897) Le suicide: étude sociologique. Félix Alcan Editeur

Durry G (2001) La sélection de la clientèle par l'assureur : aspects juridiques. Risques 45:65–71

Dwivedi M, Malik HS, Omkar S, Monis EB, Khanna B, Samal SR, Tiwari A, Rathi A (2021) Deep learning-based car damage classification and detection. In: Advances in Artificial Intelligence and Data Engineering: Select Proceedings of AIDE 2019, Springer, pp 207–221

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, vol 1104.3913, pp 214–226

Dwoskin E (2018) Facebook is rating the trustworthiness of its users on a scale from zero to one. Washington Post 21-08

Eco U (1992) Comment voyager avec un saumon. Grasset

Edgeworth FY (1922) Equal pay to men and women for equal work. The Economic Journal 32(128):431–457

Edwards J (1932) Ten years of rates and rating bureaus in ontario, applied to automobile insurance. Proceedings of Casualty Actuarial Society 19:22—64

Eidelson B (2015) Discrimination and disrespect. Oxford University Press

Eidinger E, Enbar R, Hassner T (2014) Age and gender estimation of unfiltered faces. IEEE Transactions on information forensics and security 9(12):2170–2179

Eisen R, Eckles DL (2011) Insurance economics. Springer

Ekeland I (1995) Le chaos. Flammarion

England P (1994) Neoclassical economists' theories of discrimination. In: Equal employment opportunity: Labor market discrimination and public policy, Aldine de Gruyter, pp 59–70

Ensmenger N (2015) "beards, sandals, and other signs of rugged individualism": masculine culture within the computing professions. Osiris 30(1):38–65

Epstein L, King G (2002) The rules of inference. The University of Chicago Law Review pp 1–133

Erwin C, Williams JK, Juhl AR, Mengeling M, Mills JA, Bombard Y, Hayden MR, Quaid K, Shoulson I, Taylor S, et al. (2010) Perception, experience, and response to genetic discrimination in huntington disease: The international respond-hd study. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 153(5):1081–1093

Erwin PG (1995) A review of the effects of personal name stereotypes. Representative Research in Social Psychology

European Commission (1995) Directive 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Communities 38(281):31–50

of the European Union C (2018) Proposal for a council directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disablility, age or sexual orientation. Proceedings of Council of the European Union 11015/08

Ewald F (1986) Histoire de l'Etat providence: les origines de la solidarité. Grasset

Eze EC (1997) Race and the enlightenment: A reader. Wiley

Fagyal Z (2010) Accents de banlieue. Aspects prosodiques du français populaire en contact avec les langues de l'immigration, L'Harmattan

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96(456):1348–1360

Farbmacher H, Huber M, Laffértls L, Langen H, Spindler M (2022) Causal mediation analysis with double machine learning. The Econometrics Journal 25(2):277–300

Farebrother RW (1976) Further results on the mean square error of ridge regression. Journal of the Royal Statistical Society Series B (Methodological) pp 248–250

Feder A, Oved N, Shalit U, Reichart R (2021) `causalm`: Causal model explanation through counterfactual language models. Computational Linguistics 47(2):333–386

Feeley MM, Simon J (1992) The new penology: Notes on the emerging strategy of corrections and its implications. Criminology 30(4):449–474

Feinberg J (1970) Justice and personal desert. In: Feinberg J (ed) Doing and Deserving

Feine J, Gnewuch U, Morana S, Maedche A (2019) Gender bias in chatbot design. In: International Workshop on Chatbot Research and Design, Springer, pp 79–93

Feiring E (2009) reassessing insurers' access to genetic information: genetic privacy, ignorance, and injustice. Bioethics 23(5):300–310

Feld SL (1991) Why your friends have more friends than you do. American journal of sociology 96(6):1464–1477

Feldblum S (2006) Risk classification, pricing aspects. Encyclopedia of actuarial science

Feldblum S, Brosius JE (2003) The minimum bias procedure: A practitioner's guide. In: Proceedings of the Casualty Actuarial Society, Casualty Actuarial Society Arlington, vol 90, pp 196–273

Feldman F (1995) Desert: Reconsideration of some received wisdom. Mind 104(413):63–77

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. In: proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, vol 1412.3756, pp 259–268

Feller A, Pierson E, Corbett-Davies S, Goel S (2016) A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. The Washington Post October 17

Feller W (1957) An introduction to probability theory and its applications. Wiley

Fenton N, Neil M (2018) Risk assessment and decision analysis with Bayesian networks. CRC Press

Ferber MA, Green CA (1982a) Employment discrimination: Reverse regression or reverse logic. Working Paper, University of Illinois, Champaign

Ferber MA, Green CA (1982b) Traditional or reverse sex discrimination? a case study of a large public university. Industrial and Labor Relations Review 35(4):550–564

Fermanian JD, Guegan D (2021) Fair learning with bagging. SSRN 3969362

Finger RJ (2006) Risk classification. In: Bass I, Basson S, Bashline D, Chanzit L, Gillam W, Lotkowski E (eds) Foundations of Casualty Actuarial Science, Casualty Actuarial Society, pp 287–341

Finkelstein A, Taubman S, Wright B, Bernstein M, Gruber J, Newhouse JP, Allen H, Baicker K, Group OHS (2012) The oregon health insurance experiment: evidence from the first year. The Quarterly journal of economics 127(3):1057–1106

Finkelstein EA, Brown DS, Wrage LA, Allaire BT, Hoerger TJ (2010) Individual and aggregate years-of-life-lost associated with overweight and obesity. Obesity 18(2):333–339

Finkelstein MO (1980) The judicial reception of multiple regression studies in race and sex discrimination cases. Columbia Law Review 80(4):737–754

Firpo SP (2017) Identifying and measuring economic discrimination. IZA World of Labor

Fiscella K, Fremont AM (2006) Use of geocoding and surname analysis to estimate race and ethnicity. Health services research 41(4p1):1482–1500

Fish HC (1868) The Agent's Manual of Life Assurance. Wynkoop & Hallenbeck

Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res 20(177):1–81

Fisher FM (1980) Multiple regression in legal proceedings. Columbia Law Review 80:702

Fisher RA (1921) Studies in crop variation. i. an examination of the yield of dressed grain from broadbalk. The Journal of Agricultural Science 11(2):107–135

Fisher RA, Mackenzie WA (1923) Studies in crop variation. ii. the manurial response of different potato varieties. The Journal of Agricultural Science 13(3):311–320

Flanagan T (1985) Insurance, human rights, and equality rights in canada: When is discrimination "reasonable?". Canadian Journal of Political Science/Revue canadienne de science politique 18(4):715–737

Fleurbaey M (1996) Théories économiques de la justice. Economica

Flew A (1993) Three concepts of racism. International social science review 68(3):99

Fong C, Hazlett C, Imai K (2018) Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. The Annals of Applied Statistics 12(1):156–177

Fontaine H (2003) Driver age and road traffic accidents: what is the risk for seniors? Recherche-transports-sécurité

Fontaine KR, Redden DT, Wang C, Westfall AO, Allison DB (2003) Years of life lost due to obesity. Journal of the American Medical Association 289(2):187–193

Foot P (1967) The problem of abortion and the doctrine of the double effect. Oxford review 5

Forfar DO (2006) Life table. Encyclopedia of Actuarial Science

Fortin N, Lemieux T, Firpo S (2011) Decomposition methods in economics. In: Handbook of labor economics, vol 4, Elsevier, pp 1–102

Fourcade M (2016) Ordinalization: Lewis a. coser memorial award for theoretical agenda setting 2014. Sociological Theory 34(3):175–195

Fourcade M, Healy K (2013) Classification situations: Life-chances in the neoliberal era. Accounting, Organizations and Society 38(8):559–572

Fox ET (2013) 'Piratical Schemes and Contracts': Pirate Articles and Their Society 1660-1730. PhD Thesis, University of Exeter

François P (2022) Catégorisation, individualisation. retour sur les scores de crédit. hal 03508245

Freedman DA (1999) Ecological inference and the ecological fallacy. International Encyclopedia of the social & Behavioral sciences 6(4027-4030):1–7

Freedman DA, Berk RA (2008) Weighting regressions by propensity scores. Evaluation review 32(4):392–409

Freeman S (2007) Rawls. Routledge

Frees EW (2006) Regression models for data analysis. Encyclopedia of Actuarial Science

Frees EW, Huang F (2023) The discriminating (pricing) actuary. North American Actuarial Journal 27(1):2–24

Frees EW, Meyers G, Cummings AD (2011) Summarizing insurance scores using a gini index. Journal of the American Statistical Association 106(495):1085–1098

Frees EW, Derrig RA, Meyers G (2014a) Predictive modeling applications in actuarial science, vol 1. Cambridge University Press

Frees EW, Meyers G, Cummings AD (2014b) Insurance ratemaking and a gini index. Journal of Risk and Insurance 81(2):335–366

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences 55(1):119–139

Frezal S, Barry L (2019) Fairness in uncertainty: Some limits and misinterpretations of actuarial fairness. Journal of Business Ethics

Fricker M (2007) Epistemic injustice: Power and the ethics of knowing. Oxford University Press

Friedler SA, Scheidegger C, Venkatasubramanian S (2016) On the (im) possibility of fairness. arXiv 1609.07236

Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D (2019) A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency, pp 329–338

Friedman J (2001) Greedy function approximation: a gradient boosting machine. Annals of statistics pp 1189–1232

Friedman J, Popescu BE (2008) Predictive learning via rule ensembles. The annals of applied statistics pp 916–954

Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics 28(2):337–407

Friedman J, Hastie T, Tibshirani R, et al. (2001) The elements of statistical learning. Springer

Friedman S, Canaan M (2014) Overcoming speed bumps on the road to telematics. In: Challenges and opportunities facing auto insurers with and without usage-based programs

Frisch R, Waugh FV (1933) Partial time regressions as compared with individual trends. Econometrica pp 387–401

Froot KA, Kim M, Rogoff KS (1995) The law of one price over 700 years. National Bureau of Economic Research (NBER) 5132

Fry T (2015) A discussion on credibility and penalised regression, with implications for actuarial work. Actuaries Institute

Fryer Jr RG, Levitt SD (2004) The causes and consequences of distinctively black names. The Quarterly Journal of Economics 119(3):767–805

Gaddis SM (2017) How black are lakisha and jamal? racial perceptions from names used in correspondence audit studies. Sociological Science 4:469–489

Gadet F (2007) La variation sociale en français. Editions Ophrys

Gal Y, Ghahramani Z (2016) Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning, PMLR, pp 1050–1059

Galichon A (2016) Optimal transport methods in economics. Princeton University Press

Galindo C, Moreno P, González J, Arevalo V (2009) Swimming pools localization in colour high-resolution satellite images. In: 2009 IEEE International Geoscience and Remote Sensing Symposium, IEEE, vol 4, pp IV–510

Galles D, Pearl J (1998) An axiomatic characterization of causal counterfactuals. Foundations of Science 3:151–182

Galton F (1907) Vox populi. Nature 75(7):450–451

Gambs S, Killijian MO, del Prado Cortez MNn (2010) Show me how you move and i will tell you who you are. In: Proceedings of the 3rd ACM International Workshop on Security and Privacy in GIS and LBS

Gan G, Valdez EA (2020) Data clustering with actuarial applications. North American Actuarial Journal 24(2):168–186

Gandy OH (2016) Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage. Routledge

Garg N, Schiebinger L, Jurafsky D, Zou J (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences 115(16):E3635–E3644

Garrioch D (2011) Mutual aid societies in eighteenth-century paris. French History & Civilization 4

Gautron V, Dubourg É (2015) La rationalisation des outils et méthodes d'évaluation: de l'approche clinique au jugement actuariel. Criminocorpus Revue d'Histoire de la justice, des crimes et des peines

Gebelein H (1941) Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik 21(6):364–379

Gebru T, Krause J, Wang Y, Chen D, Deng J, Aiden EL, Fei-Fei L (2017) Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. Proceedings of the National Academy of Sciences 114(50):13108–13113

Geenens G (2014) Probit transformation for kernel density estimation on the unit interval. Journal of the American Statistical Association 109(505):346–358

Geiger A, Wu Z, Lu H, Rozner J, Kreiss E, Icard T, Goodman N, Potts C (2022) Inducing causal structure for interpretable neural networks. In: International Conference on Machine Learning, PMLR, pp 7324–7338

Gelbrich M (1990) On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. Mathematische Nachrichten 147(1):185–203

Gelman A (2009) Red state, blue state, rich state, poor state: Why Americans vote the way they do-expanded edition. Princeton University Press

Ghani N (2008) Racism. In: Schaefer RT (ed) Encyclopedia of Race, Ethnicity, and Society, Sage Publications, pp 1113–1115

Giles C (2020) Goodhart's law comes back to haunt the uk's covid strategy. Financial Times 14-5

Gini C (1912) Variabilità e mutabilità. Contributo allo studio delle distribuzioni e delle relazioni statistiche

Gino F, Pierce L (2010) Robin hood under the hood: Wealth-based discrimination in illicit customer help. Organization Science 21(6):1176–1194

Ginsburg M (1940) Roman military clubs and their social functions. In: Transactions and Proceedings of the American Philological Association, JSTOR, vol 71, pp 149–156

Gintis H (2000) Strong reciprocity and human sociality. Journal of theoretical biology 206(2):169–179

Glenn BJ (2000) The shifting rhetoric of insurance denial. Law and Society Review pp 779–808

Glenn BJ (2003) Postmodernism: the basis of insurance. Risk Management and Insurance Review 6(2):131–143

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association 102(477):359–378

Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69(2):243–268

Goldberger AS (1984) Reverse regression and salary discrimination. Journal of Human Resources pp 293–318

Goldman A (1979) Justice and Reverse Discrimination

Goldstein A, Kapelner A, Bleich J, Pitkin E (2015) Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. journal of Computational and Graphical Statistics 24(1):44–65

Gollier C (2002) La solidarite sous l'angle economique'. Revue Générale du Droit des Assurances pp 824–830

Gompertz B (1825) On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. Philosophical Transactions of the Royal Society of London (115):513–583

Gompertz B (1833) On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. frs &c. by benjamin gompertz, esq. fr s. Philosophical Transactions of the Royal Society of London (2):252–253

Good IJ (1950) Probability and the Weighing of Evidence. Griffin

Goodwin S, Voola AP (2013) Framing microfinance in australia–gender neutral or gender blind? Australian Journal of Social Issues 48(2):223–239

Gordaliza P, Del Barrio E, Fabrice G, Loubes JM (2019) Obtaining fairness using optimal transport theory. In: International Conference on Machine Learning, Proceedings of Machine Learning Research, pp 2357–2365

Gosden PHJH (1961) The friendly societies in England, 1815-1875. Manchester University Press

Gosseries A (2014) What makes age discrimination special: A philosophical look at the ecj case law. Netherlands Journal of Legal Philosophy 43:59–80

Gottlieb S (2011) Medicaid is worse than no coverage at all. Wall Street Journal 10/03

Goulet JA, Nguyen LH, Amiri S (2021) Tractable approximate gaussian inference for bayesian neural networks. J Mach Learn Res 22:251–1

Gouriéroux C (1999) The econometrics of risk classification in insurance. The Geneva Papers on Risk and Insurance Theory 24(2):119–137

Gourieroux C (1999) Statistique de l'assurance. Economica

Gourieroux C, Jasiak J (2007) The econometrics of individual risk. Princeton university press

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics pp 857–871

Gowri A (2014) The irony of insurance: Community and commodity. PhD thesis, University of Southern California

Granger CW (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica: journal of the Econometric Society pp 424–438

Graves Jr JL (2015) Great is their sin: Biological determinism in the age of genomics. The Annals of the American Academy of Political and Social Science 661(1):24–50

Greene WH (1984) Reverse regression: The algebra of discrimination. Journal of Business & Economic Statistics 2(2):117–120

Greenland S (2002) Causality theory for policy uses. In: Murray C (ed) Summary measures of population, Harvard University Press, pp 291–302

Greenwell BM (2017) pdp: an r package for constructing partial dependence plots. R Journal 9(1):421

Grobon S, Mourlot L (2014) Le genre dans la statistique publique en france. Regards croisés sur l'économie (2):73–79

Groupement des Assureurs Automobiles (2021) Plan statistique automobile, résultats généraux, voitures de tourisme. GAA

Guelman L, Guillén M (2014) A causal inference approach to measure price elasticity in automobile insurance. Expert Systems with Applications 41(2):387–396

Guelman L, Guillén M, Pérez-Marín AM (2012) Random forests for uplift modeling: an insurance customer retention case. In: International conference on modeling and simulation in engineering, economics and management, Springer, pp 123–133

Guelman L, Guillén M, Perez-Marin AM (2014) A survey of personalized treatment models for pricing strategies in insurance. Insurance: Mathematics and Economics 58:68–76

Guillen M (2006) Fraud in insurance. Encyclopedia of Actuarial Science

Guillen M, Ayuso M (2008) Fraud in insurance. Encyclopedia of Quantitative Risk Analysis and Assessment 2

Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. In: International conference on machine learning, PMLR, pp 1321–1330

Gupta S, Kamble V (2021) Individual fairness in hindsight. The Journal of Machine Learning Research 22(1):6386–6420

Guseva A, Rona-Tas A (2001) Uncertainty, risk, and trust: Russian and american credit card markets compared. American sociological review pp 623–646

Guven S, McPhail M (2013) Beyond the cost model: Understanding price elasticity and its applications. In: Casualty Actuarial Society E-Forum, Spring 2013, Citeseer

Haas D (2013) Merit, fit, and basic desert. Philosophical Explorations 16(2):226–239

Haberman S, Renshaw AE (1996) Generalized linear models and actuarial science. Journal of the Royal Statistical Society: Series D (The Statistician) 45(4):407–436

Hacking I (1990) The taming of chance. 17, Cambridge University Press

Hager WD, Zimpleman L (1982) The norris decision, its implications and application. Drake Law Review 32:913

Hale K (2021) A.i. bias caused 80% of black mortgage applicants to be denied. Forbes 09/2021

Halley E (1693) An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the city of breslaw; with an attempt to ascertain the price of annuities upon lives. Philosophical Transactions of the Royal Society 17:596–610

Halpern JY (2016) Actual causality. MiT Press

Hamilton DL, Gifford RK (1976) Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. Journal of Experimental Social Psychology 12(4):392–407

Hamilton JD (1994) Time series analysis. Princeton university press

Hand DJ (2020) Dark Data: Why What You Don't Know Matters. Princeton University Press

Hansotia BJ, Rukstales B (2002) Direct marketing for multichannel retailers: Issues, challenges and solutions. Journal of Database Marketing & Customer Strategy Management 9:259–266

Hanssens DM, Parsons LJ, Schultz RL (2003) Market response models: Econometric and time series analysis, vol 2. Springer Science & Business Media

Hara K, Sun J, Moore R, Jacobs DW, Froehlich J (2014) Tohme: detecting curb ramps in google street view using crowdsourcing, computer vision, and machine learning. Proceedings of the 27th annual ACM symposium on User interface software and technology

Harari YN (2018) 21 Lessons for the 21st Century. Random House

Harcourt BE (2008) Against prediction. University of Chicago Press

Harcourt BE (2011) Surveiller et punir à l'âge actuariel. Déviance et Société 35:163

Harcourt BE (2015a) Exposed: Desire and disobedience in the digital age. Harvard University Press

Harcourt BE (2015b) Risk as a proxy for race: The dangers of risk assessment. Federal Sentencing Reporter 27(4):237–243

Harden KP (2023) Genetic determinism, essentialism and reductionism: semantic clarity for contested science. Nature Reviews Genetics 24(3):197–204

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. Advances in neural information processing systems 29:3315–3323

Hardy GH, Littlewood JE, Pólya G, Pólya G, et al. (1952) Inequalities. Cambridge university press

Hargreaves DJ, Colman AM, Sluckin W (1983) The attractiveness of names. Human Relations 36(4):393–401

Harrington SE, Niehaus G (1998) Race, redlining, and automobile insurance prices. The Journal of Business 71(3):439–469

Harris M (1970) Referential ambiguity in the calculus of brazilian racial identity. Southwestern Journal of Anthropology 26(1):1–14

Hartigan JA (1975) Clustering algorithms. John Wiley & Sons

Harwell D, Mayes B, Walls M, Hashemi S (2018) The accent gap. The Washington Post July 19

Hastie T, Tibshirani R (1987) Generalized additive models: some applications. Journal of the American Statistical Association 82(398):371–386

Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity. Monographs on statistics and applied probability 143:143

Haugeland J (1989) Artificial intelligence: The very idea. MIT press

Havens HV (1979) Issues and needed improvements in state regulation of the insurance business. US General Acounting Office

Haykin S (1998) Neural networks: a comprehensive foundation. Prentice Hall PTR

He Y, Xiong Y, Tsai Y (2020) Machine learning based approaches to predict customer churn for an insurance company. In: 2020 Systems and Information Engineering Design Symposium (SIEDS), IEEE, pp 1–6

Heckert NA, Filliben JJ, Croarkin CM, Hembree B, Guthrie WF, Tobias P, Prinz J, et al. (2002) Handbook 151: SEMATECH e-handbook of statistical methods. NIST

Hedden B (2021) On statistical criteria of algorithmic fairness. Philosophy & Public Affairs 49(2):209–231

Hedges BA (1977) Gender discrimination in pension plans: Comment. The Journal of Risk and Insurance 44(1):141–144

Heen ML (2009) Ending jim crow life insurance rates. Northwestern Journal of Law & Social Policy 4:360

Heidari H, Krause A (2018) Preventing disparate treatment in sequential decision making. In: IJCAI, pp 2248–2254

Heimer CA (1985) Reactive Risk and Rational Action. University of California Press

Heller D (2015) High price of mandatory auto insurance in predominantly african american communities. Tech. rep., Consumer Federation of America

Hellinger E (1909) Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. Journal für die reine und angewandte Mathematik 1909(136):210–271

Hellman D (1998) Two types of discrimination: The familiar and the forgotten. California Law Review 86:315

Hellman D (2011) When is discrimination wrong? Harvard University Press

Helton JC, Davis F (2002) Illustration of sampling-based methods for uncertainty and sensitivity analysis. Risk analysis 22(3):591–622

Hendriks F (1853) Contributions to the history of insurance, and of the theory of life contingencies. Journal of the Institute of Actuaries 3(2):93–120

Henley A (2014) Abolishing the stigma of punishments served: Andrew henley argues that those who have been punished should be free from future discrimination. Criminal Justice Matters 97(1):22–23

Henriet D, Rochet JC (1987) Some reflections on insurance pricing. European Economic Review 31(4):863–885

Heras AJ, Pradier PC, Teira D (2020) What was fair in actuarial fairness? History of the Human Sciences 33(2):91–114

Hernán MA, Robins JM (2010) Causal inference

Hertz J, Krogh A, Palmer RG (1991) Introduction to the theory of neural computation. CRC Press

Hesselager O, Verrall R (2006) Reserving in non-life insurance. Encyclopedia of Actuarial Science

Higham NJ (2008) Functions of matrices: theory and computation. SIAM

Hilbe JM (2014) Modeling count data. Cambridge University Press

Hill K (2022) A dad took photos of his naked toddler for the doctor. google flagged him as a criminal. The New York Times August 25

Hill K, White J (2020) Designed to deceive: do these people look real to you? The New York Times 11(21)

Hillier R (2022) The legal challenges of insuring against a pandemic. In: Pandemics: Insurance and Social Protection, Springer, Cham, pp 267–286

Hiltzik M (2013) Yes, men should pay for pregnancy coverage and here's why. Los Angeles Times November 1st

Hirschfeld HO (1935) A connection between correlation and contingency. Mathematical Proceedings of the Cambridge Philosophical Society 31(4):520–524

Hitchcock C (1997) Probabilistic causation. Stanford Encyclopedia of Philosophy

Ho DE, Imai K, King G, Stuart EA (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political analysis 15(3):199–236

Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1):55–67

Hoffman FL (1896) Race traits and tendencies of the American Negro, vol 11. American Economic Association

Hoffman FL (1931) Cancer and smoking habits. Annals of surgery 93(1):50

Hoffman KM, Trawalter S, Axt JR, Oliver MN (2016) Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. Proceedings of the National Academy of Sciences 113(16):4296–4301

Hofmann HJ (1990) Die anwendung des cart-verfahrens zur statistischen bonitätsanalyse von konsumentenkrediten. Zeitschrift fur Betriebswirtschaft 60:941–962

Hofstede G (1995) Insurance as a product of national values. The Geneva Papers on Risk and Insurance-Issues and Practice 20(4):423–429

Holland PW (1986) Statistics and causal inference. Journal of the American statistical Association 81(396):945–960

Holland PW (2003) Causation and race. ETS Research Report Series RR-03-03

Holzer H, Neumark D (2000) Assessing affirmative action. Journal of Economic literature 38(3):483–568

Homans S, Phillips GW (1868) Tontine dividend life assurance policies. Equitable Life Assurance Society of the United States

Hong D, Zheng YY, Xin Y, Sun L, Yang H, Lin MY, Liu C, Li BN, Zhang ZW, Zhuang J, et al. (2021) Genetic syndromes screening by facial recognition technology: Vgg-16 screening model construction and evaluation. Orphanet Journal of Rare Diseases 16(1):1–8

Hooker S, Moorosi N, Clark G, Bengio S, Denton E (2020) Characterising bias in compressed models 2010.03058

Hornik K (1991) Approximation capabilities of multilayer feedforward networks. Neural networks 4(2):251–257

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural networks 2(5):359–366

Horowitz MAC (1976) Aristotle and woman. Journal of the History of Biology 9:183–213

Hosmer DW, Lemesbow S (1980) Goodness of fit tests for the multiple logistic regression model. Communications in statistics-Theory and Methods 9(10):1043–1069

Houston R (1992) Mortality in early modern scotland: the life expectancy of advocates. Continuity and Change 7(1):47–69

Hsee CK, Li X (2022) A framing effect in the judgment of discrimination. Proceedings of the National Academy of Sciences 119(47):e2205988119

Hu F (2022) Semi-supervised learning in insurance: fairness and active learning. PhD thesis, Institut polytechnique de Paris

Hubbard GN (1852) De l'organisation des sociétés de bienfaisance ou de secours mutuels et des bases scientifiques sur lesquelles elles doivent être établies. Paris, Guillaumin

Huber PJ (1964) Robust estimation of a location parameter. The Annals of Mathematical Statistics 35:73–101

Hume D (1739) A Treatise of Human Nature. Cambridge University Press

Hume D (1748) An Enquiry concerning Human Understanding. Cambridge University Press

Hunt E (2016) Tay, microsoft's ai chatbot, gets a crash course in racism from twitter. The Guardian 24(3):2016

Hunter J (1775) Inaugural disputation on the varieties of man. In: Blumenbach JF (ed) De generis humani varietate nativa

Huttegger SM (2013) In defense of reflection. Philosophy of Science 80(3):413–433

Huttegger SM (2017) The probabilistic foundations of rational learning. Cambridge University Press

Ichiishi T (2014) Game theory for economic analysis. Academic Press

Ilic L, Sawada M, Zarzelli A (2019) Deep mapping gentrification in a large canadian city using deep learning and google street view. PloS one 14(3):e0212814

Imai K (2018) Quantitative social science: an introduction. Princeton University Press

Imai K, Ratkovic M (2014) Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B: Statistical Methodology pp 243–263

Imbens GW, Rubin DB (2015) Causal inference in statistics, social, and biomedical sciences. Cambridge University Press

Ingold D, Soper S (2016) Amazon doesn't consider the race of its customers. should it? Bloomberg April 21st

Institute and Faculty of Actuaries (2021) The hidden risks of being poor: the poverty premium in insurance. Faculty of Actuaries Report

Insurance Bureau of Canada (2021) Facts of the property and casualty insurance industry in canada. Insurance Bureau of Canada

Ismay P (2018) Trust among strangers: friendly societies in modern Britain. Cambridge University Press

Iten R, Wagner J, Zeier Röschmann A (2021) On the identification, evaluation and treatment of risks in smart homes: A systematic literature review. Risks 9(6):113

Ito J (2021) Supposedly 'fair' algorithms can perpetuate discrimination. Wired 02.05.2019

Jaccard P (1901) Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. Bulletin de la Société Vaudoise de Sciences Naturelles 37:241–272

Jackson JP, Depew DJ (2017) Darwinism, democracy, and race: American anthropology and evolutionary biology in the twentieth century. Taylor & Francis

Jackson MO (2019) The human network: How your social position determines your power, beliefs, and behaviors. Vintage

Jacobs J (1894) Aesop's Fables: Selected and Told Anew. Capricorn Press

Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall

Jann B (2008) The blinder–oaxaca decomposition for linear regression models. The Stata Journal 8(4):453–479

Jargowsky PA (2005) Omitted variable bias. Encyclopedia of social measurement 2:919–924

Jarvis B, Pearlman RF, Walsh SM, Schantz DA, Gertz S, Hale-Pletka AM (2019) Insurance rate optimization through driver behavior monitoring. Google Patents 10,169,822

Jean N, Burke M, Xie M, Davis WM, Lobell DB, Ermon S (2016) Combining satellite imagery and machine learning to predict poverty. Science 353(6301):790–794

Jeffreys H (1946) An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London Series A Mathematical and Physical Sciences 186(1007):453–461

Jerry RH (2023) Understanding parametric insurance: A potential tool to help manage pandemic risk. In: Covid-19 and Insurance, Springer, pp 17–62

Jewell WS (1974) Credible means are exact bayesian for exponential families. ASTIN Bulletin: The Journal of the IAA 8(1):77–90

Jiang H, Nachum O (2020) Identifying and correcting label bias in machine learning. In: International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, pp 702–712

Jiang J, Nguyen T (2007) Linear and generalized linear mixed models and their applications, vol 1. Springer

Jo HH, Eom YH (2014) Generalized friendship paradox in networks with tunable degree-attribute correlation. Physical Review E 90(2):022809

Johannesson GT (2013) The history of Iceland. ABC-CLIO

Johnston L (1945) Effects of tobacco smoking on health. British Medical Journal 2(4411):98

Jolliffe IT (2002) Principal component analysis,. Springer

Jones EE, Nisbett RE (1971) The actor and the observer: Divergent perceptions of the causes of behavior. New York: General Learning Press.

Jones ML (2016) Ctrl + Z: The Right to Be Forgotten. New York University Press

Jordan A, Krüger F, Lerch S (2019) Evaluating probabilistic forecasts with `scoringRules`. Journal of Statistical Software 90:1–37

Jordan C (1881) Sur la serie de fourier. Camptes Rendus Hebdomadaires de l'Academie des Sciences 92:228–230

Joseph S, Castan M (2013) The international covenant on civil and political rights: cases, materials, and commentary. Oxford University Press

Jost JT, Rudman LA, Blair IV, Carney DR, Dasgupta N, Glaser J, Hardin CD (2009) The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. Research in organizational behavior 29:39–69

Jung C, Kearns M, Neel S, Roth A, Stapleton L, Wu ZS (2019a) An algorithmic framework for fairness elicitation. arXiv preprint arXiv:190510660

Jung C, Kearns M, Neel S, Roth A, Stapleton L, Wu ZS (2019b) Eliciting and enforcing subjective individual fairness. arXiv preprint arXiv:190510660 1

Jung C, Kannan S, Lee C, Pai MM, Roth A, Vohra R (2020) Fair prediction with endogenous behavior. arXiv 2002.07147

Kachur A, Osin E, Davydov D, Shutilov K, Novokshonov A (2020) Assessing the big five personality traits using real-life static facial images. Scientific reports 10(1):1–11

Kaganoff BC (1996) A dictionary of Jewish names and their history. Jason Aronson

Kahlenberg Richard D (1996) The remedy. class, race and affirmative action. New York: Basic

Kahneman D (2011) Thinking, Fast and Slow. Farrar, Straus and Giroux

Kamalich RF, Polachek SW (1982) Discrimination: Fact or fiction? an examination using an alternative approach. Southern Economic Journal pp 450–461

Kamen H (2014) The Spanish Inquisition: a historical revision. Yale University Press

Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems 33(1):1–33

Kang JD, Schafer JL (2007) Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science 22(4):523–539

Kanngiesser P, Warneken F (2012) Young children consider merit when sharing resources with others. PLOS ONE 8(8)

Kant I (1775) Über die verschiedenen Rassen der Menschen. Nicolovius Edition

Kant I (1785) Bestimmung des Begriffs einer Menschenrace. Haude und Spener, Berlin

Kant I (1795) Zum ewigen Frieden. Ein philosophischer Entwurf). Nicolovius Edition

Kantorovich LV, Rubinshtein S (1958) On a space of totally additive functions. Vestnik of the St Petersburg University: Mathematics 13(7):52–59

Karapiperis D, Birnbaum B, Brandenburg A, Castagna S, Greenberg A, Harbage R, Obersteadt A (2015) Usage-based insurance and vehicle telematics: insurance market and regulatory implications. CIPR Study Series 1:1–79

Karimi H, Khan MFA, Liu H, Derr T, Liu H (2022) Enhancing individual fairness through propensity score matching. In: 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, pp 1–10

Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2020) Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8110–8119

Kearns M, Roth A (2019) The ethical algorithm: The science of socially aware algorithm design. Oxford University Press

Kearns M, Valiant L (1989) Cryptographic limitations on learning boolean formulae and finite automata. Journal of the ACM 21(1):433–444.

Kearns M, Neel S, Roth A, Wu ZS (2018) Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 1711.05144, pp 2564–2572

Keffer R (1929) An experience rating formula. Transactions of the Actuarial Society of America 30:130–139

Keita SOY, Kittles RA, Royal CD, Bonney GE, Furbert-Harris P, Dunston GM, Rotimi CN (2004) Conceptualizing human variation. Nature genetics 36(Suppl 11):S17–S20

Kekes J (1995) The injustice of affirmative action involving preferential treatment. In: Cahn S (ed) The Affirmative Action Debate,, Routledge, pp 293–304

Kelly H (2021) A priest's phone location data outed his private life. it could happen to anyone. The Wshinghton Post 22-07-2021

Kelly M, Nielson N (2006) Age as a variable in insurance pricing and risk classification. The Geneva Papers on Risk and Insurance-Issues and Practice 31(2):212–232

Keyfitz K, Flieger W, et al. (1968) World population: an analysis of vital data. The University of Chicago Press

Kiiveri H, Speed T (1982) Structural analysis of multivariate data: A review. Sociological methodology 13:209–289

Kilbertus N, Rojas-Carulla M, Parascandolo G, Hardt M, Janzing D, Schölkopf B (2017) Avoiding discrimination through causal reasoning. arXiv 1706.02744

Kim MP, Reingold O, Rothblum GN (2018) Fairness through computationally-bounded awareness. arXiv 1803.03239

Kim PT (2017) Auditing algorithms for discrimination. University of Pennsylvania Law Review 166:189

Kimball SL (1979) Reverse sex discrimination: Manhart. American Bar Foundation Research Journal 4(1):83–139

King G, Tanner MA, Rosen O (2004) Ecological inference: New methodological strategies. Cambridge University Press

Kirkpatrick K (2017) It's not the algorithm, it's the data. Communications of the ACM 60(2):21–23

Kita K, Kidziński Ł (2019) Google street view image of a house predicts car accident risk of its resident. arXiv 1904.05270

Kitagawa EM (1955) Components of a difference between two rates. Journal of the american statistical association 50(272):1168–1194

Kitagawa EM, Hauser PM (1973) Differential mortality in the united states. In: Differential Mortality in the United States, Harvard University Press

Kitchin R (2017) Thinking critically about and researching algorithms. Information, communication & society 20(1):14–29

Kiviat B (2019) The moral limits of predictive practices: The case of credit-based insurance scores. American Sociological Review 84(6):1134–1158

Kiviat B (2021) Which data fairly differentiate? american views on the use of personal data in two market settings. Sociological Science 8:26–47

Klein R (2021) Matching rate to risk: Analysis of the availability and affordability of private passenger automobile insurance. Tech. rep., Insurance Information Institute

Klein R, Grace MF (2001) Urban homeowners insurance markets in texas: A search for redlining. Journal of Risk and Insurance pp 581–613

Kleinberg J, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. arXiv 1609.05807

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2017) Human Decisions and Machine Predictions. The Quarterly Journal of Economics 133(1):237–293

Klinker F (2010) Generalized linear mixed models for ratemaking: a means of introducing credibility into a generalized linear model setting. In: Casualty Actuarial Society E-Forum, Winter 2011 Volume 2

Klugman SA (1991) Bayesian statistics in actuarial science: with emphasis on credibility, vol 15. Springer Science & Business Media

Knowlton RE (1978) Regents of the university of california v. bakke. Arkansas Law Review 32:499

Koetter F, Blohm M, Drawehn J, Kochanowski M, Goetzer J, Graziotin D, Wagner S (2019) Conversational agents for insurance companies: from theory to practice. In: International Conference on Agents and Artificial Intelligence, Springer, pp 338–362

Kohler-Hausmann I (2018) Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. Northwestern University Law Review 113:1163

Kohlleppel L (1983) Multidimensional market signalling. Institut für Gesellschafts und Wirtschaftswissenschaften., Wirtschaftstheoretische Abteilung

Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT press

Kolmogorov A (1933) Grundbegriffe der wahrscheinlichkeitsrechnung

Komiyama J, Takeda A, Honda J, Shimao H (2018) Nonconvex optimization for regression with fairness constraints. In: International conference on machine learning, Proceedings of Machine Learning Research, pp 2737–2746

Korzybski A (1958) Science and sanity: An introduction to non-Aristotelian systems and general semantics. Institute of GS

Kosinski M (2021) Facial recognition technology can expose political orientation from naturalistic facial images. Scientific reports 11(1):1–7

Kotchen TA (2011) Historical trends and milestones in hypertension research: a model of the process of translational research. Hypertension 58(4):522–538

Kotter-Grühn D, Kornadt AE, Stephan Y (2016) Looking beyond chronological age: Current knowledge and future directions in the study of subjective age. Gerontology 62(1):86–93

Kranzberg M (1986) Technology and history:" kranzberg's laws". Technology and culture 27(3):544–560

Kranzberg M (1995) Technology and history:" kranzberg's laws". Bulletin of science, technology & society 15(1):5–13

Krasanakis E, Spyromitros-Xioufis E, Papadopoulos S, Kompatsiaris Y (2018) Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In: Proceedings of the 2018 world wide web conference, pp 853–862

Kremer E (1982) Ibnr-claims and the two-way model of anova. Scandinavian Actuarial Journal 1982(1):47–55

Krikler S, Dolberger D, Eckel J (2004) Method and tools for insurance price and revenue optimisation. Journal of Financial Services Marketing 9(1):68–79

Krippner GR (2023) Unmasked: A history of the individualization of risk. Sociological Theory p 07352751231169012

Kroll JA, Huey J, Barocas S, Felten EW, Reidenberg JR, Robinson DG, Yu H (2017) Accountable algorithms. University of Pennsylvania Law Review 165:633–705

Krüger F, Ziegel JF (2021) Generic conditions for forecast dominance. Journal of Business & Economic Statistics 39(4):972–983

Krzanowski WJ, Hand DJ (2009) ROC curves for continuous data. Crc Press

Kudryavtsev AA (2009) Using quantile regression for rate-making. Insurance: Mathematics and Economics 45(2):296–304

Kuhn M, Johnson K, et al. (2013) Applied predictive modeling, vol 26. Springer

Kuhn T (2020) Root insurance commits to eliminate bias from its car insurance rates. Business Wire August 6

Kull M, Flach P (2015) Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15, Springer, pp 68–85

Kullback S (2004) Minimum discrimination information (mdi) estimation. Encyclopedia of Statistical Sciences 7

Kullback S, Leibler RA (1951) On information and sufficiency. The Annals of Mathematical Statistics 22(1):79–86

Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. Advances in neural information processing systems 30

de La Fontaine J (1668) Fables. Barbin

Laffont JJ, Martimort D (2002) The theory of incentives: the principal-agent model. Princeton University Press

Lahoti P, Gummadi KP, Weikum G (2019) Operationalizing individual fairness with pairwise fair representations. arXiv preprint arXiv:190701439

Lambert D (1992) Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics 34(1):1–14

Lamont M, Molnár V (2002) The study of boundaries in the social sciences. Annual review of sociology pp 167–195

Lancaster R, Ward R (2002) The contribution of individual factors to driving behaviour: Implications for managing work-related road safety. HM Stationery Office

Landes X (2015) How fair is actuarial fairness? Journal of Business Ethics 128(3):519–533

Langford J, Schapire R (2005) Tutorial on practical prediction theory for classification. Journal of machine learning research 6(3)

LaPar DJ, Bhamidipati CM, Mery CM, Stukenborg GJ, Jones DR, Schirmer BD, Kron IL, Ailawadi G (2010) Primary payer status affects mortality for major surgical operations. Annals of surgery 252(3):544

Laplace PS (1774) Mémoire sur la probabilité de causes par les évenements. Mémoire de l'académie royale des sciences

de Lara L (2023) Counterfactual models for fair and explainable machine learning: A mass transportation approach. PhD thesis, Institut de Mathématiques de Toulouse

de Lara L, González-Sanz A, Asher N, Loubes JM (2021) Transport-based counterfactual models. arXiv 2108.13025

Larson J, Mattu S, Kirchner L, Angwin J (2016) How we analyzed the compas recidivism algorithm. ProPublica 23-05

Larson J, Angwin J, Kirchner L, Mattu S (2017) How we examined racial discrimination in auto insurance prices. ProPublica, April 5

Lasry JM (2015) La rencontre choc de l'assurance et du big data. Risques 103:19–24

Lauer J (2017) Creditworthy: A History of Consumer Surveillance and Financial Identity in America. Columbia University Press

Laulom S (2012) Égalité des sexes et primes d'assurances. Semaine sociale Lamy (1531):44–49

Laurent H, Rivest RL (1976) Constructing optimal binary decision trees is np-complete. Information processing letters 5(1):15–17

Law S, Paige B, Russell C (2019) Take a look around: Using street view and satellite images to estimate house prices. ACM Transactions on Intelligent Systems and Technology 10(5)

Le Gouic T, Loubes JM (2017) Existence and consistency of wasserstein barycenters. Probability Theory and Related Fields 168:901–917

Le Monde (2021) La discrimination par l'accent bientôt réprimée ? une proposition de loi adoptée jeudi à l'assemblée. Le Monde 26-11-2020

Leben D (2020) Normative principles for evaluating fairness in machine learning. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp 86–92

Lebesgue H (1918) Remarques sur les théories de la mesure et de l'intégration. Annales scientifiques de l'École Normale Supérieure 35:191–250

Ledford H (2019) Millions affected by racial bias in health-care algorithm. Nature 574(31):2

Lee BK, Lessler J, Stuart EA (2010) Improving propensity score weighting using machine learning. Statistics in medicine 29(3):337–346

Lee RD, Carter LR (1992) Modeling and forecasting us mortality. Journal of the American statistical association 87(419):659–671

Lee S, Antonio K (2015) Why high dimensional modeling in actuarial science. ASTIN, AFIR/ERM & IACA Colloquia

Leeson PT (2009) The calculus of piratical consent: the myth of the myth of social contract. Public Choice 139:443–459

Lemaire J (1985) Automobile insurance: actuarial models, vol 4. Springer Science & Business Media

Lemaire J, Park SC, Wang KC (2016) The use of annual mileage as a rating variable. ASTIN Bulletin: The Journal of the IAA 46(1):39–69

Léon PR (1993) Précis de phonostylistique: parole et expressivité/Pierre R. Léon,.. Nathan

Leshno M, Lin VY, Pinkus A, Schocken S (1993) Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural networks 6(6):861–867

Leu C (2015) Looking up. The University of Chicago Magazine May-June(15)

Leuner J (2019) A replication study: Machine learning models are capable of predicting sexual orientation from facial images. arXiv 1902.10739

Levantesi S, Pizzorusso V (2019) Application of machine learning to mortality modeling and forecasting. Risks 7(1):26

Levina E, Bickel P (2001) The earth mover's distance is the mallows distance: Some insights from statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, IEEE, vol 2, pp 251–256

Levy J (2012) Freaks of fortune. Harvard University Press

Lew EA, Garfinkel L (1979) Variations in mortality by weight among 750,000 men and women. Journal of chronic diseases 32(8):563–576

Lewis AE (2004) What group?" studying whites and whiteness in the era of "color-blindness. Sociological theory 22(4):623–646

Lewis D (1973) Counterfactuals. John Wiley & Sons

Li C, Fan X (2020) On nonparametric conditional independence tests for continuous variables. Wiley Interdisciplinary Reviews: Computational Statistics 12(3):e1489

Li F, Li F (2019) Propensity score weighting for causal inference with multiple treatments. The Annals of Applied Statistics 13:2389–2415

Li G, Braver ER, Chen LH (2003) Fragility versus excessive crash involvement as determinants of high death rates per vehicle-mile of travel among older drivers. Accident Analysis & Prevention 35(2):227–235

Li P, Liu H (2022) Achieving fairness at no utility cost via data reweighing with influence. In: International Conference on Machine Learning, Proceedings of Machine Learning Research, pp 12917–12930

Liebler CA, Porter SR, Fernandez LE, Noon JM, Ennis SR (2017) America's churning races: Race and ethnicity response changes between census 2000 and the 2010 census. Demography 54(1):259–284

Light JS (1999) When computers were women. Technology and culture 40(3):455–483

Liisa HB (1994) Aging and fatal accidents in male and female drivers. Journal of Gerontology 49(6):S286–S290

Lima M (2014) The book of trees: Visualizing branches of knowledge, vol 5. Princeton Architectural Press New York

Lin J (1991) Divergence measures based on the shannon entropy. IEEE Transactions on Information theory 37(1):145–151

Lindholm M, Richman R, Tsanakas A, Wüthrich MV (2022a) Discrimination-free insurance pricing. ASTIN Bulletin: The Journal of the IAA 52(1):55–89

Lindholm M, Richman R, Tsanakas A, Wüthrich MV (2022b) A discussion of discrimination and fairness in insurance pricing. arXiv 2209.00858

Ling CX, Li C (1998) Data mining for direct marketing: Problems and solutions. In: Conference on Knowledge Discovery and Data Mining, vol 98, pp 73–79

Lipovetsky S, Conklin M (2001) Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry 17(4):319–330

Lippert-Rasmussen K (2006) The badness of discrimination. Ethical Theory and Moral Practice 9:167–185

Lippert-Rasmussen K (2013) Discrimination. Wiley-Blackwell

Lippert-Rasmussen K (2014) Born free and equal?: A philosophical inquiry into the nature of discrimination. Oxford University Press

Lippert-Rasmussen K (2020) Making sense of affirmative action. Oxford University Press

Lippmann W (1922) Public opinion. Routledge

Lipton ZC, Chouldechova A, McAuley J (2018) Does mitigating ml's impact disparity require treatment disparity? In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp 8136–8146

Liu X (2015) No fats, femmes, or asians. Moral Philosophy and Politics 2(2):255–276

Liu X (2017) Discrimination and lookism. In: Lippert-Rasmussen K (ed) Handbook of the Ethics of Discrimination, Routledge, pp 276–286

Liu Y, Chen L, Yuan Y, Chen J (2012) A study of surnames in china through isonymy. American Journal of Physical Anthropology 148(3):341–350

Lo VS (2002) The true lift model: a novel data mining approach to response modeling in database marketing. ACM SIGKDD Explorations Newsletter 4(2):78–86

Loève M (1977) Probability Theory. Springer

Löffler M, Münstermann B, Schumacher T, Mokwa C, Behm S (2016) Insurers need to plug into the internet of things–or risk falling behind. European Insurance

Loftus JR, Russell C, Kusner MJ, Silva R (2018) Causal reasoning for algorithmic fairness. arXiv 1805.05859

Loi M, Christen M (2021) Choosing how to discriminate: navigating ethical trade-offs in fair algorithmic design for the insurance sector. Philosophy & Technology pp 1–26

Lombroso C (1876) L'uomo delinquente. Hoepli

Loos RJ, Yeo GS (2022) The genetics of obesity: from discovery to biology

L'Oréal (2022) A new geography of skin color. https://wwwlorealcom/en/articles/science-and-technology/expert-inskin/

Lorenz MO (1905) Methods of measuring the concentration of wealth. Publications of the American statistical association 9(70):209–219

Lovejoy B (2021) Linkedin breach reportedly exposes data of 92% of users, including inferred salaries. 9to5mac 06/29

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Proceedings of the 31st international conference on neural information processing systems, Curran Associates, Inc., vol 30, pp 4768–4777

Luong BT, Ruggieri S, Turini F (2011) k-nn as an implementation of situation testing for discrimination discovery and prevention. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 502–510

Lutton L, Fan A, Loury A (2020) Where banks don't lend. WBEZ

MacIntyre AC (1969) Hume on 'is' and 'ought'. In: The is-ought question, Springer, pp 35–50

MacKay DJ (1992) A practical bayesian framework for backpropagation networks. Neural computation 4(3):448–472

Macnicol J (2006) Age discrimination: An historical and contemporary analysis. Cambridge University Press

Maedche A (2020) Gender bias in chatbot design. Chatbot Research and Design p 79

Mallasto A, Feragen A (2017) Learning from uncertain curves: The 2-wasserstein metric for gaussian processes. Advances in Neural Information Processing Systems 30

Mallon R (2006) 'race': normative, not metaphysical or semantic. Ethics 116(3):525–551

Mallows CL (1972) A note on asymptotic joint normality. The Annals of Mathematical Statistics pp 508–515

Mangel M, Samaniego FJ (1984) Abraham wald's work on aircraft survivability. Journal of the American Statistical Association 79(386):259–267

Mantelero A (2013) The eu proposal for a general data protection regulation and the roots of the 'right to be forgotten'. Computer Law & Security Review 29(3):229–235

Marshall A (1890) General relations of demand, supply, and value. Principles of economics: unabridged eighth edition

Marshall A (2021) Ai comes to car repair, and body shop owners aren't happy. Wired April 13

Marshall GA (1993) Racial classifications: popular and scientific. In: Harding S (ed) The "racial" economy of science: Toward a democratic future, Indiana University Press, pp 116–125

Martin GD (1977) Gender discrimination in pension plans: Author's reply. The Journal of Risk and Insurance 44(1):145–149

Mas L (2020) A confederate flag spotted in the window of police barracks in paris. France 24 10/07

Massey DS (2007) Categorically unequal: The American stratification system. Russell Sage Foundation

Matthews BW (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure 405(2):442–451

Mayer J, Mutchler P, Mitchell JC (2016) Evaluating the privacy properties of telephone metadata. Proceedings of the National Academy of Sciences 113(20):5536–5541

Maynard A (1979) Pricing, insurance and the national health service. Journal of Social Policy 8(2):157–176

Mayr E (1982) The growth of biological thought: Diversity, evolution, and inheritance. Harvard University Press

Mazieres A, Roth C (2018) Large-scale diversity estimation through surname origin inference. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique 139(1):59–73

Mbungo R (2014) L'approche juridique internationale du phénomène de discrimination fondée sur le motif des antécédents judiciaires. Revue québécoise de droit international 27(2):59–97

McCaffrey DF, Ridgeway G, Morral AR (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological methods 9(4):403

McClenahan CL (2006) Ratemaking. Casualty Actuarial Society

McCullagh P, Nelder J (1989) Generalized linear models. Chapman & Hall

McCulloch CE, Searle SR (2004) Generalized, linear, and mixed models. John Wiley & Sons

McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5:115–133

McDonnell M, Baxter D (2019) Chatbots and gender stereotyping. Interacting with Computers 31(2):116–121

McFall L (2019) Personalizing solidarity? the role of self-tracking in health insurance pricing. Economy and society 48(1):52–76

McFall L, Meyers G, Hoyweghen IV (2020) Editorial: The personalisation of insurance: Data, behaviour and innovation. Big Data & Society 7(2)

McKinley R (2014) A history of British surnames. Routledge

McKinsey (2017) Technology, jobs and the future of work. McKinsey Global Institute

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. Annual review of sociology 27(1):415–444

Meilijson I (2006) Risk aversion. Encyclopedia of Actuarial Science

Meinshausen N, Ridgeway G (2006) Quantile regression forests. Journal of machine learning research 7(6)

de Melo-Martín I (2003) When is biology destiny? biological determinism and social responsibility. Philosophy of science 70(5):1184–1194

Memmi A (2000) Racism. University of Minnesota Press

Menezes CF, Hanson DL (1970) On the theory of risk aversion. International Economic Review pp 481–487

Mercat-Bruns M (2016) Discrimination at Work. University of California Press

Mercat-Bruns M (2020) Les rapports entre vieillissement et discrimination en droit: une fertilisation croisée utile sur le plan individuel et collectif. La Revue des Droits de l'Homme 17

Merriam-Webster (2022) Dictionary

Merrill D (2012) New credit scores in a new world: Serving the underbanked. TEDxNewWallStreet

Messenger R, Mandell L (1972) A modal search technique for predictive nominal scale multivariate analysis. Journal of the American statistical association 67(340):768–772

Meuleners LB, Harding A, Lee AH, Legge M (2006) Fragility and crash over-representation among older drivers in western australia. Accident Analysis & Prevention 38(5):1006–1010

Meyers G, Van Hoyweghen I (2018) Enacting actuarial fairness in insurance: From fair discrimination to behaviour-based fairness. Science as Culture 27(4):413–438

Michelbacher G (1926) 'moral hazard' in the field of casualty insurance. Proceedings of the Casualty Actuarial Society 13(27)

Michelson S, Blattenberger G (1984) Reverse regression and employment discrimination. Journal of Business & Economic Statistics 2(2):121–122

Milanković M (1920) Théorie mathématique des phénomènes thermiques produits par la radiation solaire. Gauthier-Villars

Miller G, Gerstein DR (1983) The life expectancy of nonsmoking men and women. Public Health Reports 98(4):343

Miller H (2015a) A discussion on credibility and penalised regression, with implications for actuarial work. Actuaries Institute

Miller MJ, Smith RA, Southwood KN (2003) The relationship of credit-based insurance scores to private passenger automobile insurance loss propensity. Actuarial Study, Epic Actuaries

Miller T (2015b) Price optimization. Commonwealth of Pennsylvania, Insurance Department August 25

Mills CW (2017) Black rights/white wrongs: The critique of racial liberalism. Oxford University Press

Milmo D (2021) Working of algorithms used in government decision-making to be revealed. The Guardian November 29

Milne J (1815) A Treatise on the Valuation of Annuities and Assurances on Lives and Survivorships: On the Construction of Tables of Mortality and on the Probabilities and Expectations of Life, vol 2. Longman, Hurst, Rees, Orme, and Brown

Minsky M, Papert S (1969) An introduction to computational geometry. Cambridge tiass, HIT 479:480

Minty D (2016) Price optimisation for insurance optimising price; destroying value. Thinkpiece Chartered Insurance Institute

Miracle JM (2016) De-anonymization attack anatomy and analysis of ohio nursing workforce data anonymization. PhD thesis, Wright State University

von Mises R (1928) Wahrscheinlichkeit Statistik und Wahrheit. Springer

von Mises R (1939) Probability, statistics and truth. Macmillan

Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: Mapping the debate. Big Data & Society 3(2):2053951716679679

Mittra J (2007) Predictive genetic information and access to life assurance: The poverty of 'genetic exceptionalism'. BioSocieties 2(3):349–373

Mollat M (2006) Les pauvres au moyen-âge, vol 11. Éditions Complexe

Molnar C (2023) A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book

Molnar C, Casalicchio G, Bischl B (2018) iml: An r package for interpretable machine learning. Journal of Open Source Software 3(26):786

Monnet J (2017) Discrimination et assurance. Journal de Droit de la Santé et de l'Assurance Maladie 16:13–19

Moodie EE, Stephens DA (2022) Causal inference: Critical developments, past and future. Canadian Journal of Statistics 50(4):1299–1320

Moon R (2014) From gorgeous to grumpy: adjectives, age and gender. Gender & Language 8(1)

Moor L, Lury C (2018) Price and the person: Markets, discrimination, and personhood. Journal of Cultural Economy 11(6):501–513

Morel P, Stalk G, Stanger P, Wetenhall P (2003) Pricing myopia. The Boston Consulting Group Perspectives

Morris DS, Schwarcz D, Teitelbaum JC (2017) Do credit-based insurance scores proxy for income in predicting auto claim risk? Journal of Empirical Legal Studies 14(2):397–423

Morrison EJ (1996) Insurance discrimination against battered women: Proposed legislative protections. Ind LJ 72:259

Moulin H (1992) An application of the shapley value to fair division with money. Econometrica pp 1331–1349

Moulin H (2004) Fair division and collective welfare. MIT press

Mowbray A (1921) Classification of risks as the basis of insurance rate making with special reference to workmen's compensation. Proceedings of the Casualty Actuarial Society

Müller R, Kornblith S, Hinton GE (2019) When does label smoothing help? Advances in neural information processing systems 32

Mundubeltz-Gendron S (2019) Comment l'intelligence artificielle va bouleverser le monde du travail dans l'assurance. L'Argus de l'Assurance 10/04

Murphy AH (1973) A new vector partition of the probability score. Journal of Applied Meteorology and Climatology 12(4):595–600

Murphy AH (1996) General decompositions of mse-based skill scores: Measures of some basic aspects of forecast quality. Monthly Weather Review 124(10):2353–2369

Murphy AH, Winkler RL (1987) A general framework for forecast verification. Monthly weather review 115(7):1330–1338

Must A, Spadano J, Coakley EH, Field AE, Colditz G, Dietz WH (1999) The disease burden associated with overweight and obesity. Journal of the American Medical Association 282(16):1523–1529

Myers RJ (1977) Gender discrimination in pension plans: Further comment. The Journal of Risk and Insurance 44(1):144–145

Nadaraya EA (1964) On estimating regression. Theory of Probability & Its Applications 9(1):141–142

Nakashima R (2018) Google tracks your movements, like it or not. Associated Press August 14

Nassif H, Kuusisto F, Burnside ES, Page D, Shavlik J, Santos Costa V (2013) Score as you lift (sayl): A statistical relational learning approach to uplift modeling. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13, Springer, pp 595–611

Nathan EB (1925) Analysed mortality: the english life no. 8a tables. Transactions of the Faculty of Actuaries 10:45–124

National Association of Insurance Commissioners (2011) A consumer's guide to auto insurance. NAIC Reports

National Association of Insurance Commissioners (2022) A consumer's guide to auto insurance. NAIC Reports

Natowicz MR, Alper JK, Alper JS (1992) Genetic discrimination and the law. American Journal of Human Genetics 50(3):465

Neal RM (1992) Bayesian training of backpropagation networks by the hybrid monte carlo method. Tech. rep., Citeseer

Neal RM (2012) Bayesian learning for neural networks, vol 118. Springer Science & Business Media

Nelson A (2002) Unequal treatment: confronting racial and ethnic disparities in health care. Journal of the national medical association 94(8):666

Neyman J, Dabrowska DM, Speed T (1923) On the application of probability theory to agricultural experiments. essay on principles. section 9. Statistical Science pp 465–472

Niculescu-Mizil A, Caruana R (2005a) Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning, pp 625–632

Niculescu-Mizil A, Caruana RA (2005b) Obtaining calibrated probabilities from boosting. In: Proc. 21st Conference on Uncertainty in Artificial Intelligence (UAI'05), AUAI Press, pp 413–420

Nielsen F (2013) Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms. IEEE Signal Processing Letters 20(7):657–660

Nielsen F, Boltz S (2011) The burbea-rao and bhattacharyya centroids. IEEE Transactions on Information Theory 57(8):5455–5466

Nielsen F, Nock R (2009) Sided and symmetrized bregman centroids. IEEE transactions on Information Theory 55(6):2882–2904

Noguéro D (2010) Sélection des risques. discrimination, assurance et protection des personnes vulnérables. Revue générale du droit des assurances 3:633–663

Nordholm LA (1980) Beautiful patients are good patients: evidence for the physical attractiveness stereotype in first impressions of patients. Social Science & Medicine Part A: Medical Psychology & Medical Sociology 14(1):81–83

Norman P (2003) Statistical discrimination and efficiency. The Review of Economic Studies 70(3):615–627

Nuruzzaman M, Hussain OK (2020) Intellibot: A dialogue-based chatbot for the insurance industry. Knowledge-Based Systems 196:105810

Oaxaca R (1973) Male-female wage differentials in urban labor markets. International Economic Review 14(3):693–709

Oaxaca RL, Ransom MR (1994) On discrimination and the decomposition of wage differentials. Journal of econometrics 61(1):5–21

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453

Ohlsson E, Johansson B (2010) Non-life insurance pricing with generalized linear models, vol 174. Springer

O'Neil C (2016) Weapons of math destruction: How big data increases inequality and threatens democracy. Crown

Ong PM, Stoll MA (2007) Redlining or risk? a spatial analysis of auto insurance rates in los angeles. Journal of Policy Analysis and Management 26(4):811–830

Opitz D, Maclin R (1999) Popular ensemble methods: An empirical study. Journal of artificial intelligence research 11:169–198

Ortiz-Ospina E, Beltekian D (2018) Why do women live longer than men? Our World in Data

Orwat C (2020) Risks of discrimination through the use of algorithms. Institute for Technology Assessment and Systems Analysis

Outreville JF (1990) The economic significance of insurance markets in developing countries. Journal of Risk and Insurance pp 487–498

Outreville JF (1996) Life insurance markets in developing countries. Journal of risk and insurance pp 263–278

Owsley C, McGwin Jr G (2010) Vision and driving. Vision research 50(23):2348–2361

Oza D, Padhiyar D, Doshi V, Patil S (2020) Insurance claim processing using rpa along with chatbot. In: Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST)

Pager D (2003) The mark of a criminal record. American journal of sociology 108(5):937–975

Pager D (2008) Marked: Race, crime, and finding work in an era of mass incarceration. University of Chicago Press

Palmore E (1978) Are the aged a minority group? Journal of the American Geriatrics Society 26(5):214–217

Pardo L (2018) Statistical inference based on divergence measures. CRC press

Parléani G (2012) Commentaire des lignes directrices de la commission européenne sur les suites de l'arrêt « test achats ». Revue générale du droit des assurances 3:563

Parry M (2016) Linear scoring rules for probabilistic binary classification. Electronic Journal of Statistics 10:1596—1607

Pasquale F (2015) The black box society: the secret algorithms that control money and information. Harvard University Press

Paugam S, Cousin B, Giorgetti C, Naudet J (2017) Ce que les riches pensent des pauvres. Seuil

Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan kaufmann

Pearl J (1998) Graphs, causality, and structural equation models. Sociological Methods & Research 27(2):226–284

Pearl J (2009a) Causal inference in statistics: An overview. Statistics surveys 3:96–146

Pearl J (2009b) Causality. Cambridge university press

Pearl J (2010) An introduction to causal inference. The International Journal of Biostatistics 6(2)

Pearl J, Mackenzie D (2018) The book of why: the new science of cause and effect. Basic books

Pedreshi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Association for Computing Machinery, KDD '08, pp 560–568

Perla F, Richman R, Scognamiglio S, Wüthrich MV (2021) Time-series forecasting of mortality rates using deep learning. Scandinavian Actuarial Journal 2021(7):572–598

Petauton P (1998) L'opération d'assurance : définitions et principes. In: Ewald F, Lorenzi JH (eds) Encyclopédie de l'assurance, Economica

Peters J, Janzing D, Schölkopf B (2017) Elements of causal inference: foundations and learning algorithms. The MIT Press

Peters T (2014) Playing God?: Genetic determinism and human freedon. Routledge

Petersen A, Müller HG (2019) Fréchet regression for random objects with Euclidean predictors. The Annals of Statistics 47(2)

Petersen F, Mukherjee D, Sun Y, Yurochkin M (2021) Post-processing for individual fairness. Advances in Neural Information Processing Systems 34:25944–25955

Pfanzagl P (1979) Conditional distributions as derivatives. The Annals of Probability 7(6):1046–1050

Pfeffermann D (1993) The role of sampling weights when modeling survey data. International Statistical Review/Revue Internationale de Statistique pp 317–337

Phelps ES (1972) The statistical theory of racism and sexism. The american economic review 62(4):659–661

Phelps JT (1895) Life Insurance Sayings. Riverside Press

Picard P (2003) Les frontières de l'assurabilité. Risques 54:65–66

Pichard M (2006) Les droits à: étude de législation française. Economica

Pisu M, Azuero A, McNees P, Burkhardt J, Benz R, Meneses K (2010) The out of pocket cost of breast cancer survivors: a review. Journal of Cancer Survivorship 4(3):202–209

Plakans A, Wetherell C (2000) Patrilines, surnames, and family identity: A case study from the russian baltic provinces in the nineteenth century. The History of the Family 5(2):199–214

Platt J, et al. (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 10(3):61–74

Plečko D, Bennett N, Meinshausen N (2021) `fairadapt`: Causal reasoning for fair data pre-processing. arXiv 2110.10200

Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ (2017) On fairness and calibration. arXiv 1709.02012

Pohle MO (2020) The murphy decomposition and the calibration-resolution principle: A new perspective on forecast evaluation. arXiv 2005.01835

Pojman LP (1998) The case against affirmative action. International Journal of Applied Philosophy 12(1):97–115

Poku M (2016) Campbell's law: implications for health care. Journal of health services research & policy 21(2):137–139

Pope DG, Sydnor JR (2011) Implementing anti-discrimination policies in statistical profiling models. American Economic Journal: Economic Policy 3(3):206–31

Porrini D, Fusco G (2020) Less discrimination, more gender inequality:: The case of the italian motor-vehicle insurance. The Journal of Risk Management and Insurance 24(1):1–11

Porter TM (2020) Trust in numbers. Princeton University Press

Powell L (2020) Risk-based pricing of proper ty and liability insurance. Journal of Insurance Regulation 1

Pradier PC (2011) (petite) histoire de la discrimination (dans les assurances). Risques 87:51–57

Pradier PC (2012) Les bénéfices terrestres de la charité. les rentes viagères des hôpitaux parisiens, 1660-1690. Histoire & mMesure 26(XXVI-2):31–76

Prince AE, Schwarcz D (2019) Proxy discrimination in the age of artificial intelligence and big data. Iowa Law Review 105:1257

Proschan MA, Presnell B (1998) Expect the unexpected from conditional expectation. The American Statistician 52(3):248–252

Puddifoot K (2021) How stereotypes deceive us. Oxford University Press

Puzzo DA (1964) Racism and the western tradition. Journal of the History of Ideas 25(4):579–586

Quinlan JR (1986) Induction of decision trees. Machine learning 1:81–106

Quinlan JR (1987) Simplifying decision trees. International journal of man-machine studies 27(3):221–234

Quinlan JR (1993) C4. 5: programs for machine learning. Elsevier

Quinzii M, Rochet JC (1985) Multidimensional signalling. Journal of mathematical economics 14(3):261–284

Racine J, Rilstone P (1995) The reverse regression problem: statistical paradox or artefact of misspecification? Canadian Journal of Economics pp 502–531

Radcliffe N (2007) Using control groups to target on predicted lift: Building and assessing uplift model. Direct Marketing Analytics Journal pp 14–21

Radcliffe N, Surry P (1999) Differential response analysis: Modeling true responses by isolating the effect of a single action. Credit Scoring and Credit Control IV

Raftery AE, Madigan D, Hoeting JA (1997) Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92(437):179–191

Ransom RL, Sutch R (1987) Tontine insurance and the armstrong investigation: a case of stifled innovation, 1868–1905. The Journal of Economic History 47(2):379–390

Rattani A, Reddy N, Derakhshani R (2017) Gender prediction from mobile ocular images: A feasibility study. In: IEEE International Symposium on Technologies for Homeland Security (HST), IEEE, pp 1–6

Rattani A, Reddy N, Derakhshani R (2018) Convolutional neural networks for gender prediction from smartphone-based ocular images. IET Biometrics 7(5):423–430

Rawls J (1971) A theory of justice: Revised edition. Harvard university press

Rawls J (2001) Justice as fairness: A restatement. Harvard University Press

Rebert L, Van Hoyweghen I (2015) The right to underwrite gender: The goods & services directive and the politics of insurance pricing. Tijdschrift Voor Genderstudies 18(4):413–431

Reichenbach H (1956) The direction of time. Berkeley, University of Los Angeles Press

Reijns T, Weurding R, Schaffers J (2021) Ethical artificial intelligence – the dutch insurance industry makes it a mandate. KPMG Insights 03/2021

Reimers CW (1983) Labor market discrimination against hispanic and black men. The review of economics and statistics pp 570–579

Reinsel GC (2003) Elements of multivariate time series analysis. Springer

Rényi A (1959) On measures of dependence. Acta mathematica hungarica 10(3-4):441–451

Rescher N (2013) How wide is the gap between facts and values? In: Studies in Value Theory, De Gruyter, pp 25–52

Resnick S (2019) A probability path. Springer

Rhynhart R (2020) Mapping the legacy of structural racism in philadelphia. Philadelphia, Office pf the Controller

Ribeiro MT, Singh S, Guestrin C (2016) " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144

Ridgeway CL (2011) Framed by gender: How gender inequality persists in the modern world. Oxford University Press

Rifkin R, Klautau A (2004) In defense of one-vs-all classification. The Journal of Machine Learning Research 5:101–141

Riley JG (1975) Competitive signalling. Journal of economic theory 10(2):174–186

Rink FT (1805) Ansichten aus Immanuel Kant's Leben. Göbbels und Unzer

Rivera LA (2020) Employer decision making. Annual review of sociology 46:215–232

Robbins LA (2015) The pernicious problem of ageism. Generations 39(3):6–9

Robertson T, FT W, Dykstra R (1988) Order Restricted Statistical Inference. John Wiley & Sons, New York

Robinson PM (1988) Root-n-consistent semiparametric regression. Econometrica pp 931–954

Robinson WS (1950) Ecological correlations and the behavior of individuals. American Sociological Review 15(3):351–357

Robnik-Šikonja M, Kononenko I (1997) An adaptation of relief for attribute estimation in regression. In: Machine learning: Proceedings of the fourteenth international conference (ICML'97), vol 5, pp 296–304

Robnik-Šikonja M, Kononenko I (2003) Theoretical and empirical analysis of relieff and rrelieff. Machine learning 53(1):23–69

Robnik-Šikonja M, Kononenko I (2008) Explaining classifications for individual instances. IEEE Transactions on Knowledge and Data Engineering 20(5):589–600

Rodríguez Cardona D, Janssen A, Guhr N, Breitner MH, Milde J (2021) A matter of trust? examination of chatbot usage in insurance business. In: Proceedings of the 54th Hawaii International Conference on System Sciences, p 556

Rodríguez-Cuenca B, Alonso MC (2014) Semi-automatic detection of swimming pools from aerial high-resolution images and lidar data. Remote Sensing 6(4):2628–2646

Roemer JE (1996) Theories of distributive justice. Harvard University Press

Roemer JE (1998) Equality of opportunity. Harvard University Press

Roemer JE, Trannoy A (2016) Equality of opportunity: Theory and measurement. Journal of Economic Literature 54(4):1288–1332

Rolski T, Schmidli H, Schmidt V, Teugels JL (2009) Stochastic processes for insurance and finance. John Wiley & Sons

Rosen J (2011) The right to be forgotten. Stan L Rev Online 64:88

Rosenbaum P (2005) Observational study. Encyclopedia of statistics in behavioral science

Rosenbaum P (2018) Observation and experiment. Harvard University Press

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

Rosenberg NA (2011) A population-genetic perspective on the similarities and differences among worldwide human populations. Human biology 83(6):659

Rosenblatt F (1961) Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Tech. rep., Cornell Aeronautical Lab Inc Buffalo NY

Rosenthal JS (2006) First Look At Rigorous Probability Theory, A. World Scientific Publishing Company

Ross SM (1972) Introduction to probability models. Academic press

Roth K, Lucchi A, Nowozin S, Hofmann T (2017) Stabilizing training of generative adversarial networks through regularization. Advances in neural information processing systems 30

Rothschild-Elyassi G, Koehler J, Simon J (2018) Actuarial Justice, John Wiley & Sons, Ltd, chap 14, pp 194–206

Rothstein WG (2003) Public health and the risk factor: A history of an uneven medical revolution, vol 3. Boydell & Brewer

Rouvroy A, Berns T, Carey-Libbrecht L (2013) Algorithmic governmentality and prospects of emancipation. Réseaux 177(1):163–196

Royal A, Walls M (2019) Flood risk perceptions and insurance choice: Do decisions in the floodplain reflect overoptimism? Risk Analysis 39(5):1088–1104

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology 66(5):688

Rubinow I (1936) State pool plans and merit rating. Law and Contemporary Problems 3(1):65–88

Rubinstein A (2012) Economic fables. Open book publishers

Rubinstein Y, Brenner D (2014) Pride and prejudice: Using ethnic-sounding names and inter-ethnic marriages to identify labour market discrimination. Review of Economic Studies 81(1):389–425

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1(5):206–215

Rudin W (1966) Real and Complex Analysis. McGraw-hill New York

Ruillier J (2004) Quatre petits coins de rien du tout. Bilboquet

Rule NO, Ambady N (2010) Democrats and republicans can be differentiated from their faces. PloS one 5(1):e8733

Rumelhart DE, Hinton GE, Williams RJ (1985) Learning internal representations by error propagation. Tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. nature 323(6088):533–536

Rundle AG, Bader MD, Richards CA, Neckerman KM, Teitler JO (2011) Using google street view to audit neighborhood environments. American journal of preventive medicine 40(1):94–100

Rupke N, Lauer G (2018) Johann Friedrich Blumenbach: race and natural history, 1750–1850. Routledge

Russell C, Kusner M, Loftus C, Silva R (2017) When worlds collide: integrating different counterfactual assumptions in fairness. In: Advances in neural information processing systems, NIPS Proceedings, vol 30, pp 6414–6423

Sabbagh D (2007) Equality and transparency: A strategic perspective on affirmative action in American law. Springer

Saks S (1937) Theory of the integral. Monografie Matematyczne 7

Sakurada M, Yairi T (2014) Anomaly detection using autoencoders with nonlinear dimensionality reduction. In: Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis, pp 4–11

Salimi B, Howe B, Suciu D (2020) Database repair meets algorithmic fairness. ACM SIGMOD Record 49(1):34–41

Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S (2008) Global sensitivity analysis: the primer. John Wiley & Sons

Samadi S, Tantipongpipat U, Morgenstern JH, Singh M, Vempala S (2018) The price of fair pca: One extra dimension. Advances in neural information processing systems 31

Sanche F, Roberge I (2023) La question de la semaine sur le casier judiciaire et les assurances. Radio Canada January 31

Sandel MJ (2020) The tyranny of merit: What's become of the common good? Penguin UK

Santambrogio F (2015) Optimal transport for applied mathematicians. Birkäuser, NY 55(58-63):94

Santosa F, Symes WW (1986) Linear inversion of band-limited reflection seismograms. SIAM journal on scientific and statistical computing 7(4):1307–1330

Sarmanov O (1963) Maximum correlation coefficient (nonsymmetric case). Selected translations in mathematical statistics and probability 2:207–210

Schanze E (2013) Injustice by generalization: notes on the test-achats decision of the european court of justice. German Law Journal 14(2):423–433

Schapire RE (1990) The strength of weak learnability. Machine learning 5:197–227

Schapire RE (2013) Explaining adaboost. Empirical Inference: Festschrift in Honor of Vladimir N Vapnik pp 37–52

Schauer F (2006) Profiles, probabilities, and stereotypes. Harvard University Press

Schauer F (2017) Statistical (and non-statistical) discrimination. In: Lippert-Rasmussen K (ed) Handbook of the Ethics of Discrimination, Routledge, pp 42–53

Schilling E (2006) Accuracy and precision. Encyclopedia of Statistical Sciences p 25

Schlesinger A, O'Hara KP, Taylor AS (2018) Let's talk about race: Identity, chatbots, and ai. In: Proceedings of the 2018 chi conference on human factors in computing systems, pp 1–14

Schmeiser H, Störmer T, Wagner J (2014) Unisex insurance pricing: consumers' perception and market implications. The Geneva Papers on Risk and Insurance-Issues and Practice 39(2):322–350

Schmidt KD (2006) Prediction. Encyclopedia of Actuarial Science

Schneier B (2015) Data and Goliath: The hidden battles to collect your data and control your world. WW Norton & Company

Schweik SM (2009) The ugly laws. New York University Press

Scikit Learn (2017) Probability calibration. https://scikit-learnorg/stable/modules/calibrationhtml

Scism L (2019) New york insurers can evaluate your social media use—if they can prove why it's needed. The Wall Street Journal January 30

Scism L, Maremont M (2010a) Inside deloitte's life-insurance assessment technology. Wall Street Journal November 19

Scism L, Maremont M (2010b) Insurers test data profiles to identify risky clients. Wall Street Journal November 19

Scutari M, Panero F, Proissl M (2022) Achieving fairness with a simple ridge penalty. Statistics and Computing 32(5):77

Seelye KQ (1994) Insurability for battered women. New York Times May 12

Segall S (2013) Equality and opportunity. Oxford University Press

Seicshnaydre SE (2007) Is the road to disparate impact paved with good intentions: Stuck on state of mind in antidiscrimination law. Wake Forest L Rev 42:1141

Selbst AD, Barocas S (2018) The intuitive appeal of explainable machines. Fordham Law Review 87:1085

Seligman D (1983) Insurance and the price of sex. Fortune February 21st

Seresinhe CI, Preis T, Moat HS (2017) Using deep learning to quantify the beauty of outdoor places. Royal Society open science 4(7):170170

Shadish WR, Luellen JK (2005) Quasi-experimental designs. Encyclopedia of statistics in behavioral science

Shannon CE, Weaver W (1949) The mathematical theory of communication. Urbana, Illinois: University of Illinois Press

Shapley LS (1953) A value for n-person games. In: Kuhn HW, Tucker AW (eds) Contributions to the Theory of Games II, Princeton University Press, Princeton, pp 307–317

Shapley LS, Shubik M (1969) Pure competition, coalitional power, and fair division. International Economic Review 10(3):337–362

Shikhare S (2021) Next generation ltc - life insurance underwriting using facial score model. In: Insurance Data Science conference

Siddiqi N (2012) Credit risk scorecards: developing and implementing intelligent credit scoring, vol 3. John Wiley & Sons

Silver N (2012) The signal and the noise: Why so many predictions fail-but some don't. Penguin

Simon J (1987) The emergence of a risk society-insurance, law, and the state. Socialist Review (95):60–89

Simon J (1988) The ideological effects of actuarial practices. Law & Society Review 22:771

Singer P (2011) Practical ethics. Cambridge university press

Slovic P (1987) Perception of risk. Science 236(4799):280–285

Small ML, Pager D (2020) Sociological perspectives on racial discrimination. Journal of Economic Perspectives 34(2):49–67

Smith A (1759) The theory of moral sentiments. Penguin

Smith CS (2021) A.i. here, there, everywhere. New York Times (February 23)

Smith GC, Pell JP (2003) Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. Bmj 327(7429):1459–1461

Smythe HH (1952) The eta: a marginal japanese caste. American Journal of Sociology 58(2):194–196

Sokolova M, Japkowicz N, Szpakowicz S (2006) Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: Australasian joint conference on artificial intelligence, Springer, pp 1015–1021

Sollich P, Krogh A (1995) Learning with ensembles: How overfitting can be useful. Advances in neural information processing systems 8

Solow RM (1957) Technical change and the aggregate production function. The review of Economics and Statistics pp 312–320

Solution ICD (2020) How to increase credit score

Sorensen TA (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. Biol Skar 5:1–34

Spedicato GA, Dutang C, Petrini L (2018) Machine learning methods to perform pricing optimization. a comparison with standard glms. Variance 12(1):69–89

Speicher T, Ali M, Venkatadri G, Ribeiro FN, Arvanitakis G, Benevenuto F, Gummadi KP, Loiseau P, Mislove A (2018) Potential for discrimination in online targeted advertising. In: Conference on fairness, accountability and transparency, Proceedings of Machine Learning Research, pp 5–19

Spence M (1974) Competitive and optimal responses to signals: An analysis of efficiency and distribution. Journal of Economic theory 7(3):296–332

Spence M (1976) Informational aspects of market structure: An introduction. The Quarterly Journal of Economics pp 591–597

Spender A, Bullen C, Altmann-Richer L, Cripps J, Duffy R, Falkous C, Farrell M, Horn T, Wigzell J, Yeap W (2019) Wearables and the internet of things: Considerations for the life and health insurance industry. British Actuarial Journal 24:e22

Spiegelhalter DJ, Dawid AP, Lauritzen SL, Cowell RG (1993) Bayesian analysis in expert systems. Statistical science pp 219–247

Spirtes P, Glymour C, Scheines R (1993) Discovery algorithms for causally sufficient structures. In: Causation, prediction, and search, Springer, pp 103–162

Squires G (2011) Redlining To Reinvestment. Temple University Press

Squires GD (2003) Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas. Journal of Urban Affairs 25(4):391–410

Squires GD, Chadwick J (2006) Linguistic profiling: A continuing tradition of discrimination in the home insurance industry? Urban Affairs Review 41(3):400–415

Squires GD, DeWolfe R (1981) Insurance redlining in minority communities. The Review of Black Political Economy 11(3):347–364

Stark L, Stanhaus A, Anthony DL (2020) "i don't want someone to watch me while i'm working": Gendered views of facial recognition technology in workplace surveillance. Journal of the Association for Information Science and Technology 71(9):1074–1088

Steensma C, Loukine L, Orpana H, Lo E, Choi B, Waters C, Martel S (2013) Comparing life expectancy and health-adjusted life expectancy by body mass index category in adult canadians: a descriptive study. Population health metrics 11(1):1–12

Stein A (1994) Will health care reform protect victims of abuse-treating domestic violence as a public health issue. Human Rights 21:16

Stenholm S, Head J, Aalto V, Kivimäki M, Kawachi I, Zins M, Goldberg M, Platts LG, Zaninotto P, Hanson LM, et al. (2017) Body mass index as a predictor of healthy and disease-free life expectancy between ages 50 and 75: a multicohort study. International journal of obesity 41(5):769–775

Stephan Y, Sutin AR, Terracciano A (2015) How old do you feel? the role of age discrimination and biological aging in subjective age. PloS one 10(3):e0119293

Stevenson M (2018) Assessing risk assessment in action. Minnesota Law Review 103:303

Steyerberg E, Eijkemans M, Habbema J (2001) Application of shrinkage techniques in logistic regression analysis: a case study. Statistica Neerlandica 55(1):76–88

Stone DA (1993) The struggle for the soul of health insurance. Journal of Health Politics, Policy and Law 18(2):287–317

Stone P (2007) Why lotteries are just. Journal of Political Philosophy 15(3):276–295

Štrumbelj E, Kononenko I (2010) An efficient explanation of individual classifications using game theory. The Journal of Machine Learning Research 11:1–18

Štrumbelj E, Kononenko I (2014) Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems 41:647–665

Struyck N (1740) Inleiding tot de algemeene geographie. Tirion 1740:231

Struyck N (1912) Les oeuvres de Nicolas Struyck (1687-1769): qui se rapportent au calcul des chances, à la statistique général, z la statistique des décès et aux rentes viagèter. Société générale néerlandaise d'assurances sur la vie et de rentes viagères

Stuart EA (2010) Matching methods for causal inference: A review and a look forward. Statistical Science 25(1):1

Suresh H, Guttag JV (2019) A framework for understanding sources of harm throughout the machine learning life cycle. arXiv 1901.10002

Surowiecki J (2004) The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Doubleday & Co

Sutton W (1874) On the method used by dr. price in the construction of the northampton mortality table. Journal of the Institute of Actuaries 18(2):107–122

Swauger S (2021) The next normal: Algorithms will take over college, from admissions to advising. Washington Post (November 12)

Sweeney L (2013) Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. Queue 11(3):10–29

Szalavitz M (2017) Why do we think poor people are poor because of their own bad choices. The Guardian July 5

Szepannek G, Lübke K (2021) Facing the challenges of developing fair risk scoring models. Frontiers in artificial intelligence 4:681915

Tajfel H, Turner JC, Worchel S, Austin WG, et al. (1986) Psychology of intergroup relations. Chicago: Nelson-Hall pp 7–24

Tajfel HE (1978) Differentiation between social groups: Studies in the social psychology of intergroup relations. Academic Press

Tanaka K (2012) Surnames and gender in japan: Women's challenges in seeking own identity. Journal of Family History 37(2):232–240

Tang S, Zhang X, Cryan J, Metzger MJ, Zheng H, Zhao BY (2017) Gender bias in the job market: A longitudinal analysis. Proceedings of the ACM on Human-Computer Interaction 1(CSCW):1–19

Tasche D (2008) Validation of internal rating systems and pd estimates. In: The analytics of risk model validation, Elsevier, pp 169–196

Taylor A, Sadowski J (2015) How companies turn your facebook activity into a credit score. The Nation May 27

Taylor S (2015) Price optimization ban. Government of the District of Columbia, Department of Insurance August 25

Telles E (2014) Pigmentocracies: Ethnicity, race, and color in Latin America. UNC Press Books

Tharwat A (2021) Classification assessment methods. Applied computing and informatics 17(1):168–192

The Zebra (2022) Car insurance rating factors by state. https://wwwthezebracom/

Theobald CM (1974) Generalizations of mean square error applied to ridge regression. Journal of the Royal Statistical Society: Series B (Methodological) 36(1):103–106

Theodoridis S (2015) Machine learning: a Bayesian and optimization perspective. Academic press

Thiery Y, Van Schoubroeck C (2006) Fairness and equality in insurance classification. The Geneva Papers on Risk and Insurance-Issues and Practice 31(2):190–211

Thomas G (2012) Non-risk price discrimination in insurance: market outcomes and public policy. The Geneva Papers on Risk and Insurance-Issues and Practice 37:27–46

Thomas G (2017) Loss coverage: Why insurance works better with some adverse selection. Cambridge University Press

Thomas L, Crook J, Edelman D (2002) Credit scoring and its applications. SIAM

Thomas RG (2007) Some novel perspectives on risk classification. The Geneva Papers on Risk and Insurance-Issues and Practice 32(1):105–132

Thomson JJ (1976) Killing, letting die, and the trolley problem. The Monist 59(2):204–217

Thomson W (1883) Electrical units of measurement. Popular lectures and addresses 1(73)

Thornton SM, Pan S, Erlien SM, Gerdes JC (2016) Incorporating ethical considerations into automated vehicle control. IEEE Transactions on Intelligent Transportation Systems 18(6):1429–1439

Tian J, Pearl J (2002) A general identification condition for causal effects. In: Proceedings of the Eighteenth National Conference on Artificial Intelligence, MIT Press, pp 567–573

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58(1):267–288

Tilcsik A (2021) Statistical discrimination and the rationalization of stereotypes. American Sociological Review 86(1):93–122

Topkis DM (1998) Supermodularity and complementarity. Princeton university press

Torkamani A, Wineinger NE, Topol EJ (2018) The personal and clinical utility of polygenic risk scores. Nature Reviews Genetics 19(9):581–590

Torous W, Gunsilius F, Rigollet P (2021) An optimal transport approach to causal inference. arXiv 2108.05858

Traag V, Waltman L (2022) Causal foundations of bias, disparity and fairness. arXiv 2207.13665

Tribalat M (2016) Statistiques ethniques, Une querelle bien française. L'Artilleur

Tsamados A, Aggarwal N, Cowls J, Morley J, Roberts H, Taddeo M, Floridi L (2021) The ethics of algorithms: key problems and solutions. AI & Society pp 1–16

Tsybakov AB (2009) Introduction to nonparametric estimation. Springer Verlag

Tufekci Z (2018) Facebook's surveillance machine. New York Times 19:1

Tukey JW (1961) Curves as parameters, and touch estimation. In: Neyman J (ed) Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, vol 1, pp 681–694

Tuppat J, Gerhards J (2021) Immigrants' first names and perceived discrimination: A contribution to understanding the integration paradox. European Sociological Review 37(1):121–135

Turner R (2015) The way to stop discrimination on the basis of race. Stanford Journal of Civil Rights & Civil Liberties 11:45

Tweedie MCK (1984) An index which distinguishes between some important exponential families. Statistics: applications and new directions (Calcutta, 1981) pp 579–604

Tzioumis K (2018) Demographic aspects of first names. Scientific data 5(1):1–9

Uotinen V, Rantanen T, Suutama T (2005) Perceived age as a predictor of old age mortality: a 13-year prospective study. Age and Ageing 34(4):368–372

Upton G, Cook I (2014) A dictionary of statistics 3e. Oxford university press

US Census (2012) Frequently occurring surnames from census 2000, census report data file a: Top 1000 names. Genealogy Data

Van der Vaart AW (2000) Asymptotic statistics. Cambridge university press

Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW (2019) Calibration: the achilles heel of predictive analytics. BMC medicine 17(1):1–7

Van Deemter K (2010) Not exactly: In praise of vagueness. Oxford University Press

Van Lancker W (2020) Automating the welfare state: Consequences and challenges for the organisation of solidarity. In: Shifting Solidarities, Springer, pp 153–173

Van Parijs P (2002) Linguistic justice. Politics, Philosophy & Economics 1(1):59–74

Van Schaack D (1926) The part of the casualty insurance company in accident prevention. The Annals of the American Academy of Political and Social Science 123(1):36–40

Vandenhole W (2005) Non-discrimination and equality in the view of the UN human rights treaty bodies. Intersentia nv

Varga TV, Kozodoi N (2021) `fairness`, algorithmic fairness r package. R Vignette

Vassy JL, Christensen KD, Schonman EF, Blout CL, Robinson JO, Krier JB, Diamond PM, Lebo M, Machini K, Azzariti DR, et al. (2017) The impact of whole-genome sequencing on the primary care and outcomes of healthy adult patients: a pilot randomized trial. Annals of internal medicine 167(3):159–169

Verboven K (2011) Introduction: Professional collegia: Guilds or social clubs? Ancient Society pp 187–195

Verma S, Rubin J (2018) Fairness definitions explained. In: 2018 ieee/acm international workshop on software fairness (fairware), IEEE, pp 1–7

Verrall R (1996) Claims reserving and generalised additive models. Insurance: Mathematics and Economics 19(1):31–43

Viaene S, Dedene G, Derrig RA (2005) Auto claim fraud detection using bayesian learning neural networks. Expert systems with applications 29(3):653–666

Vidoni P (2003) Prediction and calibration in generalized linear models. Annals of the Institute of Statistical Mathematics 55:169–185

Villani C (2003) Topics in optimal transportation, vol 58. American Mathematical Society

Villani C (2009) Optimal transport: old and new, vol 338. Springer

Villazor RC (2008) Blood quantum land laws and the race versus political identity dilemma. Californial Law Review 96:801

Viswanathan KS (2006) Demutualization. Encyclopedia of Actuarial Science

Vogel R, Bellet A, Clémen S, et al. (2021) Learning fair scoring functions: Bipartite ranking under roc-based fairness constraints. In: International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, pp 784–792

Voicu I (2018) Using first name information to improve race and ethnicity classification. Statistics and Public Policy 5(1):1–13

Volkmer S (2015) Notice regarding unfair discrimination in rating: optimization. State of California, Department of Insurance February 18

Von Neumann J (1955) Method in the physical sciences. Collected Works 6:491–498

Von Neumann J, Morgenstern O (1953) Theory of games and economic behavior. Princeton university press

Žliobaite I (2015) On the relation between accuracy and fairness in binary classification. arXiv 1505.05723

Žliobaite I, Custers B (2016) Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. Artificial Intelligence and Law 24(2):183–201

Wachter S, Mittelstadt B (2019) A right to reasonable inferences: re-thinking data protection law in the age of big data and ai. Columbia Business Law Review p 494

Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harvard Journal of Law & Technology 31:841

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113(523):1228–1242

Waldron H (2013) Mortality differentials by lifetime earnings decile: Implications for evaluations of proposed social security law changes. Social Security Bulletin 73:1

Wallis KF (2014) Revisiting francis galton's forecasting competition. Statistical Science pp 420–424

Wang DB, Feng L, Zhang ML (2021) Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. Advances in Neural Information Processing Systems 34:11809–11820

Wang Y, Kosinski M (2018) Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Journal of personality and social psychology 114(2):246

Wang Y, Yao H, Zhao S (2016) Auto-encoder based dimensionality reduction. Neurocomputing 184:232–242

Wasserman L (2000) Bayesian model selection and model averaging. Journal of mathematical psychology 44(1):92–107

Wasserstein LN (1969) Markov processes over denumerable products of spaces, describing large systems of automata. Problemy Peredachi Informatsii 5(3):64–72

Watkins-Hayes C, Kovalsky E (2016) The discourse of deservingness. The Oxford handbook of the social science of poverty 1

Watson DS, Gultchin L, Taly A, Floridi L (2021) Local explanations via necessity and sufficiency: Unifying theory and practice. Uncertainty in Artificial Intelligence pp 1382–1392

Watson GS (1964) Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A pp 359–372

Weber M (1904) Die protestantische ethik und der „geist "des kapitalismus. Archiv für Sozialwissenschaft und Sozialpolitik 20:1–54

Weed DL (2005) Weight of evidence: a review of concept and methods. Risk Analysis: An International Journal 25(6):1545–1557

Weisberg HI, Tomberlin TJ (1982) A statistical perspective on actuarial methods for estimating pure premiums from cross-classified data. Journal of Risk and Insurance pp 539–563

Welsh AH, Cunningham RB, Donnelly CF, Lindenmayer DB (1996) Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecological Modelling 88(1-3):297–308

Westreich D (2012) Berkson's bias, selection bias, and missing data. Epidemiology 23(1):159

Wheatley M (2013) The fairness challenge. FCA Financial Conduct Authority Speeches October 10

White RW, Doraiswamy PM, Horvitz E (2018) Detecting neurodegenerative disorders from web search signals. NPJ digital medicine 1(1):8

Wiehl DG (1960) Build and Blood Pressure. Society of Actuaries

van Wieringen WN (2015) Lecture notes on ridge regression. arXiv 1509.09169

Wiggins B (2013) Managing Risk, Managing Race: Racialized Actuarial Science in the United States, 1881–1948. University of Minnesota PhD thesis

Wikipedia (2023) Data. Wikipedia, The Free Encyclopedia

Wilcox C (1937) Merit rating in state unemployment compensation laws. The American Economic Review pp 253–259

Wilkie D (1997) Mutuality and solidarity: assessing risks and sharing losses. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 352(1357):1039–1044

Williams BA, Brooks CF, Shmargad Y (2018) How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. Journal of Information Policy 8:78–115

Williams G (2017) Discrimination and obesity. In: Lippert-Rasmussen K (ed) Handbook of the Ethics of Discrimination, Routledge, pp 264–275

Williams JE, Bennett SM (1975) The definition of sex stereotypes via the adjective check list. Sex roles 1(4)

Willson K (2009) Name law and gender in iceland. UCLA: Center for the Study of Women

Wilson EB, Worcester J (1943) The determination of ld 50 and its sampling error in bio-assay. Proceedings of the National Academy of Sciences 29(2):79–85

Wing-Heir L (2015) Price optimization in ratemaking. State of Alaska, Department of Commerce, Community and Economic Development Bulletin B 15-12

Winter RA (2000) Optimal insurance under moral hazard. Handbook of insurance pp 155–183

Winterfeldt D, Edwards W (1986) Decision analysis and behavioral research

Witten IH, Frank E, Hall MA, Pal CJ, DATA M (2016) Practical machine learning tools and techniques. Morgan Kaufmann

Wod I (1985) Weight of evidence: A brief survey. Bayesian statistics 2:249–270

Wolff MJ (2006) The myth of the actuary: life insurance and frederick l. hoffman's race traits and tendencies of the american negro. Public Health Reports 121(1):84–91

Wolffhechel K, Fagertun J, Jacobsen UP, Majewski W, Hemmingsen AS, Larsen CL, Lorentzen SK, Jarmer H (2014) Interpretation of appearance: The effect of facial features on first impressions and personality. PloS one 9(9):e107721

Wolpert DH (1992) Stacked generalization. Neural networks 5(2):241–259

Wolthuis H (2004) Heterogeneity. Encyclopedia of Actuarial Science pp 819–821

Woodhams C, Williams M, Dacre J, Parnerkar I, Sharma M (2021) Retrospective observational study of ethnicity-gender pay gaps among hospital and community health service doctors in england. BMJ open 11(12):e051043

Worham L (1985) Insurance classification: too important to be left to the actuaries. University of Michigan Journal of Law 19:349

Works R (1977) Whatever's fair-adequacy, equity, and the underwriting prerogative in property insurance markets. Nebraska Law Review 56:445

Wortham L (1986) The economics of insurance classification: The sound of one invisible hand clapping. Ohio State Law Journal 47:835

Wright S (1921) Correlation and causation. Journal of Agricultural Research 20

Wu Y, Zhang L, Wu X, Tong H (2019) Pc-fairness: A unified framework for measuring causality-based fairness. Advances in Neural Information Processing Systems 32

Wu Z, D'Oosterlinck K, Geiger A, Zur A, Potts C (2022) Causal proxy models for concept-based model explanations. arXiv preprint arXiv:220914279

Wüthrich MV, Merz M (2008) Stochastic claims reserving methods in insurance. John Wiley & Sons

Wüthrich MV, Merz M (2022) Statistical foundations of actuarial learning and its applications, vol 3822407. Springer Nature

Yang TC, Chen VYJ, Shoff C, Matthews SA (2012) Using quantile regression to examine the effects of inequality across the mortality distribution in the us counties. Social science & medicine 74(12):1900–1910

Yao J (2016) Clustering in general insurance pricing. In: Frees E, Meyers G, Derrig R (eds) Predictive modeling applications in actuarial science, Cambridge University Press, pp 159–79

Yeung K (2018a) Algorithmic regulation: a critical interrogation. Regulation & Governance 12(4):505–523

Yeung K (2018b) A study of the implications of advanced digital technologies (including ai systems) for the concept of responsibility within a human rights framework. MSI-AUT (2018) 5

Yinger J (1998) Evidence on discrimination in consumer markets. Journal of Economic perspectives 12(2):23–40

Yitzhaki S, Schechtman E (2013) The Gini methodology: a primer on a statistical methodology. Springer

Young IM (1990) Justice and the Politics of Difference. Princeton University Press

Young RK, Kennedy AH, Newhouse A, Browne P, Thiessen D (1993) The effects of names on perception of intelligence, popularity, and competence. Journal of Applied Social Psychology 23(21):1770–1788

Zack N (2014) Philosophy of science and race. Routledge

Zadrozny B, Elkan C (2001) Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: ICML, Citeseer, vol 1, pp 609–616

Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 694–699

Zafar MB, Valera I, Rodriguez MG, Gummadi KP (2017) Fairness constraints: Mechanisms for fair classification. arXiv 1507.05259:962–970

Zafar MB, Valera I, Gomez-Rodriguez M, Gummadi KP (2019) Fairness constraints: A flexible approach for fair classification. The Journal of Machine Learning Research 20(1):2737–2778

Zafar SY, Abernethy AP (2013) Financial toxicity, part i: a new name for a growing problem. Oncology 27(2):80

Zelizer VAR (2017) Morals and markets: The development of life insurance in the United States. Columbia University Press

Zelizer VAR (2018) Morals and markets. Columbia University Press

Zenere A, Larsson EG, Altafini C (2022) Relating balance and conditional independence in graphical models. Physical Review E 106(4):044309

Zhang J, Bareinboim E (2018) Fairness in decision-making—the causal explanation formula. In: Thirty-Second AAAI Conference on Artificial Intelligence

Zhang L, Wu Y, Wu X (2016) A causal framework for discovering and removing direct and indirect discrimination. arXiv 1611.07509

Zhou ZH (2012) Ensemble methods: foundations and algorithms. CRC press

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) 67(2):301–320

Zuboff S (2019) The age of surveillance capitalism: The fight for a human future at the new frontier of power. Public Affairs