Machine Learning for Insurers and Actuaries

Arthur Charpentier

2025

SCOR FOUNDATION FOR SCIENCE

Learning, with an actuarial perspective

(lecture 3)

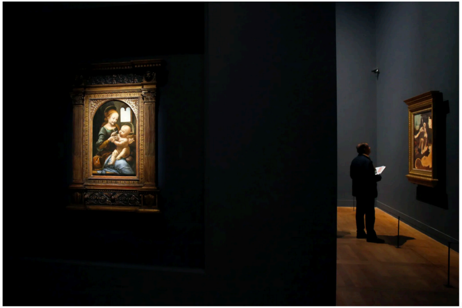
Causal Claim

The New York Times

Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

Share full article



See Cramer (2019) and PMAP 8141

How does we know if X causes Y?

X causes Y if...

...we intervene and change X without changing anything else...

...and Y changes

But in many applications, we can't do that...

Causal Perspective, Two Types of Data

- Causality is the relationship between cause and effect.
- In contrast to correlation, which measures the strength of a relationship between two variables, causality seeks to understand whether one event causes another.
- Example:

  If a drug treatment improves patient health, this is a causal relationship.
  If there is a statistical association between ice cream sales and drowning accidents, this is correlation (but not necessarily causation).
- Understanding causal relationships is essential in ML for model interpretability, decision-making, and counterfactual reasoning.
- Prediction vs. Causal Inference:

  Prediction: ML models typically focus on predicting outcomes based on observed data.
  Causal Inference: ML can go beyond prediction by identifying causal relationships, which is crucial for intervention and decision-making.

Causal Perspective, Two Types of Data

- Causal Models are necessary when we want to:

  Understand the effect of interventions or changes (e.g., how a policy change will affect an outcome).

  Estimate counterfactuals, such as "What would have happened if...?"

- In the context of personalized medicine, marketing, or economics, knowing the causal effect of actions (e.g., a drug, a marketing campaign) is more valuable than simple predictions.
- Correlation measures the strength and direction of a linear relationship between two variables, but it does not imply causality.
- Causality goes beyond correlation to explain how one variable directly influences another.
- Example:

  Correlation: There is a strong correlation between the number of hours studied and exam scores.

Causal Perspective, Two Types of Data

Causality: We hypothesize that studying more causes better performance on the exam.

- Spurious correlations can occur when a third variable is involved, which makes a relationship appear causal when it is not.
- In ML, causal inference allows us to establish true cause-effect relationships, while correlation alone can be misleading.
- A counterfactual is a hypothetical scenario describing what would have happened if a different action or event had occurred.
- In causal inference, counterfactuals allow us to estimate the effect of an intervention:

$$Y_{\text{treatment}} = \text{Outcome if treatment applied}$$

$$Y_{\text{control}} = \text{Outcome if no treatment applied}$$

- Example:

Causal Perspective, Two Types of Data

A patient receives a new drug. The counterfactual asks, "What would have happened if the patient did not receive the drug?"
The difference between the actual outcome and the counterfactual outcome represents the causal effect.

- Counterfactual reasoning helps answer "what if" questions and is crucial for understanding causal effects in ML.
- Causal Inference is the process of drawing conclusions about causal relationships from data.
- Causal Models include:

  Structural Causal Models (SCMs): Represent causal relationships using directed acyclic graphs (DAGs).
  Potential Outcomes Framework: Defines counterfactuals and causal effects using treatment and control groups.

Causal Perspective, Two Types of Data

- Example: In a healthcare study, a causal model can help estimate the effect of a drug on patient outcomes while controlling for confounding variables.
- Interventions: Once we know the causal structure, we can simulate the effects of interventions (e.g., changing a treatment or policy).
- Applications in ML:
  - Counterfactual Reasoning: Predicting the effect of actions or interventions on the system.
  - Reinforcement Learning: Estimating the impact of actions taken by the agent on future outcomes.
- Causality seeks to understand how one variable influences another, beyond mere correlation.
- Counterfactuals allow us to reason about what would have happened under different conditions, enabling causal effect estimation.
- Causal Inference in ML:

Causal Perspective, Two Types of Data

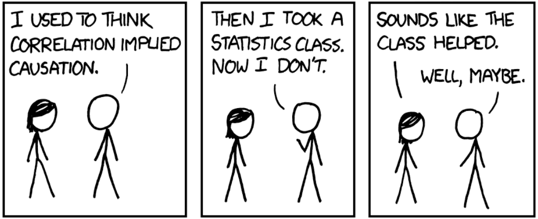
Intervention: Estimating the effect of potential interventions (e.g., treatment, marketing strategies).
Prediction: Making predictions that account for potential causal effects, not just correlations.

- Importance in ML:

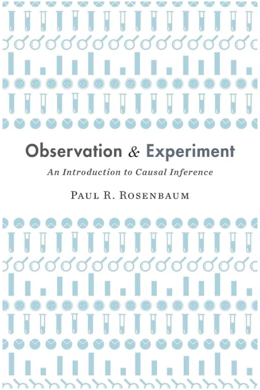
  Many ML applications (e.g., personalized recommendations, policy decisions) require understanding causal relationships.
  Causal models provide a powerful framework to go beyond prediction and allow for actionable insights.

Two Types of Data

"It is often said, 'You cannot prove causality with statistics.' One of my professors, Frederick Mosteller, liked to counter, 'You can only prove causality with statistics.' (...) The title, 'Observation and Experiment,' marks the modern distinction between randomized experiments and observational studies," Rosenbaum (2018)



Correlation, Randall Munroe, 2009 https://xkcd.com/552/



Three Types of Reasoning

"Ladder of causation" from Pearl et al. (2009)

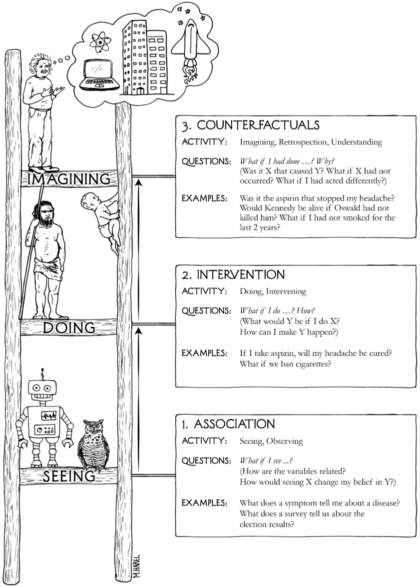
3. **Counterfactuals**
(Imagining, "*what if I had done...*")

2. **Intervention**
(Doing, "*what if I do...*")

1. **Association**
(Seeing, "*what if I see...*")

Picture source: Pearl and Mackenzie (2018)

What would be the impact of a treatment T on a variable of interest Y?



Case 1 - Intervention

"No causation without manipulation," Holland (1986)

→ Randomized Control Trial (RCT)

- Check that key demographics and other confounders are balanced
- Find difference in average outcome in treatment and control groups
- Use statistical significance to test for effects

RCT considered a Golden Standard

See Jonas Salk's polio vaccine in the 50's, Meldrum (1998)

But doesn't fix attrition problem

Case 1 - Intervention

If the study is too short, the effect might not be detectable yet; if the study is too long, attrition becomes a problem
(people might drop out because of the treatment, or because they got/didn't get into the control group)

- Hawthorne effect, observing people makes them behave differently
- John Henry effect, control group works hard to prove they're as good as the treatment group
- Spillover effect, control groups naturally pick up what the treatment group is getting

see also Yeh et al. (2018)


RESEARCH

OPEN ACCESS

Check for updates

Parachute use to prevent death and major trauma when jumping from aircraft: randomized controlled trial

Robert W Yeh,[1] Linda R Valsdottir,[1] Michael W Yeh,[2] Changyu Shen,[1] Daniel B Kramer,[1] Jordan B Strom,[1] Eric A Secemsky,[1] Joanne L Healy,[1] Robert M Domeier,[3] Dhruv S Kazi,[1] Brahmajee K Nallamothu[4] On behalf of the PARACHUTE Investigators

[1]Richard A and Susan F Smith Center for Outcomes Research in Cardiology, Beth Israel Deaconess Medical Center, Harvard Medical School, 375 Longwood Avenue, Boston, MA 02215, USA

[2]David Geffen School of Medicine, University of California, Los Angeles, CA, USA

[3]Department of Emergency Medicine, University of Michigan and Saint Joseph Hospital, Ann Arbor, MI, USA

[4]Michigan Integrated Center for Health Analytics and Medical Prediction, Department of Internal Medicine and Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, MI, USA

Correspondence to: R W Yeh ryeh@bidmc.harvard.edu (or @rwyeh on Twitter)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2018;363:k5094
http://dx.doi.org/10.1136/bmj.k5094

Accepted: 22 November 2018

ABSTRACT

OBJECTIVE
To determine if using a parachute prevents death or major traumatic injury when jumping from an aircraft.

DESIGN
Randomized controlled trial.

SETTING
Private or commercial aircraft between September 2017 and August 2018.

PARTICIPANTS
92 aircraft passengers aged 18 and over were screened for participation. 23 agreed to be enrolled and were randomized.

INTERVENTION
Jumping from an aircraft (airplane or helicopter) with a parachute versus an empty backpack (unblinded).

MAIN OUTCOME MEASURES
Composite of death or major traumatic injury (defined by an Injury Severity Score over 15) upon impact with the ground measured immediately after landing.

RESULTS
Parachute use did not significantly reduce death or major injury (0% for parachute *v* 0% for control; P>0.9). This finding was consistent across multiple subgroups. Compared with individuals screened but not enrolled, participants included in the study were on aircraft at significantly lower altitude (mean of 0.6 m for participants *v* mean of 9146 m for non-participants; P<0.001) and lower velocity (mean of 0 km/h *v* mean of 800 km/h; P<0.001).

CONCLUSIONS
Parachute use did not reduce death or major traumatic injury when jumping from aircraft in the first randomized evaluation of this intervention. However, the trial was only able to enroll participants on small stationary aircraft on the ground, suggesting cautious extrapolation to high altitude jumps. When beliefs regarding the effectiveness of an intervention exist in the community, randomized trials might selectively enroll individuals with a lower perceived likelihood of benefit, thus diminishing the applicability of the results to clinical practice.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Parachutes are routinely used to prevent death or major traumatic injury among individuals jumping from aircraft, but their efficacy is based primarily on biological plausibility and expert opinion

No randomized controlled trials of parachute use have yet been attempted, presumably owing to a lack of equipoise

WHAT THIS STUDY ADDS

This randomized trial of parachute use found no reduction in death or major injury compared with individuals jumping from aircraft with an empty backpack

Lack of enrolment of individuals at high risk could have influenced the results of the trial

Introduction

Parachutes are routinely used to prevent death or major traumatic injury among individuals jumping from aircraft. However, evidence supporting the efficacy of parachutes is weak and guideline recommendations for their use are principally based on biological plausibility and expert opinion.[1 2] Despite this widely held yet unsubstantiated belief of efficacy, many studies of parachutes have suggested injuries related to their use in both military and recreational settings,[3 4] and parachutist injuries are formally recognized in the World Health Organization's ICD-10 (international classification of diseases, 10th revision).[5] This could raise concerns for supporters of evidence-based medicine, because numerous medical interventions believed to be useful have ultimately failed to show efficacy when subjected to properly executed randomized clinical trials.[6 7]

Previous attempts to evaluate parachute use in a randomized setting have not been undertaken owing to both ethical and practical concerns. Lack of equipoise could inhibit recruitment of participants in such a trial. However, whether pre-existing beliefs about the efficacy of parachutes would, in fact, impair the enrolment of participants in a clinical trial has not been formally evaluated. To address these important gaps in evidence, we conducted the first randomized clinical trial of the efficacy of parachutes in reducing death and major injury when jumping from an aircraft.

Methods

Study protocol

Between September 2017 and August 2018, individuals were screened for inclusion in the PArticipation in RAndomized trials Compromised by widely Held beliefs aboUt lack of Treatment Equipoise (PARACHUTE) trial. Prospective participants were approached and screened by study investigators on commercial or private aircraft.

For the commercial aircraft, travel was related to trips the investigators were scheduled to take for business or personal reasons unrelated to the present study. Typically, passengers seated close to the study investigator (typically not known acquaintances) would be approached mid-flight, between the time of initial seating and time of exiting the aircraft. The

Case 2a - Double difference method (or difference of differences)

"Difference in differences" (DID), studying the differential effect of a treatment on a 'treatment group' versus a 'control group' in a natural experiment, Angrist and Pischke (2009)

Example minimum wages and employment, Card and Krueger (1994) and Imai (2022)

What happens if you raise the minimum wage?

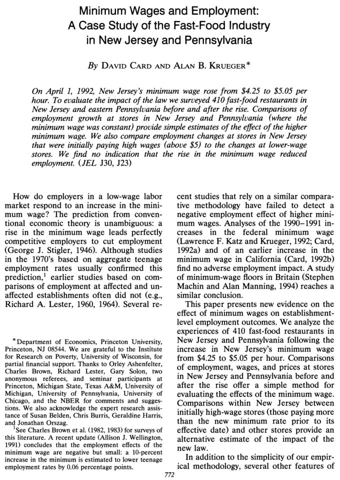
Economic theory says there should be fewer jobs

New Jersey in 1992 \$4.25 $\rightarrow$ \$5.05

Average number of jobs per fast food restaurant in NJ

$$\begin{cases} \text{before (NJ)} & : 20.44 \\ \text{after (NJ)} & : 21.03 \end{cases}$$

$\Delta = 0.59$: Is this the causal effect?



Minimum Wages and Employment:
A Case Study of the Fast-Food Industry
in New Jersey and Pennsylvania

By David Card and Alan B. Krueger*

On April 1, 1992, New Jersey's minimum wage rose from \$4.25 to \$5.05 per hour. To evaluate the impact of the law we surveyed 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise. Comparisons of employment growth at stores in New Jersey and Pennsylvania (where the minimum wage was constant) provide simple estimates of the effect of the higher minimum wage. We also compare employment changes at stores in New Jersey that were initially paying high wages (above \$5) to the changes at lower-wage stores. We find no indication that the rise in the minimum wage reduced employment. (*JEL* J30, J23)

How do employers in a low-wage labor market respond to an increase in the minimum wage? The prediction from conventional economic theory is unambiguous: a rise in the minimum wage leads perfectly competitive employers to cut employment (George J. Stigler, 1946). Although studies in the 1970's based on aggregate teenage employment rates usually confirmed this prediction,[1] earlier studies based on comparisons of employment at affected and unaffected establishments often did not (e.g., Richard A. Lester, 1960, 1964). Several recent studies that rely on a similar comparative methodology have failed to detect a negative employment effect of higher minimum wages. Analyses of the 1990–1991 increases in the federal minimum wage (Lawrence F. Katz and Krueger, 1992; Card, 1992a) and of an earlier increase in the minimum wage in California (Card, 1992b) find no adverse employment impact. A study of minimum-wage floors in Britain (Stephen Machin and Alan Manning, 1994) reaches a similar conclusion.

This paper presents new evidence on the effect of minimum wages on establishment-level employment outcomes. We analyze the experiences of 410 fast-food restaurants in New Jersey and Pennsylvania following the increase in New Jersey's minimum wage from \$4.25 to \$5.05 per hour. Comparisons of employment, wages, and prices at stores in New Jersey and Pennsylvania before and after the rise offer a simple method for evaluating the effects of the minimum wage. Comparisons within New Jersey between initially high-wage stores (those paying more than the new minimum rate prior to its effective date) and other stores provide an alternative estimate of the impact of the new law.

In addition to the simplicity of our empirical methodology, several other features of

*Department of Economics, Princeton University, Princeton, NJ 08544. We are grateful to the Institute for Research on Poverty, University of Wisconsin, for partial financial support. Thanks to Orley Ashenfelter, Charles Brown, Richard Lester, Gary Solon, two anonymous referees, and seminar participants at Princeton, Michigan State, Texas A&M, University of Michigan, University of Pennsylvania, University of Chicago, and the NBER for comments and suggestions. We also acknowledge the expert research assistance of Susan Belden, Chris Burris, Geraldine Harris, and Jonathan Orszag.

[1] See Charles Brown et al. (1982, 1983) for surveys of this literature. A recent update (Allison J. Wellington, 1991) concludes that the employment effects of the minimum wage are negative but small: a 10-percent increase in the minimum is estimated to lower teenage employment rates by 0.06 percentage points.

772

Case 2a - Double difference method (or difference of differences)

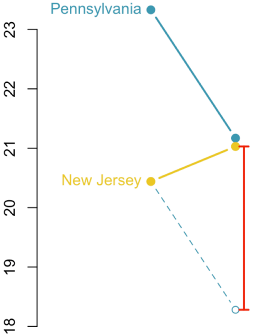
	pre	post
control	a (never treated)	b (never treated)
treatment	c (not yet treated)	d (treated)

	pre	post	Δ
Pennsylvania	$a = 23.33$	$b = 21.17$	$a - b$
New Jersey	$c = 20.44$	$d = 21.03$	$c - d$

Causal effect

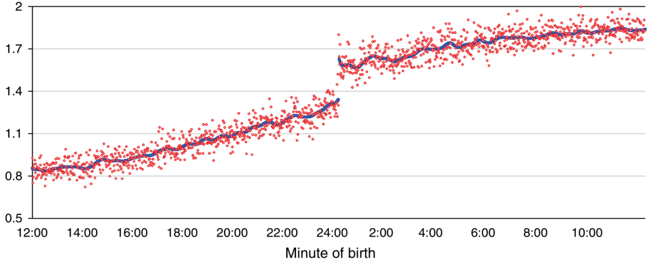
$$\Delta = \begin{cases} (d-c)-(b-a) = \quad (0.59)-(-2.16) = 2.76 \\ (d-b)-(c-a) = (-2.89)-(-0.14) = 2.76 \end{cases}$$

	pre	post
control	a	$a+\beta$
treatment	$a+\gamma$	$a+\gamma+\beta+\Delta$



Cas 2b - Regression Discontinuity

"We find that additional health insurance coverage induces substantial extensions in length of hospital stay for mother and newborn. However, remaining in the hospital longer has no effect on readmissions or mortality, and the estimates are precise. Our results suggest that for uncomplicated births, minimum insurance mandates incur substantial costs without detectable health benefits. ," Almond and Doyle Jr (2011)



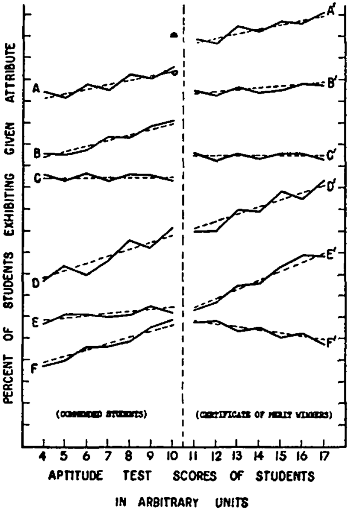


FIG. 1. Hypothetical outcomes of a regression-discontinuity analysis.

Cas 2b - Regression Discontinuity

Does extra time in the hospital improve health outcomes?

See also Howe et al. (2016), that estimate the effect of playing Pokémon GO on the number of steps taken daily up to six weeks after installation of the game.

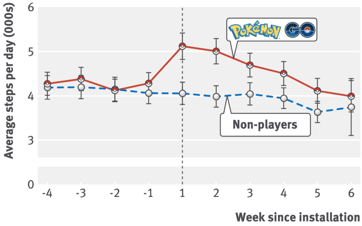


Fig 1 | Average number of daily steps and 95% confidence intervals by week before and after installation of Pokémon GO (median 8 July 2016)

"Regression discontinuity design" (RDD), Thistlethwaite and Campbell (1960) or Imbens and Lemieux (2008)

See also Imai (2022)

Case 3 - Potential Outcomes and counterfactuals

	Gender	Name	Treatment			Outcome (Weight)				Height	⋯
			t_i	0	1	y_i	$y_i^\star(0)$	$y_i^\star(1)$	TE	x_i	⋯
1	H	Alex	0	☑	☐	75	75	?	?	172	⋯
2	F	Betty	1	☐	☑	52	?	52	?	161	⋯
3	F	Beatrix	1	☐	☑	57	?	57	?	163	⋯
4	H	Ahmad	0	☑	☐	78	78	?	?	183	⋯

Treatment effect is

$$TE_i = y_i(1) - y_i(0) \text{ but in real life } TE_i = \begin{cases} y_i(1) - ??? \\ ??? - y_i(0) \end{cases}$$

Individual-level effects are impossible to observe! There are no individual counterfactuals.

Case 3 - Potential Outcomes and counterfactuals

Consider averages ?

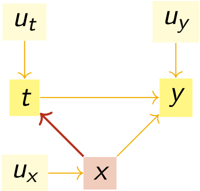
$$\overline{TE} = \overline{Y}(1) - \overline{Y}(0)$$

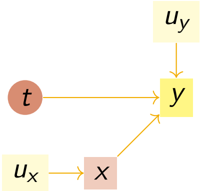
Comparing average outcomes only works if groups that received/didn't receive treatment look the same

See Causal model from Neyman-Rubin Neyman (1923), Rubin (1973, 1974), see also Sekhon (2009) and textbooks Angrist and Pischke (2009, 2014).

Case 3 - Potential Outcomes and Counterfactuals

Sewall Wright (see Wright (1921a,b, 1934)) introduced directed graphs to represent probabilistic cause and effect relationships among a set of variables





When you $do(t)$, delete all arrows into t confounders don't influence treatment.

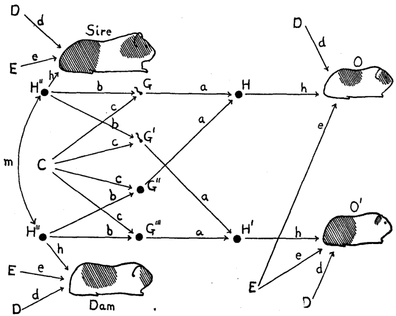


FIGURE 2.—A diagram illustrating the relations between two mated individuals and their progeny. H, H', H'' and H''' are the genetic constitutions of the four individuals. G, G', G'' and G''' are four germ-cells. E and D represent tangible external conditions and chance irregularities as factors in development. C represents chance at segregation as a factor in determining the composition of the germ-cells. Path coefficients are represented by small letters.

Case 4 - “matching”

- Nearest neighbor matching (NN)

Nearest neighbor matching (1-1)
Find untreated observations that are very close/similar to treated observations based on confounders
Lots of mathy ways to measure distance
Use Optimal Transport instead

- Inverse probability weighting (IPW)

Predict the probability of assignment to treatment using a model (logistic regression, probit regression, machine learning)
Then use propensity scores to weight observations by how ”weird” they are
Observations with high probability of treatment who don’t get it (and vice versa) have higher weight

Non-Independence and Causal Graphs

Definition 3.1: Directed acyclic graph, DAG (or causal graph)

A directed acyclic graph (DAG) $\mathcal{G}$ is a directed graph with no directed cycles.

Definition 3.2: 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 $\boldsymbol{x}_c$

$$\boldsymbol{x}_1 \perp_{\mathcal{G}} \boldsymbol{x}_2 \mid \boldsymbol{x}_c \;\Rightarrow\; \boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2 \mid \boldsymbol{X}_c.$$

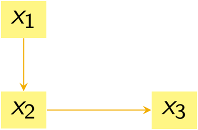
Non-Independence and Causal Graphs

Proposition 3.1: Probabilistic graphical model

If $\boldsymbol{X}$ satisfies the (global) Markov property with respect to $\mathcal{G}$

$$\mathbb{P}[x_1,\cdots,x_n]=\prod_{i=1}^{n}\mathbb{P}[x_i|\text{parents}(x_i)]$$

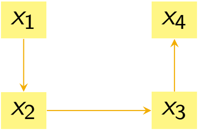
where parents(x_i) are nodes with edges directed towards x_i



Path from x_1 to x_3 is blocked by x_2, i.e., $x_1 \perp_{\mathcal{G}} x_3 \mid x_2$, or $X_1 \perp\!\!\!\perp X_3 \mid X_2$. From the chain rule,

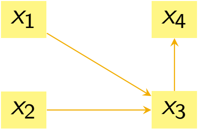
$$\mathbb{P}[x_1,x_2,x_3]=\mathbb{P}[x_1]\times\mathbb{P}[x_2|x_1]\times\underbrace{\mathbb{P}[x_3|x_2,x_1]}_{\mathbb{P}[x_3|x_2]}$$

Non-Independence and Causal Graphs



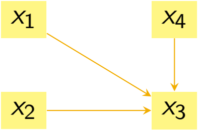
From the chain rule, for the causal graph on the left (top),

$$\mathbb{P}[x_1, x_2, x_3, x_4] = \mathbb{P}[x_1] \times \mathbb{P}[x_2|x_1] \times \mathbb{P}[x_3|x_2] \times \mathbb{P}[x_4|x_3]$$



From the chain rule, for the causal graph on the left (middle),

$$\mathbb{P}[x_1, x_2, x_3, x_4] = \mathbb{P}[x_1] \times \mathbb{P}[x_2] \times \mathbb{P}[x_3|x_1, x_2] \times \mathbb{P}[x_4|x_3]$$



From the chain rule, for the causal graph on the left (bottom),

$$\mathbb{P}[x_1, x_2, x_3, x_4] = \mathbb{P}[x_1] \times \mathbb{P}[x_2] \times \mathbb{P}[x_3|x_1, x_2, x_4] \times \mathbb{P}[x_4]$$

Non-Independence and Causal Graphs

$\mathbb{P}[Y \in \mathcal{A} | X = x]$: how $Y \in \mathcal{A}$ is likely to occur if X happened to be equal to x
Therefore, it is an observational statement.

$P[Y \in \mathcal{A} | \text{do}(X = x)]$: how $Y \in \mathcal{A}$ is likely to occur if X is set to x
It is here an intervention statement.

Non-Independence and Causal Graphs

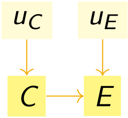
Definition 3.3: Structural Causal Models (SCM)

In a simple causal graph, with two nodes C (the cause) and E (the effect), the causal graph is $C \to E$, and the mathematical interpretation can be summarized in two assignments

$$\begin{cases} C = h_c(U_C) \\ E = h_e(C, U_E), \end{cases}$$

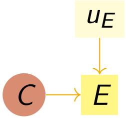
where U_C and U_E are two independent random variables, $U_C \perp\!\!\!\perp U_E$.

(a) observation



$$\begin{cases} C = h_c(U_C) \\ E = h_e(C, U_E) \end{cases}$$

(b) intervention



$$\begin{cases} C = c \quad (\text{or do}(C = c)) \\ E_c^\star = h_e(c, U_E) \end{cases}$$

Causal Inference & Observational Data

- Propensity score

> The "propensity" describes how likely a unit is to have been treated, given its covariate values. The stronger the confounding of treatment and covariates, and hence the stronger the bias in the analysis of the naive treatment effect, the better the covariates predict whether a unit is treated or not. By having units with similar propensity scores in both treatment and control, such confounding is reduced. W

"The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates," Rosenbaum and Rubin (1983)

Suppose observed data are $\{(\boldsymbol{x}_i, a_i, y_i)\}_{i=1}^n$ drawn i.i.d (independent and identically distributed) from unknown distribution $\mathbb{P}$, where $A \in \{0, 1\}$, denotes either "control" (placebo) or "treated" (medicine).

Let $Y(a)$ (or $Y(x, a)$) denote "potential outcomes" (under control and treatment),

Causal Inference & Observational Data

In many application, the quantity of interest is TE (or $TE(x)$) the treatment effect, $TE = Y(1) - Y(0)$

	Gender	Name	Treatment			Outcome (Weight)				Height	$\cdots$
			t_i	0	1	y_i	$y^\star_{i,T\leftarrow 0}$	$y^\star_{i,T\leftarrow 1}$	TE	x_i	$\cdots$
1	H	Alex	0	☑	☐	75	75	64	11	172	$\cdots$
2	F	Betty	1	☐	☑	52	67	52	15	161	$\cdots$
3	F	Beatrix	1	☐	☑	57	71	57	14	163	$\cdots$
4	H	Ahmad	0	☑	☐	78	78	61	17	183	$\cdots$

Different notations are used $y(1)$ 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).
When $a_i = 1$ is observed, and $\boldsymbol{x}_i$,

$$\begin{cases} \text{observation} & : \; y_i(1) \\ \text{counterfactual} & : \; y_i(0) \end{cases}$$

Causal Inference & Observational Data

Following Holland (1986), given a "treatment" T (here A), the average treatment effect on outcome y is

$$\tau = \text{ATE} = \mathbb{E}\big[Y(1) - Y(0)\big],$$

and following Wager and Athey (2018), given a treatment a, the conditional average treatment effect on outcome y, given some covariates $\boldsymbol{x}$,is

$$\tau(\boldsymbol{x}) = \text{CATE}(\boldsymbol{x}) = \mathbb{E}\big[Y(1) - Y(0)\big|\boldsymbol{X} = \boldsymbol{x}\big].$$

Causal Inference & Observational Data

Given a dataset, $(y_i, a_i, \boldsymbol{x}_i)$, the sample average treatment effect on outcome y is

$$\widehat{\tau} = \text{SATE} = \frac{1}{n}\sum_{i=1}^{n}\big[y_i(1) - y_i(0)\big] = \frac{1}{n}\sum_{i=1}^{n} y_i(1) - \frac{1}{n}\sum_{i=1}^{n} y_i(0),$$

difference in the average outcome between two scenarios: everyone is treated vs. nobody is treated

the sample average treatment effect for the treated on outcome y is

$$\text{SATT} = \frac{1}{n_1}\sum_{i=1}^{n} t_i\big[y_i(1) - y_i(0)\big]$$

which is the sample version of $\mathbb{E}\big[Y(1) - Y(0)\big| T = 1\big]$

Causal Inference & Randomized Experiments

Classical regression, $\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}]=\mu(\boldsymbol{X})=\boldsymbol{x}^\top\boldsymbol{\beta}$
Can we interpret coefficients as causal effects?
Suppose $Y_i(t)=\alpha+\beta t+\varepsilon_i$, $t\in\{0,1\}$ and $\mathbb{E}[\varepsilon]=0$. Here

$$Y_i(1)-Y_i(0)=\beta,\ \forall i,\ \text{i.e. constant additive unit causal effect}$$

Suppose heterogeneous treatment effect, $Y_i(t)=\alpha+\beta_i t+\varepsilon_i$,

$$Y_i(t)=\alpha+\beta t+\big(\beta_i-\beta\big)\cdot t+\varepsilon_i$$

$$\mathbb{E}\big[Y_i(1)-Y_i(0)\big]=\mathbb{E}\big[\beta_i\big]=\beta,\ \forall i,$$

Strict exogeneity assumption, $\mathbb{E}\big[\varepsilon_i\big|T_1,\cdots T_n\big]=0$
The least squares estimate $\widehat{\beta}$ is unbiased for β
Randomization of treatment: $(Y_i(0),Y_i(1))\perp\!\!\!\perp T_1\cdots,T_n\ \forall i$
Random sampling of units: $(Y_i(0),Y_i(1))$ i.i.d.

Causal Inference & Randomized Experiments

Least squares estimators are

$$\widehat{\alpha} = \frac{1}{n_0}\sum_{i=1}^{n}(1-t_i)\cdot y_i \text{ and } \widehat{\beta} = \frac{1}{n_1}\sum_{i=1}^{n} t_i\cdot y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-t_i)\cdot y_i$$

that are unbiased estimators of $\mathbb{E}[Y(0)]$ and $\mathbb{E}[Y(1)-Y(0)]$
What about the variance ?
In the homoskedastic case, $\text{Var}[\boldsymbol{\varepsilon}|T_1,\cdots,T_n]=\sigma^2\mathbb{I}$ then

$$\text{Var}[\widehat{\beta}|T_1,\cdots,T_n] = n\frac{\sigma^2}{\text{Var}[T]}$$

In the heteroskedastic case, $\text{Var}[\varepsilon|T=t]=\sigma_t^2$ then the estimated variance is biased,

$$\text{bias} = \mathbb{E}\left[n\frac{\widehat{\sigma}^2}{\text{Var}[T]}\right] - \left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_0^2}{n_0}\right)$$

Causal Inference & Randomized Experiments

left = under complete randomization / right = true variance
bias is zero when homoskedasticity assumption holds, and design is balanced ($n_0 = n_1$

Causal Inference & Observational Data

In observational data, there is no randomized treatment assignment,

$$(Y(0), Y(1)) \not\perp\!\!\!\perp A, \quad \text{confounding}$$

but the treatment assignment mechanism is often unknown
(probably observed and unobserved confounders)

- Identification: How much can you learn about the estimand if you had an infinite amount of data?
- Statistical Inference: How much can you learn about the estimand from a finite sample?

Classical assumptions for identification

- Identification: How much can you learn about the estimand if you had an infinite amount of data?
- Statistical Inference: How much can you learn about the estimand from a finite sample?

Causal Inference & Observational Data

- Strongly ignorable treatment assignment

Treatment assignment is said to be strongly ignorable if the potential outcomes are independent of treatment (A) conditional on background variables $\boldsymbol{X}$

$$(Y(0), Y(1)) \perp\!\!\!\perp A \mid \boldsymbol{X}$$

- Balancing score

Following Rubin (1973, 1974), a balancing score $b(\boldsymbol{X})$ is a function of the observed covariates $\boldsymbol{X}$ such that the conditional distribution of $\boldsymbol{X}$ given $b(\boldsymbol{X})$ is the same for treated ($A = 1$) and control ($A = 0$) units

$$A \perp\!\!\!\perp \boldsymbol{X} \mid b(\boldsymbol{X})$$

Causal Inference & Observational Data

- Propensity score

$$e(\boldsymbol{x}) = \mathbb{P}(A = 1|\boldsymbol{A} = \boldsymbol{x})$$

As proved in Rosenbaum and Rubin (1983),

- the propensity score $e(\boldsymbol{x})$ is a balancing score
- if treatment assignment is strongly ignorable given $\boldsymbol{x}$ then, it is also strongly ignorable given any balancing function (specifically, given the propensity score)

$$\big(Y(0), Y(1)\big) \perp\!\!\!\perp A \mid e(\boldsymbol{X}).$$

Causal Inference & Observational Data

- Horvitz -Thompson theory

> One very early weighted estimator is the Horvitz–Thompson estimator of the mean. When the sampling probability is known, from which the sampling population is drawn from the target population, then the inverse of this probability is used to weight the observations. This approach has been generalized to many aspects of statistics under various frameworks. In particular, there are weighted likelihoods, weighted estimating equations, and weighted probability densities from which a majority of statistics are derived. W

Suppose observed data are $\{(\boldsymbol{X}_i, A_i, Y_i)\}_{i=1}^{n}$ drawn i.i.d (independent and identically distributed) from unknown distribution $\mathbb{P}$, where $A \in \{0, 1\}$.

Causal Inference & Observational Data

Suppose observed data are $\{(\boldsymbol{X}_i, A_i, Y_i)\}_{i=1}^n$ drawn i.i.d (independent and identically distributed) from unknown distribution $\mathbb{P}$, where $A \in \{0, 1\}$.

On can derive an Inverse Probability Weighted Estimator (*IPWE*)

- $\mu_a = \mathbb{E}\left[\dfrac{\mathbf{1}_{A=a}Y}{p(A=a|\boldsymbol{X})}\right]$ where $p(a|\boldsymbol{x}) = \mathbb{P}(A=a|\boldsymbol{X}=\boldsymbol{x}) = \dfrac{\mathbb{P}(A=a,\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(\boldsymbol{X}=\boldsymbol{x})}$
- estimate $p(a|\boldsymbol{x})$ with $\hat{p}_n(a|\boldsymbol{x})$, using any propensity model (e.g., logistic regression model)
- $\hat{\mu}_{a,n}^{IPWE} = \dfrac{1}{n}\sum_{i=1}^n \dfrac{y_i \mathbf{1}_{A_i=a}}{\hat{p}_n(a_i|\boldsymbol{x}_i)}$

Causal Inference & Observational Data

We make the following assumptions.

(A1) Consistency: $Y = Y(A)$

(A2) No un-measured confounders: $\{Y(0), Y(1)\} \perp\!\!\!\perp A|\boldsymbol{X}$.

More formally, for each bounded and measurable functions f and g,

$$\mathbb{E}_{(A,Y)}\left[f(Y(\boldsymbol{X},A))\,g(A)\mid\boldsymbol{X}\right] = \mathbb{E}_{Y}\left[f(Y(\boldsymbol{X},A))\mid\boldsymbol{X}\right]\cdot\mathbb{E}_{A}\left[g(A)\mid\boldsymbol{X}\right].$$

This means that treatment assignment is based solely on covariate data and independent of potential outcomes.

(A3) Positivity: $\mathbb{P}(A=a|\boldsymbol{X}=\boldsymbol{x}) = \mathbb{E}_A[\mathbf{1}(A=a)\mid\boldsymbol{X}=\boldsymbol{x}] > 0$ for all a and $\boldsymbol{x}$.

Causal Inference & Observational Data

$$\mathbb{E}\left[Y^*(a)\right] \overset{\text{from (A1)}}{=} \mathbb{E}_{(X,Y)}\left[Y(X,a)\right] = \mathbb{E}_{(X,A,Y)}\left[\frac{Y\mathbf{1}(A=a)}{P(A=a|X)}\right]$$

$$\mathbb{E}_{(X,Y)}\left[Y(X,a)\right] = \mathbb{E}_X\left[\mathbb{E}_Y\left[Y(X,a)\,|\,X\right]\right].$$

then simply (by (A3) $\mathbb{E}_A[\mathbf{1}(A=a)\,|\,\boldsymbol{X}] > 0$)

$$\mathbb{E}_Y\left[Y(\boldsymbol{X},a)\,|\,X\right] = \frac{\mathbb{E}_Y\left[Y(\boldsymbol{X},a)\,|\,X\right]\,\mathbb{E}_A[\mathbf{1}(A=a)\,|\,\boldsymbol{X}]}{\mathbb{E}_A[\mathbf{1}(A=a)\,|\,\boldsymbol{X}]} = \frac{\mathbb{E}_{(A,Y)}\left[Y(\boldsymbol{X},a)\mathbf{1}(A=a)\,|\,\boldsymbol{X}\right]}{\mathbb{E}[\mathbf{1}(A=a)\,|\,\boldsymbol{X}]}$$

i.e.

$$\mathbb{E}_Y\left[Y(\boldsymbol{X},a)\,|\,\boldsymbol{X}\right] = \mathbb{E}_{(A,Y)}\left[\frac{Y(\boldsymbol{X},a)\mathbf{1}(A=a)}{\mathbb{E}[\mathbf{1}(A=a)\,|\,\boldsymbol{X}]}\,\middle|\,\boldsymbol{X}\right]$$

The Inverse Probability Weighted Estimator (*IPWE*) is known to be unstable if some estimated propensities are too close to 0 or 1 (see calibration issues).

Causal Inference & Observational Data

Augmented Inverse Probability Weighted Estimator ($AIPWE$), Cao et al. (2009)

$$
\begin{aligned}
\hat{\mu}^{AIPWE}_{a,n} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i \mathbf{1}_{A_i=a}}{\hat{p}_n(A_i|X_i)} - \frac{\mathbf{1}_{A_i=a}-\hat{p}_n(A_i|X_i)}{\hat{p}_n(A_i|X_i)}\hat{Q}_n(X_i,a)\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbf{1}_{A_i=a}}{\hat{p}_n(A_i|X_i)}Y_i + (1-\frac{\mathbf{1}_{A_i=a}}{\hat{p}_n(A_i|X_i)})\hat{Q}_n(X_i,a)\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Q}_n(X_i,a)\right) + \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{1}_{A_i=a}}{\hat{p}_n(A_i|X_i)}\left(Y_i - \hat{Q}_n(X_i,a)\right)
\end{aligned}
$$

here we need a regression estimator $\hat{Q}_n(\boldsymbol{x},a)$ to predict outcome Y based on covariates $\boldsymbol{X}$ and treatment A, for some subject i.

This approach is said to by "doubly robust" (with a second order bias)

Post Stratification and Weights

Inspired from techniques used in sampling theory, use post-stratification techniques, which is standard when dealing with a "biased sample".

The regression function is defined a

$$\mu(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{X}=\boldsymbol{x}] = \mathbb{E}\big[\mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{X}=\boldsymbol{x},A]\big] = \int_{\mathcal{A}} \mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{X}=\boldsymbol{x},A=a]\mathrm{d}\mathbb{P}[A=a].$$

Following Moodie and Stephens (2022), the later can be written

$$\mu(\boldsymbol{x}) = \int_{\mathcal{A}} \mathbb{E}_{\mathbb{P}}[Y\cdot W|\boldsymbol{X}=\boldsymbol{x},A=a]\mathrm{d}\mathbb{P}[A=a|\boldsymbol{X}=\boldsymbol{x}] = \mathbb{E}_{\mathbb{P}}[Y\cdot W|\boldsymbol{X}=\boldsymbol{x}],$$

where W is a version of the Radon-Nikodym derivative

$$W = \frac{\mathrm{d}\mathbb{P}[A=a]}{\mathrm{d}\mathbb{P}[A=a|\boldsymbol{X}=\boldsymbol{x}]},$$

corresponding to the change of measure that will give independence between $\boldsymbol{X}$ and A.

Post Stratification and Weights

- Properties of W

We have the following interesting property: let W be a version of the Radon-Nikodym derivative

$$W = \frac{\mathrm{d}\mathbb{P}[A=a]}{\mathrm{d}\mathbb{P}[A=a|\boldsymbol{X}=\boldsymbol{x}]},$$

then $\mathbb{E}_{\mathbb{P}}[W]=1$, $\mathbb{E}_{\mathbb{P}}[A\cdot W]=\mathbb{E}_{\mathbb{P}}[A]$ and $\mathbb{E}_{\mathbb{P}}[\boldsymbol{X}\cdot W]=\mathbb{E}_{\mathbb{P}}[\boldsymbol{X}]$.
As proved in Fong et al. (2018),

$$\mathbb{E}_{\mathbb{P}}[W]=\iint w\mathrm{d}\mathbb{P}[A=a,\boldsymbol{X}=\boldsymbol{x}]=\iint w\mathrm{d}\mathbb{P}[A=a|\boldsymbol{X}=\boldsymbol{x}]\mathrm{d}\mathbb{P}[\boldsymbol{X}=\boldsymbol{x}]$$

that can be written

$$\mathbb{E}_{\mathbb{P}}[W]=\iint \frac{\mathrm{d}\mathbb{P}[A=a]}{\mathrm{d}\mathbb{P}[A=a|\boldsymbol{X}=\boldsymbol{x}]}\mathrm{d}\mathbb{P}[A=a|\boldsymbol{X}=\boldsymbol{x}]\mathrm{d}\mathbb{P}[\boldsymbol{X}=\boldsymbol{x}],$$

Post Stratification and Weights

and therefore

$$\mathbb{E}_{\mathbb{P}}[W] = \iint \mathrm{d}\mathbb{P}[A = a]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}] = 1.$$

Similarly

$$\mathbb{E}_{\mathbb{P}}[A \cdot W] = \iint sw\mathrm{d}\mathbb{P}[A = a, \boldsymbol{X} = \boldsymbol{x}] = \iint sw\mathrm{d}\mathbb{P}[A = a|\boldsymbol{X} = \boldsymbol{x}]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}],$$

and

$$\mathbb{E}_{\mathbb{P}}[A \cdot W] = \iint s\mathrm{d}\mathbb{P}[A = a]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}] = \int \mathbb{E}_{\mathbb{P}}[S]\mathrm{d}\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}_{\mathbb{P}}[S].$$

In statistics, this Radon-Nikodym derivative is related to the propensity score, as discussed in Freedman and Berk (2008), Li and Li (2019) and Karimi et al. (2022).

What is Interpretability in Machine Learning?

- Interpretability refers to the ability to understand the internal workings of a machine learning model.
- A model is interpretable if a human can understand why it makes a certain prediction or decision.
- Example: In a decision tree, we can trace the path from the root to a leaf node to see how a prediction is made.
- Key Question: How do we interpret the model's decision-making process?
- Interpretability is crucial for trust, debugging, and ensuring ethical use of ML models.
- Explainability is the process of providing human-understandable explanations for a model's prediction.
- Unlike interpretability, explainability does not necessarily mean understanding the inner workings of the model, but rather being able to explain its output in a way that makes sense to users.

What is Interpretability in Machine Learning?

- Example: In deep learning, an explanation could be highlighting the most important features for a given prediction, using techniques like LIME or SHAP.
- Key Question: How do we explain a model's prediction to a non-expert user?
- Explainability is critical for building trust, accountability, and fairness in AI systems.
- Trust: Users are more likely to trust models that provide clear, understandable reasons for their decisions.
- Accountability: In high-stakes domains like healthcare, finance, or law, understanding how a model arrived at a decision is crucial for accountability.
- Bias Detection: Transparent models help detect and correct biases in predictions.
- Regulation: Increasingly, governments and organizations are requiring explanations for automated decisions (e.g., the -GDPR- "right to explanation").
- Model Debugging: Interpretability helps developers understand and fix issues with models, especially when they make unexpected decisions.

What is Interpretability in Machine Learning?

- Interpretability:
  - Focuses on how easily we can understand the -internal mechanics- of a model.
  - Example: Linear regression has high interpretability because we can easily inspect the coefficients.
- Explainability:
  - Focuses on -explaining the output- of a model in human-understandable terms.
  - Example: A neural network's output can be explained using techniques like -LIME- or -SHAP-, which provide local explanations.
- While these concepts overlap, interpretability is about the model itself, while explainability is about making its outputs accessible to humans.
- Tradeoff-: Often, more complex models (e.g., deep neural networks) are less interpretable but can be made more explainable through techniques.
- For Interpretable Models:

What is Interpretability in Machine Learning?

- Simple models like -decision trees-, -linear regression-, and -logistic regression- are inherently interpretable.
- Visual tools (e.g., -partial dependence plots-) help visualize how features influence model predictions.

- For Explainable Models:
  - LIME (Local Interpretable Model-agnostic Explanations): Explains individual predictions by approximating the model locally with simpler, interpretable models.
  - SHAP (Shapley Additive Explanations): Provides a unified measure of feature importance by distributing the "credit" for a prediction across features.
  - Feature Importance: Quantifies how much each feature contributes to the model's output.
- These methods make black-box models like deep learning more transparent without sacrificing predictive performance.

Interpretability

"On a collection of additional 60 images, the classifier predicts "Wolf" if there is snow (or light background at the bottom), and "Husky" otherwise, regardless of animal color, position, pose, etc.," Ribeiro et al. (2016)

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

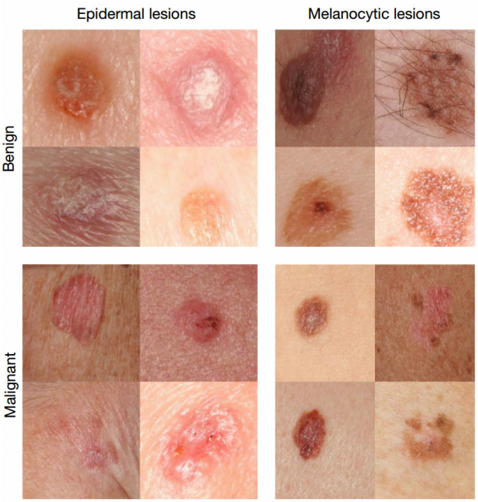
	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

Interpretability

Esteva et al. (2017) and Winkler et al. (2019) for skin cancer detection classifiers based on deep neural networks

"*So in the set of biopsy images, if an image had a ruler in it, the algorithm was more likely to call a tumor malignant, because the presence of a ruler correlated with an increased likelihood a lesion was cancerous*," Daily Beast (2017)



Interpretability

Using https://cloud.google.com/vision/, we have a "jaguar" (left) "lepoard" (right).

(see also Charpentier (2021))

Interpretability

Taxonomy of explainability

- Global vs. local: Describe model as a whole or around an observation.
- Model-specific vs. model-agnostic: Some methods are tailored to specific model classes (linear regression, tree-based), others work for all types of models.
- Intrinsic versus post-hoc: Simple models like a linear regression can be interpreted intrinsically, while complex models require post-hoc analysis of fitted model.
- Model-agnostic methods are always post-hoc
- Model-agnostic methods can also be applied to intrinsically interpretable models
- We won't make difference between "explainable", "interpretable", "intelligible"

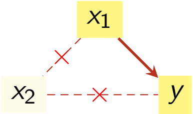
Interpretability

Definition 3.4: Ceteris paribus, Marshall (1890)

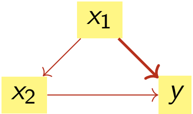
Ceteris paribus (or more precisely *ceteris paribus sic stantibus*) is a Latin phrase, meaning "all other things being equal" or "other things held constant."

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.
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).

(*ceteris paribus*)



(*mutatis mutandis*)



Interpretability

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 3.5: 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 $(X_1, X_2)^\top$

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

Interpretability

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

Interpretability

independent components and the same marginal distributions (in the sense that $X_2^\perp = X_2$, almost surely, and $X_1^\perp \overset{\mathcal{L}}{=} X_1$, and $X_1^\perp \perp\!\!\!\perp X_2$), then $\mathbb{E}_{Y|X_2^\perp|X_1^\perp}[Y|x_1^*] = \mu_2$, and

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

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.

Interpretability

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 *ceteris paribus* estimate of the effect of x_1 on the prediction is

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

(the average on the right being the empirical counterpart of the expected value on the left) 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 $\mathcal{V}_\epsilon(x_1^*) = \{i : |x_{i,1} - x_1^*| \leq \epsilon\}$ 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 "$\mathbb{E}_{Y|X_1^{\perp}}[m(X_1^{\perp}, X_2^{\perp})|x_1^*]$" do not not have

Interpretability

measure-theoretic foundations, but they will be useful to highlight that in some cases, metrics and mathematical objects "pretend" that explanatory variables are independent.

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,

Interpretability

Definition 3.6: VI_j or "variable permutation VI_j", Fisher et al. (2019)

Given a loss function ℓ and a model m, the importance of the j-th variable is

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

and the empirical version is

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

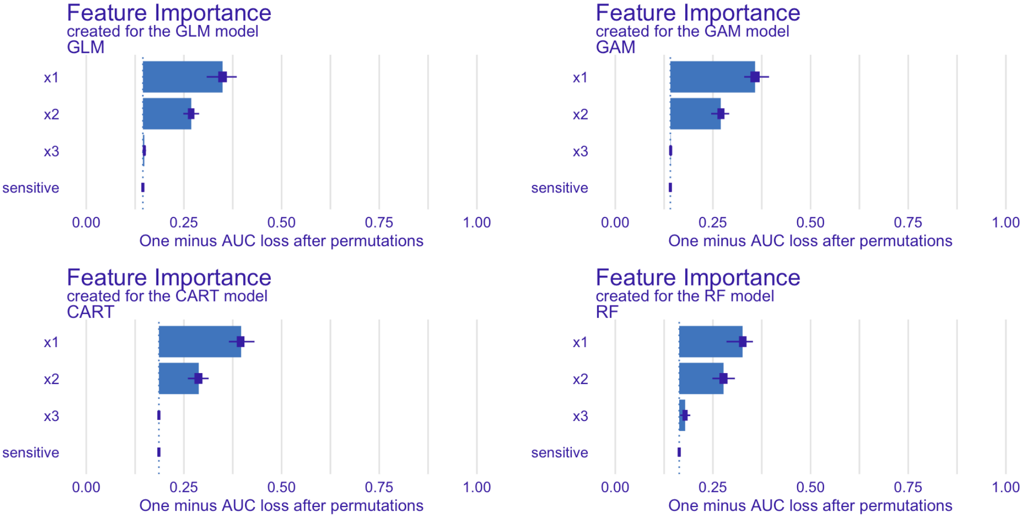
for some permutation $\tilde{\boldsymbol{x}}_j$ or $\boldsymbol{x}_j$.

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 the variable-importance function `variable_importance` from the `DALEX` package (see Biecek and Burzykowski (2021) for

Interpretability

more details). By default, the loss considered is the one associated with $1 - \text{AUC}$ 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.
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 as some "drop-out loss of AUC", and therefore, as a measure of variable importance. One could also use `FeatureImp` from the `iml` R package, based on Molnar (2023).

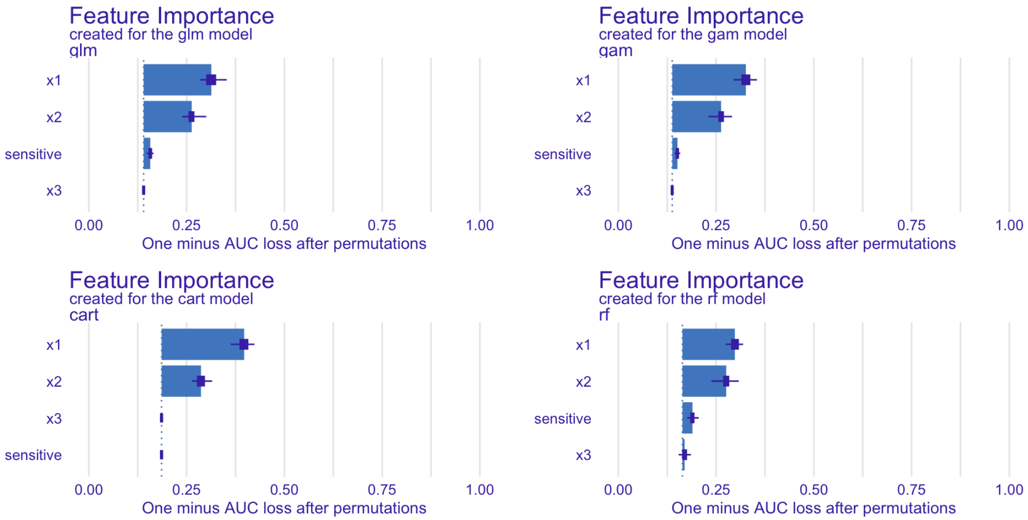
Interpretability



Interpretability

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

Interpretability



Variable importance for different models trained on `toydata2`, with the sensitive attribute s, with four variables, x_1, x_2, x_3 and s.
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, coined "*ceteris-paribus profile*" in Biecek and Burzykowski (2021),

Definition 3.7: Ceteris Paribus profile $z \mapsto m_{\boldsymbol{x}^*,j}(z)$ Goldstein et al. (2015)

Given $\boldsymbol{x}^* \in \mathcal{X}$, define on $\mathcal{X}_j$

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

Interpretability

Here, it is a *ceteris-paribus* profile in the sense that x_j^* changes (and takes variable value z) while all other components remain unchanged. 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 3.8: $dm_j^{\text{cp}}(\boldsymbol{x}^*)$

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

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

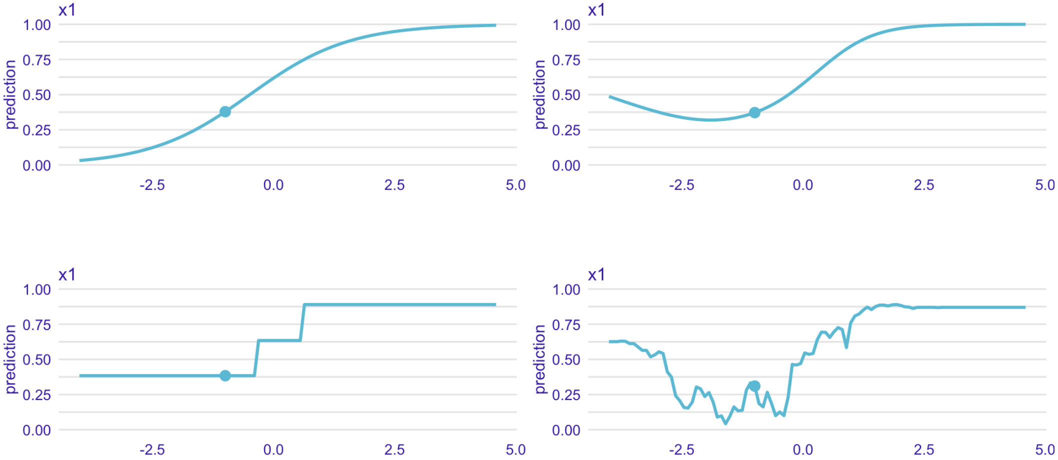
Definition 3.9: $\widehat{dm}_j^{\text{cp}}(\boldsymbol{x}^*)$

The empirical mean absolute deviation associated with the j-th variable, at $\boldsymbol{x}^*$, is

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

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, $z \mapsto m_{\boldsymbol{x}^*,1}(z)$.
$z \mapsto m_{\boldsymbol{x}^*,1}(z)$ associated with `Andrew` (when $(\boldsymbol{x}^\star, s^\star) = (-1, 8, -2, \texttt{A})$) and
$z \mapsto m_{\boldsymbol{x}^*,1}(z)$ associated with `Barbara` (when $(\boldsymbol{x}^\star, s^\star) = (1, 4, 2, \texttt{B})$). Bullet points indicate the values $m_{\boldsymbol{x}^*,1}(x_1^*)$ for `Andrew` and `Barbara`. 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.

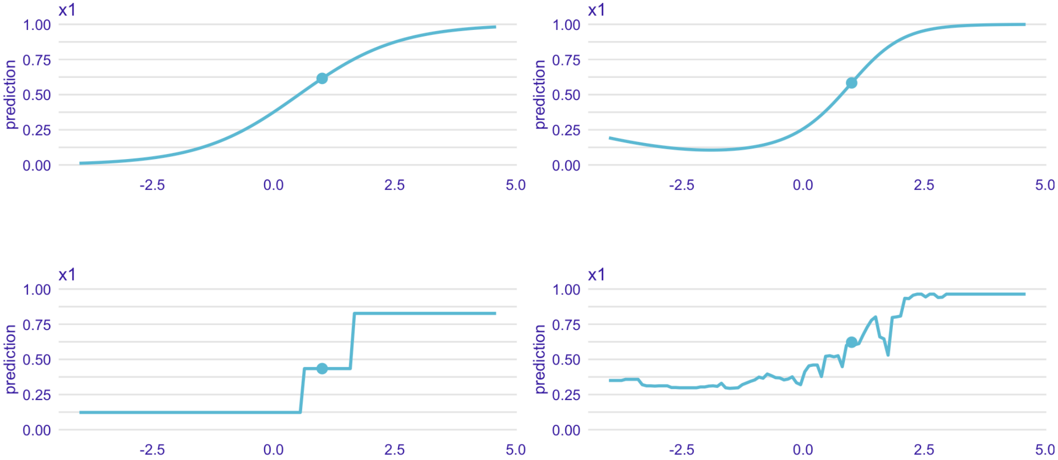
Interpretability



Interpretability

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

Interpretability



Interpretability

"*ceteris-paribus profiles*" for `Barbara` for different models trained on `toydata2`, here $\boldsymbol{z}^\star = (\boldsymbol{x}^\star, s^\star) = (1, 4, 2, \texttt{B})$.
For a standard linear model, observe that we can write

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

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 $\boldsymbol{x}^*$

$$v_j(\boldsymbol{x}^*) = m(x_1^*, \cdots, , x_{j-1}^*, x_j^*, x_{j+1}^*, \cdots, x_k^*) - \mathbb{E}_{X_j^\perp}[m(x_1^*, \cdots, x_{j-1}^*, X_j, x_{j+1}^*, \cdots, x_k^*)],$$

Interpretability

so that

$$m(\boldsymbol{x}^*) = \mathbb{E}\big[m(\boldsymbol{X})\big] + \sum_{j=1}^{k} v_j(\boldsymbol{x}^*),$$

and for the linear model $v_j(\boldsymbol{x}^*) = \beta_j(x_j^* - \mathbb{E}_{X_j^\perp|\boldsymbol{X}_{-j}}[X_j^\perp|\boldsymbol{X}_{-j} = \boldsymbol{x}_{-j}^*])$, and $\widehat{v}_j(\boldsymbol{x}^*) = \widehat{\beta}_j(x_j^* - \overline{x}_j)$.
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.,

$$v_j(\boldsymbol{x}^*) = \begin{cases} \mathbb{E}\big[m(\boldsymbol{X})\big|x_1^*, \cdots, x_k^*\big] - \mathbb{E}_{X_j^\perp|\boldsymbol{X}_{-j}}\big[m(\boldsymbol{X})\big|x_1^*, \cdots, x_{j-1}^*, x_{j+1}^*, \cdots, x_k^*\big] \\ \mathbb{E}\big[m(\boldsymbol{X})\big|\boldsymbol{x}^*\big] - \mathbb{E}_{X_j^\perp|\boldsymbol{X}_{-j}}\big[m(\boldsymbol{X})\big|\boldsymbol{x}_{-j}^*\big]. \end{cases}$$

Interpretability

Definition 3.10: $\gamma_j^{\text{bd}}(x^*)$, Biecek and Burzykowski (2021)

The breakdown contribution of the j-th variable, at $\boldsymbol{x}^*$, is

$$\gamma_j^{\text{bd}}(\boldsymbol{x}^*) = v_j(\boldsymbol{x}^*) = \mathbb{E}\big[m(\boldsymbol{X})\big|\boldsymbol{x}^*\big] - \mathbb{E}_{X_j^\perp|\boldsymbol{X}_{-j}}\big[m(\boldsymbol{X})\big|\boldsymbol{x}^*_{-j}\big].$$

"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 $j-1$ variables equal to their values in $\boldsymbol{x}^*$," Biecek and Burzykowski (2021)

We can rewrite the contribution of the j-th variable, at $\boldsymbol{x}^*$,

$$v_j(\boldsymbol{x}^*) = \begin{cases} \mathbb{E}\big[m(\boldsymbol{X})\big|x_1^*,\cdots,x_k^*\big] - \mathbb{E}_{X_j^\perp|\boldsymbol{X}_{-j}}\big[m(\boldsymbol{X})\big|x_1^*,\cdots,x_{j-1}^*,x_{j+1}^*,\cdots,x_k^*\big] \\ \mathbb{E}\big[m(\boldsymbol{X})\big|\boldsymbol{x}^*\big] - \mathbb{E}_{X_j^\perp|\boldsymbol{X}_{-j}}\big[m(\boldsymbol{X})\big|\boldsymbol{x}^*_{-j}\big]. \end{cases}$$

Interpretability

Definition 3.11: $\Delta_{j|S}(\boldsymbol{x}^*)$

The contribution of the j-th variable, at $\boldsymbol{x}^*$, conditional on a subset of variables, $S \subset \{1, \cdots, k\} \backslash \{j\}$, is

$$\Delta_{j|S}(\boldsymbol{x}^*) = \mathbb{E}_{\boldsymbol{X}^{\perp}_{S\cup\{j\}}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}^*_{S\cup\{j\}}\right] - \mathbb{E}_{\boldsymbol{X}^{\perp}_{S}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}^*_{S}\right],$$

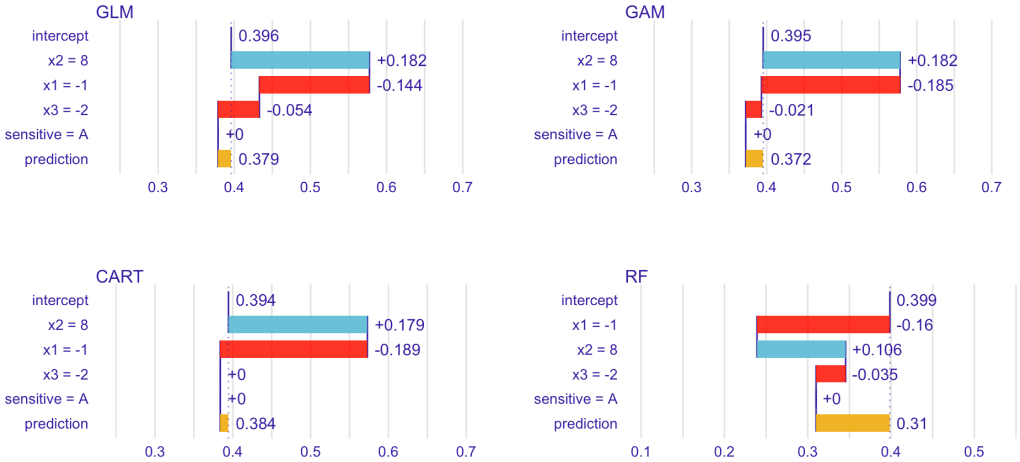
so that $v_j(\boldsymbol{x}^*) = \Delta_{j|\{1,2,\cdots,k\}\backslash\{j\}} = \Delta_{j|-j}$.

On the `toydata2` dataset, we can compute contributions of x_1, x_2 and x_3 for two individuals, `Andrew` and `Barbara`, using `type = "break_down"` in the `predict_parts` function of the `DALEX` R package. For `Andrew` 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

Interpretability

-0.19 to -0.14 on the prediction. Here s has no impact, since we consider models trained without the sensitive attribute.

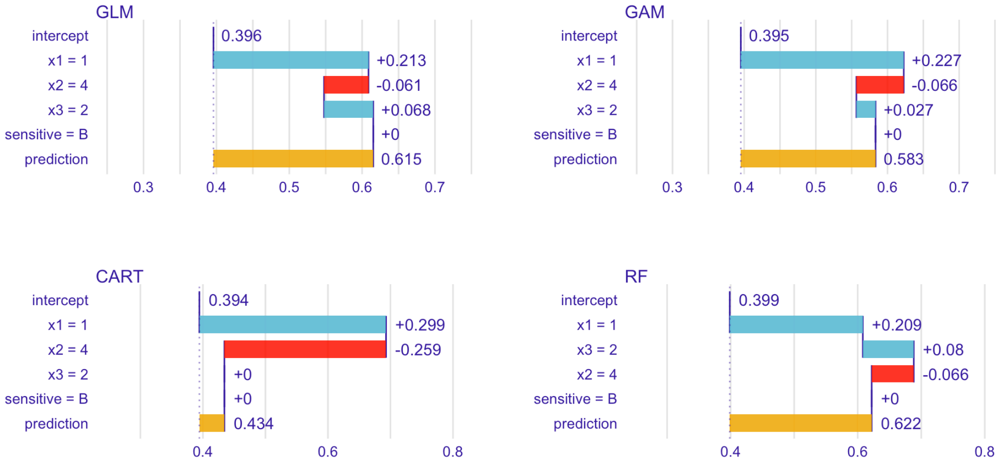
Interpretability



Interpretability

Breakdown decomposition $\widehat{\gamma}_j^{\text{bd}}(\boldsymbol{z}^*_{\texttt{A}})$ for `Andrew` for different models trained on `toydata2`, here $\boldsymbol{z}^\star_{\texttt{A}} = (\boldsymbol{x}^\star_{\texttt{A}}, s^\star) = (-1, 8, -2, \texttt{A})$.

Interpretability



Interpretability

Breakdown decomposition $\widehat{\gamma}_j^{\text{bd}}(\boldsymbol{z}_{\text{B}}^*)$ for `Barbara` for different models trained on `toydata2` (here $\boldsymbol{z}^\star = (\boldsymbol{x}^\star, s^\star) = (1, 4, 2, \texttt{B})$).

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

Interpretability

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 follows,

$$\phi_j(\mathcal{V}) = \frac{1}{k} \sum_{S\subseteq\{1,\dots,k\}\backslash\{j\}} \frac{|S|!\,(k-|S|-1)!}{k!} \left(\mathcal{V}\left(S\cup\{j\}\right) - \mathcal{V}(S)\right),$$

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, that we can rewrite

$$\phi_j(\mathcal{V}) = \frac{1}{\text{number of players}} \sum_{\text{coalitions including } j} \frac{\text{marginal contribution of } j \text{ to coalition}}{\text{number of coalitions excluding } j}.$$

Interpretability

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**": $\displaystyle\sum_{j=1}^{k}\phi_j(\mathcal{V})=\mathcal{V}(\{1,\dots,k\})$,
- "**symmetry**": if $\mathcal{V}(S\cup\{j\})=\mathcal{V}(S\cup\{j'\})\ \forall S$, then $\phi_j=\phi_{j'}$,
- "**dummy**" (or "**null player**"): if $\mathcal{V}(S\cup\{j\})=\mathcal{V}(S)\ \forall S$, then $\phi_j=0$,
- "**additivity**": if $\mathcal{V}^{(1)}$ and $\mathcal{V}^{(2)}$ have decomposition $\phi(\mathcal{V}^{(1)})$ and $\phi(\mathcal{V}^{(2)})$, then $\mathcal{V}^{(1)}+\mathcal{V}^{(2)}$ has decomposition $\phi(\mathcal{V}^{(1)}+\mathcal{V}^{(2)})=\phi(\mathcal{V}^{(1)})+\phi(\mathcal{V}^{(2)})$
- "**Linearity**" will be obtained if we add $\phi(\lambda\cdot\mathcal{V})=\lambda\cdot\phi(\mathcal{V})$.

Interpretability

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 $\phi_j(\mathcal{V}_{\boldsymbol{x}})$.
Here, we will use $\mathcal{V}_{\boldsymbol{x}^*}(S) = \mathbb{E}_{\boldsymbol{X}_S^\perp}\big[m(\boldsymbol{X})\big|\boldsymbol{x}_S^*\big]$, as value function, for any set S of variables, so that $\Delta_{j|S}(\boldsymbol{x}^*) = \mathcal{V}_{\boldsymbol{x}^*}(S \cup \{j\}) - \mathcal{V}_{\boldsymbol{x}^*}(S)$

Definition 3.12: Shapley contributions $\gamma_j^{\text{shap}}(\boldsymbol{x}^*)$

The Shapley contribution of the j-th variable, at $\boldsymbol{x}^*$, is

$$\gamma_j^{\text{shap}}(\boldsymbol{x}^*) = \frac{1}{k} \sum_{S \subseteq \{1,\dots,k\} \setminus \{j\}} \binom{k-1}{|S|}^{-1} \Delta_{j|S}(\boldsymbol{x}^*) = \phi_j(\mathcal{V}_{\boldsymbol{x}^*}).$$

Interpretability

Interestingly, for a linear regression with k uncorrelated features, and mean centered,

$$m(\boldsymbol{x}^*) = \underbrace{\beta_0}_{=\mathbb{E}\left[m(\boldsymbol{X})\right]} + \underbrace{\beta_1 x_1^*}_{\gamma_1^{\text{shap}}(\boldsymbol{x}^*)} + \underbrace{\beta_2 x_2^*}_{\gamma_2^{\text{shap}}(\boldsymbol{x}^*)} + \cdots + \underbrace{\beta_k x_k^*}_{\gamma_k^{\text{shap}}(\boldsymbol{x}^*)},$$

as discussed in Aas et al. (2021).
More generally, these contributions satisfy the following properties

- "**local accuracy**": $\displaystyle\sum_{j=1}^{k} \gamma_j^{\text{shap}}(\boldsymbol{x}^*) = m(\boldsymbol{x}^*) - \mathbb{E}\left[m(\boldsymbol{X})\right]$
- "**symmetry**": if j and k are interchangeable, $\gamma_j^{\text{shap}}(\boldsymbol{x}^*) = \gamma_k^{\text{shap}}(\boldsymbol{x}^*)$
- "**dummy**": if X_j does not contribute in the model, $\gamma_j^{\text{shap}}(\boldsymbol{x}^*) = 0$.

Interpretability

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, $k=2$, $\gamma_1^{\text{shap}}(\boldsymbol{x}^*)=\Delta_{1|2}(\boldsymbol{x}^*)=\gamma_1^{\text{bd}}(\boldsymbol{x}^*)$. 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 $\gamma_j^{\text{shap}}(\boldsymbol{x}^*)\approx m(\boldsymbol{x}_i^{*+})-m(\boldsymbol{x}_i^{*-})$, and therefore

$$\widehat{\gamma}_j^{\text{shap}}(\boldsymbol{x}^*)=\frac{1}{s}\sum_{i\in\{1,\cdots,n\}} m(\boldsymbol{x}_i^{*+})-m(\boldsymbol{x}_i^{*-}),$$

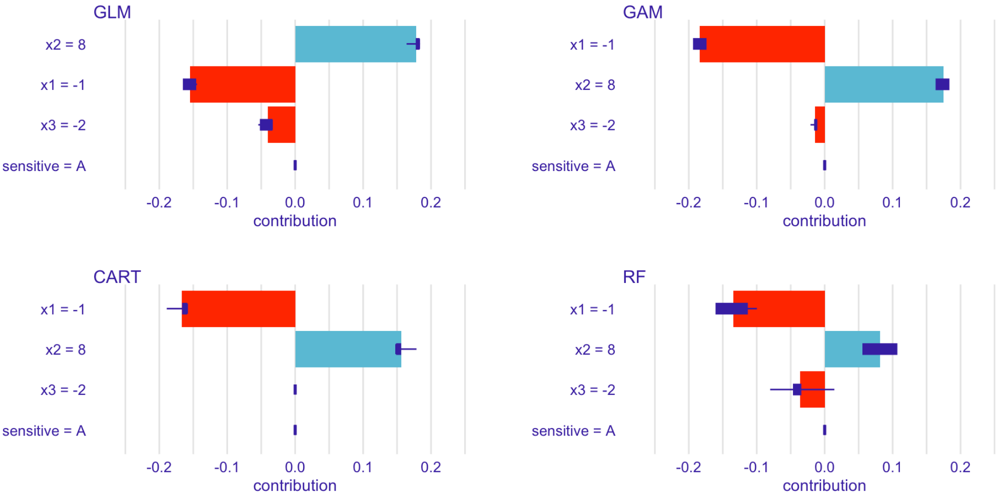
Interpretability

(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`), obtained using option `type = "shap"` in 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 opposite for `Barbara`.

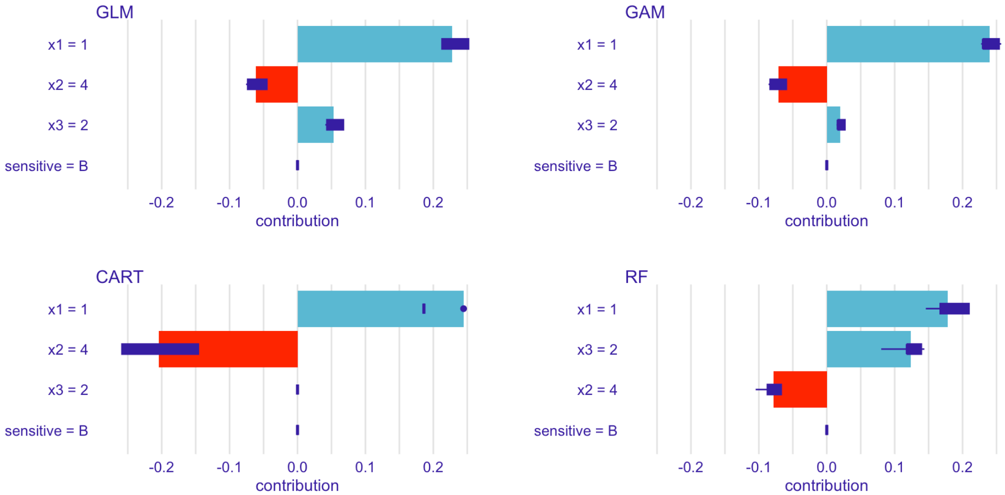
Interpretability



Interpretability

Shapley contributions $\widehat{\gamma}_j^{\text{shap}}(\boldsymbol{z}_{\texttt{A}}^\star)$ for `Andrew` for different models trained on `toydata2`, here $\boldsymbol{z}^\star = (\boldsymbol{x}^\star, s^\star) = (-1, 8, -2, \texttt{A})$.

Interpretability



Interpretability

Shapley contributions $\widehat{\gamma}_j^{\text{shap}}(\boldsymbol{z}_{\text{B}}^\star)$ for `Barbara` for different models trained on `toydata2`, here $\boldsymbol{z}^\star = (\boldsymbol{x}^\star, s^\star) = (1, 4, 2, \text{B}))$.
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 3.13: Shapley contribution $\overline{\gamma}_j^{\text{shap}}$

The contribution of the j-th variable is

$$\overline{\gamma}_j^{\text{shap}} = \frac{1}{n}\sum_{i=1}^{n} \gamma_j^{\text{shap}}(\boldsymbol{x}_i).$$

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,"

Interpretability

Definition 3.14: Shapley interaction $\gamma^{\text{shap}}_{i,j}(x^*)$

The interaction contribution between the i-th and the j-th variable, at $\boldsymbol{x}^*$, is

$$\gamma_{i,j}(\boldsymbol{x}^*) = \sum_{S\subseteq\{1,\dots,k\}\setminus\{i,j\}} \frac{|S|!\,(k-|S|-2)!}{2\ k!}\Delta_{i,j|S}(\boldsymbol{x}^*)$$

where

$$\begin{aligned}\Delta_{i,j|S}(\boldsymbol{x}^*) \quad &= \quad \mathbb{E}_{\boldsymbol{X}^{\perp}_{S\cup\{i,j\}}}\big[m(\boldsymbol{X})\big|\boldsymbol{x}^*_{S\cup\{i,j\}}\big] - \mathbb{E}_{\boldsymbol{X}^{\perp}_{S\cup\{j\}}}\big[m(\boldsymbol{X})\big|\boldsymbol{x}^*_{S\cup\{j\}}\big] \\ &- \quad \mathbb{E}_{\boldsymbol{X}^{\perp}_{S\cup\{i\}}}\big[m(\boldsymbol{X})\big|\boldsymbol{x}^*_{S\cup\{i\}}\big] + \mathbb{E}_{\boldsymbol{X}^{\perp}_{S}}\big[m(\boldsymbol{X})\big|\boldsymbol{x}^*_{S}\big].\end{aligned}$$

The “partial dependence plot,” formally defined and coined in Friedman (2001), is simply the average of “ceteris paribus profiles,”

Interpretability

Definition 3.15: 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

$$p_j(x_j^*) = \mathbb{E}_{X_j^\perp}\big[m(\boldsymbol{X})|x_j^*\big],$$

and the empirical version is

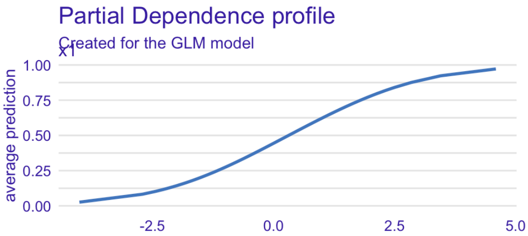
$$\widehat{p}_j(x_j^*) = \frac{1}{n}\sum_{i=1}^n m(x_j^*, \boldsymbol{x}_{i,-j}) = \frac{1}{n}\sum_{i=1}^n \underbrace{m_{\boldsymbol{x}_i,j}(x_j^*)}_{\text{ceteris paribus}} .$$

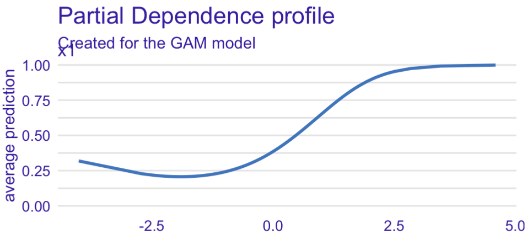
See Greenwell (2017) for the implementation in R, with the `pdp` package. One can also use `type = "partial"` in the `predict_parts` function of the `DALEX` package, as in Biecek and Burzykowski (2021).

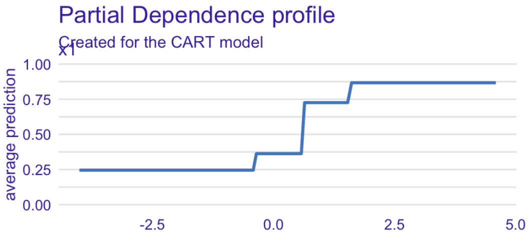
Interpretability

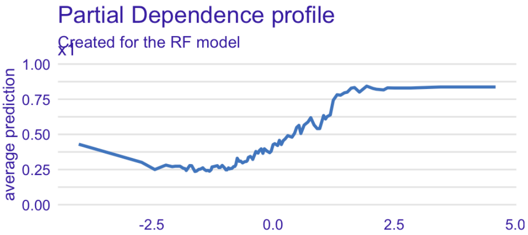
We can visualize $\widehat{p}_1$ (associated with variable x_1) in dataset `toydata2`, the average of $m(x_j^*, \boldsymbol{x}_{i,-j})$ when $i = 1, \cdots, n$, including all $m(x_j^*, \boldsymbol{x}_{i,-j})$'s.

Interpretability





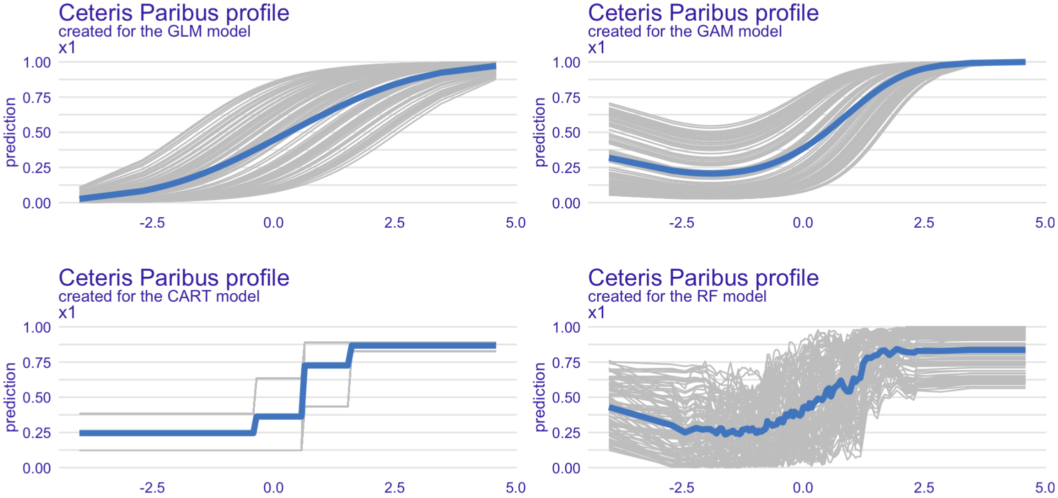




Interpretability

Partial dependence profile $\widehat{p}_1$ associated with variable x_1, for four different models trained on `toydata2`.

Interpretability



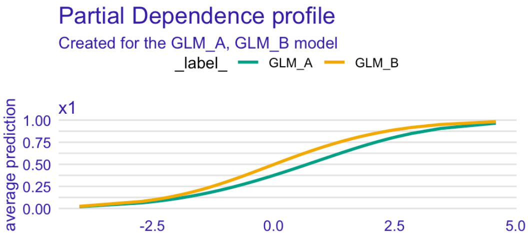
Interpretability

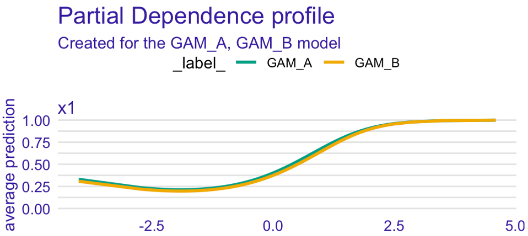
Partial dependence profile $\widehat{p}_1$ associated with variable x_1, seen as the average of ceteris paribus profiles $m(x_j^*, \boldsymbol{x}_{i,-j})$'s (in gray) for different models trained on `toydata2`.
Interestingly, instead of the sum over the n predictions, subsums can be considered, with respect to some criteria.
Sums over $s_i = \texttt{A}$ or $s_i = \texttt{B}$ are considered,

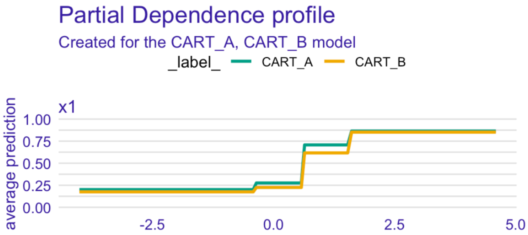
$$\widehat{p}_j^{\texttt{A}}(x_j^*) = \frac{1}{n_{\texttt{A}}} \sum_{i:s_i==\texttt{A}} m(x_j^*, \boldsymbol{x}_{i,-j}) \text{ and } \widehat{p}_j^{\texttt{B}}(x_j^*) = \frac{1}{n_{\texttt{B}}} \sum_{i:s_i==\texttt{B}} m(x_j^*, \boldsymbol{x}_{i,-j}).$$

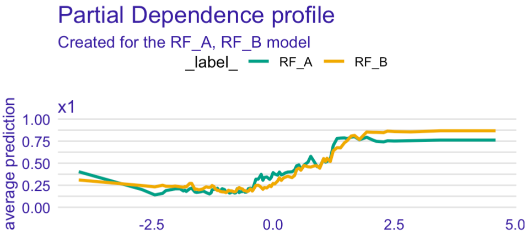
On the `toydata2` data, the three variables j (namely x_1^*, x_2^* and x_3^*) are used.
If x_3^* has a very flat impact, and no influence on the outcome, one should observe that $\widehat{p}_j^{\texttt{A}}(x_3^*)$ and $\widehat{p}_j^{\texttt{B}}(x_3^*)$ are significantly different.

Interpretability





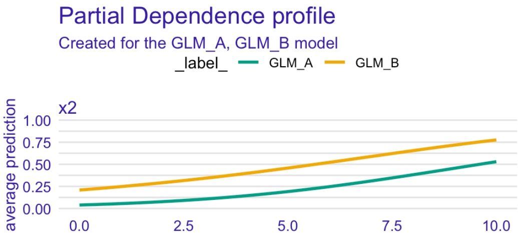


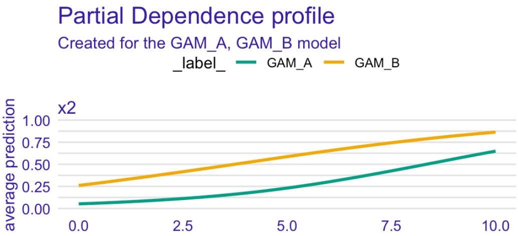


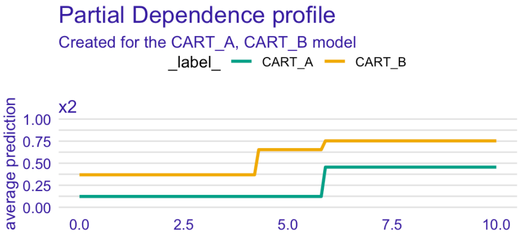
Interpretability

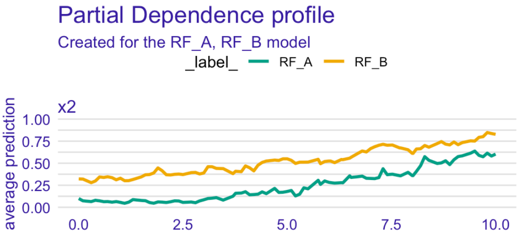
Partial dependence profiles $\widehat{p}_1^{\mathtt{A}}$ and $\widehat{p}_1^{\mathtt{B}}$, for x_1, when the sensitive attribute s is either `A` or `B`, as the average of subgroups (s_i being either `A` or `B`) for different models trained on `toydata2`.

Interpretability





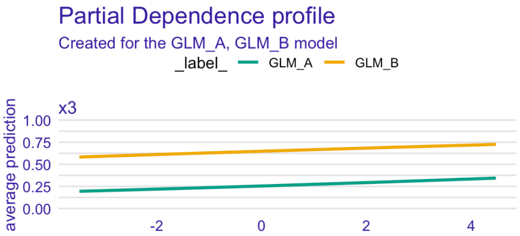


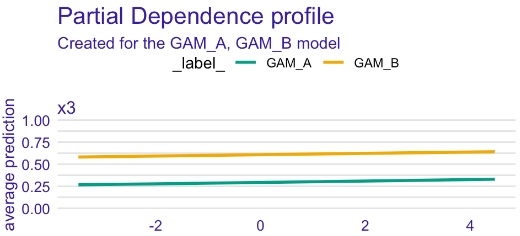


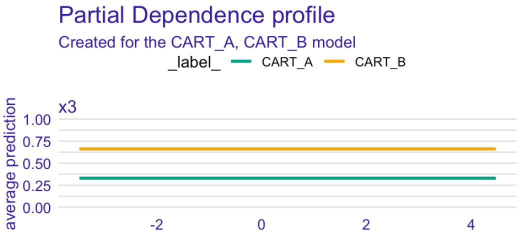
Interpretability

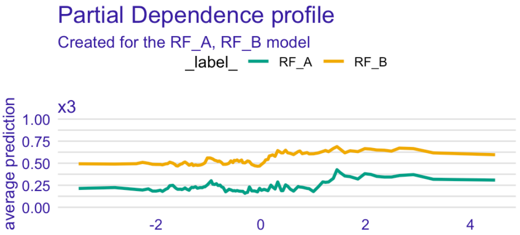
Partial dependence profiles $\widehat{p}_2^{\mathtt{A}}$ and $\widehat{p}_2^{\mathtt{B}}$, for x_2, when the sensitive attribute s is either `A` or `B`, as the average of subgroups (s_i being either `A` or `B`) for different models trained on `toydata2`.

Interpretability









Interpretability

Partial dependence profiles $\widehat{p}_3^{\mathtt{A}}$ and $\widehat{p}_3^{\mathtt{B}}$, for x_3, when the sensitive attribute s is either $\mathtt{A}$ or $\mathtt{B}$, as the average of subgroups (s_i being either $\mathtt{A}$ or $\mathtt{B}$) for different models trained on `toydata2`.

But instead of those *ceteris paribuss* dependence plots, it could be interesting to consider some local versions, or *mutatis mutandis* dependence plots. Apley and Zhu (2020) introduced the "local dependence plot" and the "accumulated local plot," defined as follows,

Interpretability

Definition 3.16: Local Dependence Plot $\ell_j(x_j^*)$ and $\widehat{\ell}_j(x_j^*)$

The local dependence plot is defined as

$$\ell_j(x_j^*) = \mathbb{E}_{X_j}\big[m(\boldsymbol{X})|x_j^*\big]$$

$$\widehat{\ell}_j(x_j^*) = \frac{1}{\text{card}(V(x_j^*))}\sum_{i\in V(x_j^*)} m(x_j^*,\boldsymbol{x}_{i,-j}) \text{ where } V(x_j^*) = \big\{i : d(x_{i,j},x_j^*) \leq \epsilon\big\},$$

$$\text{or } \widetilde{\ell}_j(x_j^*) = \frac{1}{\sum_i \omega_i(x_j^*)}\sum_{i=1}^n \omega_i(x_j^*) m(x_j^*,\boldsymbol{x}_{i,-j}) \text{ where } \omega_i(x_j^*) = K_h(x_j^* - x_{i,j}),$$

for a smooth version, for some kernel K_h.

Apley and Zhu (2020) suggested to use, instead,

Interpretability

Definition 3.17: Accumulated Local $a_j(x_j^*)$, Apley and Zhu (2020)

$$a_j(x_j^*) = \int_{-\infty}^{x_j^*} \mathbb{E}_{X_j}\left[\frac{\partial m(x_j, \boldsymbol{X}_{-j})}{\partial x_j}\Big| x_j\right] \mathrm{d}x_j.$$

The following estimate was considered

Definition 3.18: Accumulated Local function $\widehat{a}_j(x_j^*)$

$$\widehat{a}_j(x_j^*) = \alpha + \sum_{u=1}^{k_j^*} \frac{1}{n_u} \sum_{u:x_{i,j}\in(a_{u-1},a_u]} \left[m(a_k, \boldsymbol{x}_{i,-j}) - m(a_{k-1}, \boldsymbol{x}_{i,-j})\right],$$

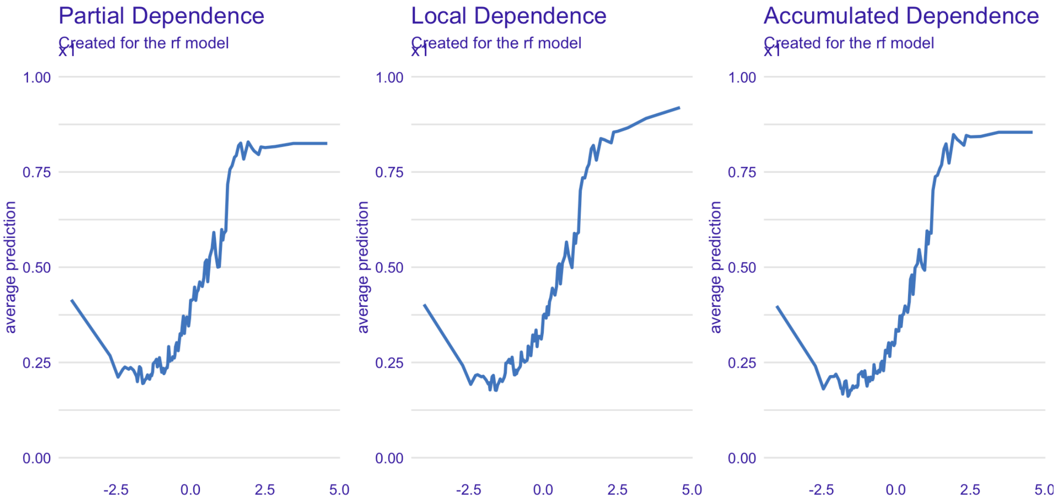
(where α is some normalization constant, since $\mathbb{E}[\widehat{a}_j(X_j)] = 0$).

Interpretability

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 the `predict_parts` function, 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"`,

See 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`.

Interpretability



Simpson's Paradox

Under-identification corresponds to the case where the true model would be $y_i = b_0 + \boldsymbol{x}_1^\top \boldsymbol{x}_1 + \boldsymbol{x}_2^\top \boldsymbol{x}_2 + \varepsilon_i$, but the estimated model is $y_i = b_0 + \boldsymbol{x}_1^\top \boldsymbol{b}_1 + \eta_i$ (in other words, the variables $\boldsymbol{x}_2$ are not used in the regression). The maximum likelihood estimator of $\boldsymbol{b}_1$ is (with the classical matrix writing in econometrics, such as Davidson et al. (2004) or Charpentier et al. (2018))

$$
\begin{aligned}
\widehat{\boldsymbol{b}}_1 &= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{y} \\
&= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top [\boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}] \\
&= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_1 \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{\beta}_2 + (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta}_1 + \underbrace{(\boldsymbol{X}_1' \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{\beta}_2}_{\boldsymbol{\beta}_{12}} + \underbrace{(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{\varepsilon}}_{\nu_i}
\end{aligned}
$$

(see previously)

Simpson's Paradox

With a simple regression model

$$\widehat{b}_1 = \frac{\widehat{\text{cov}}[x_1, y]}{\widehat{\text{Var}}[x_1]} = \frac{\widehat{\text{cov}}[x_1, \beta_0 + \beta_1 x_1 + \beta x_2 + \varepsilon]}{\widehat{\text{Var}}[x_1]}$$

and

$$\widehat{b}_1 = \beta_1 \cdot \underbrace{\frac{\widehat{\text{cov}}[x_1, x_1]}{\widehat{\text{Var}}[x_1]}}_{=1} + \beta_2 \cdot \frac{\widehat{\text{cov}}[x_1, x_2]}{\widehat{\text{Var}}[x_1]} + \underbrace{\frac{\widehat{\text{cov}}[x_1, \varepsilon]}{\widehat{\text{Var}}[x_1]}}_{=0} = \beta_1 + \beta_2 \cdot \frac{\widehat{\text{cov}}[x_1, x_2]}{\widehat{\text{Var}}[x_1]}$$

Simpson's Paradox

A classical example if from Bickel et al. (1975), graduate admissions at U.C. Berkeley

	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%

Simpson's Paradox

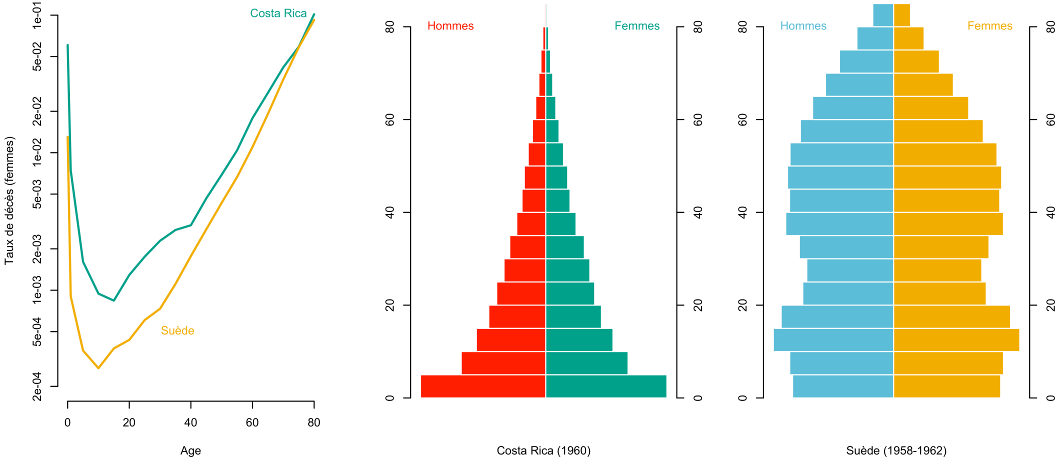
See also survivor's on the Titanic

	Total	Femmes	Hommes
third class passengers	$181/709 \sim$ **25.5%**	$106/216 \sim 49.1\%$	$75/493 \sim 15.2\%$
crew member	$211/890 \sim 23.7\%$	$20/\ \ 23 \sim$ **86.9%**	$191/867 \sim$ **22.0%**

Mathematically, there's no real paradox, in the sense that

$$\frac{a_1}{c_1} < \frac{a_2}{c_2} \text{ et } \frac{b_1}{d_1} < \frac{b_2}{d_2} \not\Leftrightarrow \frac{a_1+b_1}{c_1+d_1} < \frac{a_2+b_2}{c_2+d_2}$$

Simpson's Paradox



Transfer learning

- **Transfer Learning** is a machine learning technique where a model trained on one task is reused or adapted to a different but related task.
- The key idea is to transfer knowledge gained from solving one problem to another, which can significantly reduce the time and data required for training on a new task.
- Transfer learning is especially useful in scenarios where:
  - Labeled data is scarce for the target task.
  - Training a model from scratch would be computationally expensive.
- E.g. a model trained to recognize cats in images can be adapted to recognize dogs by fine-tuning the model on a smaller dataset of dog images.
- Transfer learning typically involves two stages:
  - Pre-training: A model is trained on a large dataset for a source task (e.g., image classification using ImageNet).

Transfer learning

   - Fine-tuning: The pre-trained model is then adapted to the target task by adjusting its weights based on a smaller dataset related to the new task.
- Example:
   - Pre-train a deep neural network on ImageNet for general object recognition.
   - Fine-tune the pre-trained model on a smaller dataset of medical images to detect specific conditions (e.g., lung cancer).
- The success of transfer learning depends on the similarity between the source and target tasks.
- **Inductive Transfer Learning**: The source and target tasks are different, but the model is adapted to learn a new task using the knowledge from the source task.
- **Transductive Transfer Learning**: The source and target tasks are the same, but the source and target datasets differ. The model is adapted to handle variations in data distribution.

Transfer learning

- **Unsupervised Transfer Learning**: The source task is learned with unlabelled data, and knowledge is transferred to a supervised task.
- **Domain Adaptation**: A special case of transfer learning where the task remains the same, but the source and target domains differ (e.g., different sensor types or data distributions).
- Examples:
  - Inductive: A model for detecting cars can be adapted to detect trucks.
  - Transductive: A model trained on photos taken in sunny weather may need to adapt to handle photos taken in cloudy weather.
- **Computer Vision**:
- Pre-trained models like ResNet or VGG are used to solve a wide range of tasks such as facial recognition, object detection, and medical imaging.
- **Natural Language Processing (NLP)**:

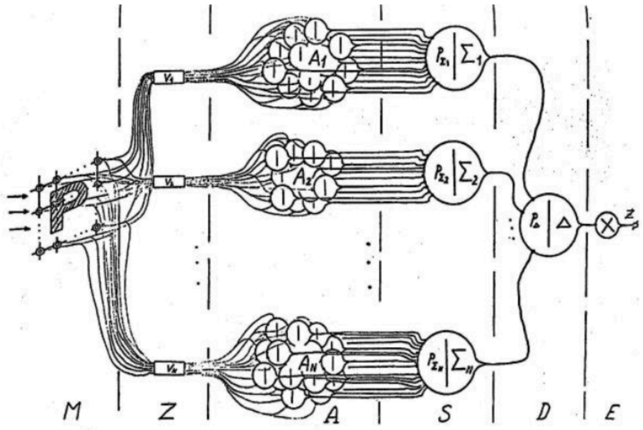
Transfer learning

- ◦ Models like BERT, GPT, and T5 are pre-trained on large text corpora and fine-tuned for tasks like sentiment analysis, text classification, and machine translation.
- **Healthcare**:
- ◦ Transfer learning is used to train models for tasks like diagnosing diseases from medical images when labeled data is scarce.
- **Reinforcement Learning**:
- ◦ Transfer learning is used to transfer knowledge across different environments or tasks in reinforcement learning, enabling faster learning and generalization.
- ◦ Example: Fine-tuning a pre-trained image classification model on a smaller dataset of rare diseases to improve diagnostic accuracy.

Transfer learning in Machine Learning Literature

> **Transfer learning**
> Transfer learning (TL) is a technique in machine learning (ML) in which knowledge learned from a task is re-used in order to boost performance on a related task. W

Transfer learning in Machine Learning Literature



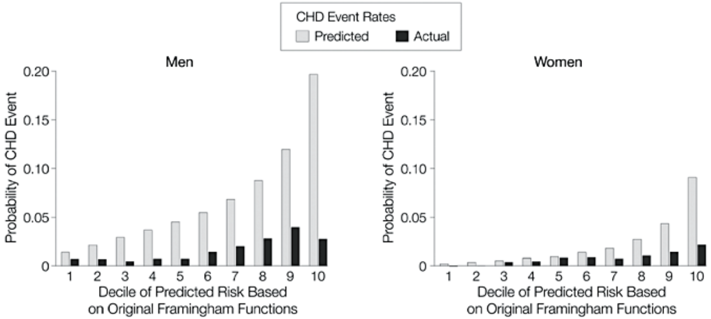
Source: Bozinovski and Fulgosi (1976), The influence of pattern similarity and transfer learning

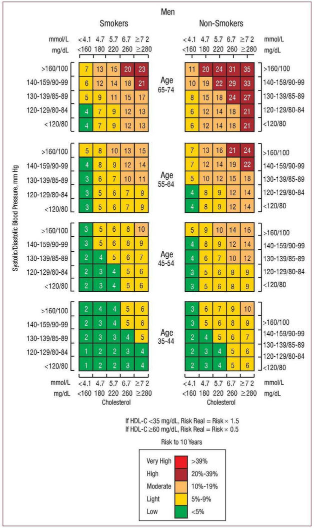
Transfer learning in Machine Learning Literature

- Framingham coronary heart disease (CHD) risk score, Wilson et al. (1987, 1998); D'Agostino et al. (2001)

6 risk factors: age, BP, smoking, diabetes, total cholesterol (TC), and high-density lipoprotein cholesterol (HDL-C)

Framingham (U.S.) participants are of European descent what if we use it on Chinese people ?, Liu et al. (2004)

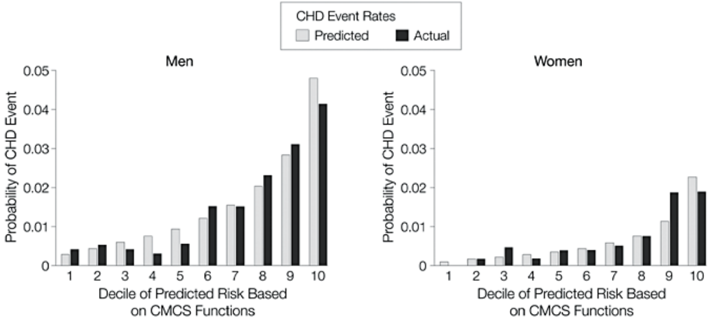




Transfer learning in Machine Learning Literature

- Framingham coronary heart disease (CHD) risk score, Liu et al. (2004)

Refitted on Chinese population,
Chinese Multi-provincial Cohort Study (CMCS)



Risk Factors	CMCS β	Framingham β
Age	0.07	0.05
Age squared	NA	NA
Blood pressure		
Optimal	−0.51	0.09
Normal		
High normal	0.21	0.42
Stage 1 hypertension	0.33	0.66
Stage 2-4 hypertension	0.77	0.90
TC, mg/dL		
<160	−0.51	−0.38
160-199		
200-239	0.07	0.57
240-279	0.32	0.74
≥280	0.52	0.83
HDL-C, mg/dL		
<35	−0.25	0.61
35-44	0.01	0.37
45-49		
50-59	−0.07	0.00
≥60	−0.40	−0.46
Diabetes	0.09	0.53
Smoking	0.62	0.73

Climate, Finance and Insurance

As mentioned in Intergovernmental Panel on Climate Change, page 594

"What does the accuracy of a climate model's simulation of past or contemporary climate say about the accuracy of its projections of climate change? This question is just beginning to be addressed, exploiting the newly available ensembles of models..." Randall et al. (2007)

A standard financial disclaimer, see e.g.,

"Past performance is no guarantee of future returns," Brain (2010)

or in insurance (about wildfire losses in California)

"Looking backward has become less effective in predicting the future," Frazier (2021)

"History Doesn't Repeat Itself, but It Often Rhymes," Mark Twain (1874)

Motivation, statistics, *rebus sic stantibus*

Statistics : *clausula rebus sic stantibus* ("with things thus standing")

> Statistics commonly deals with random samples. A random sample can be thought of as a set of objects that are chosen randomly. More formally, it is "a sequence of independent, identically distributed random data points". (...) Independent and identically distributed random variables are often used as an assumption, which tends to simplify the underlying mathematics. In practical applications of statistical modeling, however, the assumption may or may not be realistic W

Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space,

Let $y_1, y_2, \cdots, y_n$ be n i.i.d. samples of a random variable Y distributed by $\mathbb{P}$

Motivation, statistics, *rebus sic stantibus*

An important concept in actuarial science is the return period.

"**1.0.1. Conditions.** The aim of a statistical theory of extreme values is to analyze observed extremes and to forecast further extremes. (...) The essential condition in the analysis is the clausula rebus sic stantibus," Emil Gumbel (1958), Statistics of Extremes, page 1.

- *rebus sic stantibus* is Latin for "with things thus standing" ("in gelijkblijvende omstandigheden" or "les choses demeurant en l'état")
- *clausula rebus sic stantibus* is the legal doctrine allowing for a contract or a treaty to become inapplicable because of a fundamental change of circumstances,
- maxim *omnis conventio intelligitur rebus sic stantibus* for "every convention is understood with circumstances as they stand", by the Italian jurist Scipione Gentili (1563–1616).

Motivation, statistics, *rebus sic stantibus*

"The distribution from which the extremes have been drawn and its parameters must remain constant in time (or space), or the influence that time (or space) exercises upon them must be taken into account or eliminated (...) This assumption, made in most statistical work, is hardly ever realized." Emil Gumbel (1958), Statistics of Extremes, page 1.

"**1.0.3. The Flood Problem**. Similar stationary time series may easily be obtained for annual droughts, largest precipitations, snowfalls, maxima and minima of atmospheric pressures and temperatures, and other meteorological phenomena." Emil Gumbel (1958), Statistics of Extremes, page 4.

Gumbel (1941a,b) discussed "the return period of flood flows", term used in Fuller (1914) Hazen (1930), on flood flows.

Motivation, statistics, *rebus sic stantibus*

Definition 3.19: Geometric distribution

The probability that the first occurrence of success requires k independent trials, each with success probability p, the probability that the k-th trial is the first success is

$$\mathbb{P}(X = k) = (1 - p)^{k-1} p$$

for $k = 1, 2, 3, 4, \cdots$. And then, $\mathbb{E}_{\mathbb{P}}[X] = p^{-1}$.

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	$\cdots$	Y_{k-1}	Y_k
1	1	1	1	1	1	1	$\cdots$	1	1
0	0	0	0	0	0	0	$\cdots$	0	0

Motivation, climate change

A personal take on science and society

World view

By Gavin Schmidt

Why 2023's heat anomaly is worrying scientists

Climate models struggle to explain why planetary temperatures spiked suddenly. More and better data are urgently needed.

When I took over as the director of NASA's Goddard Institute for Space Studies, I inherited a project that tracks temperature changes since 1880. Using this trove of data, I've made climate predictions at the start of every year since 2016. It's humbling, and a bit worrying, to admit that no year has confounded climate scientists' predictive capabilities more than 2023 has.

For the past nine months, mean land and sea surface

"If the anomaly does not stabilize by August, then the world will be in uncharted territory."

from stratospheric water vapour, and the ramping up of solar activity in the run-up to a predicted solar maximum. But these factors explain, at most, a few hundredths of a degree in warming (Schoeberl, M. R. *et al. Geophys. Res. Lett.* **50**, e2023GL104634; 2023). Even after taking all plausible explanations into account, the divergence between expected and observed annual mean temperatures in 2023 remains about 0.2 °C — roughly the gap between the previous and current annual record.

There is one more factor that could be playing a part. In 2020, new regulations required the shipping industry to use cleaner fuels that reduce sulfur emissions. Sulfur compounds in the atmosphere are reflective and influence several properties of clouds, thereby having

Climate, how to predict in "uncharted territory", Schmidt (2024)?

Motivation, climate change

> A wildfire (or forest fire, bushfire) is an unplanned, uncontrolled and unpredictable fire in an area of combustible vegetation. W

Climate risk in California (U.S.)

"Why is it illegal in California to consider climate-informed catastrophe models when setting wildfire insurance premiums?" Frazier (2021)

Some general context:

California Code Of Regulations, title 10, Chapter 5 (Insurance Commissioner), § 2644 ("Determination of Reasonable Rates")

Cal. Code Regs. tit. 10 § 2644.4 (Projected Losses)

"Projected losses" means the insurer's historic losses per exposure, adjusted by catastrophe adjustment, as prescribed in section 2644.5.

Motivation, climate change

Cal. Code Regs. tit. 10 § 2644.5 (Catastrophe Adjustment)

In those insurance lines and coverages where catastrophes occur, the catastrophic losses of any one accident year in the recorded period are replaced by a loading based on a multi-year, long-term average of catastrophe claims. The number of years over which the average shall be calculated shall be at least 20 years for homeowners multiple peril fire, and at least 10 years for private passenger auto physical damage. Where the insurer does not have enough years of data, the insurer's data shall be supplemented by appropriate data. The catastrophe adjustment shall reflect any changes between the insurer's historical and prospective exposure to catastrophe due to a change in the mix of business. There shall be no catastrophe adjustment for private passenger auto liability.

Transfer learning and domain adaptation ($\mathbb{P}_s \neq \mathbb{P}_t$)

"Traditional machine learning is characterized by training data and testing data having the same input feature space and the same data distribution. When there is a difference in data distribution between the training data and test data, the results of a predictive learner can be degraded," Furht et al. (2016)

- **notations**

Consider some training (source) sample $\mathcal{D}_s = \{(\boldsymbol{x}_{s,i}, y_{s,i})\}$ and some test (target) sample $\mathcal{D}_t = \{(\boldsymbol{x}_{t,i})\}$, both being i.i.d., with distributions $\mathbb{P}_s$ and $\mathbb{P}_t$.
In a regression problem, $y = m(\boldsymbol{x}) + \varepsilon$, i.e. $m(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{E} = \boldsymbol{x}]$
Consider a parametric model, $m(\boldsymbol{x}|\boldsymbol{\theta}))$, for some $\boldsymbol{\theta} \in \Theta$.
Classical empirical risk minimization (ERM) leads to

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\Theta}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_{s,i}, m(\boldsymbol{x}_{s,i}|\boldsymbol{\theta})\big) \right\}$$

Transfer learning and domain adaptation ($\mathbb{P}_s \neq \mathbb{P}_t$)

If $\mathbb{P}_s = \mathbb{P}_t$, $\widehat{\boldsymbol{\theta}}$ is said to be consistent Shimodaira (2000). Otherwise...
Importance weighted empirical risk minimization (IWERM) is

$$\tilde{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\Theta}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \boxed{\frac{\mathbb{P}_s(\boldsymbol{x}_{s,i})}{\mathbb{P}_t(\boldsymbol{x}_{s,i})}} \ell(y_{s,i}, m(\boldsymbol{x}_{s,i}|\boldsymbol{\theta})) \right\}$$

which is now consistent.
One can define adaptative importance weighted empirical risk minimization (AIWERM)

$$\tilde{\boldsymbol{\theta}}_\gamma \in \underset{\boldsymbol{\theta}\in\Theta}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \boxed{\left(\frac{\mathbb{P}_s(\boldsymbol{x}_{s,i})}{\mathbb{P}_t(\boldsymbol{x}_{s,i})}\right)^\gamma} \ell(y_{s,i}, m(\boldsymbol{x}_{s,i}|\boldsymbol{\theta})) \right\},$$

$\gamma \in [0,1]$ is the flattening parameter,

$$\begin{cases} \gamma = 0, & \text{ordinary ERM} \\ \gamma = 1, & \text{IWERM} \end{cases}$$

Transfer learning and domain adaptation ($\mathbb{P}_s \neq \mathbb{P}_t$)

One could consider regularlized importance weighted empirical risk minimization (RIWERM)

$$\tilde{\boldsymbol{\theta}}_\lambda \in \underset{\boldsymbol{\theta}\in\Theta}{\text{argmin}}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}_s(\boldsymbol{x}_{s,i})}{\mathbb{P}_t(\boldsymbol{x}_{s,i})}\ell(y_{s,i}, m(\boldsymbol{x}_{s,i}|\boldsymbol{\theta})) + \lambda\mathcal{P}(\boldsymbol{\theta})\right\},$$

for some penalty function $\mathcal{P}(\boldsymbol{\theta})$ (classically $\|\boldsymbol{\theta}\|_{\ell_1}$ (lasso) or $\|\boldsymbol{\theta}\|_{\ell_2}$ (ridge) types of penalty), and $\lambda \geq 0$.

Transfer learning and domain adaptation ($\mathbb{P}_s \neq \mathbb{P}_t$)

- Application in a regression context

Polynomial regression model,

$$\mathbb{P}_{x,\boldsymbol{\theta}} \sim \mathcal{N}(P_\beta(x), \sigma^2) \text{ and } \boldsymbol{\theta} = (\beta, \sigma^2), \text{ for some polynomial } P_\beta$$

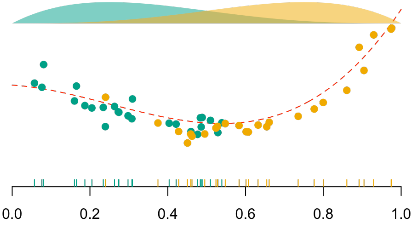
i.e., $y = \beta_0 + \beta_1 x + \cdots + \beta_k x^k + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Suppose that the "true" distribution is

$$\mathbb{Q}_x \sim \mathcal{N}(Q(x), 1)$$

e.g., $Q(x) = -(2x - 1/2) + (2x - 1/2)^3$
Suppose also that

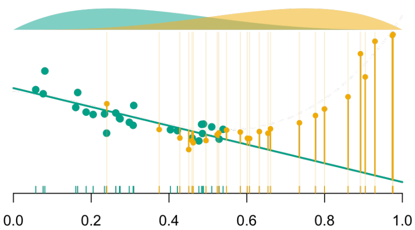
$$\begin{cases} \text{source}: & \pi_s \sim \mathcal{B}(a_s, b_s) \\ \text{target}: & \pi_t \sim \mathcal{B}(a_t, b_t) \end{cases}$$

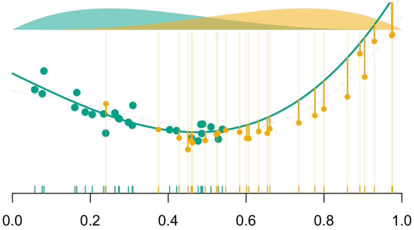


Transfer learning and domain adaptation ($\mathbb{P}_s \neq \mathbb{P}_t$)

Linear model (mis-specified) and cubic model (well-specified)

$$\max_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}) = \max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - P_{\beta(x_i)})^2$$

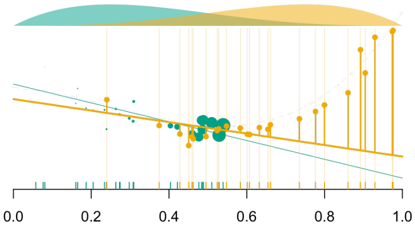


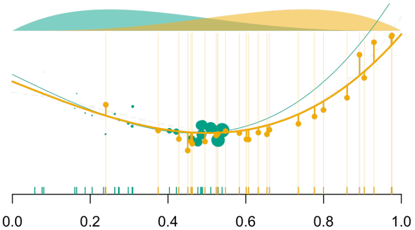


Transfer learning and domain adaptation ($\mathbb{P}_s \neq \mathbb{P}_t$)

Linear model (mis-specified) and cubic model (well-specified)

$$\max_{\boldsymbol{\theta}} \log \mathcal{L}_{\omega}(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x}) = \max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \boxed{\omega(x_i)} \log p(y|\boldsymbol{x},\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \boxed{\frac{x_i^{a_t}(1-x_i)^{b_t}}{x_i^{a_s}(1-x_i)^{s_t}}} (y_i - P_{\beta(x_i)})^2$$





Calibration and conformal prediction

- **Calibration** is the process of adjusting the probability estimates output by a model to better reflect the true likelihood of an event.
- Many machine learning models, such as logistic regression or SVMs, output predicted probabilities that may not correspond to actual frequencies.
- Well-calibrated models make reliable probability predictions: If a model predicts "0.8" for class A, then in 80% of cases, class A should be the correct label.
- Example: If a model predicts a probability of 0.7 for an event, but the true event occurs only 50% of the time when the model predicts 0.7, the model is not well-calibrated.
- Platt Scaling:

  A logistic regression model is fit on the model's output probabilities to transform the predictions into calibrated probabilities.
  Commonly used for SVMs.
- Isotonic Regression:

Calibration and conformal prediction

A non-parametric method that fits a step function to the predicted probabilities and adjusts them accordingly.
Suitable when the number of training examples is large.

- Beta Calibration:

  A generalization of Platt Scaling that uses the Beta distribution for calibration, useful for both binary and multi-class classification.

  These methods help correct overconfidence or underconfidence in probabilistic predictions.

- Improved Decision-Making:

  Well-calibrated probabilities lead to better decision-making, especially in high-stakes domains like healthcare or finance.

  Reliability of Predictions:

  Calibration ensures that the predicted probabilities represent actual event frequencies, helping in risk assessment and uncertainty quantification.

Calibration and conformal prediction

- Example: In medical diagnosis, a model predicting a 90% probability of a disease needs to be trusted to match actual outcomes with that 90% frequency.
- Application: In weather forecasting, better-calibrated probabilities can improve the reliability of predictions like "rain tomorrow" or "storm risk."
- **Conformal Prediction** (CP) is a framework for generating prediction sets (or intervals) that provide a guarantee on the coverage probability.
- Given a new prediction, CP allows us to form a set of possible outcomes, along with a confidence level (e.g., 95%).
- Unlike traditional models that output a single prediction, conformal prediction outputs a set of predictions that likely contains the true outcome.
- Example: A regression model might output a prediction interval of [3.5, 4.2] with 95% confidence, meaning the true value will fall within that range 95% of the time.

Calibration and conformal prediction

- Key Idea: Conformal prediction provides a confidence guarantee that the true label will lie within the prediction set with a specified probability.
- Non-Parametric:

  CP can be applied to any machine learning algorithm, whether it's linear regression, neural networks, or random forests.
- Calibration of the Prediction Sets:

  The prediction sets are constructed based on past data, using a non-conformity measure to assess how well the model's prediction fits the existing data.
  Example: If the model's prediction is too different from past instances, the set may be expanded to include more possibilities.
- Guarantees: If the model is properly calibrated, the prediction set will contain the true label with the specified probability (e.g., 95%).
- Medical Applications:

Calibration and conformal prediction

Conformal prediction can be used in medical diagnostics, providing a confidence interval for disease probabilities, helping doctors make better-informed decisions.

- Finance:

  CP can be used in financial risk management to provide confidence intervals for market predictions (e.g., stock prices).

- Machine Learning Applications:

  Can be used with any machine learning model to create confidence intervals for regression tasks or prediction sets for classification tasks.

- Example: In a self-driving car system, conformal prediction can provide a confidence set for potential obstacles, increasing safety.
- Comparison to Traditional Methods: Conformal prediction provides reliable uncertainty estimates, unlike traditional models which may only output point estimates.

Calibration (when "probabilities" are badly assessed)

"Guo et al. (2017) have shown that modern neural networks are poorly calibrated and over-confident despite having better performance," Müller et al. (2019) or "deep neural networks tend to be overconfident and poorly calibrated after training," Wang et al. (2021)

Calibration (when "probabilities" are badly assessed)

Global balance,

$$\mathbb{E}\big[Y-\widehat{s}(\boldsymbol{X})\big]=\mathbb{E}\big[\mu(\boldsymbol{x})-\widehat{s}(\boldsymbol{X})\big]=0.$$

Economically, if $\widehat{s}(\boldsymbol{x})$ is the price, the portfolio is self-financing (for random losses Y).
Empirical global balance (in-sample)

$$\sum_{i=1}^{n}\big[y_i-\widehat{s}(\boldsymbol{x}_i)\big]=0.$$

Marginal balance,

$$\begin{cases}\mathbb{E}\big[Y-\widehat{s}(\boldsymbol{X})\mid X_j\big]=\mathbb{E}\big[\mu(\boldsymbol{x})-\widehat{s}(\boldsymbol{X})\mid X_j\big]=0\\ \mathbb{E}\big[Y-\widehat{s}(\boldsymbol{X})\mid \boldsymbol{X}\big]=\mathbb{E}\big[\mu(\boldsymbol{x})-\widehat{s}(\boldsymbol{X})\mid \boldsymbol{X}\big]=0\end{cases}$$

Economically, subgroups $\boldsymbol{x}$ are self-financing (for random losses Y).

Calibration (when "probabilities" are badly assessed)

Well-calibration (or "marginal balance", w.r.t. $\widehat{s}(\boldsymbol{x})$)

$$\mathbb{E}\big[Y-\widehat{s}(\boldsymbol{X}) \mid \widehat{s}(\boldsymbol{X})\big]=\mathbb{E}\big[\mu(\boldsymbol{x})-\widehat{s}(\boldsymbol{X}) \mid \widehat{s}(\boldsymbol{X})\big]=0.$$

Economically, price-based subgroups $\widehat{s}(\boldsymbol{x})$ are self-financing (for random losses Y).

Proposition 3.2: Well-calibration

The true regression function $\eta(\boldsymbol{x})=\mathbb{E}[Y \mid \boldsymbol{X}=\boldsymbol{x}]$ is well-calibrated, and so is the expected value, $\mathbb{E}[Y]$.

Calibration (when "probabilities" are badly assessed)

Definition 3.20: Recalibration

Given a model $s : \mathcal{X} \to \mathbb{R}$, the following re-calibration step gives an auto-calibrated regression function

$$s_{\text{rcb}}(\boldsymbol{x}) = \mathbb{E}[Y \mid s(\boldsymbol{X}) = s(\boldsymbol{x})]$$

Calibration (when "probabilities" are badly assessed)

In many applications, we need to properly assess $\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})$

> model calibration can be also used to refer to Bayesian inference about the value of a model's parameters, given some data set, or more generally to any type of fitting of a statistical model. As Philip Dawid puts it, "*a forecaster 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.*"
> W, see Dawid (1982).

Prediction $\widehat{Y}$ of Y is a well-calibrated prediction if $\mathbb{E}_{\mathbb{P}}[Y|\widehat{Y}=p]=\widehat{y}$, for all $p\in(0,1)$.

"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," Silver (2012)

"we desire that the estimated class probabilities are reflective of the true underlying probability of the sample," Kuhn and Johnson (2013)

Calibration (when ”probabilities” are badly assessed)

“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,” von Mises (1928, 1939).

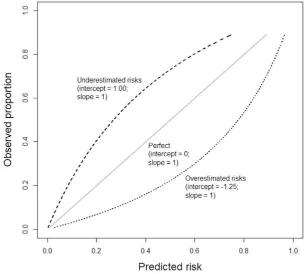
Calibration (when "probabilities" are badly assessed)

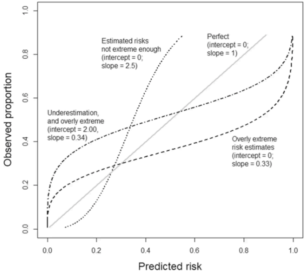
As explained in Van Calster et al. (2019), "among patients with an estimated risk of 20%, we expect 20 in 100 to have or to develop the event".

- If 40 out of 100 in this group are found to have the disease, the risk is underestimated,
- If we observe that in this group, 10 out of 100 have the disease, we have overestimated the risk.

Hosmer-Lemeshow test, from Hosmer Jr et al. (2013) (logistic regression), and Bier score, from Brier et al. (1950) and Murphy (1973).

Function plotted in psychological papers Keren (1991).





Calibration (when "probabilities" are badly assessed)

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}\big(\hat{s}(\boldsymbol{x}_i) - y_i\big)^2$$

Calibration curve is defined as

$$g : \begin{cases}[0,1] \to [0,1] \\ p \mapsto g(p) := \mathbb{E}[Y \mid \widehat{s}(\boldsymbol{X}) = p]\end{cases}$$

The g function for a well-calibrated model $\widehat{s}$ is the identity function $g(p) = p$.

- Quantile Bins

Set $\widehat{y}_i = \hat{s}(\boldsymbol{x}_i)$, sorted $\widehat{y}_1 \leq \widehat{y}_2 \leq \cdots \leq \widehat{y}_n$, partition $\mathcal{I}_1, \cdots, \mathcal{I}_{10}$ of $\{1, 2, \cdots, n\}$.
As in Pakdaman Naeini et al. (2015), consider scatter plot

$$(u_, v_k), \text{ where } u_k = \frac{1}{n_k}\sum_{i\in\mathcal{I}_k}\widehat{y}_i \text{ and } v_k = \frac{1}{n_k}\sum_{i\in\mathcal{I}_k} y_i$$

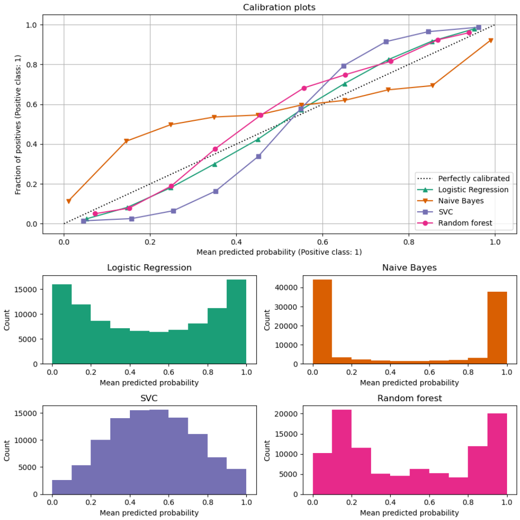
Calibration (when "probabilities" are badly assessed)

Wilks (1990), Pakdaman Naeini et al. (2015) and Kumar et al. (2019) considered quantile-based bins : $\overline{g}$ is the continuous piecewise linear function, interpolating linearly between the points

$$\{(\overline{s}_k,\overline{y}_k)\} \text{ where } k=1,\cdots,10,$$

$$\overline{s}_k=\frac{10}{n}\sum_{i\in I_k}\widehat{s}(\boldsymbol{x}_i) \text{ and } \overline{y}_k=\frac{10}{n}\sum_{i\in I_k}y_i,$$

$$I_k=\left\{i:\left\lceil\frac{k-1}{10}\cdot n\right\rceil\leq \text{rank}(\widehat{s}(\boldsymbol{x}_i))\leq\left\lfloor\frac{k}{10}\cdot n\right\rfloor\right\}$$



Calibration (when "probabilities" are badly assessed)

- Local Regression

Given sample $\{(\boldsymbol{x}_i, y_i)\}$ and score $\widehat{s}$, consider a **local regression** of y's against $\widehat{s}(\boldsymbol{x})$'s, as in Loader (2006), see Austin and Steyerberg (2019); Denuit et al. (2021). E.g.

$$\widehat{g}(p) := \frac{\displaystyle\sum_{i=1}^{n} K_h(p - \widehat{s}(\boldsymbol{x}_i)) \cdot y_i}{\displaystyle\sum_{i=1}^{n} K_h(p - \widehat{s}(\boldsymbol{x}_i))}, \ \forall p \in [0,1],$$

based on Nadaraya (1964); Watson (1964), for some kernel K and some bandwidth h. One could also consider some kernel based local regression (of degree 1 or 2), as suggested in Denuit et al. (2021).

Calibration (when "probabilities" are badly assessed)

- Isotonic Regression

Since g should be increasing, quite naturally, we could consider an **isotonic regression** of y's against $\widehat{s}(\boldsymbol{x})$'s, as in Kruskal (1964), see Niculescu-Mizil and Caruana (2005), $\widetilde{g}$ is the continuous piecewise linear function, interpolating linearly between the points $(\widehat{s}(\boldsymbol{x}_i),\hat{y}_i)$, where $\widehat{s}(\boldsymbol{x}_i)$'s are sorted,

$$\widetilde{g}(p) := \begin{cases} \hat{y}_1 & \text{if } p \leq \widehat{s}(\boldsymbol{x}_1) \\ \hat{y}_i + \dfrac{p-\widehat{s}(\boldsymbol{x}_i)}{\widehat{s}(\boldsymbol{x}_{i+1})-\widehat{s}(\boldsymbol{x}_i)}(\hat{y}_{i+1}-\hat{y}_i) & \text{if } \widehat{s}(\boldsymbol{x}_i) \leq x \leq \widehat{s}(\boldsymbol{x}_{i+1}) \\ \hat{y}_n & \text{if } x \geq \widehat{s}(\boldsymbol{x}_n) \end{cases}$$

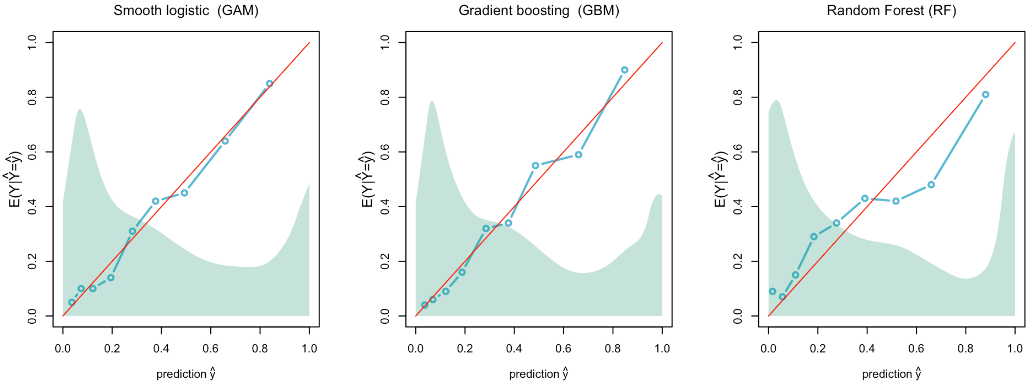
where

$$\min_{\hat{y}_1,\cdots,\hat{y}_n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \text{ subject to } \hat{y}_i \leq \hat{y}_j \text{ for all } (i,j)\in E,$$

$E=\{(i,j):\widehat{s}(\boldsymbol{x}_i)\leq \widehat{s}(\boldsymbol{x}_j)\}$ specifies the partial ordering of the observed inputs $\widehat{s}(\boldsymbol{x}_i)$.

Calibration (when "probabilities" are badly assessed)

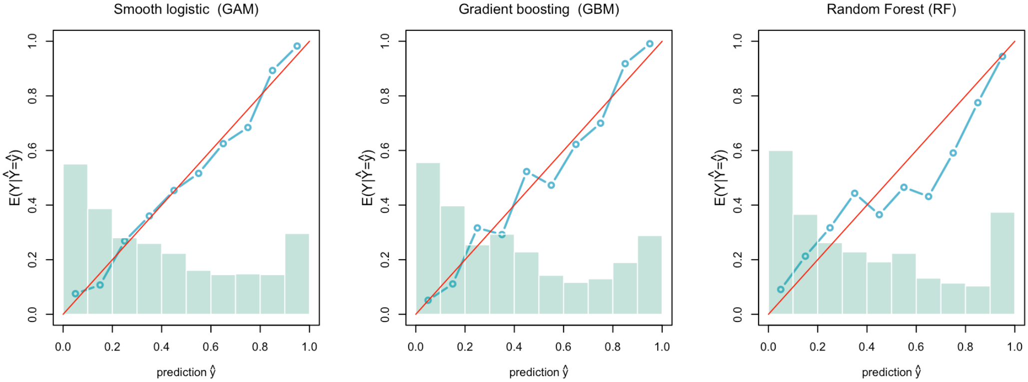
Calibration scatterplot per quantile bins



(see also Fernandes Machado et al. (2024a,b))

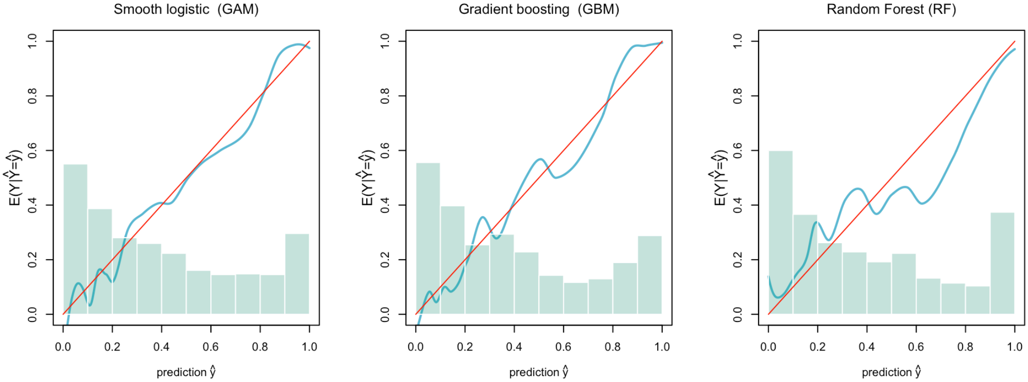
Calibration (when "probabilities" are badly assessed)

Local regression scatterplot per bins, $[0; 0.1)$, $[0.1; 0.2)$, $[0.2; 0.3)$, $[0.3; 0.4)$, etc



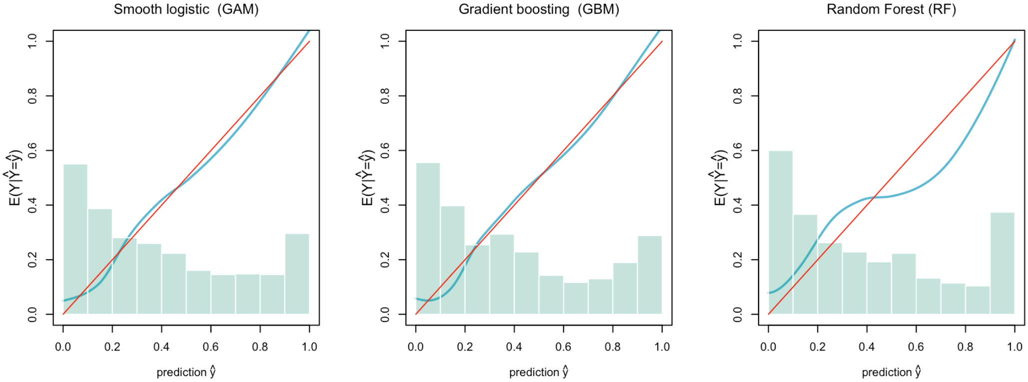
Calibration (when "probabilities" are badly assessed)

Calibration scatterplot per local regression (small bandwidth)



Calibration (when "probabilities" are badly assessed)

Local regression scatterplot per local regression (larger bandwidth)



Calibration (when ”probabilities” are badly assessed)

A standard metric for assessing calibration is Brier score (see Gupta et al. (2021); Kull et al. (2017); Platt et al. (1999); Rahimi et al. (2020)), from Brier (1950):

Brier score (MSE), $$\boxed{\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}\big(\hat{s}(\mathbf{x}_i) - y_i\big)^2.}$$

Austin and Steyerberg (2019) and Zhang et al. (2020) proposes the **Integrated Calibration Index** (ICI) based on the calibration curve,

Integrated Calibration Index, $$\boxed{\mathrm{ICI} = \frac{1}{n}\sum_{i=1}^{n}\mid \hat{s}(\mathbf{x}_i) - \hat{g}(\hat{s}(\mathbf{x}_i)) \mid .}$$

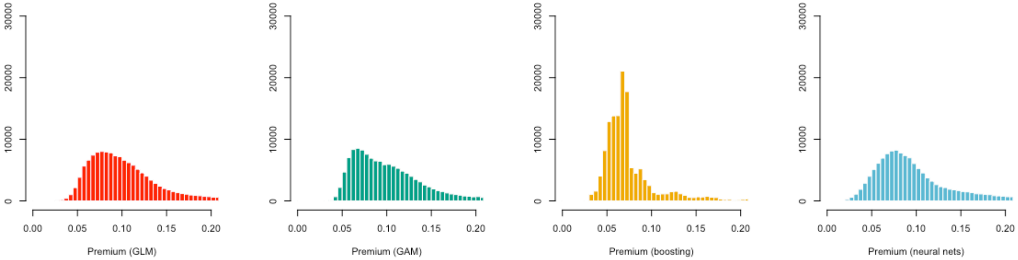
Local Calibration Score, $$\boxed{\mathrm{LCS} = \frac{1}{n}\sum_{i=1}^{n}\big(\hat{s}(\mathbf{x}_i) - \hat{g}(\hat{s}(\mathbf{x}_i))\big)^2.}$$

Application in Motor Insurance

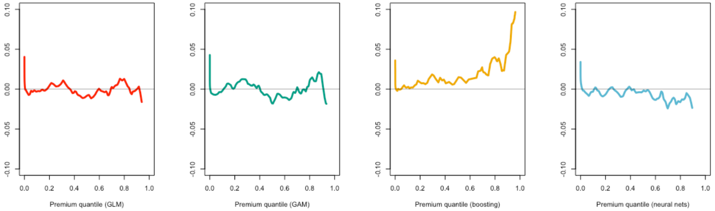
Consider claims (annual) frequency, corrected from the exposure, `freMTPL2freq` from `CASDataset` package, as in Denuit et al. (2021).

	$\widehat{m}^{\text{glm}}$	$\widehat{m}^{\text{gam}}$	$\widehat{m}^{\text{bst}}$
average $\widehat{m}(\mathbf{x})$'s	0.1087	0.1092	0.0820
10% quantile	0.0605	0.0598	0.0498
90% quantile	0.1682	0.1713	0.1244

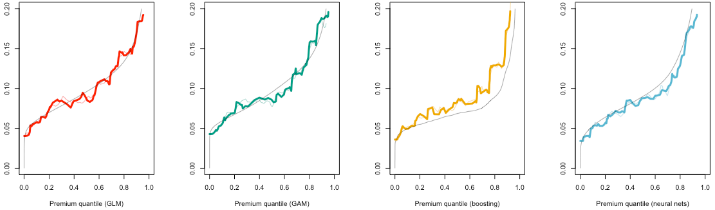
Application in Motor Insurance



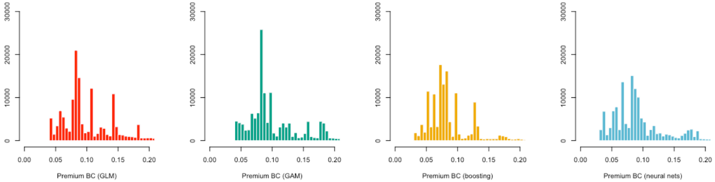
Application in Motor Insurance



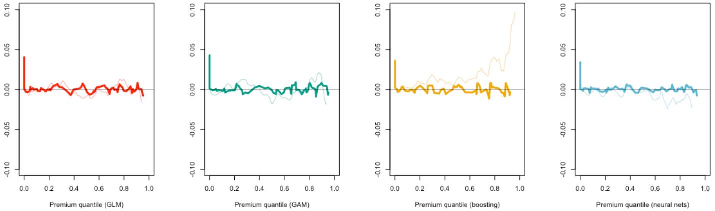
Evolution of $p \mapsto \mathbb{E}[Y|\widehat{m}(\boldsymbol{X}) = p]$ and $u \mapsto \mathbb{E}[Y|\widehat{m}(\boldsymbol{X}) = F_{\widehat{m}}^{-1}(u)]$



Application in Motor Insurance



Recalibrated models



Conformal Prediction

> **Conformal prediction**
> Conformal prediction (CP) is a machine learning framework for uncertainty quantification that produces statistically valid prediction regions (prediction intervals) for any underlying point predictor only assuming exchangeability of the data. W

Need for probablistic predictions, see Vovk et al. (2005) or Da Veiga (2024)
Even if model predictions can be very close, the intervals may heavily vary depending on the underlying assumptions used to build them
Somehow, classical problem, discussed with various underlying ideas.

Conformal Prediction

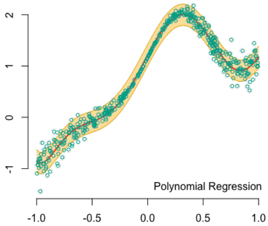
- central limit theorems that are available for some models (polynomial regression, local-averaging methods, ...)

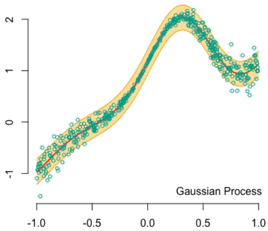
  see polynomial regression

  theoretical guarantees (possibly only asymptotic)

  model may be wrong
- Bayesian paradigm (Gaussian processes or Bayesian neural networks more
- see Gaussian process (and posterior distribution)

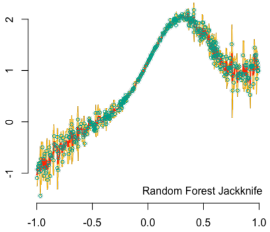
  the influence of the prior is not negligible recently, ...)

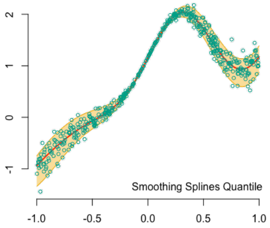




Conformal Prediction

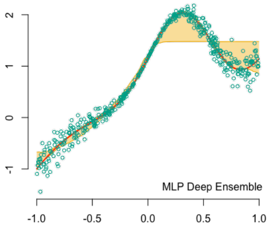
- resampling methods (bootstrap, cross-validation, leave-one-out, jackknife,
- see random forests, with bootstrap and out-of-bag resampling method
- we may lack theoretical guarantees that they provide valid intervals
- quantile regression, where the model is trained to specifically learn quantiles of the target conditional distribution instead of the mean
- see smoothing splines,
- could be not very smooth

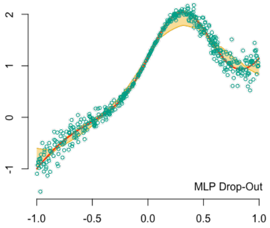




Conformal Prediction

- heuristic approaches in the neural network community (deep ensembles, drop-out, ...)
- see deep ensemble, and drop-out
- theoretical guarantees may be even more lacking





Conformal Prediction

Definition 3.21: Prediction interval (or band)

Given $\{(y_i, \boldsymbol{x}_i)\}$, a **prediction interval** $\widehat{\mathcal{C}}_n$ with error level $\alpha \in (0,1)$ is a function

$$\widehat{\mathcal{C}}_n : \mathcal{X} \to \text{subsets of } \mathcal{Y}$$

built from an i.i.d. sample $\{(Y_i, \boldsymbol{X}_i)\}$, from $\mathbb{P}$ such that, for any new $(Y_{n+1}, \boldsymbol{X}_{n+1}) \sim \mathbb{P}$, we have

$$\mathbb{P}\big[Y_{n+1} \in \widehat{\mathcal{C}}_n(\boldsymbol{X}_{n+1})\big] \geq 1 - \alpha.$$

coverage is guaranteed in average over all random draws of training data
It must be distribution-free, i.e. the coverage guarantee (1) must hold without assumptions on the data generating process
It must be valid in a non-asymptotic framework (for any n)

Conformal Prediction

Definition 3.22: Exchangeability

Random variables $Z_1, \cdots, Z_n$ are **exchangeable** if for any permutation σ of $\{1, 2, \cdots, n\}$

$$(Z_1, Z_2, \cdots, Z_n) \overset{\mathcal{L}}{=} (Z_{\sigma(1)}, Z_{\sigma(2)}, \cdots, Z_{\sigma(n)})$$

- if $Z_1, \cdots, Z_n$ are i.i.d., then $Z_1, \cdots, Z_n$ are exchangeable,
- if $Z_1, \cdots, Z_n | \Psi$ are i.i.d. conditional on Ψ (conditional independence), then $Z_1, \cdots, Z_n$ are exchangeable,
- if $Z_1, \cdots, Z_n$ sampled uniformly from a finite set, then $Z_1, \cdots, Z_n$ are exchangeable,
- if $X_1, \cdots, X_n$ are exchangeable and if $Z_i = \psi(X_i)$, then $Z_1, \cdots, Z_n$ are exchangeable,

Conformal Prediction

Definition 3.23: Exchangeability

Given $\boldsymbol{z} = (z_1, \cdots, z_n)$, and $\tau \in (1/n, 1)$,

$$\text{quantile}_\tau(\boldsymbol{z}) = \inf_{x \in \mathbb{R}} \{\widehat{F}_n(x) \geq \tau\} \text{ where } \widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(z_i \leq x),$$

corresponding to the $\lceil n\tau \rceil$ smallest value of vector $\boldsymbol{z}$.

Proposition 3.3: Exchangeability and quantiles

If $Z_1, \cdots, Z_n$ are exchangeable random variables, then for any i and any $\tau \in [0, 1]$,

$$\mathbb{P}\big[Z_i \leq \text{quantile}_\tau(\boldsymbol{Z})\big] \geq \tau$$

Conformal Prediction

Thus, if $Z_1, \cdots, Z_n$,

$$\mathbb{P}\big[Z_{n+1} \leq \underbrace{\text{quantile}_\tau(\boldsymbol{Z}, Z_{n+1})}^{\text{depends on } Z_{n+1}}\big] \geq \tau$$

To illustrate suppose that $Y_1, \cdots, Y_n, Y_{n+1}$ are i.i.d. Gaussian variables. Since

$$\sqrt{\frac{n}{n+1}} \cdot \frac{Y_{n+1} - \overline{Y}_n}{\widehat{s}_n} \sim T(n-1) \text{ where } \overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i \text{ and } \widehat{s}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2$$

so that $\mathbb{P}\big[Y_{n+1} \leq \widehat{q}_n\big] \geq 1 - \alpha$ where

$$\widehat{q}_n = \overline{Y}_n + \widehat{s}_n \sqrt{\frac{n+1}{n}} \cdot F^{-1}_{Std(n-1)}(1-\alpha).$$

Observe that

$$\boxed{Z_{n+1} \leq \text{quantile}_\tau(\boldsymbol{Z}, Z_{n+1})} \iff \boxed{Z_{n+1} \leq \text{quantile}_{\tau\frac{n+1}{n}}(\boldsymbol{Z})}$$

Conformal Prediction

Suppose that $\widehat{\mathcal{C}}_n(\boldsymbol{x}) = \big[\widehat{\mu}_n(\boldsymbol{x}) \pm \widehat{q}_n\big]$, then

$$\mathbb{P}\big[Y_{n+1} \in \widehat{\mathcal{C}}_n(\boldsymbol{X}_{n+1})\big] = \mathbb{P}\big[\,\big|Y_{n+1} - \widehat{\mu}_n(\boldsymbol{X}_{n+1})\big| \leq \widehat{q}_n\big] = \mathbb{P}\big[R_{n+1} \leq \widehat{q}_n\big] \geq 1 - \alpha,$$

where R_i denote absolute residuals, $R_i = \big|Y_i - \widehat{\mu}_n(\boldsymbol{X}_i)\big|$.
We cannot use $\widehat{q}_n = \text{quantile}_{(1-\alpha)\frac{n+1}{n}}(R_1, \cdots, R_n)$ because $R_1, \cdots, R_n, R_{n+1}$ are not exchangeable...

Split Conformal Prediction

Classically, split the training data $\mathcal{D}_n$ into a **proper training set** $\mathcal{D}_n^t$ and a **hold-out calibration set** $\mathcal{D}_n^c$ (disjoints), with n_t and n_c observations, respectively.
Define R_i denote absolute residuals, $R_i = \big|Y_i - \widehat{\mu}_{n_t}(\boldsymbol{X}_i)\big|$ on the calibration dataset.
Conditional on the training dataset, R_i's (in the calibration dataset) are independent, therefore, they are exchangeable...
Thus, if $\boldsymbol{R}_c$ is the set of absolute residuals on the calibration dataset (compared with prediction on the training dataset $\big|Y_i - \widehat{\mu}_{n_t}(\boldsymbol{X}_i)\big|$),

$$\mathbb{P}\left[R_{n+1} \leq \text{quantile}_{(1-\alpha)\frac{n_c+1}{n_c}}(\boldsymbol{R}_c)\middle|\mathcal{D}_n^t\right] \geq 1-\alpha$$

i.e.

$$\widehat{C}_n^{\text{split}}(\boldsymbol{x}) = \widehat{\mu}_{n_t}(\boldsymbol{x}) \pm \widehat{q}_{n_c}$$

$$\mathbb{P}\left[Y_{n+1} \in \boxed{\widehat{C}_n^{\text{split}}(\boldsymbol{X}_{n+1})}\middle|\mathcal{D}_n^t\right] \geq 1-\alpha$$

where $\widehat{q}_{n_c} = \text{quantile}_{(1-\alpha)\frac{n_c+1}{n_c}}(\boldsymbol{R}_c)$

Split Conformal Prediction

Proposition 3.4: (Coverage for split conformal, Vovk et al. (2005))

If $(Y_1, \boldsymbol{X}_1), \cdots, (Y_n, \boldsymbol{X}_n), (Y_{n+1}, \boldsymbol{X}_{n+1})$ are exchangeable, the split conformal interval satisfies

$$\widehat{\mathcal{C}}_n^{\text{split}}(\boldsymbol{x}) = \widehat{\mu}_{n_t}(\boldsymbol{x}) \pm \text{quantile}_{(1-\alpha)\frac{n_c+1}{n_c}}(\boldsymbol{R}_c)$$

$$\mathbb{P}\Big[Y_{n+1} \in \boxed{\widehat{\mathcal{C}}_n^{\text{split}}(\boldsymbol{X}_{n+1})} \Big| \mathcal{D}_n^t\Big] \geq 1 - \alpha.$$

The interval

$$\widehat{\mathcal{C}}_n^{\text{split}}(\boldsymbol{x}) = \widehat{\mu}_{n_t}(\boldsymbol{x}) \pm \text{quantile}_{(1-\alpha)\frac{n_c+1}{n_c}}(\boldsymbol{R}_c)$$

is natural, and can be compared to the “naive” interval

$$\widehat{\mathcal{C}}_n^{\text{naive}}(\boldsymbol{x}) = \widehat{\mu}_n(\boldsymbol{x}) \pm \text{quantile}_{(1-\alpha)\frac{n+1}{n}}(\boldsymbol{R})$$

Split Conformal Prediction

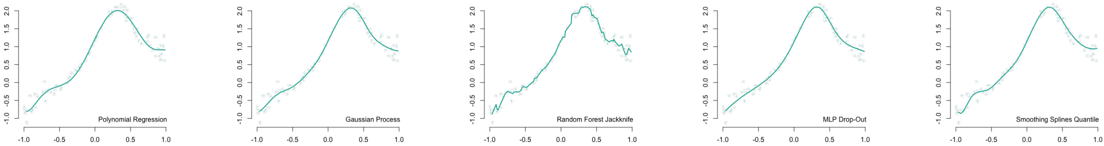
We can write

$$\mathbb{P}\Big[Y_{n+1} \in \widehat{C}_n^{\text{split}}(\boldsymbol{X}_{n+1})\Big] = \mathbb{E}\left[\mathbb{P}\Big[Y_{n+1} \in \widehat{C}_n^{\text{split}}(\boldsymbol{X}_{n+1})\Big|\mathcal{D}_n^t\Big]\right] \geq 1-\alpha.$$

This type of guarantee is called **marginal coverage**, in the sense that the probability has been marginalized over all the randomness. Split conformal prediction thus satisfies also a marginal coverage guarantee.

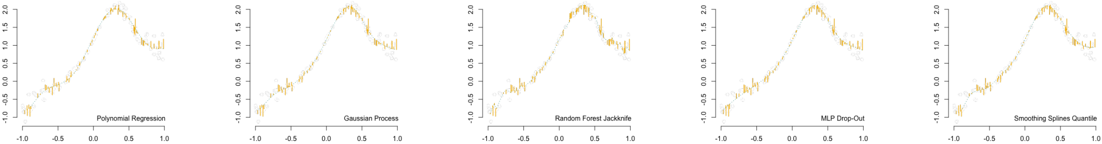
Split Conformal Prediction

Split data to create a proper training set, a calibration set, and keep the test set (by randomly splitting the data set)



Split Conformal Prediction

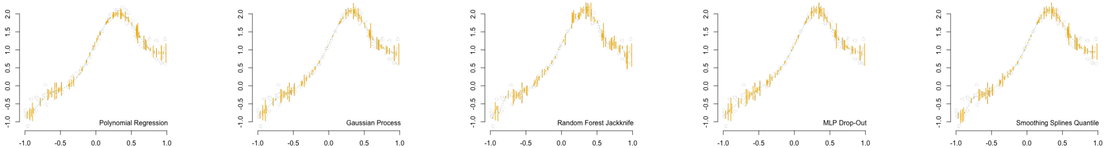
On the proper training set, learn $\widehat{m}$



Split Conformal Prediction

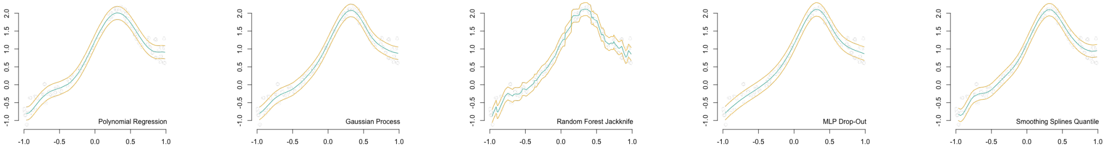
On the calibration set, predict with $\widehat{m}$, $R_i = y_i - \widehat{m}(\boldsymbol{x}_i)$ and consider $|R_i|$ ("conformity scores")

Compute their (1α) empirical quantile, $\text{quantile}_{(1-\alpha)\frac{n+1}{n}}(\boldsymbol{R})$

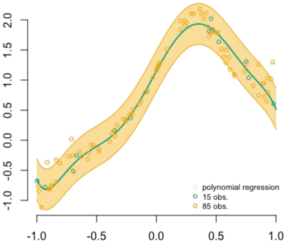


Split Conformal Prediction

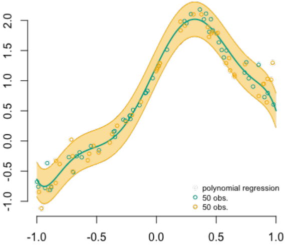
On the test set, predict with $\widehat{m}$ and add $\pm\text{quantile}_{(1-\alpha)}(\boldsymbol{R})$



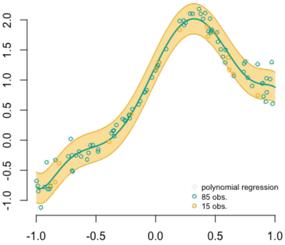
Split Conformal Prediction



93.3% coverage

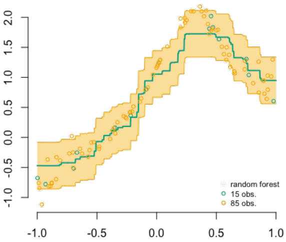


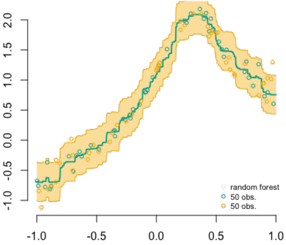
91.8% coverage

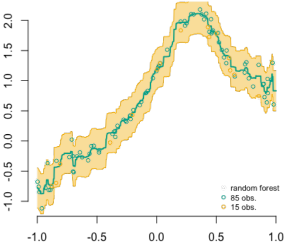


93.1% coverage

Split Conformal Prediction







86.9% coverage   94.4% coverage   95.5% coverage

Cross Validation and Conformal Prediction

Instead of

$$\widehat{\mathcal{C}}_n^{\text{naive}}(\boldsymbol{x}) = \widehat{\mu}_n(\boldsymbol{x}) \pm \text{quantile}_{(1-\alpha)\frac{n+1}{n}}(\boldsymbol{R})$$

write, as in Barber (2024)

$$\widehat{\mathcal{C}}_n^{\text{naive}}(\boldsymbol{x}) = \widehat{\mu}_n(\boldsymbol{x}) \pm \widehat{Q}_{n,\alpha}^{+}(\boldsymbol{R})$$

(empirical $\frac{n+1}{n}(1-\alpha)$ quantile from sample $\boldsymbol{R} = \{R_1, \cdots, R_n\}$).
And instead of

$$\widehat{\mathcal{C}}_n^{\text{naive}}(\boldsymbol{x}) = \widehat{\mu}_n(\boldsymbol{x}) \pm \widehat{Q}_{n,\alpha}^{+}(\{|Y_i - \widehat{\mu}_n(\boldsymbol{X}_i))$$

why not consider a Jacknife (leave-one-out) version

$$\widehat{\mathcal{C}}_n^{\text{jack}}(\boldsymbol{x}) = \widehat{\mu}_n(\boldsymbol{x}) \pm \widehat{Q}_{n,\alpha}^{+}(\{|Y_i - \widehat{\mu}_{-i}(\boldsymbol{X}_i)|\})$$

Cross Validation and Conformal Prediction

$$\widehat{C}_n^{\text{jack}}(\boldsymbol{x}) = \left[\widehat{\mu}_n(\boldsymbol{x}) - \widehat{Q}^+_{n,\alpha}(\{|Y_i - \widehat{\mu}_{-i}(\boldsymbol{X}_i)|\})\,;\widehat{\mu}_n(\boldsymbol{x}) + \widehat{Q}^+_{n,\alpha}(\{|Y_i - \widehat{\mu}_{-i}(\boldsymbol{X}_i)|\})\right]$$

or

$$\widehat{C}_n^{\text{jack}}(\boldsymbol{x}) = \left[\widehat{Q}^+_{n,\alpha}(\{\widehat{\mu}_n(\boldsymbol{x}) - |Y_i - \widehat{\mu}_{-i}(\boldsymbol{X}_i)|\})\,;\,\widehat{Q}^+_{n,\alpha}(\{\widehat{\mu}_n(\boldsymbol{x}) + |Y_i - \widehat{\mu}_{-i}(\boldsymbol{X}_i)|\})\right]$$

Unfortunarly, no theoretical coverage guarantee without additional assumptions.
Better idea

$$\widehat{C}_n^{\text{jack+}}(\boldsymbol{x}) = \left[\widehat{Q}^-_{n,1-\alpha}(\{\widehat{\mu}_{-i}(\boldsymbol{x}) - |Y_i - \widehat{\mu}_{-i}(\boldsymbol{X}_i)|\})\,;\,\widehat{Q}^+_{n,\alpha}(\{\widehat{\mu}_{-i}(\boldsymbol{x}) + |Y_i - \widehat{\mu}_{-i}(\boldsymbol{X}_i)|\})\right]$$

where $\widehat{Q}^+_{n,\alpha}$ is the $\lceil(1-\alpha)(n+1)\rceil$-th ordered observation, $\widehat{Q}^+_{n,1-\alpha}$ is the $\lfloor\alpha(n+1)\rfloor$-th one.

Cross Validation and Conformal Prediction

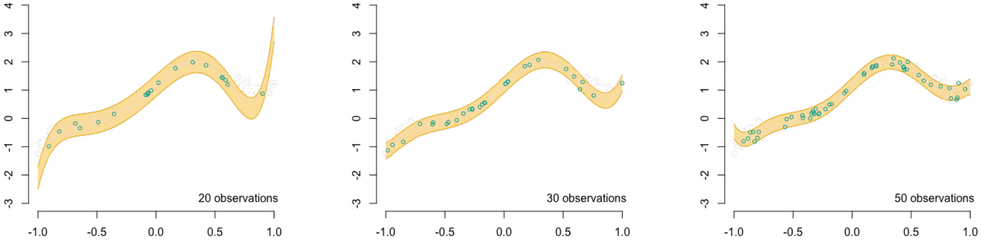
Proposition 3.5: Coverage for jackknife+ conformal, Barber et al. (2021)

If $(Y_1, \boldsymbol{X}_1), \cdots, (Y_n, \boldsymbol{X}_n), (Y_{n+1}, \boldsymbol{X}_{n+1})$ are exchangeable, the jackknife+ conformal interval satisfies

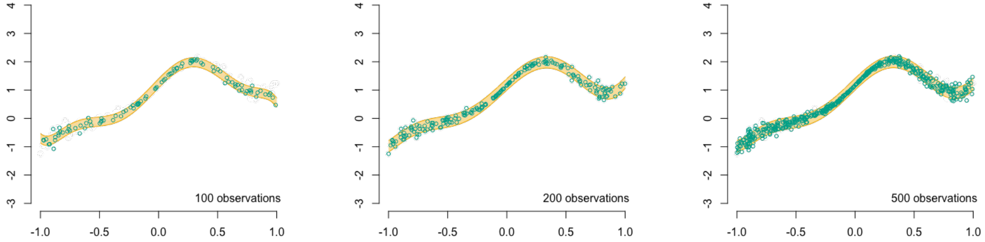
$$\mathbb{P}\Big[Y_{n+1} \in \widehat{\mathcal{C}}_n^{\text{jack+}}(\boldsymbol{X}_{n+1})\Big] \geq 1 - 2\alpha.$$

$$\Big[\widehat{Q}^{-}_{n,1-\alpha}(\{\widehat{\mu}_{-i}(\boldsymbol{X}_{n+1}) - |Y_i - \widehat{\mu}_{-i}(\boldsymbol{X}_i)|\}); \widehat{Q}^{+}_{n,\alpha}(\{\widehat{\mu}_{-i}(\boldsymbol{X}_{n+1}) + |Y_i - \widehat{\mu}_{-i}(\boldsymbol{X}_i)|\})\Big]$$

Cross Validation and Conformal Prediction



Cross Validation and Conformal Prediction



Cross Validation and Conformal Prediction

Similarly, consider a K**-fold cross-validation**, $\bigcup_{k=1}^{K}\mathcal{I}_k=\{1,2,\cdots,n\}$, for $i\in\mathcal{I}_k$,

$$\widehat{C}_n^{\text{cv-}K+}(\boldsymbol{x})=\left[\widehat{Q}^-_{n,1-\alpha}(\{\widehat{\mu}^{\text{cv}}_{-\mathcal{I}_k}(\boldsymbol{x})-|Y_i-\widehat{\mu}^{\text{cv}}_{-\mathcal{I}_k}(\boldsymbol{X}_i)|\})\,;\,\widehat{Q}^+_{n,\alpha}(\{\widehat{\mu}^{\text{cv}}_{-\mathcal{I}_k}(\boldsymbol{x})+|Y_i-\widehat{\mu}^{\text{cv}}_{-\mathcal{I}_k}(\boldsymbol{X}_i)|\})\right].$$

Proposition 3.6: Coverage for K-fold cross-validation+ conformal, Barber et al. (2021)

If $(Y_1,\boldsymbol{X}_1),\cdots,(Y_n,\boldsymbol{X}_n),(Y_{n+1},\boldsymbol{X}_{n+1})$ are exchangeable, the cv-K fold+ conformal interval satisfies

$$\mathbb{P}\Big[Y_{n+1}\in\widehat{C}_n^{\text{cv-}K+}(\boldsymbol{X}_{n+1})\Big]\geq 1-2\alpha-\min\left\{\frac{2(1-K^{-1})}{nK^{-1}+1},\frac{1-Kn^{-1}}{K+1}\right\}\geq 1-2\alpha-\sqrt{2n^{-1}}.$$

$$\left[\widehat{Q}^-_{n,1-\alpha}(\{\widehat{\mu}^{\text{cv}}_{-\mathcal{I}_k}(\boldsymbol{x})-|Y_i-\widehat{\mu}^{\text{cv}}_{-\mathcal{I}_k}(\boldsymbol{X}_i)|\});\widehat{Q}^+_{n,\alpha}(\{\widehat{\mu}^{\text{cv}}_{-\mathcal{I}_k}(\boldsymbol{x})+|Y_i-\widehat{\mu}^{\text{cv}}_{-\mathcal{I}_k}(\boldsymbol{X}_i)|\})\right]$$

Conformal Prediction, Going Further

Definition 3.24: Stabilitiy

A predictive approach is (out-of-sample) stable if it satisfies

$$\mathbb{P}\Big[\Big|\widehat{\mu}_n(\boldsymbol{X}_{n+1})-\widehat{\mu}_{-i}(\boldsymbol{X}_{n+1})\Big|\leq\varepsilon\Big]\geq 1-\alpha,\ \forall i=1,\cdots,n.$$

or if

$$\mathbb{E}\Big[\Big|\widehat{\mu}_n(\boldsymbol{X}_{n+1})-\widehat{\mu}_{-i}(\boldsymbol{X}_{n+1})\Big|\Big]\leq\beta,\ \forall i=1,\cdots,n.$$

Linear regression is stable (unless $k\sim n$), as well as Ridge, Lasso, and Bagging.

What is an “actuary”?

- “actuarial” ?

“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. Actuaries guide insurance companies in making decisions about **large categories that have the effect of attributing to the entire category certain characteristics that are probabilistically indicated by membership in the category, but that still may not be possessed by a particular member of the category**,” Schauer (2006).

PROFILES

PROBABILITIES

AND

STEREOTYPES

FREDERICK SCHAUER

The Belknap Press of Harvard University Press
Cambridge, Massachusetts
London, England

generalization is the stock in trade of the insurance industry. Indeed, the insurance industry has its own name for this kind of decisionmaking. To be an *actuary* is to be a specialist in generalization, and actuaries engage in a form of decisionmaking that is sometimes called *actuarial.* Actuaries guide insurance companies in making decisions about large categories (teenage males living in northern New Jersey) that have the effect of attributing to the entire category certain characteristics (carelessness in driving) that are probabilistically indicated by membership in the category, but that still may not be possessed by a particular member of the category (this *particular* teenage male living in northern New Jersey).

Occasionally the actuarial generalizations of the insurance industry become controversial. One example is the use of generalizations about the comparative safety of different neighborhoods as a basis for setting the rates for homeowners’ insurance or determining the willing-

What is an "actuarial model" (as in most actuarial textbooks)?

- linear regression on categories - "**segmentation**"

$+\beta_3$ ceteris paribus

$$\widehat{y}(\text{man}) = \beta_0 + \beta_1 \mathbf{1}_{\text{urban}} + \beta_2 \mathbf{1}_{\text{young}} + \boxed{\beta_3}\, \mathbf{1}_{\text{man}} = \widehat{y}(\text{woman}) + \boxed{\beta_3}$$

- Poisson regression (frequency) on categories, or not

$$\widehat{y}(\text{man}) = \exp\left[\beta_0 + \beta_1 \mathbf{1}_{\text{urban}} + \beta_2 \boxed{\mathbf{1}_{\text{young}}} + \boxed{\beta_3}\, \mathbf{1}_{\text{man}}\right] = \widehat{y}(\text{woman}) \cdot \boxed{\exp[\beta_3]}$$

$\times e^{\beta_3}$ ceteris paribus

$$\widehat{y}(\text{man}) = \exp\left[\beta_0 + \beta_1 \mathbf{1}_{\text{urban}} + \beta_2 \boxed{\text{age}} + \boxed{\beta_3}\, \mathbf{1}_{\text{man}}\right] = \widehat{y}(\text{woman}) \cdot \boxed{\exp[\beta_3]}$$

If β_3 small, $e^{\beta_3} \approx 1 + \beta_3$, i.e. "$\beta_3 = 0.2$" $\longleftrightarrow$ "+20% for men"

Thus "**interpretation**" is simple (if we do not discuss what "ceteris paribus" means).

Why could there be a problem?

- **Econometrics** is dead, long live "**artificial intelligence**"
- "**Machine learning**" context, i.e. black boxes, with less intuitive interpretation
- "**Big data**" context, i.e. easy to get proxies for protected/sensitive variables

y	urban	age	race
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

y	urban	age	zip	lastname	model	credit
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮

It is possible to predict the "race" based on non-protected variables, e.g. names and geolocation, see "Bayesian Improved Surname Geocoding (BISG)", Elliott et al. (2009), Imai and Khanna (2016)

"OK, let's not use race, but should we use zip code, which of course is a proxy for race in our segregated society?," O'Neil (2016).

Where could there be a problem?

Ratemaking is an issue, but also **underwriting**,

"**Redlining**", for loans, but also insurance, Kerner (1968)

"use of a **red line around the questionable areas on territorial maps** centrally located in the Underwriting Division for ease of reference by all Underwriting personnel [...] mark off certain areas * * * to denote a lack of interest in business arising in these areas In New York these are called K.O. areas meaning **knock-out areas**; in Boston they are called **redline districts. Same thing – don't write the business**s."

to requests for information reveal clearly that business in certain geographic territories is restricted. For example, one underwriting guide states:

"An underwriter should be aware of the following situations in his territory:

1. The blighted areas.
2. The redevelopment operations.
3. Peculiar weather conditions which might make for a concentration of windstorm or hail losses.
4. The economic makeup of the area.
5. The nature of the industries in the area, etc.

"This knowledge can be gathered by drives through the area, by talking to and visiting agents, and by following local newspapers as to incidents of crimes and fires. A good way to keep this information available and up to date is by *the use of a red line* around the questionable areas on territorial maps centrally located in the Underwriting Division for ease of reference by all Underwriting personnel." (Italics added.)

A New York City insurance agent at our hearings put it more pointedly:

"[M]ost companies mark off certain areas * * * to denote a lack of interest in business arising in these areas In New York these are called K.O. areas—meaning knock-out areas; in Boston they are called redline districts. Same thing—don't write the business."

What is a "actuarial fairness"?

- "**Actuarial fairness**" ?

... "on an **actuarially fair** basis; that is, if the costs of medical care are a random variable with mean m, the company will charge a premium m, and agree to indemnify the individual for all medical costs," Arrow (1963).

"**actuarially fair premiums**" = "**expected losses**"

of the insured risk, see also Frezal and Barry (2020).

THE AMERICAN ECONOMIC REVIEW

VOLUME LIII DECEMBER 1963 NUMBER 5

UNCERTAINTY AND THE WELFARE ECONOMICS OF MEDICAL CARE

By KENNETH J. ARROW*

the latter. Suppose, therefore, an agency, a large insurance company plan, or the government, stands ready to offer insurance against medical costs on an actuarially fair basis; that is, if the costs of medical care are a random variable with mean m, the company will charge a premium m, and agree to indemnify the individual for all medical costs. Under these circumstances, the individual will certainly prefer to take out a policy and will have a welfare gain thereby.

Will this be a social gain? Obviously yes, if the insurance agent is suffering no social loss. Under the assumption that medical risks on different individuals are basically independent, the pooling of them reduces the risk involved to the insurer to relatively small proportions.

"governments must recognise that there is a difference between unfair discrimination and insurers differentiating prices according to risk," Swiss Re (2015), cited in Meyers and Van Hoyweghen (2018)

What is a “actuarial fairness”?

"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."[11]

“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,” Casey et al. (1976), cited in Walters (1981)

So "actuarial fairness" has to do with "accuracy"?

Following Arrow (1963), "**actuarially fair premiums**" = "**expected losses**"

- but still, there is no "**law of one price**" in insurance, Froot et al. (1995)

$\rightarrow$ with different models and different portfolio, we can have two different premiums

- estimating "**expected losses**" means maximizing "**accuracy**"

average losses / empirical losses

$$\boxed{\overline{y}} = \underset{\gamma\in\mathbb{R}}{\text{argmin}}\left\{\boxed{\sum_{i=1}^n (y_i-\gamma)^2}\right\} \text{ or } \boxed{\mathbb{E}[Y]} = \underset{\gamma\in\mathbb{R}}{\text{argmin}}\left\{\boxed{\sum_y (y-\gamma)^2\mathbb{P}[Y=y]}\right\}$$

least squares

i.e. we want to minimize the error between observed loses y and predictions $\widehat{y}$.

with binary observations $y\in\{0,1\}$, hard to assess if $\widehat{y}=12.2486\%$ is accurate or not...

Discrimination? Individual vs. Group Treatment

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

- "**Ten Oever**" judgement (*Gerardus Cornelis Ten Oever v Stichting Bedrijfspensioenfonds voor het Glazenwassers – en Schoonmaakbedrijf*, in April 1993), the Advocate General Van Gerven (1993) 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**," from Britz (2008) ("**Generalisierungsunrecht**")

→ Actuarial pricing is essentially discriminatory... and unfair.

“At the core of insurance business lies discrimination”.

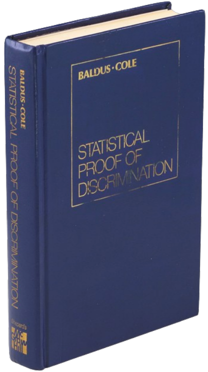
- ”What is unique about insurance is that **even statistical discrimination which by definition is absent of any malicious intentions, poses significant moral and legal challenges**. Why? Because on the one hand, policy makers would like insurers to treat their insureds equally, without discriminating 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. ” Avraham (2017)

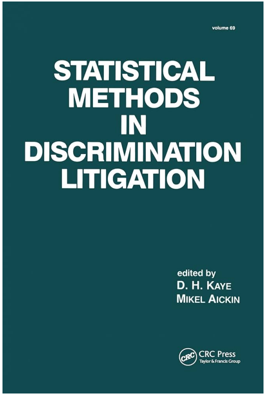
- “Technology is neither good nor bad; nor is it neutral,” Kranzberg (1986)

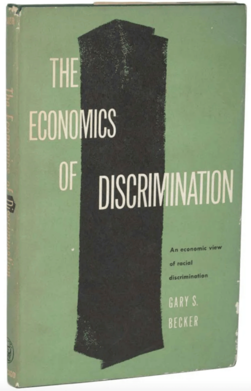
- “Machine learning won’t give you anything like gender neutrality ‘for free’ that you didn’t explicitly ask for,” Kearns and Roth (2019)

Quantifying discrimination, isn't it an old problem?

See Becker (1957) or Baldus and Cole (1980), among (many) others.







Several papers over the past 15 years revisited various notions and concepts.

Is there a (simple) way to quantify unfairness ?

- classical fairness concept are related to so called "**group fairness**", where we have a statistical (overall perspective),
- in some problems, we focus on discrimination in "continuous outcomes",
  - $\widehat{m}(\boldsymbol{x}_i, s_i) \in [0, 1]$ (score) that could also be denoted $\widehat{y}_i$
  - $\widehat{m}(\boldsymbol{x}_i, s_i) \in \mathbb{R}_+$ (premium) that could also be denoted $\widehat{y}_i$

  $\rightarrow$ classical in insurance modeling
- in some problems, we focus on discrimination in binary decisions $\widehat{y}_i \in \{0, 1\}$, usually obtained as
  - $\widehat{y}_i = \mathbf{1}(\widehat{m}(\boldsymbol{x}_i, s_i) > \text{threshold}) \in \{0, 1\}$ (class) that could also be denoted

  $\rightarrow$ classical in computer science

Several definitions of “fairness” or “non-discriminatory”

Definition 3.25: Fairness through unawareness, Dwork et al. (2012)

A model m satisfies the fairness through unawareness criteria, with respect to sensitive attribute $s \in \mathcal{S}$ if $m : \mathcal{X} \to \mathcal{Y}$.

“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,” Apfelbaum et al. (2010)

Several definitions of “fairness” or “non-discriminatory”

Definition 3.26: Four definitions of cultural fairness, Darlington (1971)

A test ($\widehat{y}$) is considered "**culturally fair**" if it fits the appropriate equation

$$\begin{cases} \text{Cor}[S,\widehat{Y}] = \text{Cor}[S,Y]/\text{Cor}[Y,\widehat{Y}] \\ \text{Cor}[S,\widehat{Y}] = \text{Cor}[S,Y] \\ \text{Cor}[S,\widehat{Y}] = \text{Cor}[S,Y]\cdot\text{Cor}[Y,\widehat{Y}] \\ \text{Cor}[S,\widehat{Y}] = 0 \end{cases}$$

See also Thorndike (1971), Linn and Werts (1971), following Cleary (1968).

Several definitions of "fairness" or "non-discriminatory"

Definition 3.27: 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 $(\boldsymbol{X}, S, Y)$.

For classifiers, one might ask for independence $\widehat{Y} \perp\!\!\!\perp S$ (where $\widehat{y}$ is a class), as Darlington (1971).

Several definitions of "fairness" or "non-discriminatory"

Definition 3.28: Demographic Parity, Calders and Verwer (2010), Corbett-Davies et al. (2017)

A decision function $\widehat{y}$ – or a classifier m_t, taking values in $\{0,1\}$ – satisfies demographic parity, with respect to some sensitive attribute S if (equivalently)

$$\begin{cases}\mathbb{P}[\widehat{Y}=1|S=\mathtt{A}]=\mathbb{P}[\widehat{Y}=1|S=\mathtt{B}]=\mathbb{P}[\widehat{Y}=1]\\ \mathbb{E}[\widehat{Y}|S=\mathtt{A}]=\mathbb{E}[\widehat{Y}|S=\mathtt{B}]=\mathbb{E}[\widehat{Y}]\\ \mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{A}]=\mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{B}]=\mathbb{P}[m_t(\boldsymbol{Z})=1].\end{cases}$$

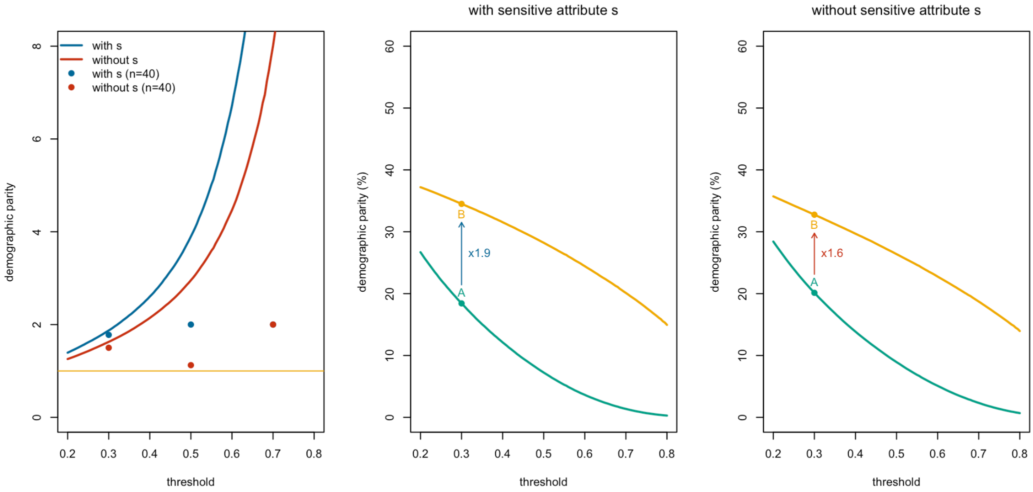
Several definitions of “fairness” or “non-discriminatory”

	unaware (without s)				aware (with s)			
	GLM	GAM	CART	RF	GLM	GAM	CART	RF
$n = 1000$, various t, ratio $\mathbb{P}[\widehat{Y}=1\|S=\texttt{B}]/\mathbb{P}[\widehat{Y}=1\|S=\texttt{A}]$								
$t = 30\%$	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

(`dem_parity` from R package `fairness`)

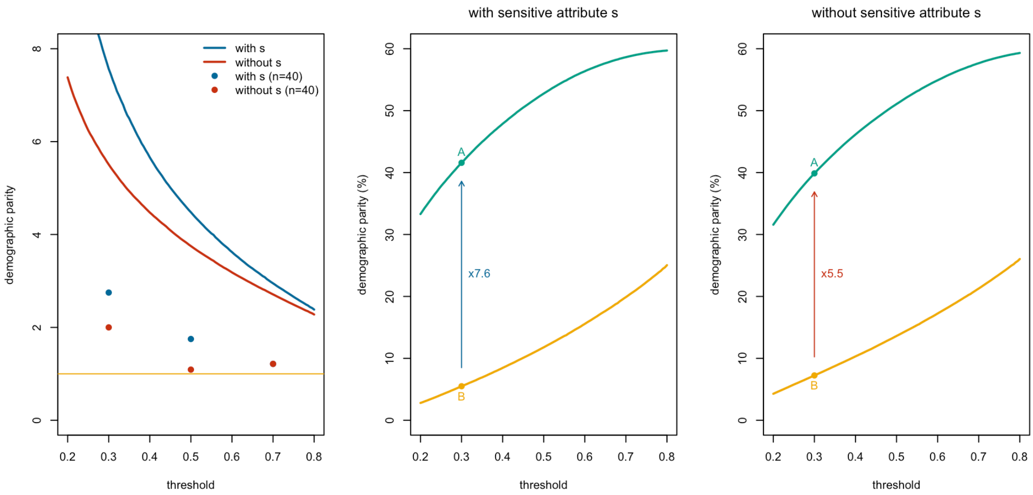
On the left-hand side, evolution of the ratio ratio $\mathbb{P}[\widehat{Y}=1|S=\texttt{B}]/\mathbb{P}[\widehat{Y}=1|S=\texttt{A}]$.
The horizontal line (at $y=1$) corresponds to perfect demographic parity.
In the middle $t \mapsto \mathbb{P}[m_t(\boldsymbol{X}) > t|S=\texttt{B}]$ and $t \mapsto \mathbb{P}[m_t(\boldsymbol{X}) > t|S=\texttt{A}]$ on the model with s, and on the right-hand side without s.

Several definitions of “fairness” or “non-discriminatory”



On the left-hand side, evolution of the ratio ratio $\mathbb{P}[\widehat{Y}=1|S=\text{B}]/\mathbb{P}[\widehat{Y}=1|S=\text{A}]$.

Several definitions of “fairness” or “non-discriminatory”



On the left-hand side, evolution of the ratio ratio $\mathbb{P}[\widehat{Y}=0|S=\mathtt{A}]/\mathbb{P}[\widehat{Y}=0|S=\mathtt{B}]$

Several definitions of “fairness” or “non-discriminatory”

Definition 3.29: Weak Demographic Parity

A decision function $\widehat{y}$ satisfies weak demographic parity if

$$\mathbb{E}[\widehat{Y}|S=\texttt{A}]=\mathbb{E}[\widehat{Y}|S=\texttt{B}].$$

Definition 3.30: Strong Demographic Parity

A decision function $\widehat{y}$ satisfies demographic parity if $\widehat{Y} \perp\!\!\!\perp S$, i.e., for all A,

$$\mathbb{P}[\widehat{Y}\in\mathcal{A}|S=\texttt{A}]=\mathbb{P}[\widehat{Y}\in\mathcal{A}|S=\texttt{B}],\ \forall\mathcal{A}\subset\mathcal{Y}.$$

Several definitions of “fairness” or “non-discriminatory”

Proposition 3.7

A model m satisfies the strong demographic parity property if and only if

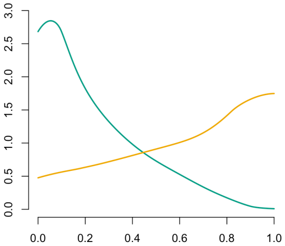
$$d_{\mathrm{TV}}(\mathbb{P}_{m|\mathtt{A}},\mathbb{P}_{m|\mathtt{B}}) = d_{\mathrm{TV}}(\mathbb{P}_{\mathtt{A}},\mathbb{P}_{\mathtt{B}}) = 0.$$

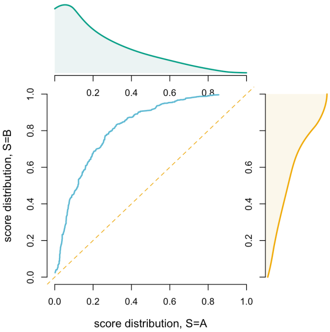
$d_{\mathrm{TV}}(\mathbb{P}_{m|\mathtt{A}},\mathbb{P}_{m|\mathtt{B}})$ could be seen as a measure of “unfairness”, but for a non-binary sensitive attribute, a more general definition is necessary (see Denis et al. (2021)).

Proposition 3.8

A model m satisfies is strongly fair if and only if $W_2(\mathbb{P}_{\mathtt{A}},\mathbb{P}_{\mathtt{B}}) = 0$.

Several definitions of “fairness” or “non-discriminatory”

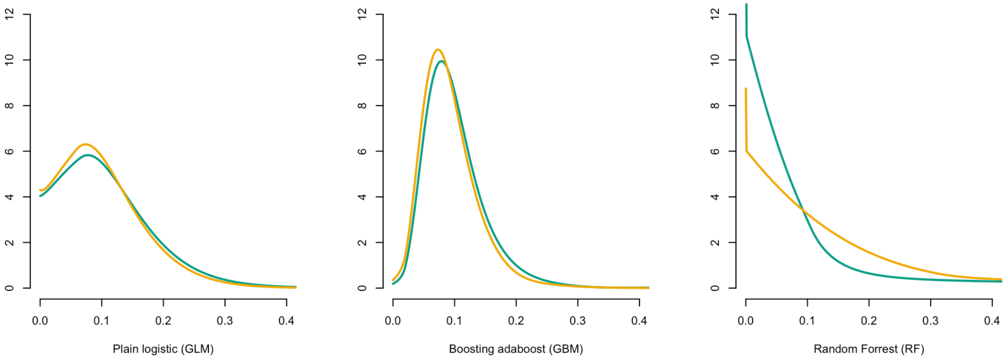





```
> model_glm = glm(y~x1+
    x2+x3, data=
    toydata2, family=
    binomial)
> pred_y_glm = predict(
    model_glm, type="
    response")
> sA = pred_y_glm[
    toydata2$sensitive
    =="A"]
> library(transport)
> wasserstein1d(sA,sB)
[1] 0.3860795
```


Several definitions of “fairness” or “non-discriminatory”

On the `FrenchMotor` dataset, consider GLM, GBM and RF for claim occurence




```
> wasserstein1d(lA,lB)
[1] 0.007220468
```

```
> wasserstein1d(bA,bB)
[1] 0.008895917
```

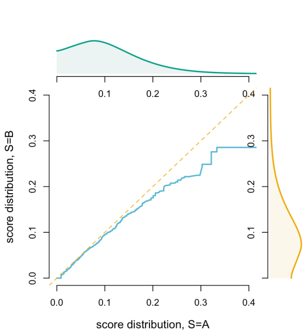
```
> wasserstein1d(fA,fB)
[1] 0.01001088
```

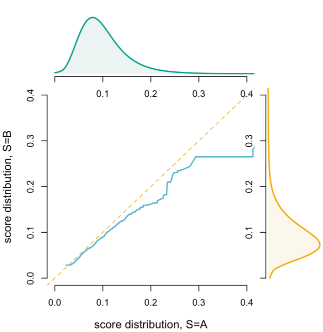
Several definitions of “fairness” or “non-discriminatory”

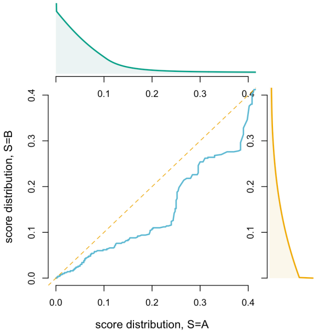
```
> wasserstein1d(lA,lB)
[1] 0.007220468
```

```
> wasserstein1d(bA,bB)
[1] 0.008895917
```

```
> wasserstein1d(fA,fB)
[1] 0.01001088
```







Several definitions of "fairness" or "non-discriminatory"

Definition 3.31: Unfairness, Denis et al. (2021); Chzhen and Schreuder (2022)

Given a model m, let $\mathbb{P}_m$ denote the distribution of $m(\boldsymbol{X}, S)$ and $\mathbb{P}_{m|s}$ denote the conditional distribution of $m(\boldsymbol{X}, S)$ given $S = s$, define

$$\begin{cases} \mathcal{U}_{\mathrm{TV}}(m) = \max_{s\in\{\mathtt{A},\mathtt{B}\}} \{d_{\mathrm{TV}}(\mathbb{P}_m, \mathbb{P}_{m|s}) \text{ or } \sum_{s\in\{\mathtt{A},\mathtt{B}\}} d_{\mathrm{TV}}(\mathbb{P}_m, \mathbb{P}_{m|s}) \\ \mathcal{U}_{\mathrm{KS}}(m) = \max_{s\in\{\mathtt{A},\mathtt{B}\}} \{d_{\mathrm{KS}}(\mathbb{P}_m, \mathbb{P}_{m|s})\} \text{ or } \sum_{s\in\{\mathtt{A},\mathtt{B}\}} d_{\mathrm{KS}}(\mathbb{P}_m, \mathbb{P}_{m|s}) \\ \mathcal{U}_{\mathrm{W}_k}(m) = \max_{s\in\{\mathtt{A},\mathtt{B}\}} \{W_k(\mathbb{P}_m, \mathbb{P}_{m|s})\} \text{ or } \sum_{s\in\{\mathtt{A},\mathtt{B}\}} W_k(\mathbb{P}_m, \mathbb{P}_{m|s}) \end{cases}$$

In the original version, Chzhen and Schreuder (2022) suggested to use the one on the right.

Several definitions of “fairness” or “non-discriminatory”

Those measures characterize strong demographic parity,

Proposition 3.9: Strong Demographic Parity

A model m is strongly fair if and only if $\mathcal{U}(m) = 0$.

Separation and Equalized Odds

Definition 3.32: 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 $(\boldsymbol{X}, S, Y)$.

Separation and Equalized Odds

Definition 3.33: 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

$$\begin{cases}\mathbb{P}[\widehat{Y}=1|S=\mathtt{A},Y=1]=\mathbb{P}[\widehat{Y}=1|S=\mathtt{B},Y=1]=\mathbb{P}[\widehat{Y}=1|Y=1]\\ \mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{A},Y=1]=\mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{B},Y=1]=\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=1],\end{cases}$$

which corresponds to parity of true positives, in the two groups, $\{\mathtt{A},\mathtt{B}\}$.

Separation and Equalized Odds

Definition 3.34: Strong Equal Opportunity

A classifier $m(\cdot)$, taking values in $\{0,1\}$, satisfies equal opportunity, with respect to some sensitive attribute S if

$$\mathbb{P}[m(\boldsymbol{X},S)\in\mathcal{A}|S=\mathtt{A},Y=1]=\mathbb{P}[m(\boldsymbol{X},S)\in\mathcal{A}|S$$

for all $\mathcal{A}\subset[0,1]$.

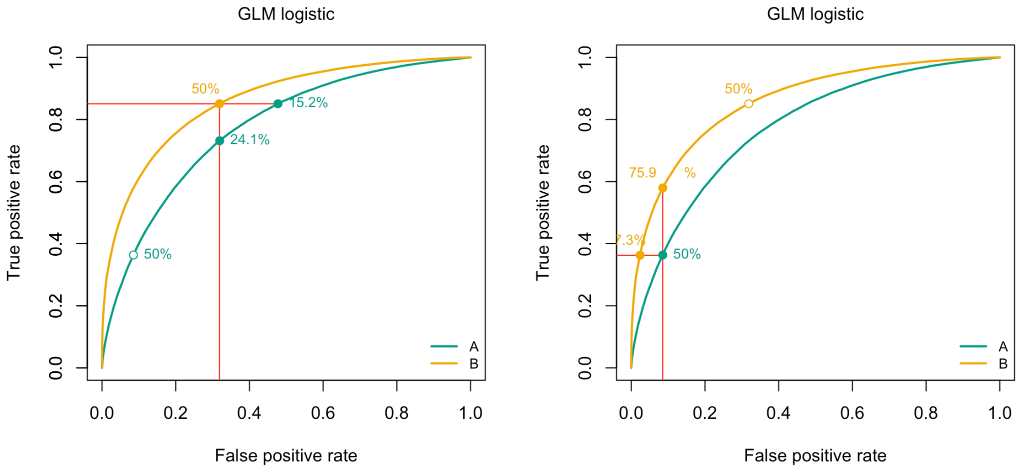
Separation and Equalized Odds

Definition 3.35: 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

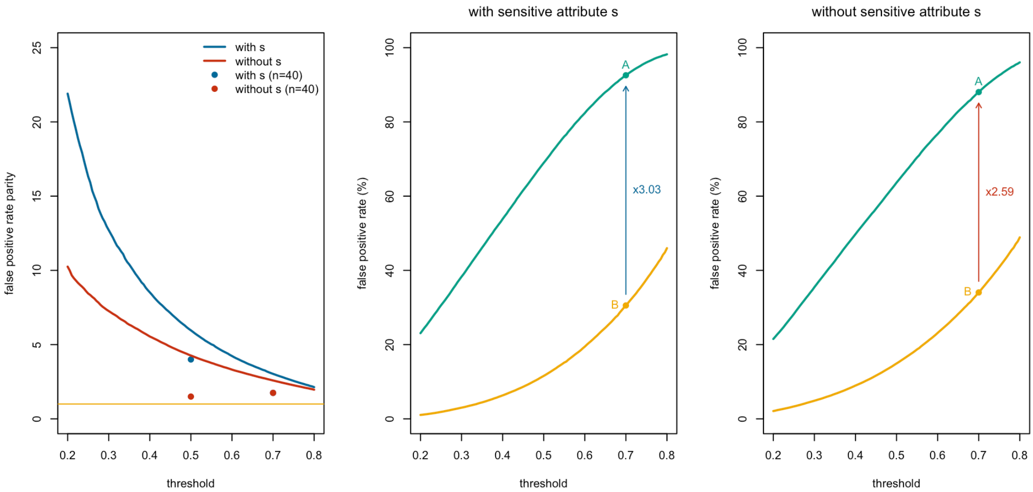
$$\begin{cases}\mathbb{P}[\widehat{Y}=1|S=\mathtt{A},Y=0]=\mathbb{P}[\widehat{Y}=1|S=\mathtt{B},Y=0]=\mathbb{P}[\widehat{Y}=1|Y=0]\\ \mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{A},Y=0]=\mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{B},Y=0]=\mathbb{P}[m_t(\boldsymbol{Z})=1|Y=0].\end{cases}$$

Separation and Equalized Odds



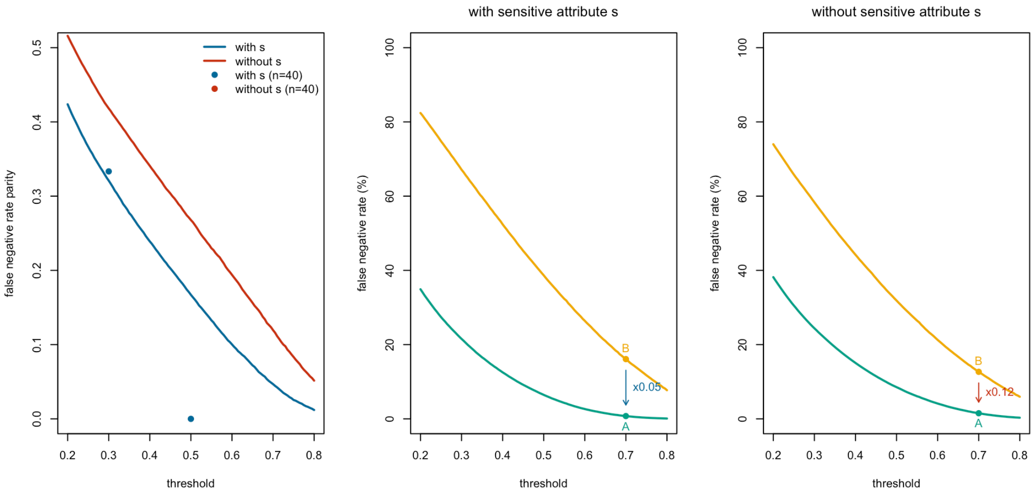
ROC curves (TPR against FPR) for the logistic regression on `toydata2`.

Separation and Equalized Odds



Evolution of the false positive rates, `fpr_parity` from `fairness`.

Separation and Equalized Odds



Evolution of the false negative rates, `fnr_parity` from `fairness`.

Separation and Equalized Odds

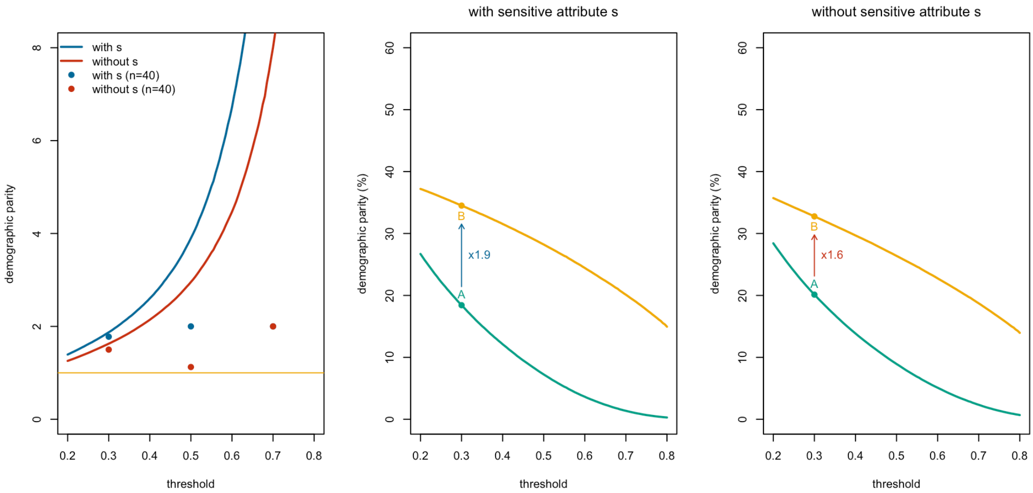
Definition 3.36: Equalized Odds, Hardt et al. (2016)

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

$$\begin{cases}\mathbb{P}[\widehat{Y}=1|S=\mathtt{A},Y=y]=\mathbb{P}[\widehat{Y}=1|S=\mathtt{B},Y=y]=\mathbb{P}[\widehat{Y}=1|Y=y],\ \forall y\in\{0,1\}\\ \mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{A},Y=y]=\mathbb{P}[m_t(\boldsymbol{Z})=1|S=\mathtt{B},Y=y],\ \forall y\in\{0,1\},\end{cases},$$

which corresponds to parity of true positive and false positive, in the two groups.

Separation and Equalized Odds



Evolution of the equalized odds metrics

Separation and Equalized Odds

One can also consider any kind of standard metrics on confusion matrices, such as ϕ (introduced in Yule (1912)), usually named "Matthews correlation coefficient"

Definition 3.37: ϕ-fairness, Chicco and Jurman (2020)

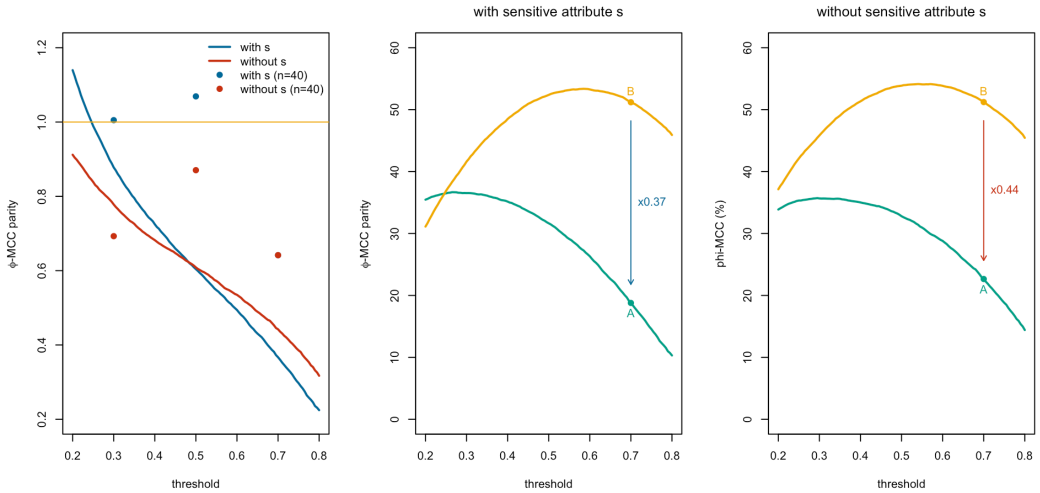
We will have ϕ-fairness if $\phi_{\texttt{A}} = \phi_{\texttt{B}}$, where ϕ_s denotes Matthews correlation coefficient for the s group,

$$\phi_s = \frac{\text{TP}_s \cdot \text{TN}_s - \text{FP}_s \cdot \text{FN}_s}{\sqrt{(\text{TP}_s + \text{FP}_s)(\text{TP}_s + \text{FN}_s) \cdot (\text{TN}_s + \text{FP}_s)(\text{TN}_s + \text{FN}_s)}},\ s \in \{\texttt{A}, \texttt{B}\}.$$

but one could consider the F_1-score (as defined in Van Rijsbergen (1979)), Fowlkes–Mallows or Jaccard indices (in Fowlkes and Mallows (1983) or Jaccard (1901)).

... or AUC as we will considered later on.

Separation and Equalized Odds



Evolution of the ϕ-fairness metric

Separation and Equalized Odds

Definition 3.38: Class Balance, Kleinberg et al. (2016)

We will have class balance in the weak sense if

$$\mathbb{E}[m(\boldsymbol{X})|Y=y,S=\texttt{A}]=\mathbb{E}[m(\boldsymbol{X})|Y=y,S=\texttt{B}],\ \forall y\in\{0,1\},$$

or in the strong sense if

$$\mathbb{P}[m(\boldsymbol{X})\in\mathcal{A}|Y=y,S=\texttt{A}]=\mathbb{P}[m(\boldsymbol{X})\in\mathcal{A}|Y=y,S=\texttt{B}],\ \forall\mathcal{A}\subset[0,1],\ \forall y\in\{0,1\}.$$

Separation and Equalized Odds

Definition 3.39: Similar Mistreatement, Zafar et al. (2019)

We will have similar mistreatment, or "*lack of disparate mistreatment*," if

$$\begin{cases}\mathbb{P}[\widehat{Y}=Y|S=\mathtt{A}]=\mathbb{P}[\widehat{Y}=Y|S=\mathtt{B}]=\mathbb{P}[\widehat{Y}=Y]\\ \mathbb{P}[m_t(\boldsymbol{X})=Y|S=\mathtt{A}]=\mathbb{P}[m_t(\boldsymbol{X})=Y|S=\mathtt{B}]=\mathbb{P}[m_t(\boldsymbol{X})=Y].\end{cases}$$

Definition 3.40: Equality of ROC curves, Vogel et al. (2021)

Let $\text{FRP}_s(t)=\mathbb{P}[m(\boldsymbol{X})>t|Y=0,S=s]$ and $\text{TPR}_s(t)=\mathbb{P}[m(\boldsymbol{X})>t|Y=1,S=s]$, where $s\in\{\mathtt{A},\mathtt{B}\}$. Set $\Delta_{TPR}(t)=\text{TPR}_{\mathtt{B}}\circ\text{TPR}_{\mathtt{A}}^{-1}(t)-t$ et $\Delta_{FRP}(t)=\text{FPR}_{\mathtt{B}}\circ\text{FPR}_{\mathtt{A}}^{-1}(t)-t$. We will have fairness with respect to ROC curves if $\|\Delta_{TPR}\|_\infty=\|\Delta_{FPR}\|_\infty=0$.

Separation and Equalized Odds

Definition 3.41: AUC Fairness, Borkan et al. (2019)

We will have AUC fairness if $\text{AUC}_{\mathtt{A}} = \text{AUC}_{\mathtt{B}}$, where AUC_s is the AUC associated with model m within the s group.

	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

Sufficiency and Calibration

Inspired by Cleary (1968), define

Definition 3.42: Sufficiency, Barocas et al. (2017)

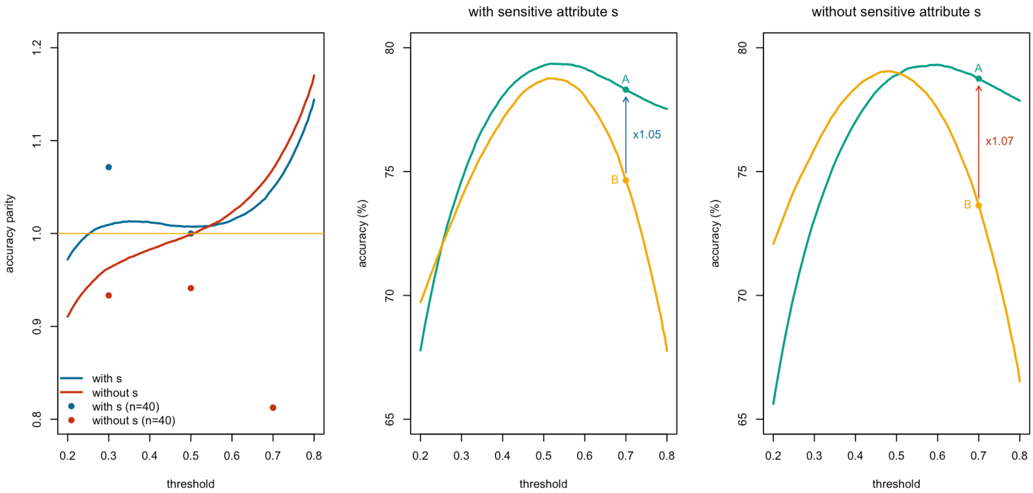
A model $m : \mathcal{Z} \to \mathcal{Y}$ satisfies the sufficiency property if $Y \perp\!\!\!\perp S \mid m(\boldsymbol{Z})$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$.

Definition 3.43: Calibration Parity, Accuracy Parity, Kleinberg et al. (2016), Zafar et al. (2019)

Calibration parity is met if

$$\mathbb{P}[Y = 1 | m(\boldsymbol{X}) = t, S = \texttt{A}] = \mathbb{P}[Y = 1 | m(\boldsymbol{X}) = t, S = \texttt{B}], \ \forall t \in [0, 1].$$

Sufficiency and Calibration



Evolution of accuracy, in groups A and B.

Sufficiency and Calibration

Definition 3.44: Good Calibration, Kleinberg et al. (2017), Verma and Rubin (2018)

Fairness of good calibration is met if

$$\mathbb{P}[Y=1|m(\boldsymbol{X})=t, S=\texttt{A}]=\mathbb{P}[Y=1|m(\boldsymbol{X})=t, S=\texttt{B}]=t,\ \forall t\in[0,1].$$

Definition 3.45: Non-Reconstruction of 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

$$\mathbb{P}[S=\texttt{A}|\boldsymbol{X}, m(\boldsymbol{X}), \widehat{Y}, Y]=\mathbb{P}[S=\texttt{B}|\boldsymbol{X}, m(\boldsymbol{X}), \widehat{Y}, Y].$$

Relaxation and Approximate Fairness

Definition 3.46: Disparate Impact, Feldman et al. (2015)

A decision function $\widehat{Y}$ has a disparate impact, for a given threshold τ, if,

$$\min\left\{\frac{\mathbb{P}[\widehat{Y}=1|S=\texttt{A}]}{\mathbb{P}[\widehat{Y}=1|S=\texttt{B}]},\frac{\mathbb{P}[\widehat{Y}=1|S=\texttt{B}]}{\mathbb{P}[\widehat{Y}=1|S=\texttt{A}]}\right\}<\tau \text{ (usually 80\%).}$$

The 80% rule was suggested by the "Technical Advisory Committee on Testing", from the State of California Fair Employment Practice Commission (FEPC) in 1971, or the 1978 "Uniform Guidelines on Employee Selection Procedures", a document used by the U.S. Equal Employment Opportunity Commission (EEOC), see Biddle (2017).

Relaxation and Approximate Fairness

We have defined (Definition 3.31) unfairness as

$$\mathcal{U}_k(m) = \max_{s\in\{\mathtt{A},\mathtt{B}\}} \{W_k(\mathbb{P}_m, \mathbb{P}_{m|s})\},$$

so that m is (strongly) fair if and only if $\mathcal{U}_k(m) = 0$.

Chzhen and Schreuder (2022) introduced the notion of Relative Improvement

Definition 3.47: ε-Approximate Fairness

Model m is ε-approximately fair if $\mathcal{U}_k(m) \leq \varepsilon \cdot \mathcal{U}_k(m^\star)$, where $m^\star$ is Bayes regressor, for some $\epsilon \geq 0$.

Three different concepts ?

$$\begin{cases} \text{Independence (Definition 3.27)}: & m(\boldsymbol{Z}) \perp\!\!\!\perp S \\ \text{Separation (Definition 3.32)}: & m(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y \\ \text{Sufficiency (Definition 3.42)}: & Y \perp\!\!\!\perp S \mid m(\boldsymbol{Z}) \end{cases}$$

- Independence assumes no differences among groups, regardless of accuracy
- Separation minimizes differences among groups by not trying to maximize accuracy
- Sufficiency maximizes accuracy by not trying to minimize differences among groups

See Kleinberg et al. (2016) or Chouldechova (2017).

Impossibility theorems

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.

$$\begin{cases} \text{Independence (Definition 3.27)}: & m(\boldsymbol{Z}) \perp\!\!\!\perp S \\ \text{Separation (Definition 3.32)}: & m(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y \\ \text{Sufficiency (Definition 3.42)}: & Y \perp\!\!\!\perp S \mid m(\boldsymbol{Z}) \end{cases}$$

Proposition 3.10

Suppose that a model m satisfies the independence condition (3.27) and the sufficiency property (3.42), with respect to a sensitive attribute s, then necessarily, $Y \perp\!\!\!\perp S$.

Therefore, unless the sensitive attribute s has no impact on the outcome y, there is no model m which satisfies independence and sufficiency simultaneously.

Impossibility theorems

From the sufficiency property , $S \perp\!\!\!\perp Y \mid m(\boldsymbol{Z})$, 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

$$\mathbb{P}[S = s, Y \in \mathcal{A}] = \mathbb{E}\big[\mathbb{P}[S = s | m(\boldsymbol{Z})] \cdot \mathbb{P}[Y \in \mathcal{A} | m(\boldsymbol{Z})]\big].$$

And from the independence property (3.42), $m(\boldsymbol{Z}) \perp\!\!\!\perp S$, we can write the first component $\mathbb{P}[S = s | m(\boldsymbol{Z})] = \mathbb{P}[S = s]$, almost surely, and therefore

$$\mathbb{P}[S = s, Y \in \mathcal{A}] = \mathbb{E}\big[\mathbb{P}[S = s] \cdot \mathbb{P}[Y \in \mathcal{A} | m(\boldsymbol{Z})]\big] = \mathbb{P}[S = s] \cdot \mathbb{P}[Y \in \mathcal{A}],$$

for all $s \in \mathcal{S}$ and $\mathcal{A} \subset \mathcal{Y}$, corresponding to the independence between S and Y.

Impossibility theorems

Proposition 3.11

Consider a classifier m_t taking values in $\mathcal{Y}=\{0,1\}$. Suppose that m_t satisfies the independence condition (3.27) and the separation property (3.32), 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 (3.27), $m_t(\boldsymbol{Z}) \perp\!\!\!\perp S$, and the separation property (3.32), $m_t(\boldsymbol{Z}) \perp\!\!\!\perp S \mid Y$, them, for $\widehat{y}\in\mathcal{Y}$ and for $s\in\mathcal{S}$,

$$\mathbb{P}[m_t(\boldsymbol{Z})=\widehat{y}]=\mathbb{P}[m_t(\boldsymbol{Z})=\widehat{y}|S=s]=\mathbb{E}\big[\mathbb{P}[m_t(\boldsymbol{Z})=\widehat{y}|Y,S=s]\big],$$

that we can write

$$\mathbb{P}[m_t(\boldsymbol{Z})=\widehat{y}]=\sum_y \mathbb{P}\big[m_t(\boldsymbol{Z})=\widehat{y}\big|Y=y,S=s\big]\cdot\mathbb{P}\big[Y=y\big|S=s\big],$$

Impossibility theorems

or

$$\mathbb{P}[m_t(\boldsymbol{Z})=\widehat{y}]=\sum_y \mathbb{P}\big[m_t(\boldsymbol{Z})=\widehat{y}\big|Y=y\big]\cdot\mathbb{P}\big[Y=y\big|S=s\big],$$

almost surely. Furthermore, we can also write

$$\mathbb{P}[m_t(\boldsymbol{Z})=\widehat{y}]=\sum_y \mathbb{P}\big[m_t(\boldsymbol{Z})=\widehat{y}\big|Y=y\big]\cdot\mathbb{P}\big[Y=y\big],$$

so that, if we combine the two expressions, we get

$$\sum_y \mathbb{P}\big[m_t(\boldsymbol{Z})=\widehat{y}\big|Y=y\big]\cdot\Big(\mathbb{P}\big[Y=y\big|S=s\big]-\mathbb{P}\big[Y=y\big]\Big)=0,$$

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

$$\mathbb{P}\big[m_t(\boldsymbol{Z})=\widehat{y}\big|Y=1\big]\cdot\Big(\mathbb{P}\big[Y=1\big|S=s\big]-\mathbb{P}\big[Y=1\big]\Big)$$

Impossibility theorems

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.

Impossibility theorems

Proposition 3.12

Consider a classifier m_t taking values in $\mathcal{Y} = \{0,1\}$. Suppose that m_t satisfies the sufficiency condition (3.42) and the separation property (3.32), with respect to a sensitive attribute s, then necessarily either $\mathbb{P}[m_t(\boldsymbol{Z}) = 1|Y = 1] = 0$ or $Y \perp\!\!\!\perp S$ or $m_t(\boldsymbol{Z}) \perp\!\!\!\perp Y$.

Suppose that m_t satisfies the sufficiency condition (3.42) and the separation property (3.32), 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

$$\mathbb{P}[Y = 1|S = s, m_t(\boldsymbol{Z}) = 1] = \frac{\mathbb{P}[m_t(\boldsymbol{Z}) = 1|Y = 1, S = s] \cdot \mathbb{P}[Y = 1|S = s]}{\mathbb{P}[m_t(\boldsymbol{Z}) = 1|S = s]},$$

Impossibility theorems

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

$$\mathbb{P}[Y=1|S=s,m_t(\boldsymbol{Z})=1]=\frac{\text{TPR}}{p_s\cdot\text{TPR}+(1-p_s)\cdot\text{FPR}}.$$

Impossibility theorems

Suppose 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

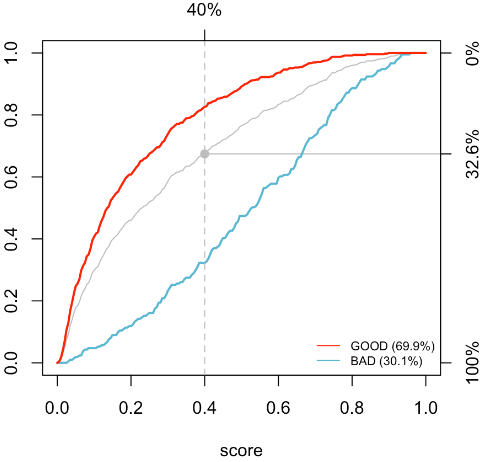
$$\frac{\text{TPR}}{p_s \cdot \text{TPR} + (1 - p_s) \cdot \text{FPR}} = \frac{\text{TPR}}{p_{s'} \cdot \text{TPR} + (1 - p_{s'}) \cdot \text{FPR}}.$$

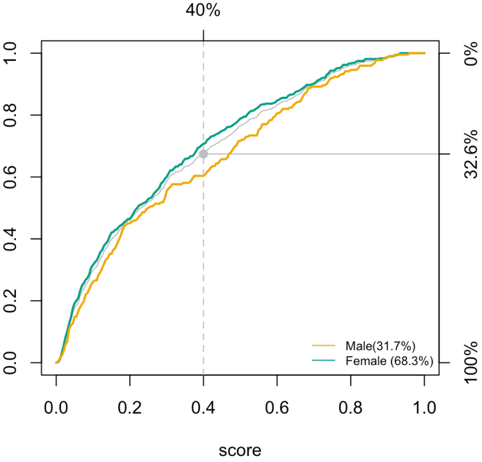
Supposes that $\text{TPR} \neq 0$ (otherwise $\text{TPR} = \mathbb{P}[m_t(\boldsymbol{Z}) = 1|Y = 1] = 0$ as stated in the proposition), then

$$(p_s - p_{s'}) \cdot \text{TPR} = (p_s - p_{s'}) \cdot \text{FPR} \neq 0,$$

and therefore $m_t(\boldsymbol{Z}) \perp\!\!\!\perp Y$.

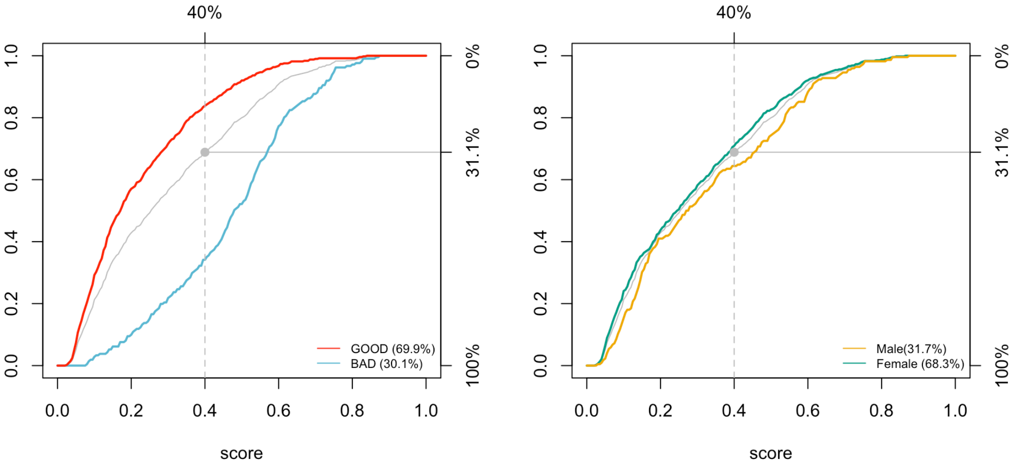
Numerical examples





Conditional distributions of scores on `GermanCredit`, logistic regression.

Numerical examples



Conditional distributions of scores on `GermanCredit`, boosting model.

Numerical examples

	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

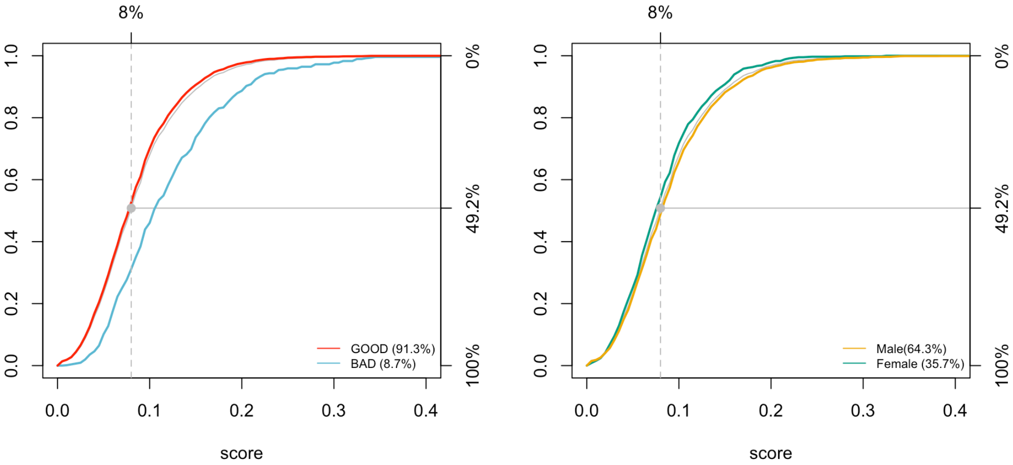
Fairness metrics on `GermanCredit`, with threshold at 20%.

Numerical examples

	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

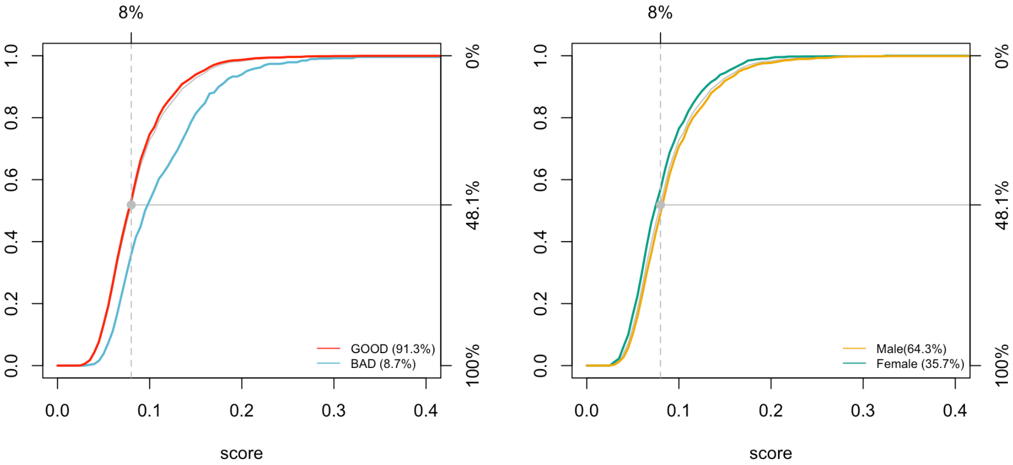
Fairness metrics on `GermanCredit`, with threshold at 40%.

Numerical examples



Conditional distributions of scores on `FrenchMotor`, from the logistic regression.

Numerical examples



Conditional distributions of scores on `FrenchMotor`, from a boosting classification.

Several definitions of "fairness" or "non-discriminatory"

demographic parity $\rightarrow$ $\mathbb{E}[\underbrace{\widehat{Y}}_{\text{score } \widehat{y}} \mid \overbrace{S = A}^{\text{sensitive}}] \overset{?}{=} \mathbb{E}[\underbrace{\widehat{Y}}_{\text{score } \widehat{y}} \mid \overbrace{S = B}^{\text{sensitive}}]$

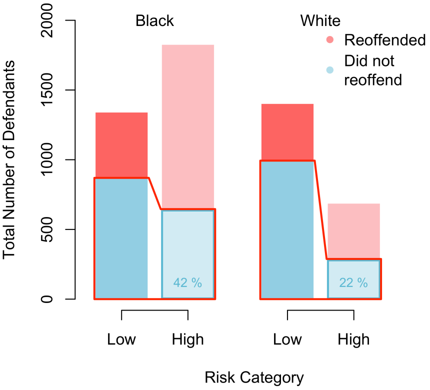
equalized odds $\rightarrow$ $\mathbb{E}[\underbrace{\widehat{Y}}_{\text{score } \widehat{y}} \mid \overbrace{Y = y}^{\text{outcome } y}, S = A] \overset{?}{=} \mathbb{E}[\underbrace{\widehat{Y}}_{\text{score } \widehat{y}} \mid \overbrace{Y = y}^{\text{outcome } y}, S = B], \forall y$

calibration $\rightarrow$ $\mathbb{E}[\overbrace{Y}^{\text{outcome } y} \mid \underbrace{\widehat{Y} = u}_{\text{score } \widehat{y}}, S = A] \overset{?}{=} \mathbb{E}[\overbrace{Y}^{\text{outcome } y} \mid \underbrace{\widehat{Y} = u}_{\text{score } \widehat{y}}, S = B], \forall u$

Isn't it a problem to have several definitions?

From Feller et al. (2016),

- for White people, among those who did not re-offend (y), 22% were wrongly classified ($\widehat{y}$),
- for Black people, among those who did not re-offend, 42% were wrongly classified,
- **Problem**, since $42\% \gg 22\%$

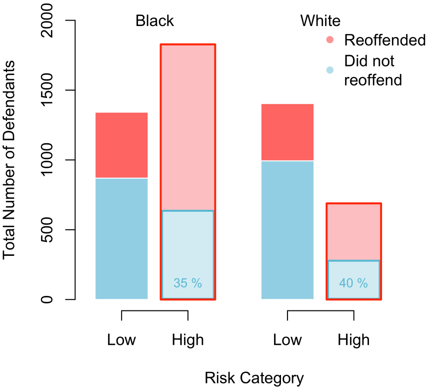


$$\mathbb{P}[\widehat{Y}=\text{high} \mid Y=\text{no}, S=\text{black}] = 42\% \overset{?}{=} \mathbb{P}[\widehat{Y}=\text{high} \mid Y=\text{no}, S=\text{white}] = 22\%,$$

Isn't it a problem to have several definitions?

From Dieterich et al. (2016),

- for White people, among those who were classified as high risk ($\widehat{y}$), 40% did not re-offend (y),
- for Black people, among those who were classified as high risk ($\widehat{y}$), 35% did not re-offend (y),
- **No problem**, since $35 \approx 40\%$



$$\mathbb{P}[Y = \text{no} \mid \widehat{Y} = \text{high}, S = \text{black}] = 35\% \overset{?}{=} \mathbb{P}[Y = \text{no} \mid \widehat{Y} = \text{high}, S = \text{white}] = 40\%.$$

Is it always possible to have a sensitive-free model (with respect to ...)?

For **decisions** ($\widehat{y} \in \{0,1\}$, e.g., "obtain a loan"),

$$\text{demographic parity} \rightarrow \mathbb{P}[\underbrace{\widehat{Y}=1}_{\text{decision } \widehat{y}} \mid S=\text{A}] \stackrel{?}{=} \mathbb{P}[\underbrace{\widehat{Y}=1}_{\text{decision } \widehat{y}} \mid S=\text{B}]$$

those decisions are usually based on **scores**, and **thresholds**

$$\text{demographic parity} \rightarrow \mathbb{E}[\underbrace{\widehat{m}(\boldsymbol{X},S)}_{\text{score } \widehat{m}} > t \mid S=\text{A}] \stackrel{?}{=} \mathbb{E}[\underbrace{\widehat{m}(\boldsymbol{X},S)}_{\text{score } \widehat{m}} > t \mid S=\text{B}]$$

One can achieve demographic parity, simply selecting **different thresholds**

$$\text{demographic parity} \rightarrow \mathbb{E}[\widehat{m}(\boldsymbol{X},S) > t_{\text{A}} \mid S=\text{A}] \stackrel{?}{=} \mathbb{E}[\widehat{m}(\boldsymbol{X},S) > t_{\text{B}} \mid S=\text{B}]$$

(with that strategy, usually impossible to achieve equalized odds)

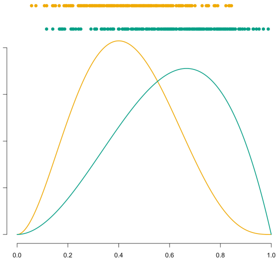
Is it always possible to have a sensitive-free model (with respect to ...)?

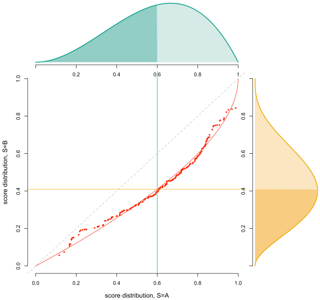
For **decisions** ($\widehat{y} \in \{0,1\}$, e.g., "obtain a loan"), we considered

$$\text{demographic parity} \rightarrow \mathbb{E}[\widehat{Y} \mid S = \mathsf{A}] \stackrel{?}{=} \mathbb{E}[\widehat{Y} \mid S = \mathsf{B}]$$

and we can consider the analogous for **scores** (possibly used to assess premiums),

$$\text{demographic parity} \rightarrow \mathbb{E}[\underbrace{\widehat{m}(\boldsymbol{X}, S)}_{\text{score } \widehat{y}} \mid S = \mathsf{A}] \stackrel{?}{=} \mathbb{E}[\underbrace{\widehat{m}(\boldsymbol{X}, S)}_{\text{score } \widehat{y}} \mid S = \mathsf{B}]$$





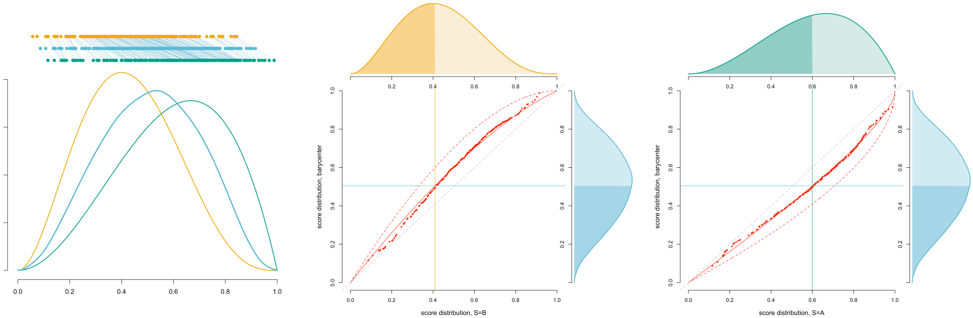
- individual in group A with a score $\widehat{y}(\mathsf{A}) = 60\%$ corresponding to quantile α (here 0.5)
- in group B, the same quantile α corresponds to $\widehat{y}(\mathsf{B}) = 40\%$

Is it always possible to have a sensitive-free model (with respect to ...)?

- To get a fair model (**neutral with respect to s**), consider an average between the two models,

score in group A with quantile α score in group B with quantile α

$$\widehat{y}^\star = \mathbb{P}[S = \text{A}] \cdot \widehat{y}(\text{A}) + \mathbb{P}[S = \text{B}] \cdot \widehat{y}(\text{B})$$



"In order to treat some persons equally, we must treat them differently"

- 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**," Knowlton (1978), cited in Lippert-Rasmussen (2020)

- In 2007, John G. Roberts of the U.S. Supreme Court submits

  "The way to stop discrimination on the basis of race is to **stop discriminating on the basis of race**," Sabbagh (2007) and Turner (2015)

See philosophical discussions about **affirmative action**, e.g., Rubenfeld (1997); Pojman (1998); Anderson (2004)

“Neutral with respect to some sensitive attribute?”

What does “**neutral with respect to** s” really means ?

We have seen that accuracy was assessed with respect to data in the portfolio,

$$\boxed{\overline{y}} = \underset{\gamma\in\mathbb{R}}{\text{argmin}}\left\{\sum_{i=1}^{n}(y_i-\gamma)^2\right\} \text{ or } \boxed{\mathbb{E}[Y]} = \underset{\gamma\in\mathbb{R}}{\text{argmin}}\left\{\sum_{y}(y-\gamma)^2\mathbb{P}[Y=y]\right\}$$

based on observations from the insurer’s portfolio. Technically, should we consider

- expected values / probabilities / independence properties based on $\mathbb{P}$ (portfolio)
- expected values / probabilities / independence properties based on $\mathbb{Q}$ (market)

(ongoing work *Why portfolio-specific fairness should fail to extend market-wide: Selection bias in insurance* with M.P. Côté & O. Côté)

Should we ask for neutrality “in the portfolio” or for some “targeted population” ?

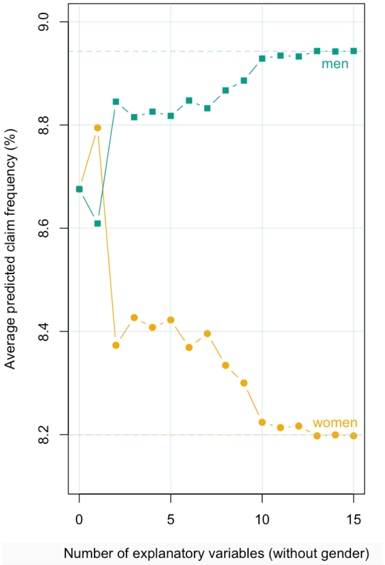
Discrimination in the data, or in the model?

On a French motor dataset, average claim frequencies are 8.94% (men) and 8.20% (women).

Consider some logistic regression to estimate annual claim frequency, on k explanatory variables excluding gender.

	men	women
$k = 0$	8.68%	8.68%
$k = 2$	8.85%	8.37%
$k = 8$	8.87%	8.33%
$k = 15$	8.94%	8.20%
empirical	8.94%	8.20%

Models simply tend to reproduce what was observed in the data (see "**is-ought**" problem, in Hume (1739)).



Discrimination in the data, or in the model?

David Hume's "**is-ought**" problem, in Hume (1739)

what **is** observed, what is **statistically normal**

$$\pi(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{X} = \boldsymbol{x}] \text{ where } \mathbb{P} \text{ is the historical probability}$$

$\neq$ what **should be**, what we expect from an **ethical norm**

$$\pi(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}^\star}[Y|\boldsymbol{X} = \boldsymbol{x}] \text{ where } \mathbb{P}^\star \text{ is some "fair" probability}$$

"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," Chollet (2021)

Classical **clausula rebus sic stantibus** ("with things thus standing") in predictive modeling (statistics and machine learning)

Discrimination in the data, or in the model?

- change the training data to de-bias (through weights) : **pre-processing**

if we can draw i.i.d. copies of a random variable X_i's, under probability $\mathbb{P}$, then

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i) \to \mathbb{E}_{\mathbb{P}}[h(X)], \text{ as } n \to \infty \text{ “\textbf{law of large numbers}”}$$

but if we want to reach $\mathbb{E}_{\mathbb{Q}}[h(X)]$, consider

$$\frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{\mathrm{d}\mathbb{Q}(x_i)}{\mathrm{d}\mathbb{P}(x_i)}}_{\text{weight } \omega_i} h(x_i) \to \mathbb{E}_{\mathbb{Q}}[h(X)], \text{ as } n \to \infty.$$

- keep the biases data, but distort the outcome : **post-processing**
- add a fairness constraint (penalty) in the optimization problem : **in-processing**
  as classical adversarial techniques, Grari et al. (2021)

Discrimination, with different perspectives

- Regulatory perspective, "**group fairness**" (discussed previously)
- Policyholders perspective, "**individual fairness**"

A decision satisfies individual fairness if "had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same."

- also named "**counterfactual fairness**" in Kusner et al. (2017), and should be related to classical causal inference problem, (conditional) average treatment effect (the "treatement" being the sensitive attribute),

  "other things being equal" ?**ceteris paribus** ? See "revolving variable" in Kilbertus et al. (2017). Consider a men ($s = \text{A}$) with height $x = 6'3$ (or 190 cm). If that person had been a women ($s = \text{B}$) would she have height $x = 6'3$?

  (hint: no, consider similar quantiles, as discussed previously, see Charpentier et al. (2023a))

What if we neither observe nor collect sensitive personal information (s) ?

September 27, 2023, the Colorado Division of Insurance exposed a new proposed regulation entitled Concerning Quantitative Testing of External Consumer Data and Information Sources, Algorithms, and Predictive Models Used for Life Insurance Underwriting for Unfairly Discriminatory Outcomes. Use of **BIFSG** (Bayesian Improved First Name Surname and Geocoding), from Elliott et al. (2009).

Consider 12 people living near Atlanta, GA (Fulton & Gwinnett counties),

	last	first	county	city	zipcode	whi	bla	his	asi
2	RADLEY	OLIVIA	Fulton	Fairburn	30213	14	83	1	0
3	BOORSE	KEISHA	Fulton	Atlanta	30331	97	0	3	0
4	MAZ	SAVANNAH	Gwinnett	Norcross	30093	5	6	76	13
5	GAULE	NATASHIA	Gwinnett	Snellville	30078	67	19	14	0
6	MCMELLEN	ISMAEL	Gwinnett	Lilburn	30047	73	15	6	3
7	WASHINGTON	BRYN	Gwinnett	Norcross	30093	0	95	3	0

(ongoing *Predicting Unobserved Multi-Class sensitive Attributes : Enhancing Calibration with Nested Dichotomies for Fairness* with A.M. Patrón Piñerez, A. Fernandes Machado, & E. Gallic)

Can we use aggregate data related to sensitive information ($\bar{s}$) ?

from Bickel et al. (1975), discussed as an illustration of "**Simpson's paradox**"

Can we use aggregate data related to sensitive information ($\bar{s}$) ?

	Total	Men	Women	Proportions
Total	$5233/12763 \sim 41\%$	$3714/8442 \sim$ **44%**	$1512/4321 \sim 35\%$	66%-34%
Top 6	$1745/4526 \sim 39\%$	$1198/2691 \sim$ **45%**	$557/1835 \sim 30\%$	59%-41%
A	$597/933 \sim 64\%$	$512/825 \sim 62\%$	$89/108 \sim$ **82%**	88%-12%
B	$369/585 \sim 63\%$	$353/560 \sim 63\%$	$17/\ 25 \sim$ **68%**	96%- 4%
C	$321/918 \sim 35\%$	$120/325 \sim$ **37%**	$202/593 \sim 34\%$	35%-65%
D	$269/792 \sim 34\%$	$138/417 \sim 33\%$	$131/375 \sim$ **35%**	53%-47%
E	$146/584 \sim 25\%$	$53/191 \sim$ **28%**	$94/393 \sim 24\%$	33%-67%
F	$43/714 \sim 6\%$	$22/373 \sim 6\%$	$24/341 \sim$ **7%**	52%-48%

Data from Bickel et al. (1975). Formalized as follows: S is the (binary) genre, $\widehat{Y}$ the admission decision, and X the program (category),

Can we use aggregate data related to sensitive information ($\bar{s}$) ?

$$\mathbb{P}[\underbrace{\widehat{Y} = \text{yes}}_{\text{overall admission}} \mid \overbrace{S = \text{men}}^{\text{sensitive}}] \geq \mathbb{P}[\widehat{Y} = \text{yes} \mid \overbrace{S = \text{women}}^{\text{sensitive}}]$$

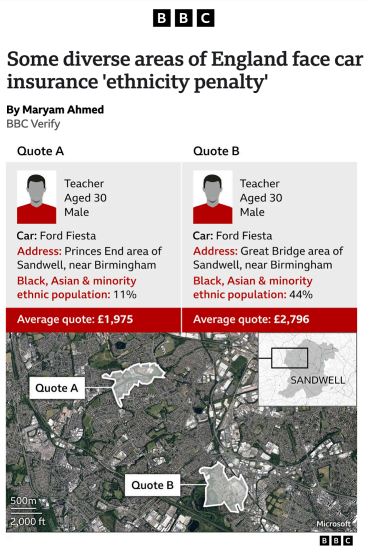
overall admission

$$\mathbb{P}[\widehat{Y} = \text{yes} \mid X = x, S = \text{men}] \leq \mathbb{P}[\widehat{Y} = \text{yes} \mid X = x, S = \text{women}], \forall x.$$

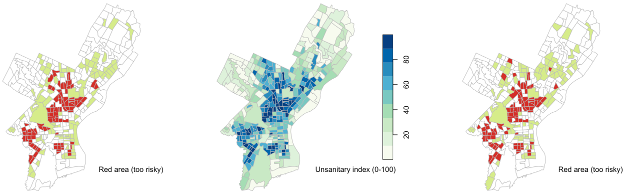
conditional on program

"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,"
Bickel et al. (1975)

Disentangling correlations



See some diverse areas of England face car insurance 'ethnicity penalty' (remove from the BBC website since)



y, x and s can easily be correlated variables

spurious correlations problem ?

Need to use causal models to avoid indirect discrimination

Multiple sensitive attributes, “robbing Peter to pay Paul”?

sensitive attribute 1

$$\mathbb{E}[\,\widehat{m}(\boldsymbol{X},S_1,S_2)\,|\,S_1=\text{A}\,] \neq \mathbb{E}[\,\widehat{m}(\boldsymbol{X},S_1,S_2)\,|\,S_1=\text{B}\,]$$

$$\mathbb{E}[\,\widehat{m}(\boldsymbol{X},S_1,S_2)\,|\,S_2=\text{C}\,] \approx \mathbb{E}[\,\widehat{m}(\boldsymbol{X},S_1,S_2)\,|\,S_2=\text{D}\,]$$

sensitive attribute 2

Distort model $\widehat{m}$ to achieve fairness with respect to S_1 $\longrightarrow$ model $\widetilde{m}$

sensitive attribute 1

$$\mathbb{E}[\,\widetilde{m}(\boldsymbol{X},S_1,S_2)\,|\,S_1=\text{A}\,] = \mathbb{E}[\,\widetilde{m}(\boldsymbol{X},S_1,S_2)\,|\,S_1=\text{B}\,]$$

$$\mathbb{E}[\,\widetilde{m}(\boldsymbol{X},S_1,S_2)\,|\,S_2=\text{C}\,] \neq \mathbb{E}[\,\widetilde{m}(\boldsymbol{X},S_1,S_2)\,|\,S_2=\text{D}\,]$$

sensitive attribute 2

Mitigation, In-Processing

In a linear regression problem, $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$. Zafar et al. (2017) suggested

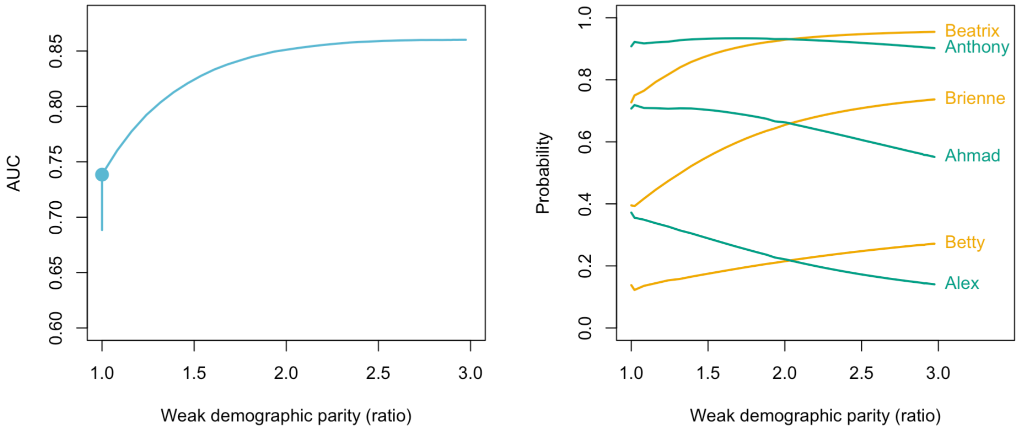
$$\beta^\star = \min_\beta \left\{ \mathbb{E}\left[\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2\right] \right\} \text{ s.t. } \left|\text{Cov}[\boldsymbol{X}\beta, S]\right| \leq c \;(\in \mathbb{R}_+).$$

	$\widehat{m}(\boldsymbol{x}, s)$, aware ← less fair				more fair →	$\widehat{m}(\boldsymbol{x})$, unaware ← less fair			more fair →
$\widehat{\beta}_0$ (Intercept)	-2.55	-2.29	-1.97	-1.51	-1.03	-2.14	-1.98	-1.78	-1.63
$\widehat{\beta}_1$ (x_1)	0.88	0.88	0.85	0.77	0.62	1.01	0.84	0.57	0.26
$\widehat{\beta}_2$ (x_2)	0.37	0.37	0.35	0.32	0.25	0.37	0.35	0.31	0.24
$\widehat{\beta}_3$ (x_3)	0.02	0.02	0.02	0.02	0.03	0.15	0.02	-0.15	-0.29
$\widehat{\beta}_{\text{B}}$ ($\mathbf{1}_{\text{B}}$)	0.82	0.44	-0.03	-0.70	-1.31	-	-	-	-

Mitigation, In-Processing

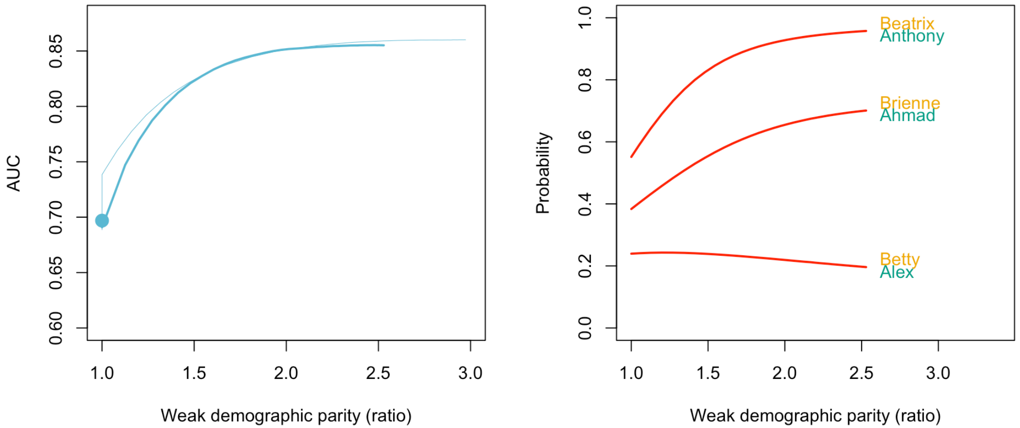
	$\widehat{m}(\boldsymbol{x}, s)$, aware					$\widehat{m}(\boldsymbol{x})$, unaware			
	$\leftarrow$ less fair			more fair $\rightarrow$		$\leftarrow$ less fair		more fair $\rightarrow$	
Betty	0.27	0.25	0.22	0.17	0.14	0.20	0.22	0.24	0.24
Brienne	0.74	0.71	0.66	0.54	0.40	0.70	0.66	0.55	0.38
Beatrix	0.95	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
$\mathbb{E}[\widehat{m}(\boldsymbol{x}_i, s_i)\|S = \texttt{A}]$	0.23	0.26	0.31	0.36	0.42	0.25	0.30	0.37	0.41
$\mathbb{E}[\widehat{m}(\boldsymbol{x}_i, s_i)\|S = \texttt{B}]$	0.67	0.65	0.61	0.53	0.42	0.64	0.61	0.54	0.41
(ratio)	$\times$2.97	$\times$2.49	$\times$2.01	$\times$1.46	$\times$1.00	$\times$2.53	$\times$2.02	$\times$1.48	$\times$1.00
AUC	0.86	0.86	0.85	0.82	0.74	0.86	0.85	0.82	0.70

Mitigation, In-Processing



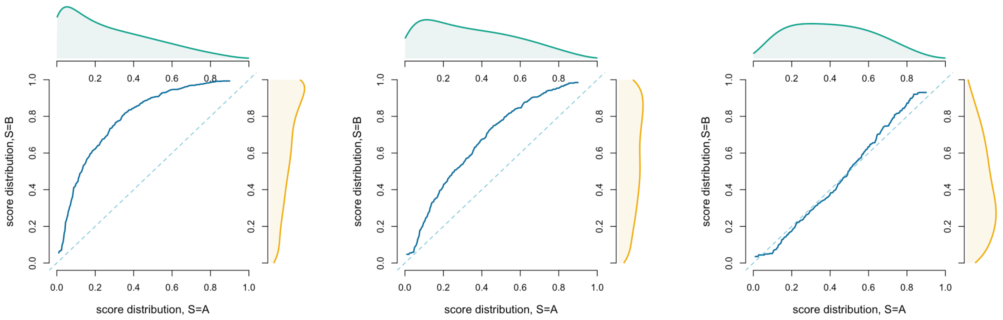
AUC of $\widehat{m}_{\widehat{\beta}_\lambda}$ and evolution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i, s_i)$ (with a logistic regression)

Mitigation, In-Processing



AUC of $\widehat{m}_{\widehat{\beta}_\lambda}$ and evolution of $\widehat{m}_{\widehat{\beta}_\lambda}(\boldsymbol{x}_i)$ (with a logistic regression)

Mitigation, In-Processing



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-hand side and high value on the right-hand side), associated with a demographic parity penalty criteria.

Mitigation, Post-Processing

Definition 3.48: Wasserstein W_2 Barycenter, Agueh and Carlier (2011)

$$\overline{\mathbb{Q}} = \underset{\mathbb{Q}}{\text{argmin}} \left\{ \sum_{i=1}^{k} \omega_i W_2(\mathbb{Q}, \mathbb{P}_i)^2 \right\},$$

For univariate distributions, the optimal transport $\mathcal{T}^\star$ is the monotone transformation.

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

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),

Mitigation, Post-Processing

Proposition 3.13: Wasserstein W_2 Barycenter,

$$\overline{\mathbb{Q}} = \left(\sum_{i=1}^{k} \omega_i \mathcal{T}^{1\to i}\right)_{\#} \mathbb{P}_1.$$

Computation of barycenters can be computationnaly difficult, Altschuler and Boix-Adsera (2021)

For univariate distributions, there is a simple expression, $\mathcal{T}^{1\to i}$ 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

Proposition 3.14: Wasserstein W_2 Barycenter, univariate distributions

$\mathcal{T}^{1\to i}$ 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

$$\overline{\mathbb{Q}} = \left(\sum_{i=1}^{n} k\omega_i \mathcal{T}^{1\to i}\right)_{\#} \mathbb{P}_1.$$

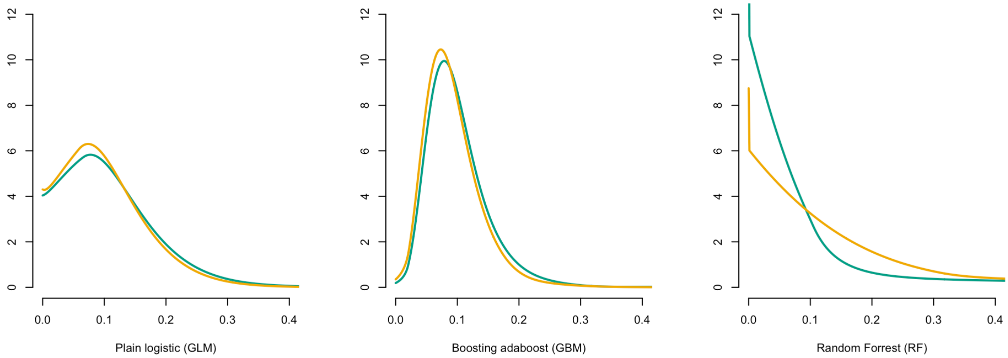
Proposition 3.15: Wasserstein W_2 Barycenter, univariate distributions

Given two scores $m(\boldsymbol{x}, s = \mathtt{A})$ and $m(\boldsymbol{x}, s = \mathtt{B})$, the “fair barycenter score” is

$$\begin{cases} m^\star(\boldsymbol{x}, s = \mathtt{A}) = \mathbb{P}[S = \mathtt{A}] \cdot m(\boldsymbol{x}, s = \mathtt{A}) + \mathbb{P}[S = \mathtt{B}] \cdot F_{\mathtt{B}}^{-1} \circ F_{\mathtt{A}}\big(m(\boldsymbol{x}, s = \mathtt{A})\big) \\ m^\star(\boldsymbol{x}, s = \mathtt{B}) = \mathbb{P}[S = \mathtt{A}] \cdot F_{\mathtt{A}}^{-1} \circ F_{\mathtt{B}}\big(m(\boldsymbol{x}, s = \mathtt{B})\big) + \mathbb{P}[S = \mathtt{B}] \cdot m(\boldsymbol{x}, s = \mathtt{B}). \end{cases}$$

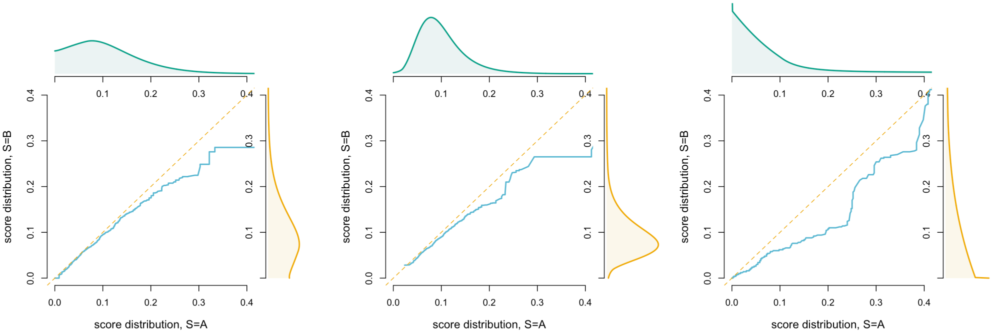
Application to FrenchMotor

If the two models are balanced, $m^\star$ is also balanced.
Annual claim occurrence (motor insurance, Charpentier et al. (2023b))
Three models (plain GLM, GBM, Random Forest)



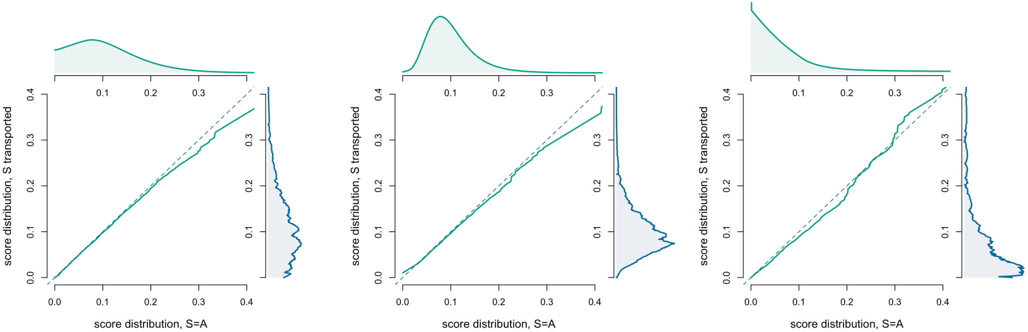
Application to FrenchMotor

Predictions are different for men $(= \texttt{A})$ and women $(S = \texttt{B})$



since $W_2 \neq 0$ consider post processing mitigation

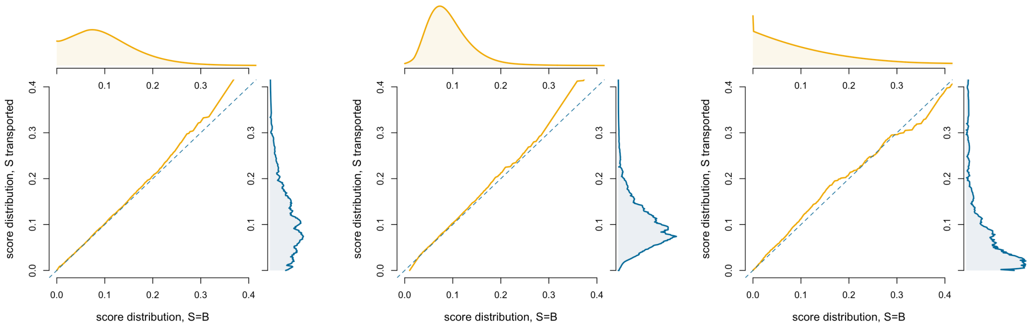
Application to FrenchMotor



Given scores $m(\boldsymbol{x}, s = \mathtt{A})$ and $m(\boldsymbol{x}, s = \mathtt{B})$, the “fair barycenter score” is

$$m^\star(\boldsymbol{x}, s = \mathtt{A}) = \mathbb{P}[S = \mathtt{A}] \cdot m(\boldsymbol{x}, s = \mathtt{A}) + \mathbb{P}[S = \mathtt{B}] \cdot F_{\mathtt{B}}^{-1} \circ F_{\mathtt{A}}\big(m(\boldsymbol{x}, s = \mathtt{A})\big)$$

Application to FrenchMotor

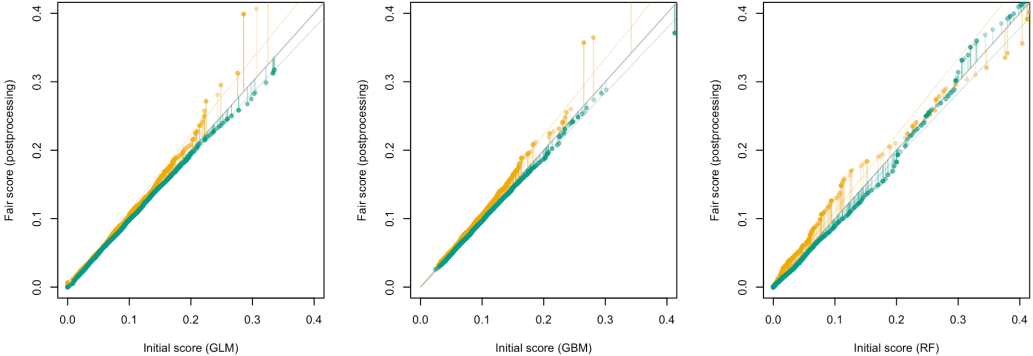


Given scores $m(\boldsymbol{x}, s = \mathtt{A})$ and $m(\boldsymbol{x}, s = \mathtt{B})$, the "fair barycenter score" is

$$m^{\star}(\boldsymbol{x}, s = \mathtt{B}) = \mathbb{P}[S = \mathtt{A}] \cdot F_{\mathtt{A}}^{-1} \circ F_{\mathtt{B}}\big(m(\boldsymbol{x}, s = \mathtt{B})\big) + \mathbb{P}[S = \mathtt{B}] \cdot m(\boldsymbol{x}, s = \mathtt{B})$$

Application to `FrenchMotor`

We can plot $\{(m(\boldsymbol{x}_i, \mathtt{A}), m^\star(\boldsymbol{x}_i, \mathtt{A})\}$ and $\{(m(\boldsymbol{x}_i, \mathtt{B}), m^\star(\boldsymbol{x}_i, \mathtt{B})\}$



Application to FrenchMotor

Numerical values, for initial occurence probability of 5%, 10% and 20%, we have

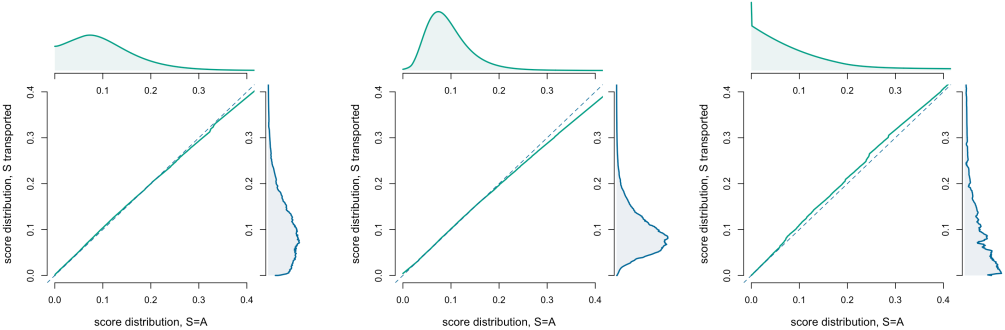
	A (men)				B (women)			
	$\times 0.94$	GLM	GBM	RF	$\times 1.11$	GLM	GBM	RF
$m(\boldsymbol{x}) = 5\%$	4.73%	4.94%	4.80%	4.42%	5.56%	5.16%	5.25%	6.15%
$m(\boldsymbol{x}) = 10\%$	9.46%	9.83%	9.66%	8.92%	11.12%	10.38%	10.49%	12.80%
$m(\boldsymbol{x}) = 20\%$	18.91%	19.50%	18.68%	18.26%	22.25%	20.77%	21.63%	21.12%

Application to FrenchMotor

We can do the same for discrimination against "old" drivers.

	A (younger < 65)				B (old > 65)			
	$\times 1.01$	GLM	GBM	RF	$\times 0.94$	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%

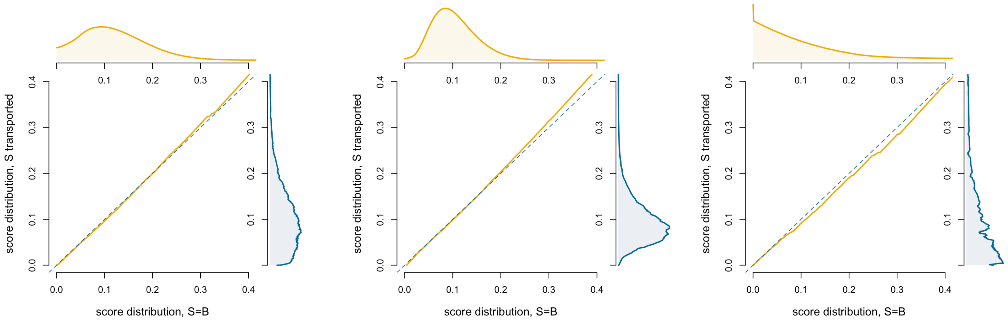
Application to FrenchMotor



Given scores $m(\boldsymbol{x}, s = \mathtt{A})$ and $m(\boldsymbol{x}, s = \mathtt{B})$, the "fair barycenter score" is

$$m^{\star}(\boldsymbol{x}, s = \mathtt{A}) = \mathbb{P}[S = \mathtt{A}] \cdot m(\boldsymbol{x}, s = \mathtt{A}) + \mathbb{P}[S = \mathtt{B}] \cdot F_{\mathtt{B}}^{-1} \circ F_{\mathtt{A}}\big(m(\boldsymbol{x}, s = \mathtt{A})\big)$$

Application to FrenchMotor

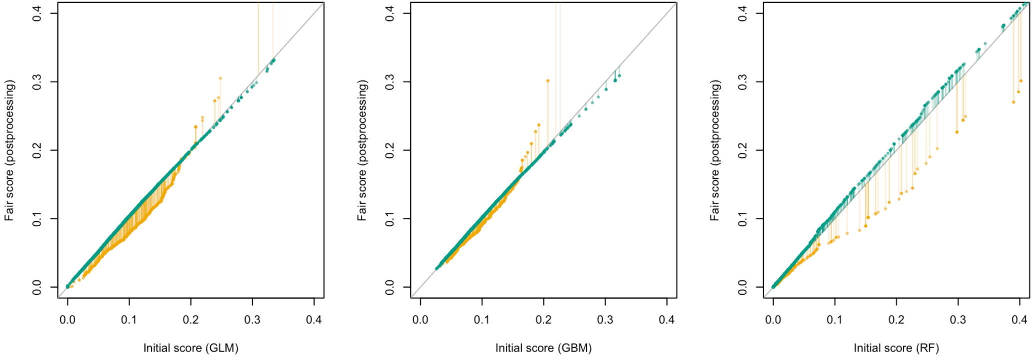


Given scores $m(\boldsymbol{x}, s = \mathtt{A})$ and $m(\boldsymbol{x}, s = \mathtt{B})$, the "fair barycenter score" is

$$m^{\star}(\boldsymbol{x}, s = \mathtt{B}) = \mathbb{P}[S = \mathtt{A}] \cdot F_{\mathtt{A}}^{-1} \circ F_{\mathtt{B}}\big(m(\boldsymbol{x}, s = \mathtt{B})\big) + \mathbb{P}[S = \mathtt{B}] \cdot m(\boldsymbol{x}, s = \mathtt{B})$$

Application to FrenchMotor

We can plot $\{(m(\boldsymbol{x}_i, \texttt{A}), m^\star(\boldsymbol{x}_i, \texttt{A})\}$ and $\{(m(\boldsymbol{x}_i, \texttt{B}), m^\star(\boldsymbol{x}_i, \texttt{B})\}$



References

Aas, K., Jullum, M., and Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502.

Agueh, M. and Carlier, G. (2011). Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924.

Almond, D. and Doyle Jr, J. J. (2011). After midnight: A regression discontinuity design in length of postpartum hospital stays. *American Economic Journal: Economic Policy*, 3(3):1–34.

Altschuler, J. M. and Boix-Adsera, E. (2021). Wasserstein barycenters can be computed in polynomial time in fixed dimension. *The Journal of Machine Learning Research*, 22(1):2000–2018.

Anderson, T. H. (2004). *The pursuit of fairness: A history of affirmative action*. Oxford University Press.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angrist, J. D. and Pischke, J.-S. (2014). *Mastering'metrics: The path from cause to effect*. Princeton university press.

Apfelbaum, E. P., Pauker, K., Sommers, S. R., and Ambady, N. (2010). In blind pursuit of racial equality? *Psychological science*, 21(11):1587–1592.

References

Apley, D. W. and 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.

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *The American Economic Review*, 53(5):941–973.

Austin, P. C. and Steyerberg, E. W. (2019). The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine*, 38:4051–4065.

Avraham, R. (2017). Discrimination and insurance. In Lippert-Rasmussen, K., editor, *Handbook of the Ethics of Discrimination*, pages 335–347. Routledge.

Azen, R. and Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological methods*, 8(2):129.

Baldus, D. C. and Cole, J. W. (1980). *Statistical proof of discrimination*. Mcgraw-Hill.

Barber, R. F. (2024). *An introduction to conformal prediction and distribution-free inference*. Columbia University.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507.

Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1:2017.

References

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago press.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404.

Biddle, D. (2017). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Routledge.

Biecek, P. and Burzykowski, T. (2021). *Explanatory model analysis: explore, explain, and examine predictive models*. CRC Press.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and 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*, pages 491–500.

Bozinovski, S. and Fulgosi, A. (1976). The influence of pattern similarity and transfer learning upon training of a base perceptron b2. In *Proceedings of Symposium Informatica*, volume 3, pages 121–126.

Brain, J. (2010). "past performance is not necessarily indicative of future results"—the proven-in-use argument and the retrospective application of modern standards. In *5th IET International Conference on System Safety 2010*, pages 1–4. IET.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

References

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Brier, G. W. et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Britz, G. (2008). *Einzelfallgerechtigkeit versus Generalisierung: verfassungsrechtliche Grenzen statistischer Diskriminierung*. Mohr Siebeck.

Calders, T. and Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21(2):277–292.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734.

Card, D. and Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food Industry in new jersey and pennsylvania. *The American Economic Review*, 84(4):772–793.

Casey, B., Pezier, J., and 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.

Charpentier, A. (2021). Le mythe de l'interprétabilité et de l'explicabilité des modèles. *Risques*, 128:109–115.

References

Charpentier, A., Flachaire, E., and Gallic, E. (2023a). Causal inference with optimal transport. In Thach, N. N., Kreinovich, V., Ha, D. T., and Trung, N. D., editors, *Optimal Transport Statistics for Economics and Related Topics*. Springer Verlag.

Charpentier, A., Flachaire, E., and Ly, A. (2018). Econometrics and machine learning. *Economie et Statistique*, 505(1):147–169.

Charpentier, A., Hu, F., and Ratz, P. (2023b). Mitigating discrimination in insurance with wasserstein barycenters. bias. In *3rd Workshop on Bias and Fairness in AI, International Workshop of ECML PKDD*.

Chicco, D. and 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.

Chzhen, E. and Schreuder, N. (2022). A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, 5(2):115–124.

References

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *arXiv*, 1701.08230.

Cramer, M. (2019). Another benefit to going to museums? you may live longer. *New York Times*, December 22.

Cunningham, S. (2021). *Causal inference*. Yale University Press.

Da Veiga, S. (2024). *Tutorial on conformal prediction and related methods*. ETICS 2024 Research School.

D'Agostino, R. B., Grundy, S., Sullivan, L. M., Wilson, P., Group, C. R. P., et al. (2001). Validation of the framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *Journal of the American Medical Association*, 286(2):180–187.

Darlington, R. B. (1971). Another look at "cultural fairness" 1. *Journal of educational measurement*, 8(2):71–82.

Davidson, R., MacKinnon, J. G., et al. (2004). *Econometric theory and methods*, volume 5. Oxford University Press New York.

Dawid, A. P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

References

De Baere, G. and Goessens, E. (2011). Gender differentiation in insurance contracts after the judgment in case c-236/09, Association Belge des Consommateurs Test-Achats asbl v. conseil des ministres. *Colum. J. Eur. L.*, 18:339.

Denis, C., Elie, R., Hebiri, M., and Hu, F. (2021). Fairness guarantee in multi-class classification. *arXiv*, 2109.13642.

Denuit, M., Charpentier, A., and Trufin, J. (2021). Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics & Economics*, 101:485–497.

Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(7.4):1.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, volume 1104.3913, pages 214–226.

Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., and Lurie, N. (2009). Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9(2):69–83.

References

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, volume 1412.3756, pages 259–268.

Feller, A., Pierson, E., Corbett-Davies, S., and 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.

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024a). From uncertainty to precision: Enhancing binary classifier performance through calibration. *arXiv preprint arXiv:2402.07790*, 2402.07790.

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024b). Probabilistic scores of classifiers, calibration is not enough. *arXiv preprint arXiv:2408.03421*.

Fisher, A., Rudin, C., and 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.

References

Fong, C., Hazlett, C., and 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.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569.

Frazier, R. (2021). California's ban on climate-informed models for wildlife insurance premiums. *Ecology L. Currents*, 48:24.

Freedman, D. A. and Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation review*, 32(4):392–409.

Frezal, S. and Barry, L. (2020). Fairness in uncertainty: Some limits and misinterpretations of actuarial fairness. *Journal of Business Ethics*, 167:127–136.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Froot, K. A., Kim, M., and Rogoff, K. S. (1995). The law of one price over 700 years. *National Bureau of Economic Research Cambridge*, 5132.

Fuller, W. E. (1914). Flood flows. *Transactions of the American Society of Civil Engineers*, 77(1):564–617.

References

Furht, B., Villanustre, F., Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). Transfer learning techniques. In *Big data technologies and applications*, pages 53–99. Springer.

Goldstein, A., Kapelner, A., Bleich, J., and 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.

Grari, V., Lamprier, S., and Detyniecki, M. (2021). Fairness without the sensitive attribute via causal variational autoencoder.

Greenwell, B. M. (2017). `pdp`: an r package for constructing partial dependence plots. *R Journal*, 9(1):421.

Gumbel, E. J. (1941a). Probability-interpretation of the observed return-periods of floods. *Eos, Transactions American Geophysical Union*, 22(3):836–850.

Gumbel, E. J. (1941b). The return period of flood flows. *The annals of mathematical statistics*, 12(2):163–190.

Gumbel, E. J. (1958). *Statistics of extremes*. Columbia university press.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

References

Gupta, K., Rahimi, A., Ajanthan, T., Sminchisescu, C., Mensink, T., and Hartley, R. I. (2021). Calibration of neural networks using splines. In *International Conference on Learning Representations (ICLR)*.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Hazen, A. (1930). *Flood flows: a study of frequencies and magnitudes*. Wiley.

Helton, J. C. and Davis, F. (2002). Illustration of sampling-based methods for uncertainty and sensitivity analysis. *Risk analysis*, 22(3):591–622.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Howe, K. B., Suharlim, C., Ueda, P., Howe, D., Kawachi, I., and Rimm, E. B. (2016). Gotta catch'em all! pokémon go and physical activity among young adults: difference in differences study. *British Medical Journal*, 355.

Hume, D. (1739). *A Treatise of Human Nature*. Cambridge University Press Archive.

Ichiishi, T. (2014). *Game theory for economic analysis*. Academic Press.

References

Imai, K. (2022). *Quantitative Social Science*. Princeton University Press.

Imai, K. and Khanna, K. (2016). Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, 24(2):263–272.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise de Sciences Naturelles*, 37:547–579.

Karimi, H., Khan, M. F. A., Liu, H., Derr, T., and Liu, H. (2022). Enhancing individual fairness through propensity score matching. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Kearns, M. and Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta psychologica*, 77(3):217–273.

Kerner, O. (1968). *Report of The National Advisory Commission on Civil Disorder*. Bantam Books.

References

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

Kim, P. T. (2017). Auditing algorithms for discrimination. *University of Pennsylvania Law Review*, 166:189.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv*, 1609.05807.

Knowlton, R. E. (1978). Regents of the university of california v. bakke. *Arkansas Law Review*, 32:499.

Kranzberg, M. (1986). Technology and history:" kranzberg's laws". *Technology and culture*, 27(3):544–560.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*, volume 26. Springer.

References

Kull, M., Filho, T. M. S., and Flach, P. (2017). Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, 11(2):5052 – 5080.

Kumar, A., Liang, P. S., and Ma, T. (2019). Verified uncertainty calibration. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076.

Li, F. and Li, F. (2019). Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 13:2389–2415.

Linn, R. L. and Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement*, 8(1):1–4.

Lipovetsky, S. and Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330.

Lippert-Rasmussen, K. (2020). *Making sense of affirmative action*. Oxford University Press.

References

Liu, J., Hong, Y., D'Agostino Sr, R. B., Wu, Z., Wang, W., Sun, J., Wilson, P. W., Kannel, W. B., and Zhao, D. (2004). Predictive value for the chinese population of the framingham chd risk assessment tool compared with the chinese multi-provincial cohort study. *Journal of the American Medical Association*, 291(21):2591–2599.

Loader, C. (2006). *Local regression and likelihood*. Springer.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Proceedings of the 31st international conference on neural information processing systems*, volume 30, pages 4768–4777. Curran Associates, Inc.

Marshall, A. (1890). General relations of demand, supply, and value. *Principles of economics: unabridged eighth edition*.

Meldrum, M. (1998). "a calculated risk": the salk polio vaccine field trials of 1954. *British Medical Journal*, 317(7167):1233–1236.

Merriam-Webster (2022). *Dictionary*. .

Meyers, G. and Van Hoyweghen, I. (2018). Enacting actuarial fairness in insurance: From fair discrimination to behaviour-based fairness. *Science as Culture*, 27(4):413–438.

References

Molnar, C. (2023). *A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book.

Moodie, E. E. and Stephens, D. A. (2022). Causal inference: Critical developments, past and future. *Canadian Journal of Statistics*, 50(4):1299–1320.

Moulin, H. (1992). An application of the shapley value to fair division with money. *Econometrica*, pages 1331–1349.

Moulin, H. (2004). *Fair division and collective welfare*. MIT press.

Müller, R., Kornblith, S., and Hinton, G. E. (2019). When does label smoothing help? *Advances in neural information processing systems*, 32.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. *Statistical Science*, 5:465–480.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

References

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Pakdaman Naeini, M., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1):2901–2907.

Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

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.

Pojman, L. P. (1998). The case against affirmative action. *International Journal of Applied Philosophy*, 12(1):97–115.

Rahimi, A., Shaban, A., Cheng, C.-A., Hartley, R., and Boots, B. (2020). Intra order-preserving functions for calibration of multi-class neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13456–13467. Curran Associates, Inc.

References

Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J., et al. (2007). Climate models and their evaluation. In *Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC (FAR)*, pages 589–662. Cambridge University Press.

Ribeiro, M. T., Singh, S., and 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*, pages 1135–1144.

Rifkin, R. and Klautau, A. (2004). In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141.

Robnik-Šikonja, M. and Kononenko, I. (1997). An adaptation of relief for attribute estimation in regression. In *Machine learning: Proceedings of the fourteenth international conference (ICML'97)*, volume 5, pages 296–304.

Robnik-Šikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1):23–69.

Robnik-Šikonja, M. and Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600.

Rosenbaum, P. (2018). *Observation and experiment*. Harvard University Press.

References

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubenfeld, J. (1997). Affirmative action. *Yale Law Journal*, 107:427.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, pages 159–183.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Sabbagh, D. (2007). *Equality and transparency: A strategic perspective on affirmative action in American law*. Springer.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global sensitivity analysis: the primer*. John Wiley & Sons.

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.

Schauer, F. (2006). *Profiles, probabilities, and stereotypes*. Harvard University Press.

Schmidt, G. (2024). Climate models can't explain 2023's huge heat anomaly — we could be in uncharted territory. *Nature*, 627:467.

References

Sekhon, J. S. (2009). Causal inference, matching, and regression discontinuity. In *CELS 2009 4th Annual Conference on Empirical Legal Studies Paper*.

Shapley, L. S. (1953). A value for n-person games. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton.

Shapley, L. S. and Shubik, M. (1969). Pure competition, coalitional power, and fair division. *International Economic Review*, 10(3):337–362.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.

Štrumbelj, E. and Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18.

Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665.

Swiss Re (2015). Life insurance risk selection: Required differentiation or unfair discrimination? *Sigma*.

Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6):309.

References

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement*, 8(2):63–70.

Turner, R. (2015). The way to stop discrimination on the basis of race. *Stanford Journal of Civil Rights & Civil Liberties*, 11:45.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):1–7.

Van Gerven, G. (1993). Case c-109/91, Gerardus Cornelis Ten Oever v. Stichting bedrijfspensioenfonds voor het glazenwassers-en schoonmaakbedrijf. *EUR-Lex*, 61991CC0109.

Van Rijsbergen, C. (1979). Information retrieval: theory and practice. In *Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems*, volume 79.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE.

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*, pages 784–792. PMLR.

von Mises, R. (1928). *Wahrscheinlichkeit Statistik und Wahrheit*. Springer.

von Mises, R. (1939). *Probability, statistics and truth*. Macmillan.

References

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Walters, M. A. (1981). Risk classification standards. In *Proceedings of the Casualty Actuarial Society*, volume 68, pages 1–18.

Wang, D.-B., Feng, L., and Zhang, M.-L. (2021). Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. *Advances in Neural Information Processing Systems*, 34:11809–11820.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Wilks, D. S. (1990). On the combination of forecast probabilities for consecutive precipitation periods. *Weather and Forecasting*, 5(4):640–650.

Wilson, P. W., Castelli, W. P., and Kannel, W. B. (1987). Coronary risk prediction in adults (the framingham heart study). *The American journal of cardiology*, 59(14):G91–G94.

References

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847.

Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., et al. (2019). Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141.

Wright, S. (1921a). Correlation and causation. *Journal of Agricultural Research*, 20.

Wright, S. (1921b). Systems of mating. i. the biometric relations between parent and offspring. *Genetics*, 6(2):111.

Wright, S. (1934). The method of path coefficients. *The annals of mathematical statistics*, 5(3):161–215.

Yeh, R. W., Valsdottir, L. R., Yeh, M. W., Shen, C., Kramer, D. B., Strom, J. B., Secemsky, E. A., Healy, J. L., Domeier, R. M., Kazi, D. S., et al. (2018). Parachute use to prevent death and major trauma when jumping from aircraft: randomized controlled trial. *bmj*, 363.

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652.

References

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *arXiv*, 1507.05259:962–970.

Zhang, J., Kailkhura, B., and Han, T. Y.-J. (2020). Mix-n-match : Ensemble and compositional methods for uncertainty calibration in deep learning. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11117–11128. PMLR.