Actuariat & Données

Arthur Charpentier (Université de Rennes 1 & UQàM)

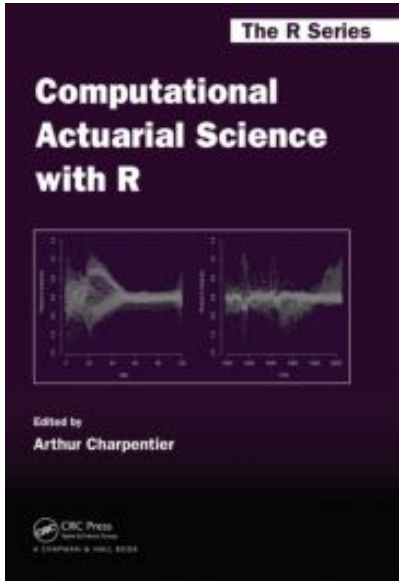
Rencontres de l'Actuariat

Generali, Paris, Mai 2016

http://freakonometrics.hypotheses.org

Actuariat & Données

Arthur Charpentier (Université de Rennes 1 & UQàM)



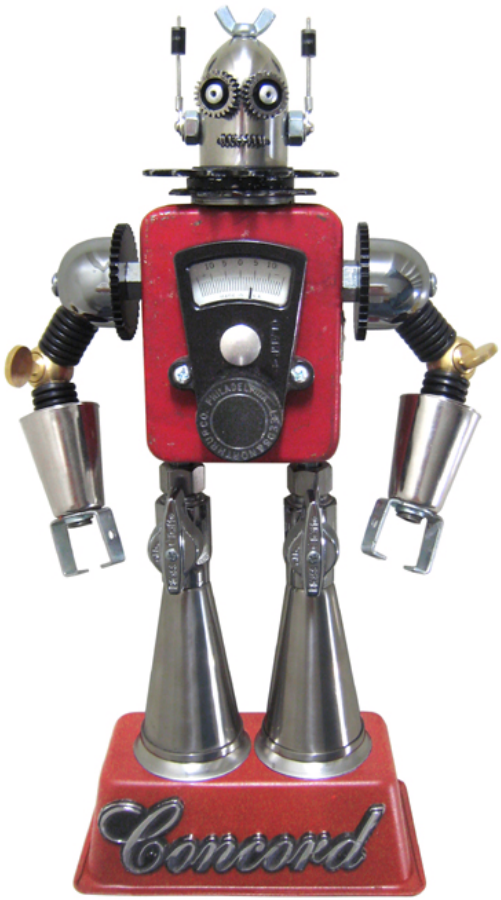
Maître de Conférence, Université de Rennes 1
Directeur des Études Data Science pour l'Actuariat, IA
Porteur de la Chaire *actinfo* (Institut Louis Bachelier)
(anciennement professeur UQàM & ENSAE Paristech
actuaire à Hong Kong, et FFSA, Paris)
PhD en Statistique (KU Leuven), Fellow Institut des Actuaires
MSc Mathématiques Financières (Paris Dauphine) & ENSAE
Éditeur du blog freakonometrics.hypotheses.org
Éditeur de Computational Actuarial Science, CRC Press

Data - Données

"People use statistics as the drunken man uses lamp posts - for support rather than illumination", Andrew Lang ou pas. Voir aussi Chris Anderson The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, 2008

1. **Enjeux autour du Big Data**
2. **Économétrie vs. Machine Learning**
3. **Données, Modèles & Actuariat**

Partie 1.
Enjeux autour du Big Data

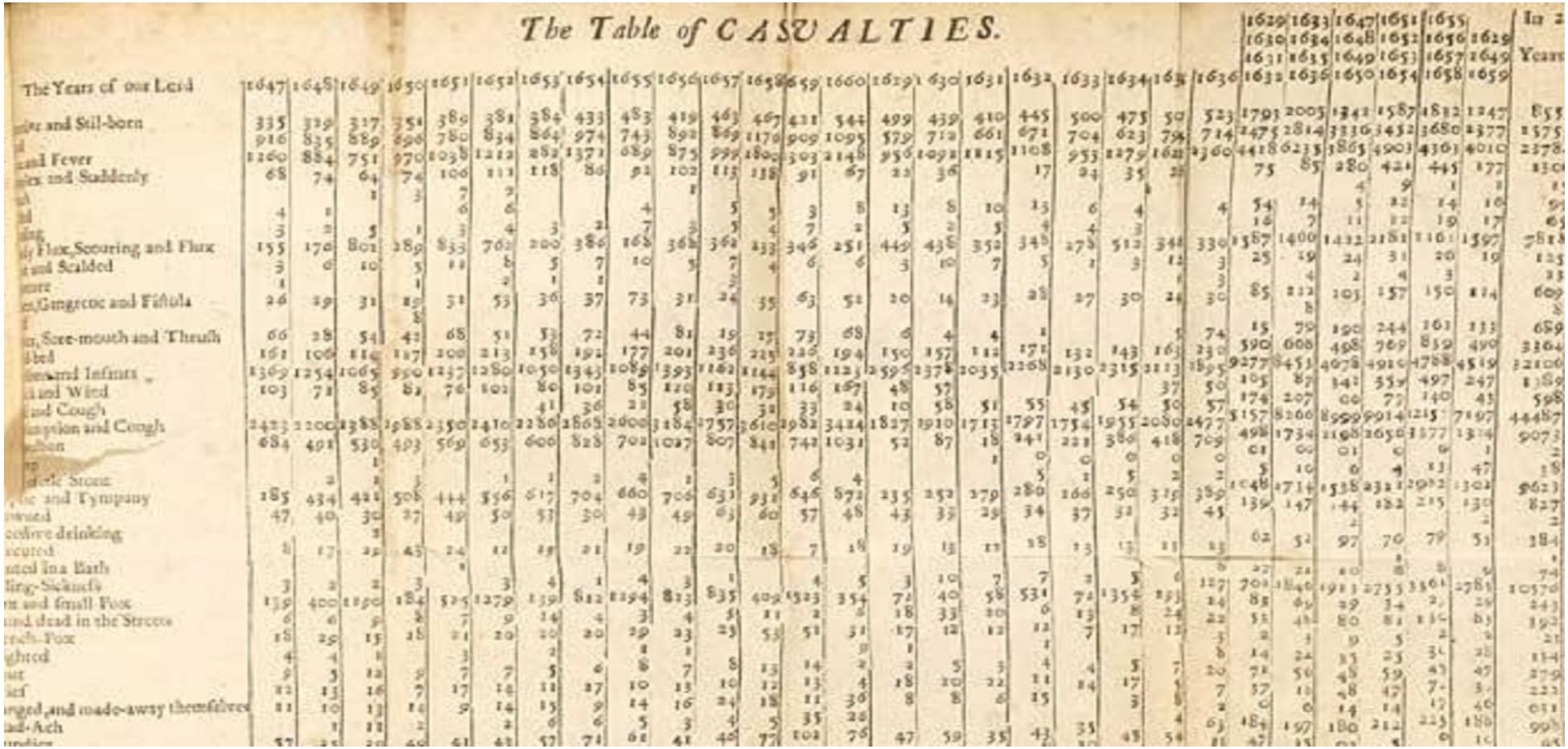


Aspects historiques autour de la donnée

Stocker des données: Bâton de comptage (*tally sticks*), depuis le Paleolithique

Les bâtons servait de mémoire stockant des quantités, voire des messages.

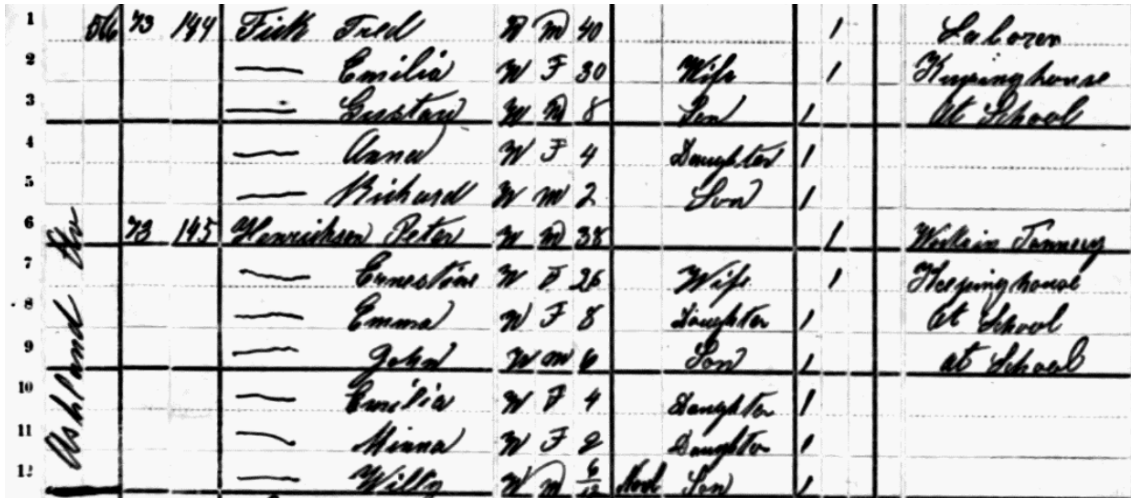
Aspects historiques autour de la donnée



Collecter des données: John Graunt a conduit une analyse de la propagation de la peste, en Europe, en 1663

Aspects historiques autour de la donnée

Manipulation des données: Herman Hollerith a créé une machine mécanographique (*Tabulating Machine*) perforant des cartes pour le recensement américain de 1880, cf 1880 Census,
n =50 million d'américains.



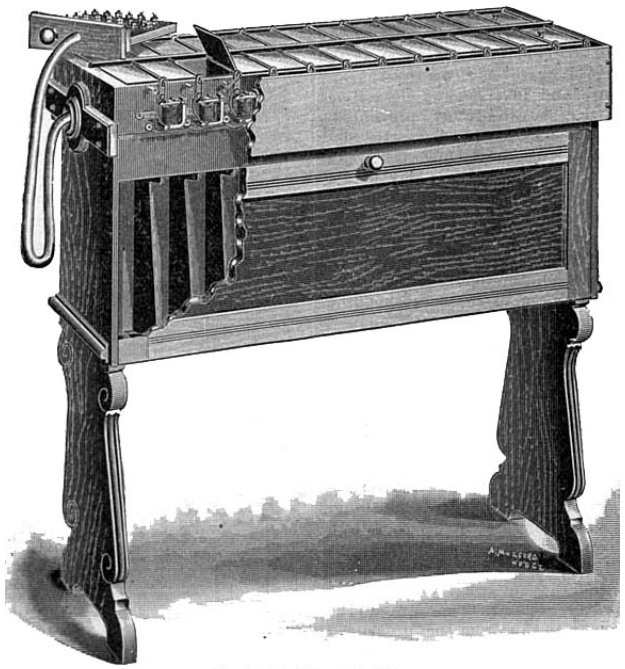


Fig. 3.—Sorting Machine.
Hollerith's Electric Sorting and Tabulating Machine.

Aspects historiques autour de la donnée

Échantillon & Sondage:

Élection présidentielle de 1936

Literary Digest Poll, 2.4 millions de lecteurs

A. Landon: 57% vs. F.D. Roosevelt: 43%

George Gallup, échantillon de 50,000 personnes

A. Landon: 44% vs. F.D. Roosevelt: 56%

Résultats réels

A. Landon: 38% vs. F.D. Roosevelt: 62%

Today

AMERICA SPEAKS

THE NATIONAL WEEKLY POLL of PUBLIC OPINION

Next Sunday

Institute Forecasts the Re-election of Franklin D. Roosevelt, Gives Him 54% of Popular Vote, Minimum of 315 Electors

Major Party Percent Is 55.7; New York in F.D.R. 'Sure' Column

Election Forecast

Election Will Test Clashing Poll Methods

Techniques d'échantillonnage, sondage, prévision à partir de petits échantillons (contre l'exhaustivité des recensements).

Aspects historiques autour de la donnée

Allen-Scott Report

Data Center Plan Called Privacy Invasion

By ROBERT S. ALLEN and PAUL SCOTT

WASHINGTON — A special White House task force is recommending the creation of a federal data center which eventually could have a comprehensive file on every man, woman and child in the country.

Now under study in inner administration circles, the still-secret report advocates the gradual transfer of all governmental records and statistics to magnetic computer tape, which would be turned over to a newly-created agency that would function as a general data center.

The computerized information would be available, at the push of a button, to a wide range of government authorities.

Estimated cost of the pro-

cial Security, census data, medical, credit and criminal reports.

"Comprehensive information of this kind, centralized in one agency," says Gallagher, "could constitute a highly dangerous dossier bank. Such an agency would be a distinct departure from our American tradition."

Subcommittee investigators have ascertained that the task force's report states that a vast accumulation of government records already is on computer tape and could be turned over to the proposed general data center immediately. Listed as among these available files are:

Internal Revenue Service — 74.2 million personal and corporate tax returns.

Defense Department — 14

the most intimate information, the investigators learned, are freely passed around among agencies. Graphically illustrative of this practice and its harsh consequences are the following two instances:

A teenager visiting Washington stayed with an uncle, at his mother's suggestion. During the night the boy was sexually assaulted by the uncle. Years later, as a Phi Beta Kappa graduate from a leading Eastern university, the boy applied for a job with the National Security Agency. During a required lie detector test he told about the assault. His frank admission cost him the desired job.

But that wasn't all. This affair, in which he was an inno-

Data Center: Le gouvernement américain a créé le premier data center en 1965, pour stocker des 175 millions d'empreintes et 742 millions de documents fiscaux.

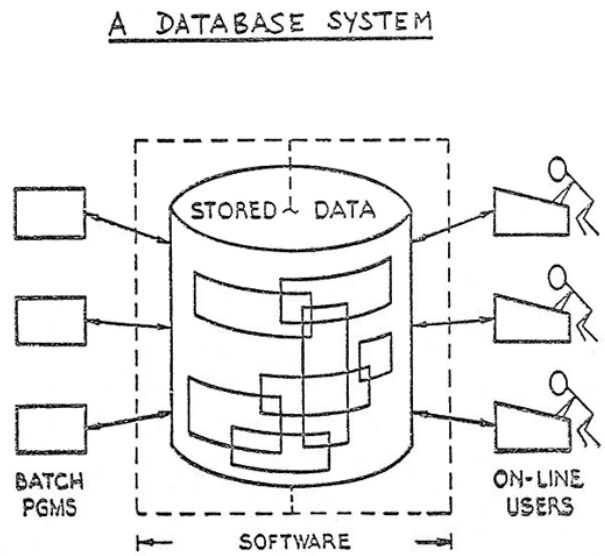
Aspects historiques autour de la donnée

Aspects historiques autour de la donnée

Manipuler des données: Notion de bases de données relationnelles, développée par Edgar F. Codd
cf Relational Model of Data for Large Shared Data Banks, Codd (1970)

Les bases sont penséees comme des matrices, avec des attributs en colonne, et les bases étant liées entre elles par un attribut commun (clé)

Début des diagrammes relationnels.

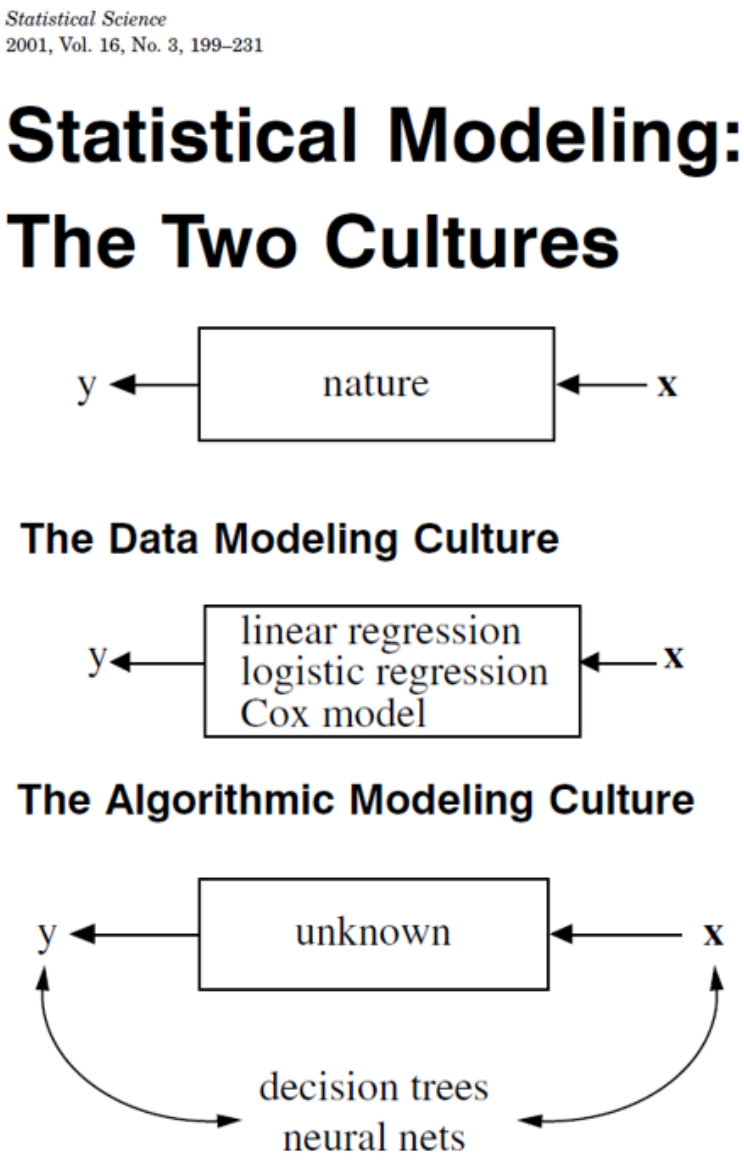


Les deux cultures

'The Two Cultures', de Breiman (2001)

- Data Modeling (statistique & économétrie)
- Algorithmic Modeling (algorithmes)

cf discussion Partie 2.



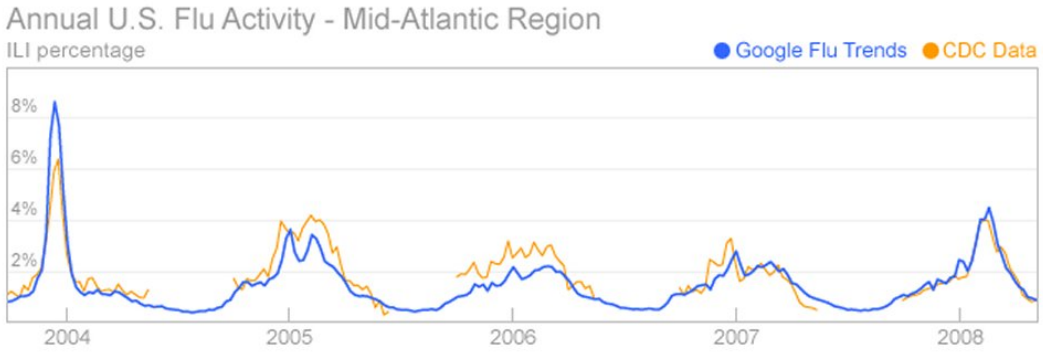
Et au XXIth siècle...

Début médiatique du Big Data à partir de Nature (2008) avant de toucher des revues plus orientées business.

Les 3V de Gartner (Volume, Variété, Vélocité), cf Gartner.

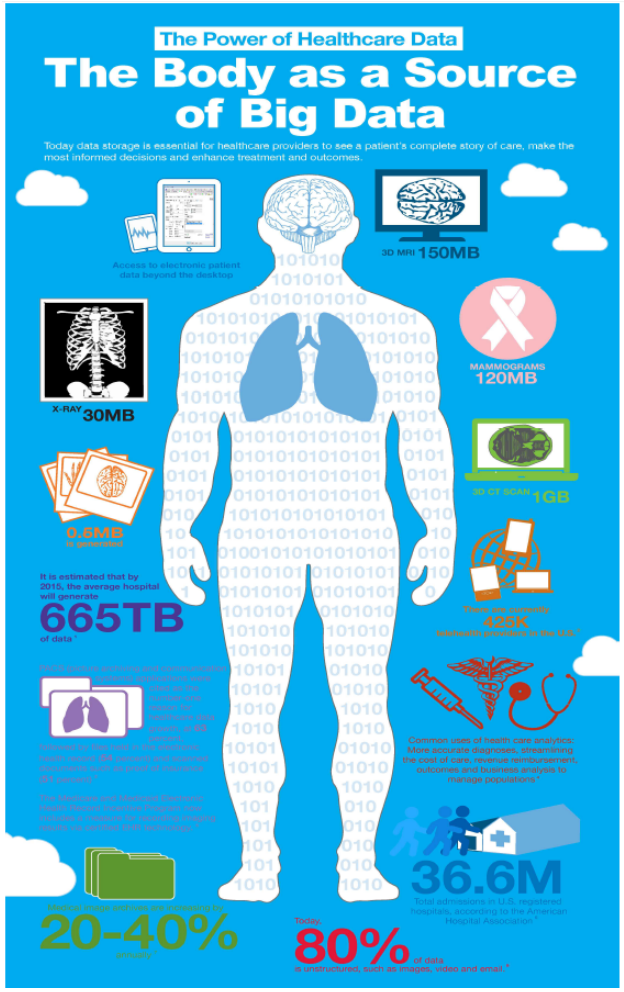
Big Data & Assurance (santé, par exemple)

Exemple: application Google Flu Trend



voir aussi Lazer *et al.* (2014)

Mais beaucoup plus avec des objets connectés.



Aspects computationnels

Comprendre le fonctionnement est devenu indispensable*

CPU Central Processing Unit, le coeur de la machine

RAM Random Access Memory, la mémoire non-persistante

HD Hard Drive, la mémoire persistante

CPU est rapide (mais de vitesse finie)

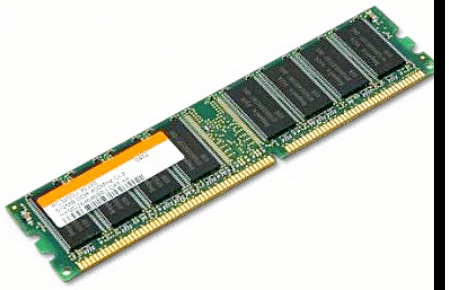
RAM est temporaire, rapide mais limitée

HD permanente, lente, et importante

Exemple Lire un fichier de 100Mb $\sim$ 1.01sec.

Exemple Lire 150 fichiers de 1Mb $\sim$ 0.9sec.

distinction entre la latence (temps entre la stimulation et la réponse, e.g. 10ms pour lire le premier bit) et la performance (nombre d'opérations par seconde, e.g. 100Mb/sec)



* merci à David Sibaï sur cette section.

Aspects computationnels

PC 'standard' :

CPU : 4 core, 1ns de latence

RAM : 32 ou 64 Gb, 100ns latence, 20Gb/sec

HD : 1 Tb, 10ms de latence, 100Mo/sec

Parcourir un fichier texte de 2Tb,

en comptant les *espaces* 2.10^{12} opérations

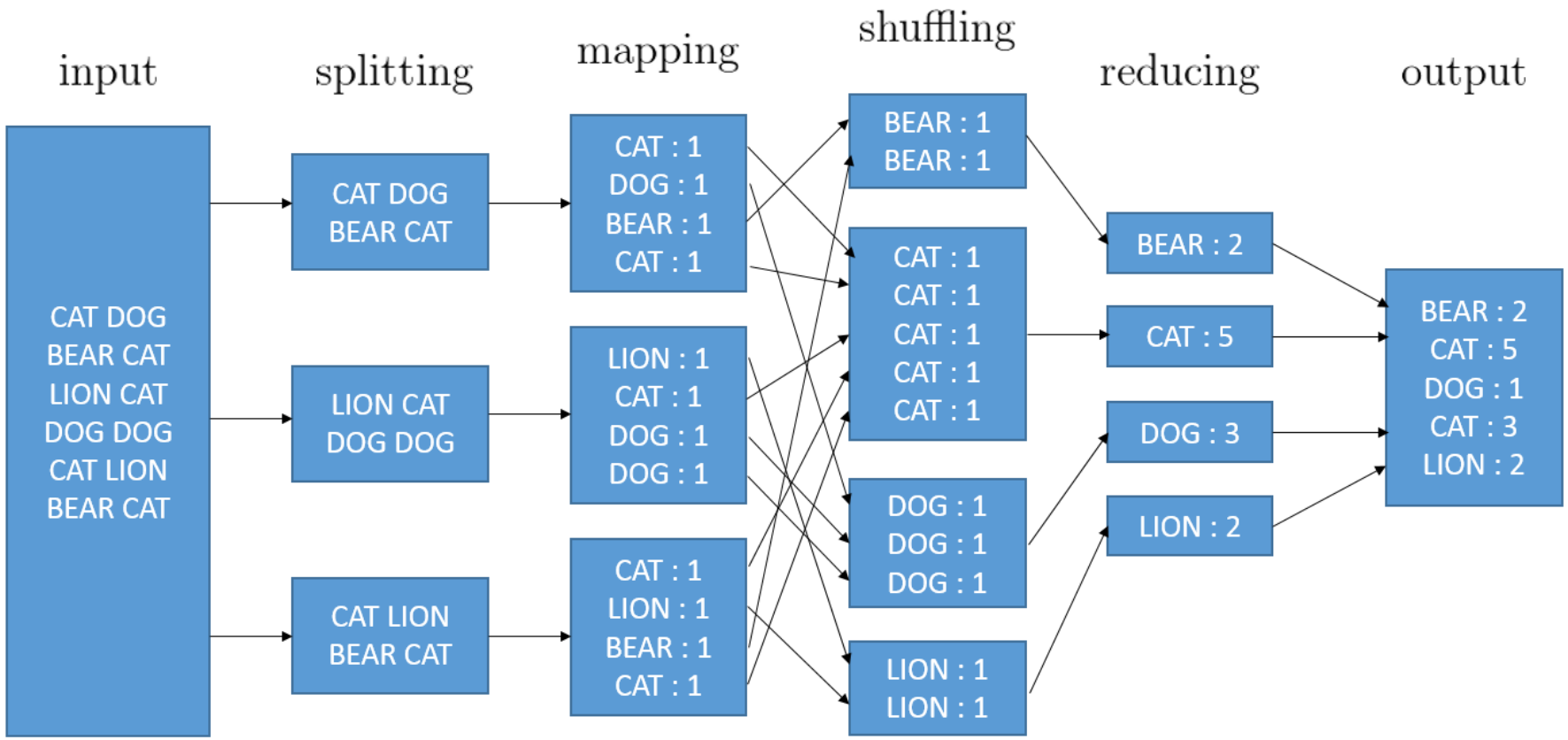
Stockage HD, 100Mb/sec $\sim 2.10^4$ sec $\sim$ 6 heures

La magie n'existe pas, il y a toujours des limites...

Solution possible: paralléliser

Aspects computationnels

La parallélisation consister à subdivisionner les tâches (map) puis à les aggréger ensuite (reduce). Cf comptage de mots dans un texte avec MapReduce



Partie 2.
Économétrie vs. Machine Learning

Les données en grande dimension ('Big Data')

Terminologie de Bühlmann & van de Geer (2011) et Koch (2013). $\boldsymbol{X}$ est une matrice $n \times p$.

Portnoy (1988) montre que l'estimateur du maximum de vraisemblance est asymptotiquement Gaussien quand $p^2/n \to 0$, lorsque $n, p \to \infty$. On a des **données massives** si $p > \sqrt{n}$.

Le concept de **sparcité** est basé non pas sur p, mais sur la taille effective (variables explicatives 'significatives'). On peut alors envisager le cas $p > n$.

La grande dimension (en p) peut faire peur à cause de la **malédiction de la dimension**, cf Bellman (1957). Le volume de la sphère unité dans $\mathbb{R}^p$ tend vers 0 lorsque $p \to \infty$: l'espace est vide.

Données et modèles

Quelle histoire raconter à partir des données $\{(y_i, \boldsymbol{x}_i)\}$, cf Freedman (2005)

- l'histoire **causale** : pour tout $\boldsymbol{x}$ on associe $m(\boldsymbol{x})$, et on rajoute un bruit ε. Le but est de retrouver $m(\cdot)$, les résidus étant définis comme la différence entre y et $m(\boldsymbol{x})$.
- l'histoire **de lois conditionnelles** : pour le modèle Gaussien, Y sachant $\boldsymbol{X} = \boldsymbol{x}$ suit une loi $\mathcal{N}(m(\boldsymbol{x}), \sigma^2)$. $m(\boldsymbol{x})$ correspond alors à l'espérance conditionnelle. Aucune hypothèse causale n'est faite ici. Il faut un modèle probabiliste ici $(\Omega, \mathcal{F}, \mathbb{P})$, et une loi conditionnelle pour $Y|\boldsymbol{X}$
- l'histoire **exploratoire des données** : On a juste des données. On veut résumer l'information contenue dans les $\boldsymbol{x}$ de manière à être le plus proche possible de y (i.e. $\min\{\ell(\boldsymbol{y}, m(\boldsymbol{x}))\}$) pour une fonction de perte ℓ.

Voir discussion dans Varian (2014)

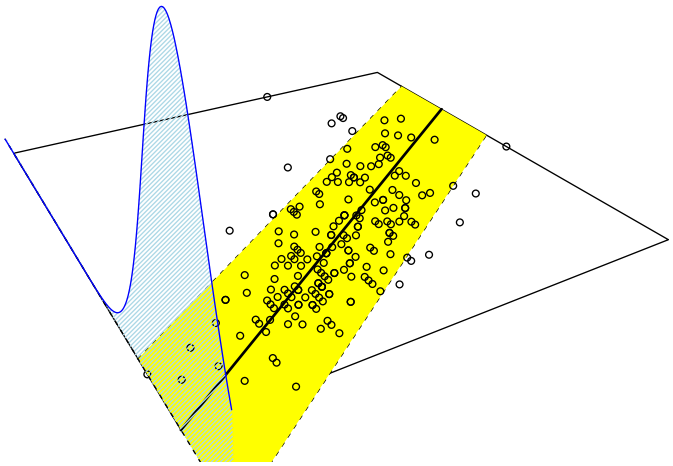
Économétrie: le modèle linéaire

Modèle linéaire, homoscédastique Gaussien

$$(Y|\boldsymbol{X}=\boldsymbol{x}) \sim \mathcal{N}(\mu(\boldsymbol{x}), \sigma^2)$$

- **linéaire**: $\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}] = \mu(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}$
- **homoscédastique**: $\mathrm{Var}[Y|\boldsymbol{X}=\boldsymbol{x}] = \sigma^2$.

Condition du Premier Ordre $\boldsymbol{X}^\mathsf{T}[\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}] = \boldsymbol{0}$
et $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$



Le modèle linéaire généralisé

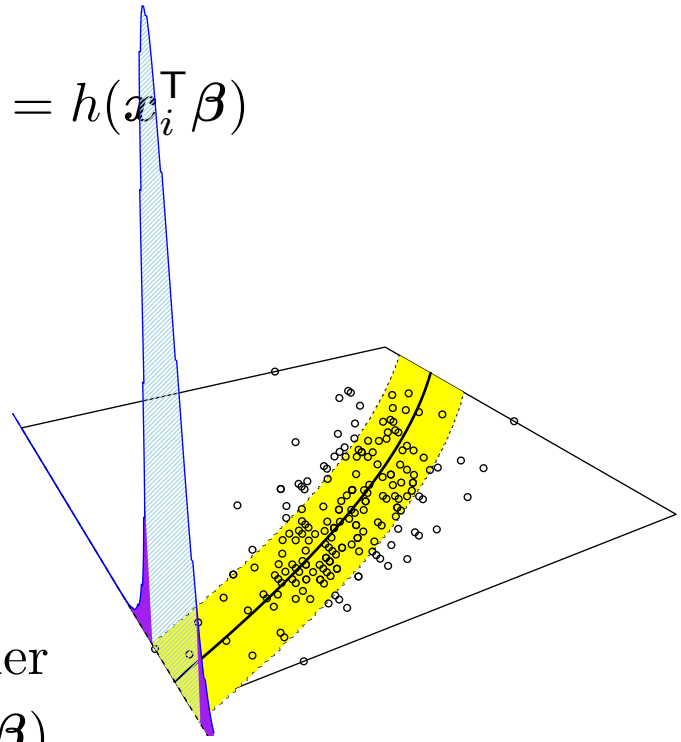
Loi (conditionnelle, $Y|\boldsymbol{X}$) dans la famille exponentielle

$$f(y_i|\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right) \text{ with } \theta_i = h(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})$$

i.e.

$$(Y|\boldsymbol{X}=\boldsymbol{x}) \sim \mathcal{L}(\theta_{\boldsymbol{x}},\varphi)$$

- **Modèle linéaire**: $[Y|\boldsymbol{X}=\boldsymbol{x}] = g^{-1}(\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta})$



e.g. $(Y|\boldsymbol{X}=\boldsymbol{x}) \sim \mathcal{P}(\exp[\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}])$.
Ici $\boldsymbol{\beta}$ vérifie $\boldsymbol{X}^\mathsf{T}\boldsymbol{W}[\boldsymbol{y}-\widehat{\boldsymbol{\mu}}]=\boldsymbol{0}$ (condition du premier ordre, maximum de vraisemblance) avec $\boldsymbol{W}=\boldsymbol{W}(\boldsymbol{\beta})$ et $\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\beta})$.

Tests et significativité

La variable x_k est significative si on rejette le test $H_0 : \beta_k = 0$ contre $H_1 : \beta_k \neq 0$.
Utilisation du **t-test de Student** basé sur $t_k = \widehat{\beta}_k / \text{se}_{\widehat{\beta}_k}$,

On calcule la **p-value** $\mathbb{P}[|T| > |t_k|]$ avec $T \sim t_\nu$ (avec souvent $\nu = p + 1$).

En grande dimension, problème du **FDR (False Discovery Ratio)**

Avec $\alpha = 5\%$, 5% des variables sont faussement significatives

Si $p = 100$ avec 5 variables (réellement) significatives, on a 5 faux positifs (en moyenne), i.e. 50% FDR, cf Benjamini & Hochberg (1995).

Apprentissage statistique (Machine Learning)

Pas de modèle probabiliste, mais une **fonction de perte**, ℓ. On se donne un ensemble de fonctions $\mathcal{M}$, $\mathcal{X} \to \mathcal{Y}$, et on pose

$$m^\star = \underset{m \in \mathcal{M}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, m(\boldsymbol{x}_i)) \right\}$$

La **fonction de perte quadratique** est appréciée parce que

$$\overline{y} = \underset{m \in \mathbb{R}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \frac{1}{n} [y_i - m]^2 \right\}$$

qui donne, dans une version probabiliste $\mathbb{E}(Y) = \underset{m \in \mathbb{R}}{\text{argmin}} \left\{ \|Y - m\|_{\ell_2}^2 \right\}$

Pour $\tau \in (0, 1)$, on obtient la **régression quantile** (cf Koenker (2005))

$$m^\star = \underset{m \in \mathcal{M}_0}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell_\tau(y_i, m(\boldsymbol{x}_i)) \right\} \text{ avec } \ell_\tau(x, y) = |(x - y)(\tau - \mathbf{1}_{x \leq y})|$$

Boosting & Apprentissage faible (weak)

On cherche à résoudre

$$m^\star = \underset{m\in\mathcal{M}}{\text{argmin}} \left\{ \sum_{i=1}^n \ell(y_i, m(\boldsymbol{x}_i)) \right\}$$

qui est un problème complexe dans un espace $\mathcal{M}$ de fonctions $\mathcal{X} \to \mathcal{Y}$ général.

On considère une procédure itérative, qui consiste à apprendre de ses erreurs,

$$m^{(k)}(\cdot) = \underbrace{m_1(\cdot)}_{\sim \boldsymbol{y}} + \underbrace{m_2(\cdot)}_{\sim \boldsymbol{\varepsilon}_1} + \underbrace{m_3(\cdot)}_{\sim \boldsymbol{\varepsilon}_2} + \cdots + \underbrace{m_k(\cdot)}_{\sim \boldsymbol{\varepsilon}_{k-1}} = m^{(k-1)}(\cdot) + m_k(\cdot).$$

Boosting & Apprentissage faible (weak)

Il est possible de reconnaître une descente de gradient. Pas

$$\underbrace{f(\boldsymbol{x}_k)}_{\langle f,\boldsymbol{x}_k\rangle} \sim \underbrace{f(\boldsymbol{x}_{k-1})}_{\langle f,\boldsymbol{x}_{k-1}\rangle} + \underbrace{(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})}_{\alpha_k} \underbrace{\nabla f(\boldsymbol{x}_{k-1})}_{\langle \nabla f,\boldsymbol{x}_{k-1}\rangle}$$

mais une version duale

$$\underbrace{f_k(\boldsymbol{x})}_{\langle f_k,\boldsymbol{x}\rangle} \sim \underbrace{f_{k-1}(\boldsymbol{x})}_{\langle f_{k-1},\boldsymbol{x}\rangle} + \underbrace{(f_k - f_{k-1})}_{a_k} \underbrace{\star}_{\langle f_{k-1},\nabla \boldsymbol{x}\rangle}$$

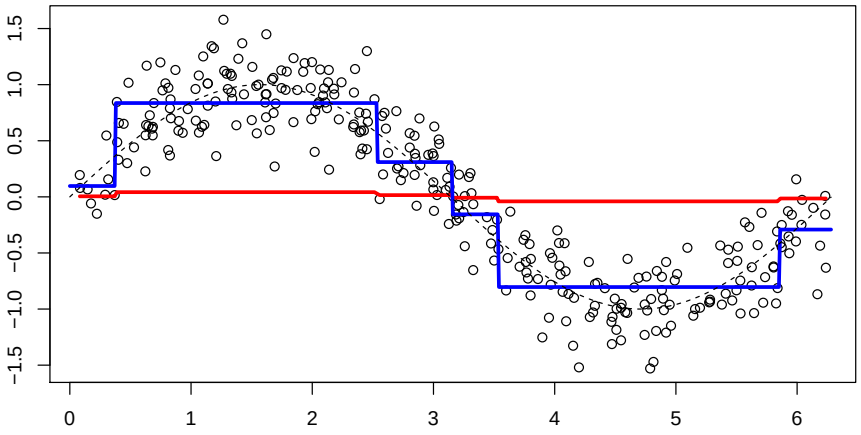
où $\star$ est un gradient, dans un espace fonctionnel. On a alors ici

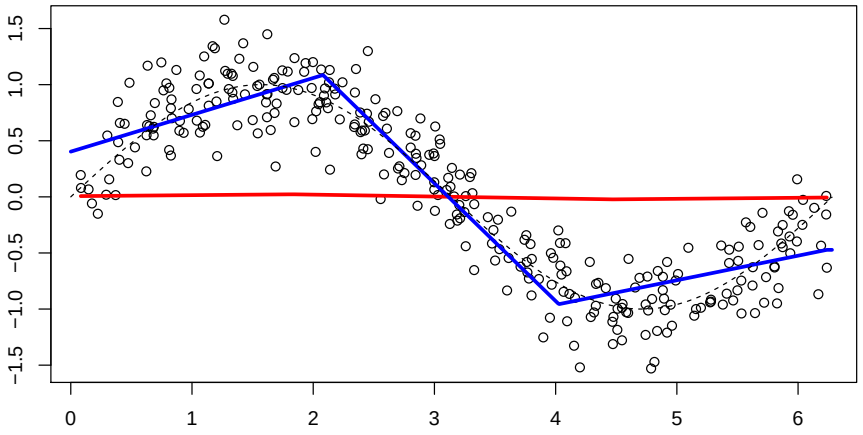
$$m^{(k)}(\boldsymbol{x}) = m^{(k-1)}(\boldsymbol{x}) + \underset{f\in\mathcal{F}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, m^{(k-1)}(\boldsymbol{x}) + f(\boldsymbol{x})) \right\}$$

où on va choisir un ensemble $\mathcal{F}$ relativement simple (faible) de fonctions en escalier (appelées *stumps*)

Boosting & Apprentissage faible (weak)

Classiquement $\mathcal{F}$ sont des fonctions en escalier, mais on peut prendre toutes sortes de fonctions (e.g. des splines linéaires)





On peut ajouter un paramètre d'**atténuation** (shrinkage) pour apprendre encore plus lentement, $\varepsilon_1 = y - \alpha \cdot m_1(\boldsymbol{x})$ avec $\alpha \in (0, 1)$, etc.

Sur-apprentissage & Pénalisation

On résout ici, pour une norme $\|\cdot\|$ (reflétant la complexité du modèle)

$$\min\left\{\sum_{i=1}^{n}\ell(y_i,\beta_0+\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta})+\lambda\|\boldsymbol{\beta}\|\right\}=\min\Big\{\text{objectif}(\boldsymbol{\beta})+\text{pénalité}(\boldsymbol{\beta})\Big\}.$$

cf Ridge. Ces estimateurs ne sont plus **sans biais**, mais ont un mse plus faible.

Considérons un échantillon i.id. $\{y_1,\cdots,y_n\}$ suivant $\mathcal{N}(\theta,\sigma^2)$, et cherchons un estimateur proportionnel à $\overline{y}$, i.e. $\widehat{\theta}=\alpha\overline{y}$. $\alpha=1$ correspond à l'estimateur du maximum de vraisemblance. Or

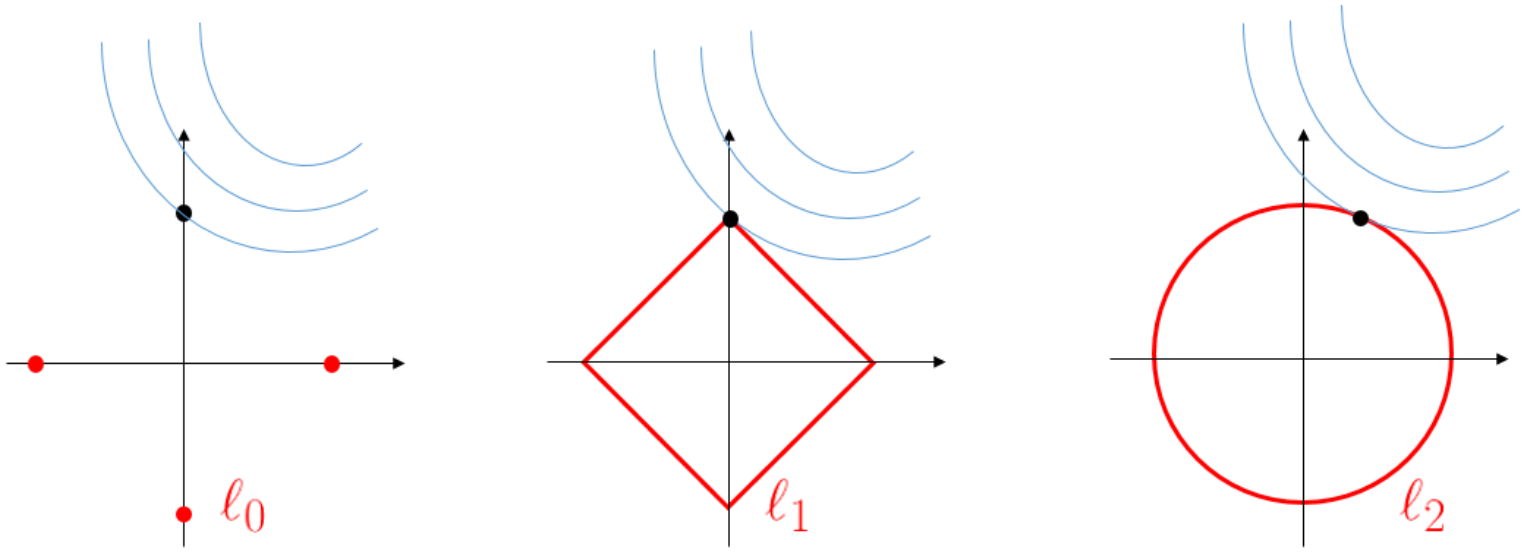
$$\text{mse}[\widehat{\theta}]=\underbrace{(\alpha-1)^2\mu^2}_{\text{biais}[\widehat{\theta}]^2}+\underbrace{\frac{\alpha^2\sigma^2}{n}}_{\text{Var}[\widehat{\theta}]}$$

et donc $\alpha^\star=\mu^2\cdot\left(\mu^2+\dfrac{\sigma^2}{n}\right)^{-1}<1.$

$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \text{argmin} \left\{ \sum_{i=1}^{n} \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\| \right\},$$

correspond au **Lagrangien** du problème d'optimisation sous contrainte

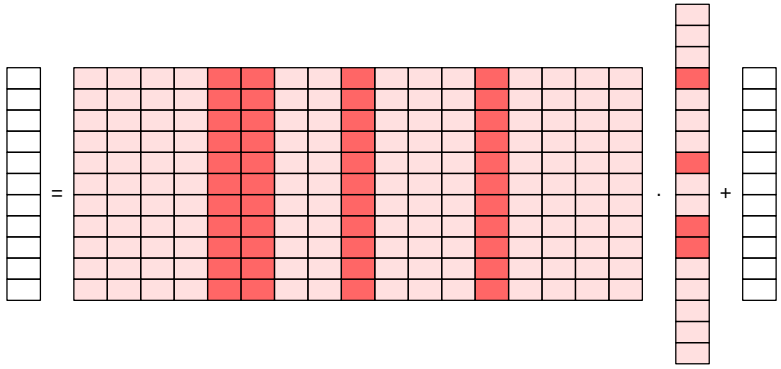
$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|\leq s}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) \right\}$$



LASSO & Modèles sparses

Souvent, p est (très) grand, mais beaucoup de covariables sont juste du bruit, $\beta_j = 0$ pour plusieurs j. Soit s le nombre de **variables significatives**, avec $s << p$, cf Hastie, Tibshirani & Wainwright (2015),

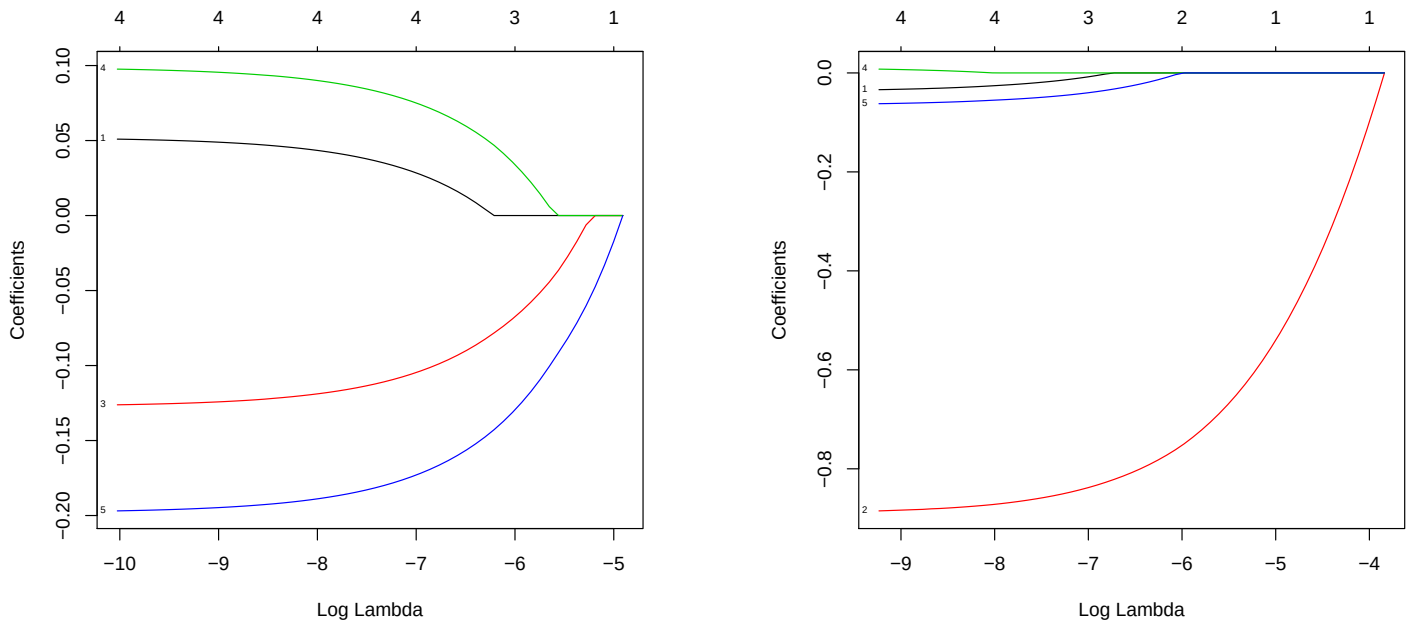
$$s = \text{card}\{\mathcal{S}\} \text{ où } \mathcal{S} = \{j; \beta_j \neq 0\}$$



Le vrai modèle est ici $y = \boldsymbol{X}_{\mathcal{S}}^{\mathsf{T}}\boldsymbol{\beta}_{\mathcal{S}} + \varepsilon$, avec $\boldsymbol{X}_{\mathcal{S}}^{\mathsf{T}}\boldsymbol{X}_{\mathcal{S}}$ de plein rang (colonne).

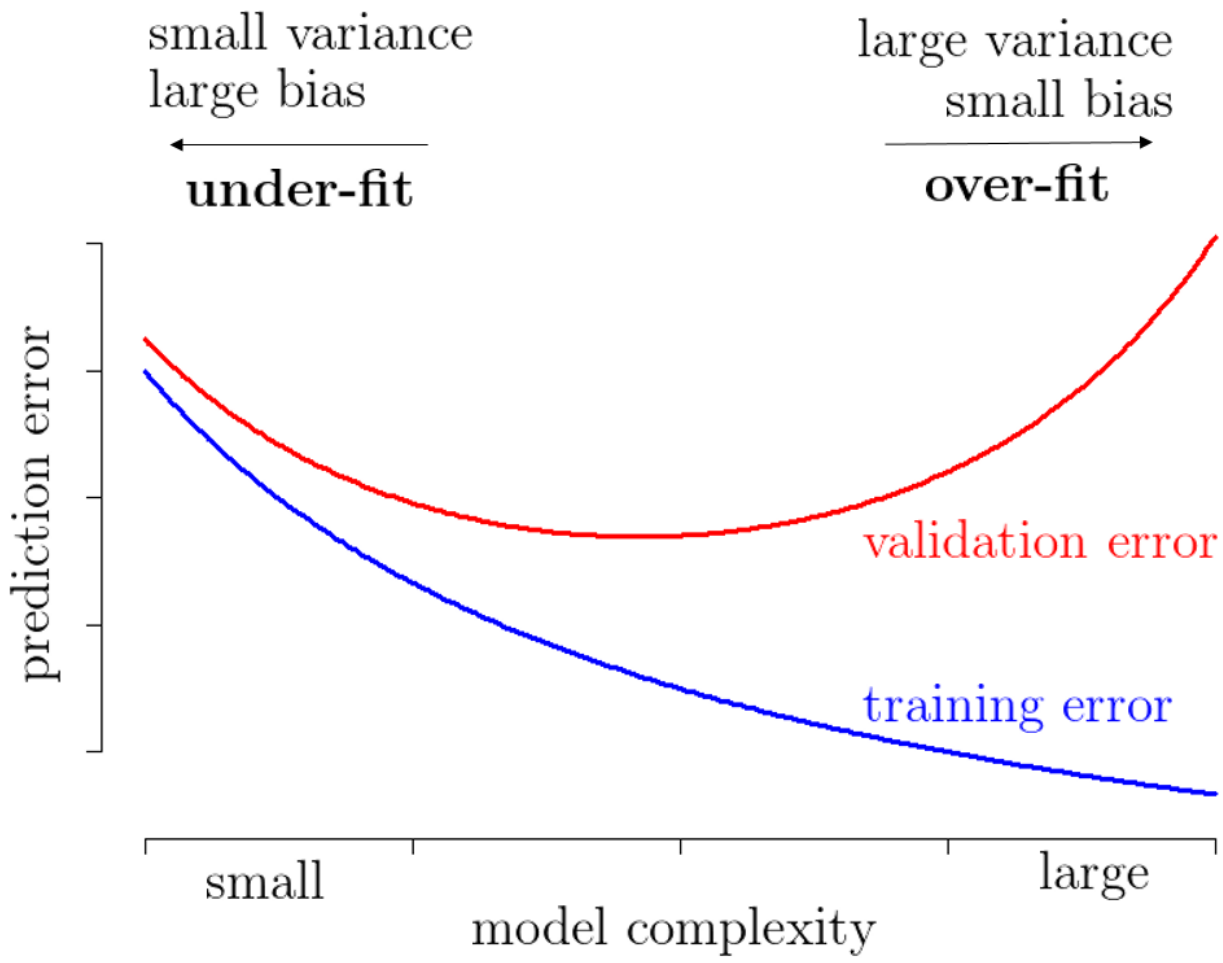
LASSO & Modèles sparses

Évolution de $\widehat{\beta}_\lambda$ en fonction de $\log \lambda$



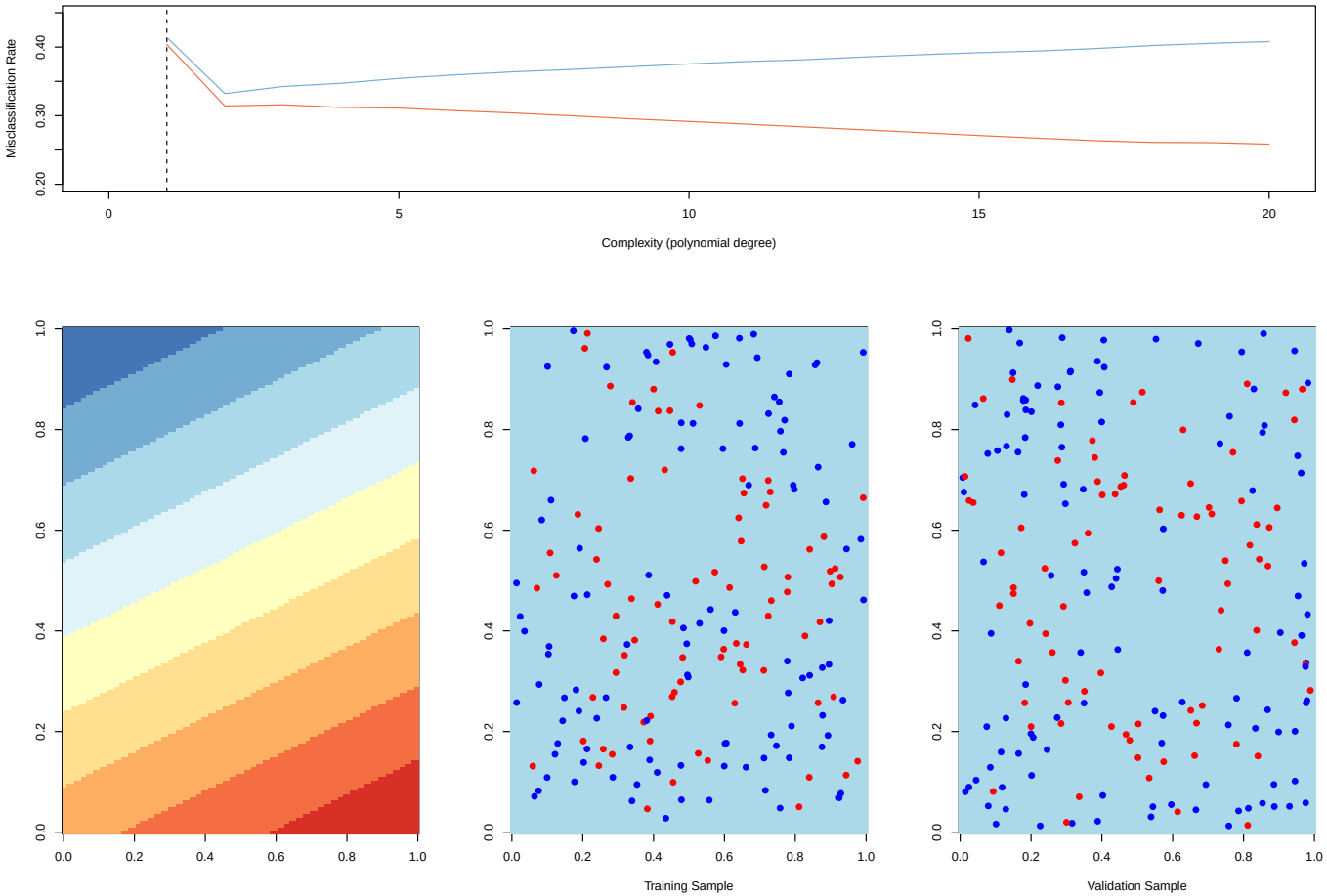
sur un modèle de fréquence dommage auto (à gauche) et RC auto (à droite).

In-Sample & Out-Sample



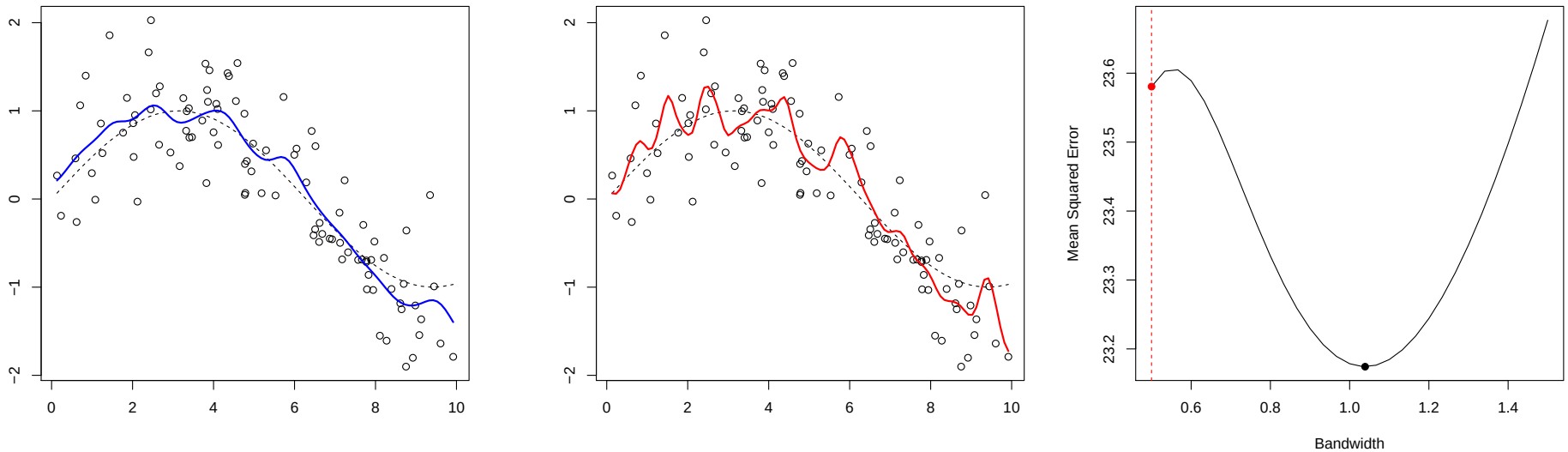
Overfit, Généralisation & Complexité

Dans une régression polynômiale, la complexité correspond au degré du polynôme



Paramètres de lissage (thinning) et validation croisée

$$\widehat{m}^{[h^\star]}(x) = \widehat{\beta}_0^{[x]} + \widehat{\beta}_1^{[x]} x \text{ avec } (\widehat{\beta}_0^{[x]}, \widehat{\beta}_1^{[x]}) = \underset{(\beta_0,\beta_1)}{\text{argmin}} \left\{ \sum_{i=1}^n \omega_{h^\star}^{[x]} [y_i - (\beta_0 + \beta_1 x_i)]^2 \right\}$$



$$\text{ou } h^\star = \text{argmin}\{\text{mse}(h)\} \text{ avec } \text{mse}(h) = \frac{1}{n} \sum_{i=1}^n \left[y_i - \widehat{m}_{(i)}^{[h]}(x_i) \right]^2$$

Économétrie vs. Machine Learning

L'aspect Big Data

- $n \to \infty$: loi du 0/1, tout est alors simplifié (soit vrai, soit faux)
- $p \to \infty$: complexité algorithmique, besoin de choisir les variables

Économétrie vs. Machine Learning

- interprétation probabiliste des modèles économétriques
  (peut induire en erreur, e.g. p-value)
  reste valide dans le cas non-i.id (séries chronologiques, panel, etc)
- machine learning permet d'avoir des modèles prédictifs généralisables

  outils algorithmiques, importance du bootstrap, de la validation-croisée, de la sélection de variable, des nonlinéaritiés et d'effets croisés, etc

Part 3.
Données, Modèles & Actuariat

L'arbitrage Privacy-Utility

Au Massachusetts, la Group Insurance Commission (GIC) gère l'assurance santé de tous les employés d'état, et a obligation de publier des données agrégées

GIC(zip, date of birth, sex, diagnosis, procedure, ...)

Sweeney (2000) a obtenu les registres de votes pour Cambridge, Massachusetts,

VOTER(name, party, ..., zip, date of birth, sex)

William Weld (ancien governeur) habite à Cambridge, et figure dans VOTER

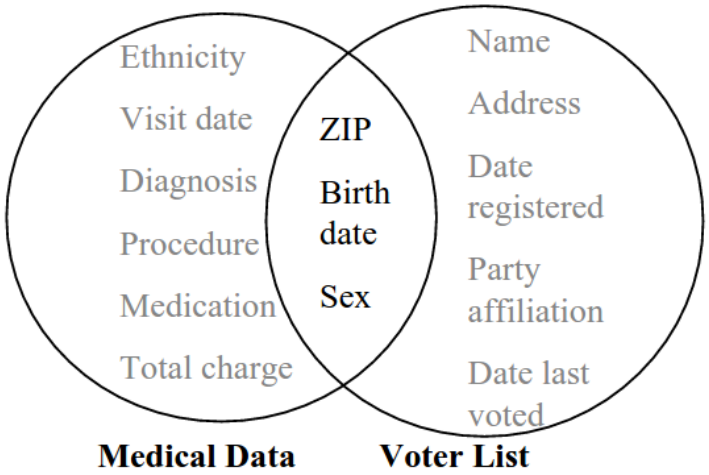


Figure 1 Linking to re-identify data

L'arbitrage Privacy-Utility

- 6 personnes dans VOTER avaient la même date de naissance
- seulement 3 étaient des hommes
- et Weld était le seul avec le zip code
- Sweeney a obtenu l'historique médical de Weld

"87% of Americans are uniquely identified by their zip code, gender and birth date", cf Sweeney (2000).

Une base de données est k-anonymisée si l'information relative à une personne ne peut être distinguée d'au moins $k-1$ autre personnes.

manière des caractéristiques $\mathbf{\Omega}$ de l'assuré, et lui réclame donc une prime pure de montant $\mathbb{E}[S]$, la même que celle qu'il réclame à tous les assurés du portefeuille. Dans ce cas, la situation est telle que présentée au Tableau 3.7.

	Assurés	Assureur
Dépense	$\mathbb{E}[S]$	$S-\mathbb{E}[S]$
Dépense moyenne	$\mathbb{E}[S]$	0
Variance	0	$\mathbb{V}[S]$

TAB. 3.7 – *Situation des assurés et de l'assureur en l'absence de segmentation.*

L'assureur prend donc l'entièreté de la variance des sinistres $\mathbb{V}[S]$ à sa charge, que celle-ci soit due à l'hétérogénéité du portefeuille, ou à la variabilité intrinsèque des montants des sinistres.

Transfert de risque en information complète

A l'autre extrême, supposons que l'assureur incorpore toute l'information $\mathbf{\Omega}$ dans la tarification. On serait alors dans la situation décrite au Tableau 3.8.

	Assurés	Assureur
Dépense	$\mathbb{E}[S\|\mathbf{\Omega}]$	$S-\mathbb{E}[S\|\mathbf{\Omega}]$
Dépense moyenne	$\mathbb{E}[S]$	0
Variance	$\mathbb{V}\Big[\mathbb{E}[S\|\mathbf{\Omega}]\Big]$	$\mathbb{V}\Big[S-\mathbb{E}[S\|\mathbf{\Omega}]\Big]$

TAB. 3.8 – *Situation des assurés et de l'assureur dans le cas où la segmentation est opérée sur base de* $\mathbf{\Omega}$.

Contrairement au cas précédent, la prime payée par un assuré prélevé au hasard dans le portefeuille est à présent une variable aléatoire: $\mathbb{E}[S|\mathbf{\Omega}]$ dépend des caractéristiques $\mathbf{\Omega}$ de cet assuré. Comme la variable aléatoire $S-\mathbb{E}[S|\mathbf{\Omega}]$ est centrée, le risque assumé par l'assureur la variance du résultat financier de l'opération d'assurance, i.e.

$$\mathbb{V}\Big[S-\mathbb{E}[S|\mathbf{\Omega}]\Big] = \mathbb{E}\Big[\Big(S-\mathbb{E}[S|\mathbf{\Omega}]\Big)^2\Big]$$

Aucune segmentation

	Assuré	Assureur
Perte	$\mathbb{E}[S]$	$S-\mathbb{E}[S]$
Perte moyenne	$\mathbb{E}[S]$	0
Variance	0	$\text{Var}[S]$

Information parfaite: $\mathbf{\Omega}$ observable

	Assuré	Assureur
Perte	$\mathbb{E}[S\|\mathbf{\Omega}]$	$S-\mathbb{E}[S\|\mathbf{\Omega}]$
Perte moyenne	$\mathbb{E}[S]$	0
Variance	$\text{Var}\Big[\mathbb{E}[S\|\mathbf{\Omega}]\Big]$	$\text{Var}\Big[S-\mathbb{E}[S\|\mathbf{\Omega}]\Big]$

$$\text{Var}[S] = \underbrace{\mathbb{E}\Big[\text{Var}[S|\mathbf{\Omega}]\Big]}_{\rightarrow\ \text{assureur}} + \underbrace{\text{Var}\Big[\mathbb{E}[S|\mathbf{\Omega}]\Big]}_{\rightarrow\ \text{assuré}}.$$

On assiste dans ce cas à un partage de la variance totale de S (c'est-à-dire du risque) entre les assurés et l'assureur, matérialisé par la formule

$$\mathbb{V}[S] = \underbrace{\mathbb{E}\Big[\mathbb{V}[S|\boldsymbol{\Omega}]\Big]}_{\rightarrow\text{assureur}} + \underbrace{\mathbb{V}\Big[\mathbb{E}[S|\boldsymbol{\Omega}]\Big]}_{\rightarrow\text{assurés}}.$$

Ainsi, lorsque toutes les variables pertinentes $\boldsymbol{\Omega}$ ont été prises en compte, l'intervention de l'assureur se limite à la part des sinistres due exclusivement au hasard; en effet, $\mathbb{V}[S|\boldsymbol{\Omega}]$ représente les fluctuations de S dues au seul hasard. Dans cette situation idéale, l'assureur mutualise le risque et il n'y a donc aucune solidarité induite entre les assurés du portefeuille: chacun paie en fonction de son propre risque.

Transfert des risques en information partielle

Bien entendu, la situation décrite au paragraphe précédent est purement théorique puisque parmi les variables explicatives $\boldsymbol{\Omega}$ nombreuses sont celles qui ne peuvent pas être observées par l'assureur. En assurance automobile par exemple, l'assureur ne peut pas observer la vitesse à laquelle roule l'assuré, son agressivité au volant, ni le nombre de kilomètres qu'il parcourt chaque année[2]. Dès lors, l'assureur ne peut utiliser qu'un sous-ensemble $\boldsymbol{X}$ des variables explicatives contenues dans $\boldsymbol{\Omega}$, i.e. $\boldsymbol{X} \subset \boldsymbol{\Omega}$. La situation est alors semblable à celle décrite au Tableau 3.9.

	Assuré	Assureur
Dépense	$\mathbb{E}[S\|\boldsymbol{X}]$	$S - \mathbb{E}[S\|\boldsymbol{X}]$
Dépense moyenne	$\mathbb{E}[S]$	0
Variance	$\mathbb{V}\Big[\mathbb{E}[S\|\boldsymbol{X}]\Big]$	$\mathbb{E}\Big[\mathbb{V}[S\|\boldsymbol{X}]\Big]$

TAB. 3.9 – *Situation de l'assuré et de l'assureur dans le cas où la segmentation est opérée sur base de $\boldsymbol{X} \subset \boldsymbol{\Omega}$.*

Il est intéressant de constater que

$$\begin{aligned}\mathbb{E}\Big[\mathbb{V}[S|\boldsymbol{X}]\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{V}[S|\boldsymbol{\Omega}]\Big|\boldsymbol{X}\Big]\Big] + \mathbb{E}\Big[\mathbb{V}\Big[\mathbb{E}[S|\boldsymbol{\Omega}]\Big|\boldsymbol{X}\Big]\Big] \\ &= \underbrace{\mathbb{E}\Big[\mathbb{V}[S|\boldsymbol{\Omega}]\Big]}_{\text{mutualisation}} + \underbrace{\mathbb{E}\Big\{\mathbb{V}\Big[\mathbb{E}[S|\boldsymbol{\Omega}]\Big|\boldsymbol{X}\Big]\Big\}}_{\text{solidarité}}. \qquad (3.22)\end{aligned}$$

Information imparfaite: $\boldsymbol{X} \subset \boldsymbol{\Omega}$ est observable

	Assuré	Assureur
Perte	$\mathbb{E}[S\|\boldsymbol{X}]$	$S - \mathbb{E}[S\|\boldsymbol{X}]$
Perte moyenne	$\mathbb{E}[S]$	0
Variance	$\text{Var}\Big[\mathbb{E}[S\|\boldsymbol{X}]\Big]$	$\mathbb{E}\Big[\text{Var}[S\|\boldsymbol{X}]\Big]$

$$\begin{aligned}\text{Var}[S] &= \mathbb{E}\Big[\text{Var}[S|\boldsymbol{X}]\Big] + \text{Var}\Big[\mathbb{E}[S|\boldsymbol{X}]\Big] \\ &= \underbrace{\underbrace{\mathbb{E}\Big[\text{Var}[S|\boldsymbol{\Omega}]\Big]}_{\text{mutualisation}} + \underbrace{\mathbb{E}\Big[\text{Var}\Big[\mathbb{E}[S|\boldsymbol{\Omega}]\Big|\boldsymbol{X}\Big]\Big]}_{\text{solidarité}}}_{\rightarrow\text{assureur}} \\ &\quad + \underbrace{\text{Var}\Big[\mathbb{E}[S|\boldsymbol{X}]\Big]}_{\rightarrow\ \text{assuré}}.\end{aligned}$$

Segmentation et mutualisation
les deux faces d'une même pièce ?

Arthur Charpentier
Professeur à l'Université du Québec, Montréal

Michel Denuit
Professeur à l'Université catholique de Louvain

Romuald Elie
Professeur à l'Université de Marne-la-Vallée

L'assurance repose fondamentalement sur l'idée que la mutualisation des risques entre des assurés est possible. Cette mutualisation, qui peut être vue comme une relecture actuarielle de la loi des grands nombres, n'a de sens qu'au sein d'une population de risques « homogènes » [Charpentier, 2011]. Cette condition (actuarielle) impose aux assureurs de segmenter, ce que confirment plusieurs travaux économiques (1). Avec l'explosion du nombre de données, et donc de variables tarifaires possibles, certains assureurs évoquent l'idée d'un tarif individuel, semblant remettre en cause l'idée même de mutualisation des risques. Entre cette force qui pousse à segmenter et la force de rappel qui tend (pour des raisons sociales mais aussi actuarielles, ou au moins de robustesse statistique (2)) à imposer une solidarité minimale entre les assurés, quel équilibre va en résulter dans un contexte de forte concurrence entre les sociétés d'assurance ?

Tarification sans segmentation

Sans segmentation, le « prix juste » d'un risque est l'espérance mathématique de la charge annuelle. C'est l'idée du théorème fondamental de la valorisation actuarielle : en moyenne, la somme des primes doit permettre d'indemniser l'intégralité des sinistres survenus dans l'année. Afin d'illustrer les différents aspects de la construction du tarif et ses conséquences, on va utiliser les données présentées dans le tableau 1 (voir p. xx), qui indique la fréquence annuelle de sinistres.

Les facteurs de risque sont ici le lieu d'habitation et l'âge de l'assuré, et on observe la fréquence de sinistre par classe. Le coût unitaire, supposé fixe, équivaut à 1 000 euros. La prime pure est alors E [*S*] = 1 000 x E [*N*]. Dans cet exemple, la prime pure sans segmentation sera de 82,30 euros.

Modèle simple, $\boldsymbol{\Omega} = \{\boldsymbol{X}_1, \boldsymbol{X}_2\}$.

Quatre modèles

$$\begin{cases} \widehat{m}_0(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}[S] \\ \widehat{m}_1(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}[S|\boldsymbol{X}_1 = \boldsymbol{x}_1] \\ \widehat{m}_2(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}[S|\boldsymbol{X}_2 = \boldsymbol{x}_2] \\ \widehat{m}_{12}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}[S|\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2] \end{cases}$$

PolNum	CalYear	Gender	Type	Category	Occupation	Age	Group1	Bonus	Poldur	Value	Adind	SubGroup2	Group2	Density
200285786	2010	Male	E	Large	Employed	48	14	40	0	32345	1	O31	O	35.43401501
200285787	2010	Male	B	Medium	Employed	30	8	-30	9	8995	0	Q29	Q	239.4551701
200285788	2010	Female	B	Large	Housewife	47	2	-50	2	9145	1	U21	U	88.29014956
200285789	2010	Female	D	Large	Self-employed	48	13	-30	15	22075	1	R21	R	275.2822626
200285790	2010	Male	C	Medium	Housewife	57	12	-50	1	24985	1	Q5	Q	99.6400095
200285791	2010	Male	D	Medium	Self-employed	21	15	50	1	12100	1	R11	R	259.0040603
200285792	2010	Male	B	Small	Employed	44	5	-40	15	9820	1	Q10	Q	169.7885554
200285793	2010	Male	F	Small	Self-employed	37	17	-50	5	28680	1	Q5	Q	99.6400095
200285794	2010	Female	C	Large	Retired	49	3	20	4	28470	0	L94	L	84.22903844
200285795	2010	Female	A	Medium	Unemployed	35	5	20	5	8590	0	L112	L	66.06668352
200285796	2010	Male	E	Large	Self-employed	50	10	-30	3	20490	1	Q10	Q	169.7885554
200285797	2010	Female	B	Medium	Housewife	31	8	140	1	8385	1	P28	P	41.2451199
200285798	2010	Female	E	Medium	Self-employed	41	11	90	3	6410	1	L47	L	66.76541883
200285799	2010	Female	A	Medium	Housewife	44	10	-30	8	8485	0	P29	P	20.86448407
200285800	2010	Male	B	Large	Retired	69	8	-40	11	9380	1	U14	U	123.0152076
200285801	2010	Male	F	Medium	Housewife	45	11	30	0	19700	0	L40	L	76.05272599
200285802	2010	Male	E	Large	Retired	53	8	-30	6	10980	1	U19	U	61.79475865
200285803	2010	Male	C	Small	Employed	47	10	-10	9	21980	0	L96	L	45.66982293
200285804	2010	Female	D	Large	Retired	46	7	-50	1	28925	1	U12	U	54.93181221
200285805	2010	Female	C	Large	Retired	67	17	-50	9	14525	1	L52	L	73.25249905

Numtppd	Numtpbi	Indtppd	Indtpbi
0	1	0	1056.0334927
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
3	1	5800.0189068	16.507641942
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

$X_{1,i}$ $X_{k,i}$ Y_i

PolNum	CalYear	Gender	Type	Category	Occupation	Age	Group1	Bonus	Poldur	Value	Adind	SubGroup2	Group2	Density
200375666	2011	Female	A	Large	Employed	46	11	50	0	42975	0	L18	L	58.91132801
200375667	2011	Male	B	Large	Unemployed	31	8	80	11	14835	0	U8	U	125.1320458
200375668	2011	Female	D	Medium	Employed	27	7	-40	13	19000	1	R30	R	296.4319078
200375670	2011	Male	B	Small	Self-employed	22	7	-10	14	33305	0	Q33	Q	129.6690079
200375672	2011	Male	B	Small	Employed	21	17	-20	14	25995	0	T25	T	28.51184808
200375674	2011	Male	C	Medium	Employed	45	19	-50	0	8320	1	N21	N	71.18027901
200375675	2011	Male	C	Medium	Housewife	51	19	30	3	8445	0	L110	L	83.90453994
200375676	2011	Male	E	Large	Self-employed	49	16	-50	3	19545	0	L58	L	64.53563007
200375677	2011	Male	C	Small	Housewife	31	11	-20	5	5030	1	Q7	Q	83.76263662
200375678	2011	Female	A	Medium	Housewife	31	9	-50	14	15480	1	P7	P	25.62227499
200375679	2011	Male	B	Large	Housewife	69	13	-50	7	29580	0	Q23	Q	205.4307964
200375682	2011	Male	A	Medium	Self-employed	43	13	140	3	3735	0	U16	U	91.54176264
200375683	2011	Female	A	Medium	Self-employed	64	18	-20	6	13670	1	O35	O	21.45273029
200375685	2011	Male	E	Large	Employed	25	8	-10	6	17315	0	O22	O	32.18545326
200375688	2011	Male	B	Small	Retired	55	7	-40	3	19410	1	R49	R	208.8164363
200375689	2011	Female	F	Medium	Self-employed	54	9	-40	14	4165	0	U12	U	54.93181221
200375690	2011	Male	D	Large	Housewife	42	9	80	0	11970	1	L125	L	44.16537902
200375692	2011	Male	E	Large	Employed	36	12	-20	7	28415	0	L48	L	71.62174491
200375693	2011	Male	F	Medium	Self-employed	26	10	-30	6	4300	0	L97	L	63.82886936
200375694	2011	Female	B	Small	Unemployed	24	6	-40	7	24005	0	M17	M	201.6569069

Marché en concurrence

Règle de choix: les assurés choisissent la prime la moins chère,

	A	B	C	D	E	F
	787.93	706.97	1032.62	907.64	822.58	**603.83**
	**170.04**	197.81	285.99	212.71	177.87	265.13
	473.15	447.58	**343.64**	410.76	414.23	425.23
	337.98	**336.20**	468.45	339.33	383.55	672.91

Marché en concurrence

Règle de choix: les assurés choisissent (au hasard) parmi les trois primes les moins chères

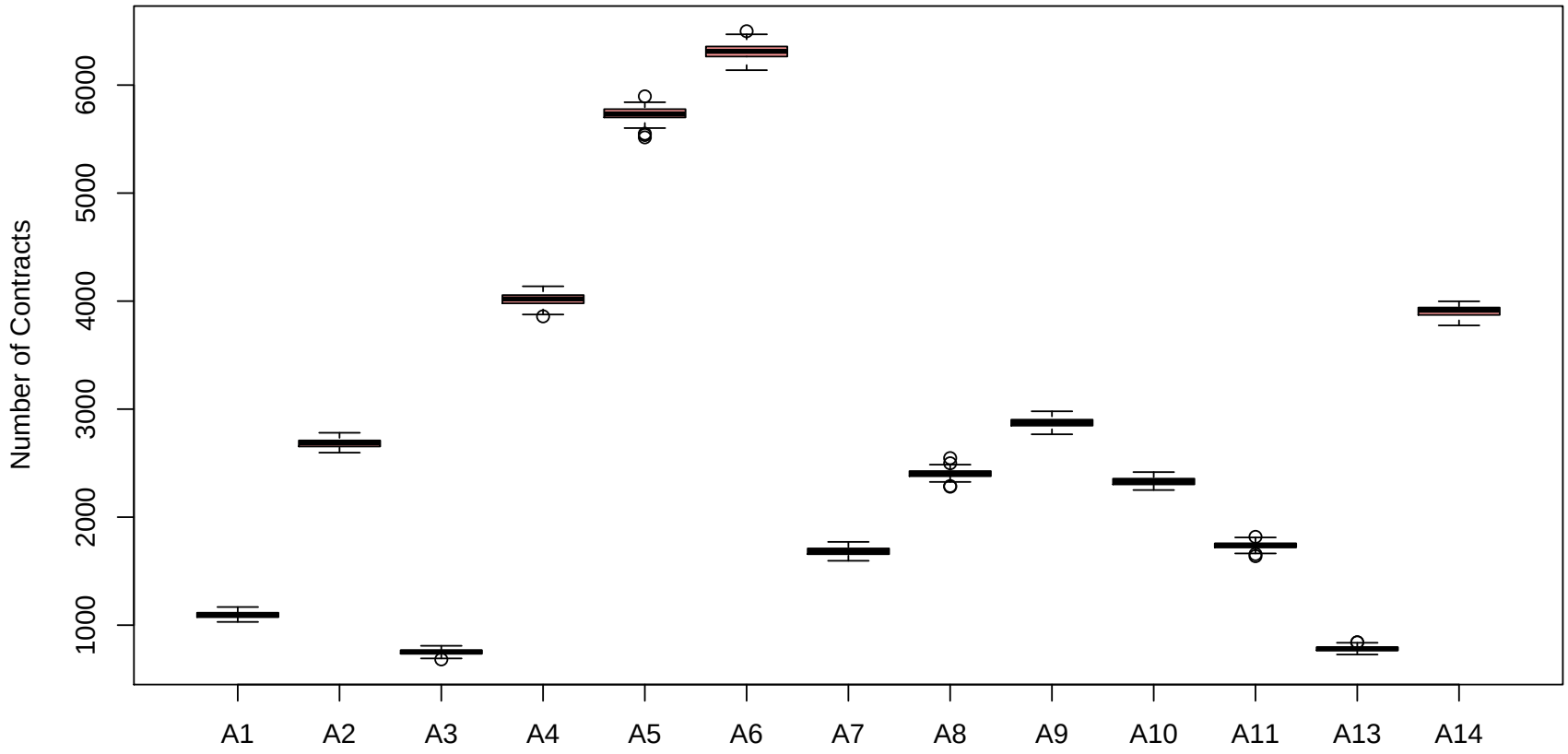
	A	B	C	D	E	F
	787.93	706.97	1032.62	907.64	822.58	603.83
	170.04	197.81	285.99	212.71	177.87	265.13
	473.15	447.58	343.64	410.76	414.23	425.23
	337.98	336.20	468.45	339.33	383.55	672.91

Marché en concurrence

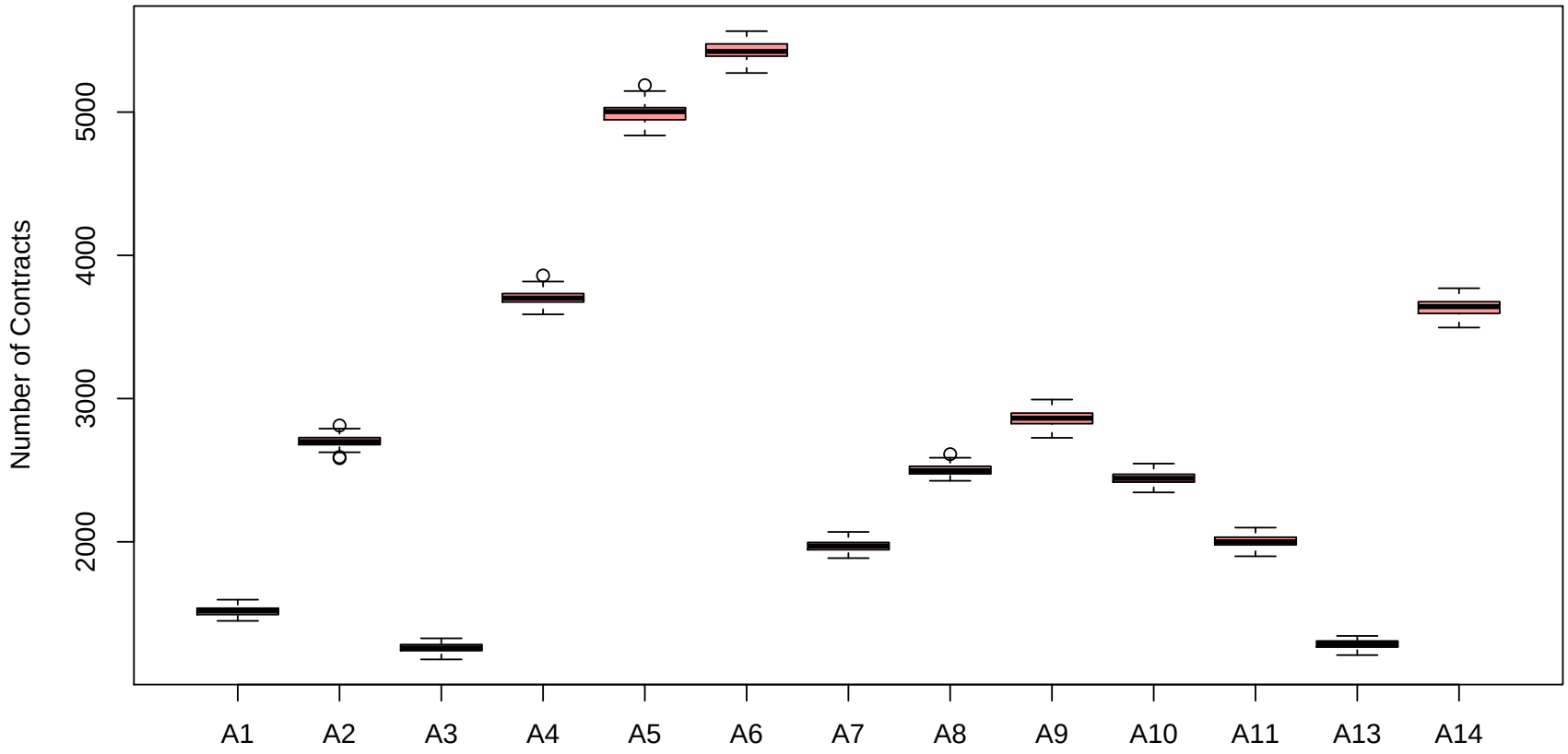
Règle de choix: les assurés se voient attribuer (au hasard) un assureur pour l'année $n-1$. L'année n, si leur assureur est parmi les 3 moins chers, il est retenu, sinon choix au hasard parmi 4.

	A	B	C	D	E	F
	787.93	706.97	1032.62	907.64	822.58	603.83
	170.04	197.81	285.99	212.71	**177.87**	265.13
	473.15	447.58	343.64	410.76	414.23	425.23
	**337.98**	336.20	468.45	339.33	383.55	672.91

Parts de marché (règle 2)

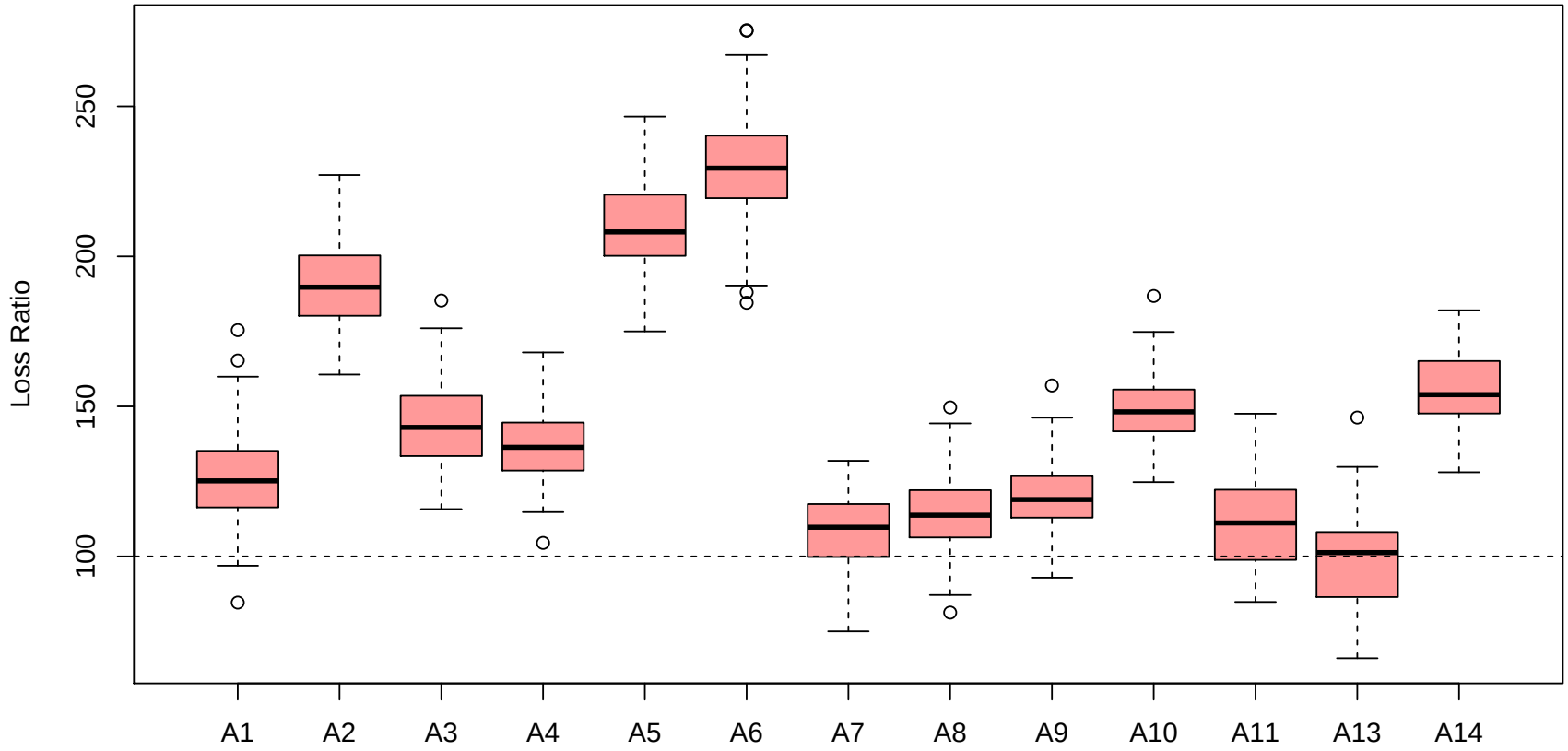


Parts de marché (règle 3)



Ratio Sinistres / Primes (règle 2)

Ratio du marché $\sim 154\%$.

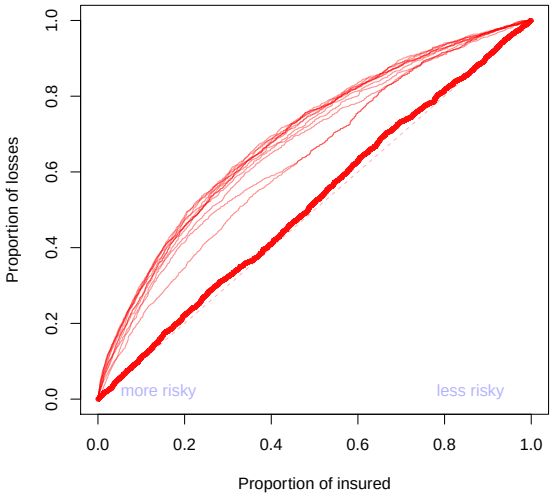


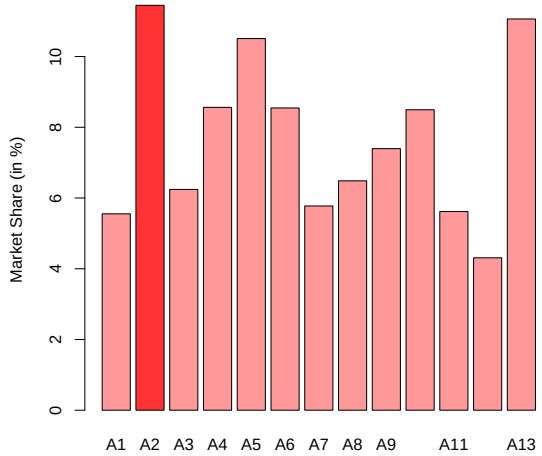
Assureur A2

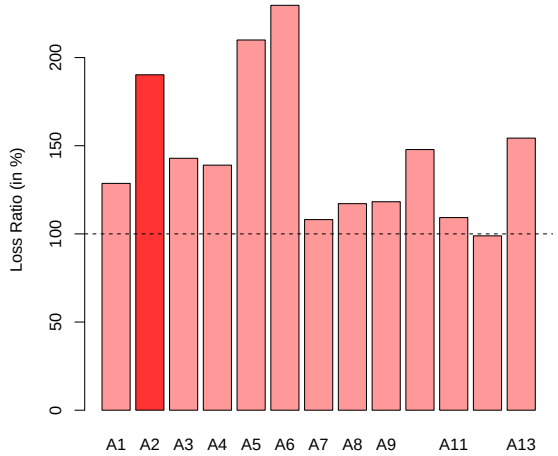
Pas de segmentation, prime unique

Remarque tous les prix ont été normalisés,

$$\pi_2 = m_2(\boldsymbol{x}_i) = \frac{1}{n}\sum_{i=1}^{n} m_j(\boldsymbol{x}_i)\ \forall j$$







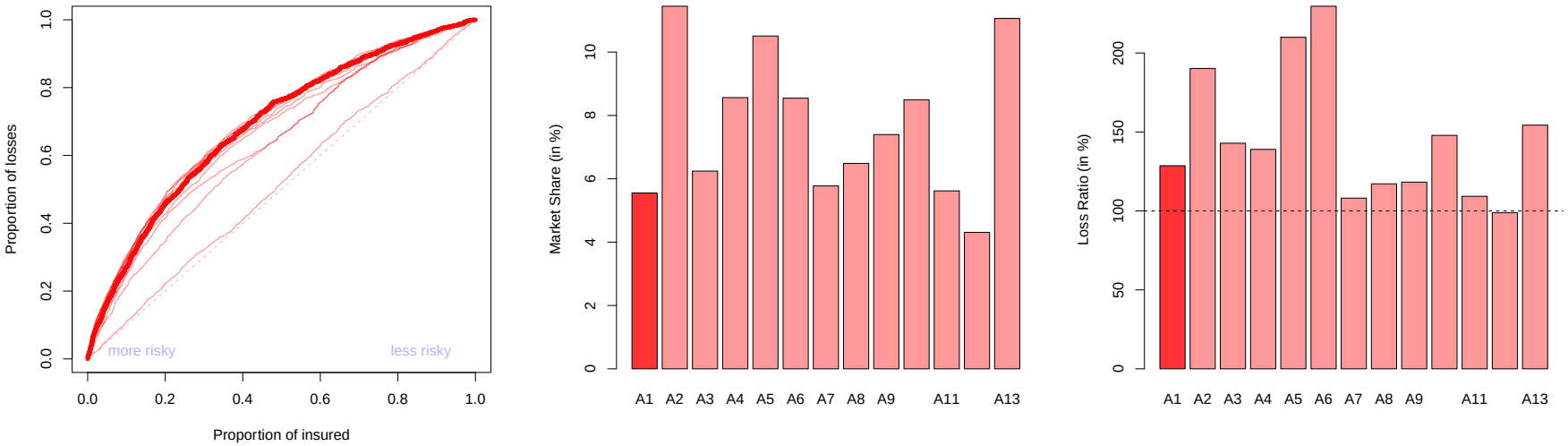
Assureur A1

Modèles GLM, fréquence RC matériel/corporel et coûts matériels

Age coupé en classe [18-30], [30-45], [45-60] et [60+], effets croisés avec occupation

Lissage manuel, SAS et Excel

Actuaire dans une mutuelle (en France)



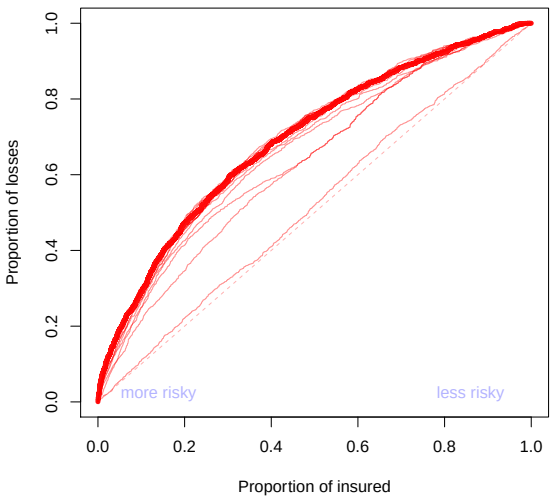
Assureur A8/A9

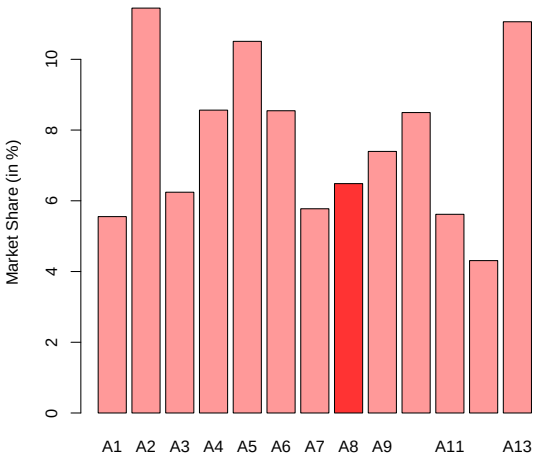
Modèles GLM, fréquence et coûts, suppression des graves (>15k)

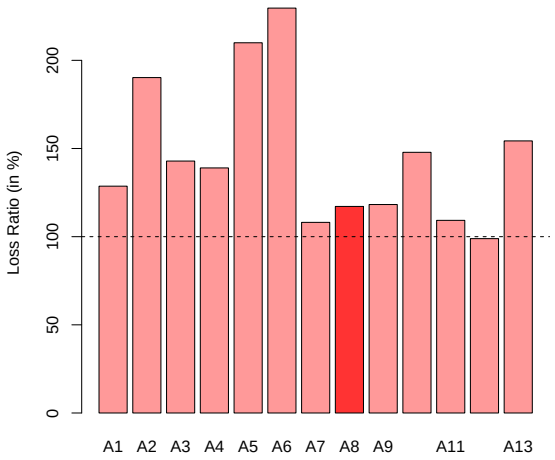
Interaction âge-genre

Utilisation d'un logiciel commercial de pricing (développé par Actuaris)

Actuaire dans une mutuelle (en France)





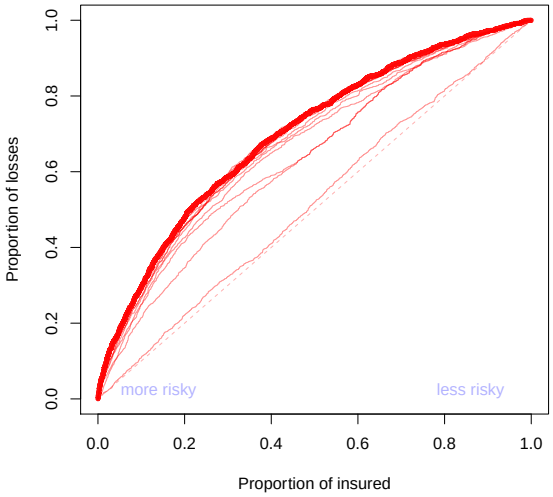


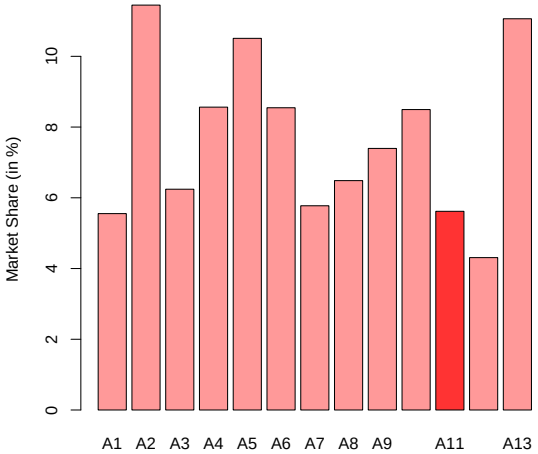
Assureur A11

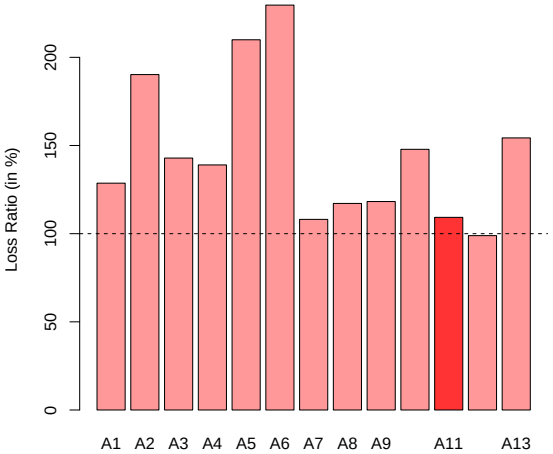
Toutes les variables sauf une, utilisation de deux modèles XGBoost (gradient boosting)

Correction pour primes négatives (plafonnées)

Programmé en Python par un actuaire dans une compagnie d'assurance (inscrit à la formation ADS).





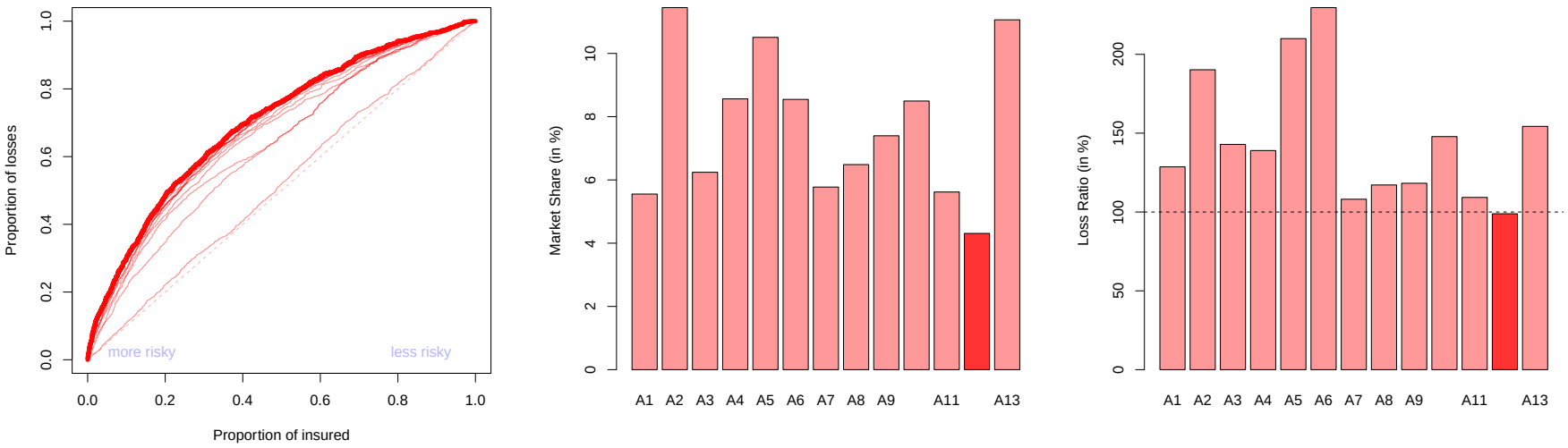


Assureur A12

Toutes les variables, utilisation de deux modèles XGBoost (gradient boosting)

Correction pour primes négatives (plafonnées)

Programmé en R par un actuaire dans une compagnie d'assurance en Europe.



Conclusion sur le Pricing Game