

Accuracy, équité, proxys, interprétabilité, causalité, modèles, données

Arthur Charpentier¹

¹ Université du Québec à Montréal

Journée IA et éthique, 2022

Agenda

- ▶ Accuracy (précision) et équité
- ▶ Interprétation et paradoxe de Simpson
- ▶ Interprétabilité
- ▶ Discrimination
- ▶ Causalité
- ▶ Interaction entre données et modèles

Accuracy : $\pi(\mathbf{x}) = \mathbb{E}_{\mathbb{P}}[Y|\mathbf{X} = \mathbf{x}]$ (\mathbb{P} probabilité historique) (is)

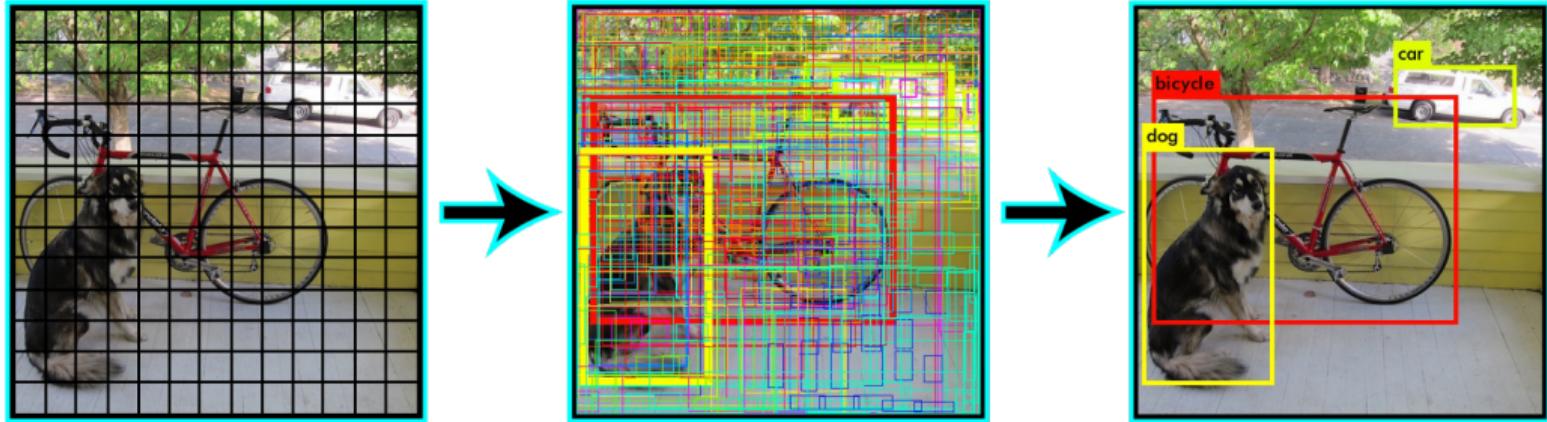
Équité : $\pi^*(\mathbf{x}) = \mathbb{E}_{\mathbb{P}^*}[Y|\mathbf{X} = \mathbf{x}]$ (\mathbb{P}^* probabilité souhaitable) (ought, Hume (1739))

Interprétabilité I



Prédire les futurs Prix Nobels (sur la base d'historique passé) ?
Idée de [Lions \(2020\)](#), photo [Le Monde \(2021\)](#)

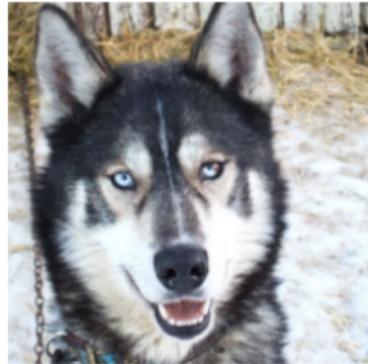
Interprétabilité II



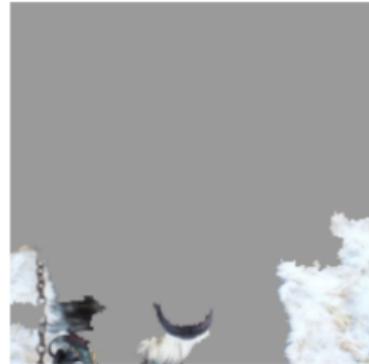
via **yolo** pour la détection d'objets (et d'animaux)

Interprétabilité III

“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.

Interprétabilité IV

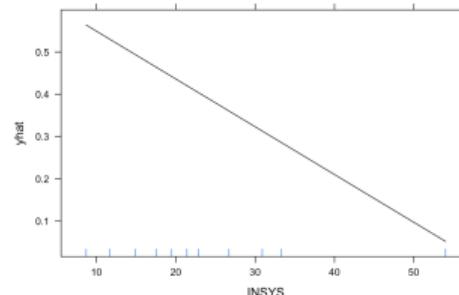
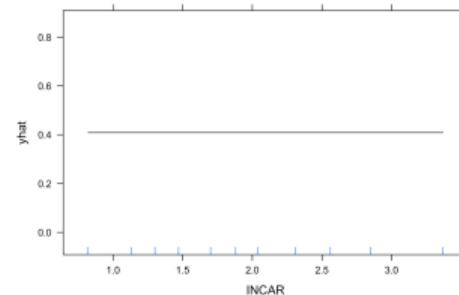
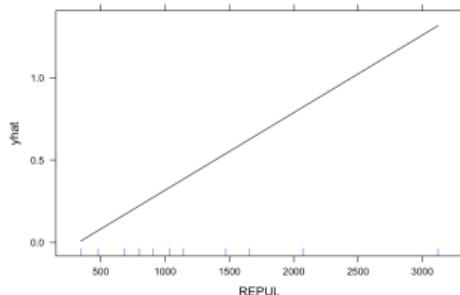
► Partial Dependence Plot

Introduit par Friedman (2001). On écrit \mathbf{x} en deux composantes : \mathbf{x}_s (variable(s) d'intérêt) et \mathbf{x}_c les autres variables, $\mathbf{x} = (\mathbf{x}_s, \mathbf{x}_c)$.

$$p(\mathbf{x}_s) = \mathbb{E}[m(\mathbf{x}_s, \mathbf{x}_c)] \text{ et } \hat{p}(\mathbf{x}_s) = \frac{1}{n} \sum_{i=1}^n \hat{m}(\mathbf{x}_s, \mathbf{x}_{i,c})$$

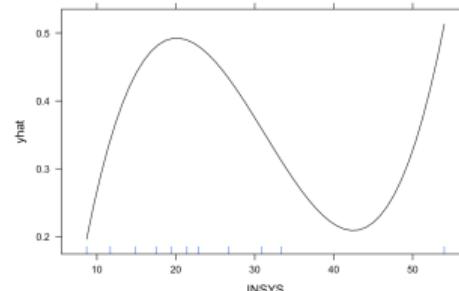
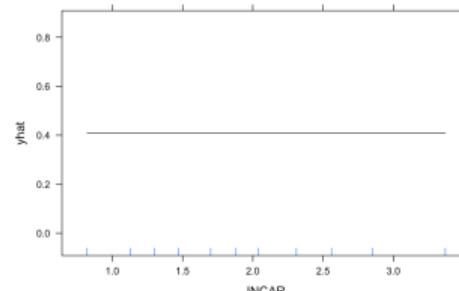
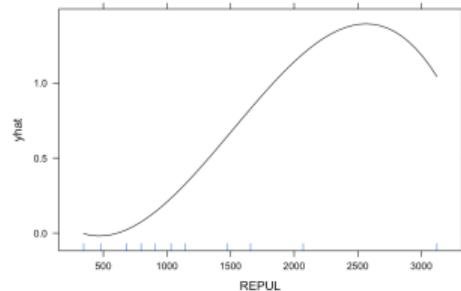
Par exemple pour un modèle linéaire ([myocarde](#) dataset),

$$\hat{m}(\mathbf{x}_s, \mathbf{x}_c) = \hat{\beta} + \mathbf{x}_s^\top \hat{\boldsymbol{\beta}}_S + \mathbf{x}_c^\top \hat{\boldsymbol{\beta}}_C$$

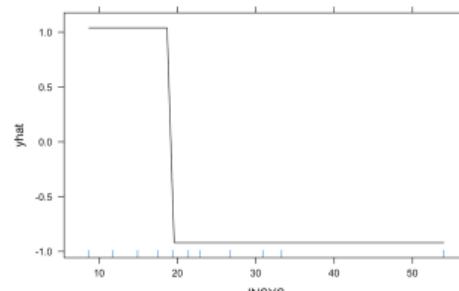
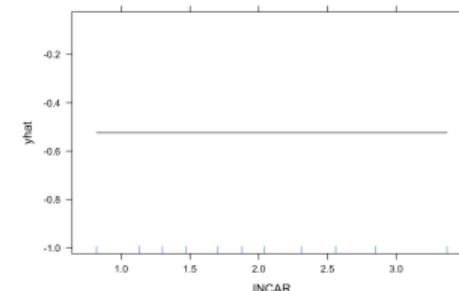
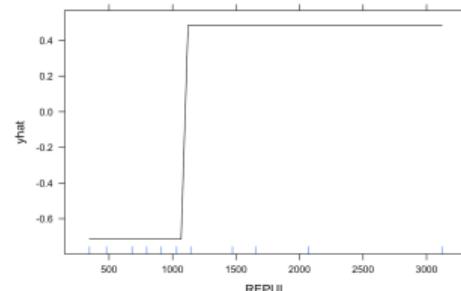


Interprétabilité V

Pour un modèle additif (GAM)

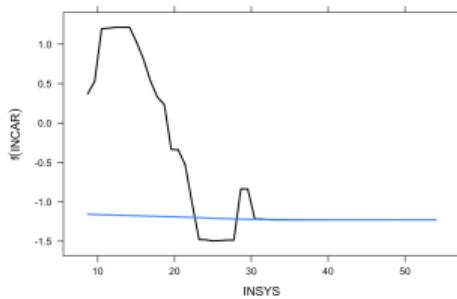
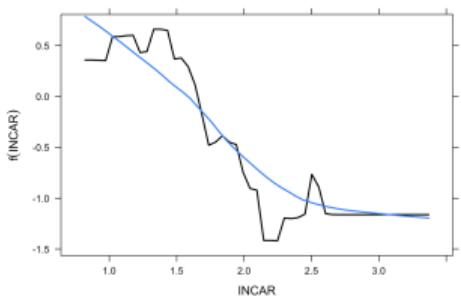
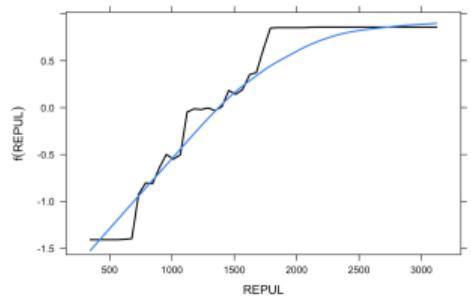
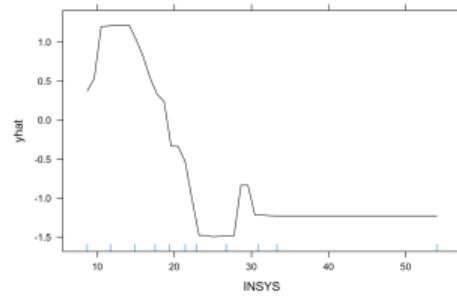
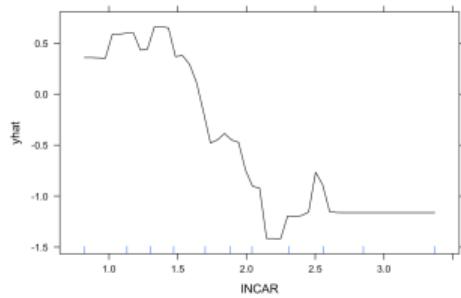
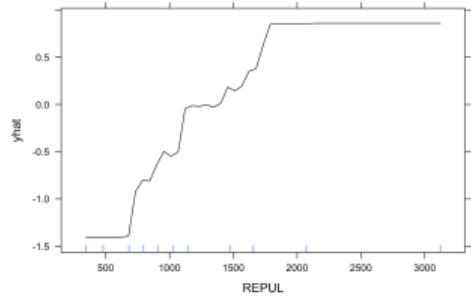


ou pour un arbre de classification



Interprétabilité VI

Pour une forêt aléatoire

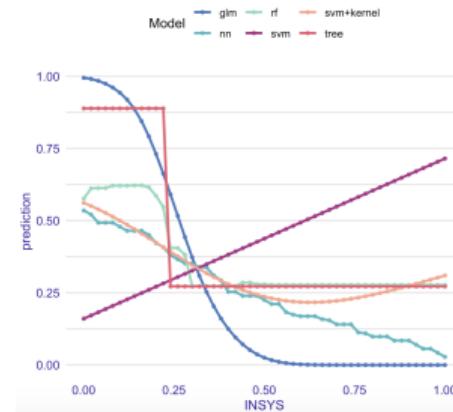
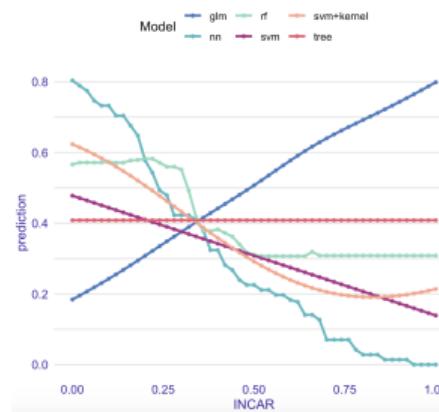
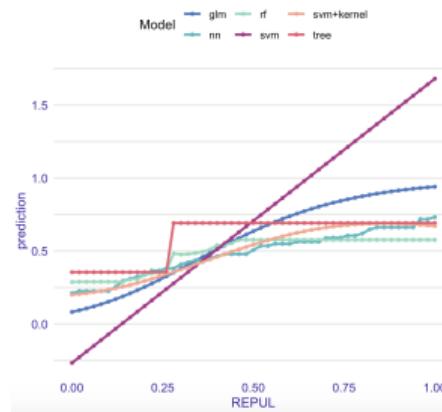


Interprétabilité VII

► Accumulated Local Effects

Introduit par Apley and Zhu (2020). Le Partial Dependence Plot de x_s est

$$p(\mathbf{x}_s) = \mathbb{E}[m(\mathbf{x}_s, \mathbf{S}_c)] \text{ et } \hat{p}(\mathbf{x}_s) = \frac{1}{n} \sum_{i=1}^n \hat{m}(\mathbf{x}_s, \mathbf{x}_{i,c})$$

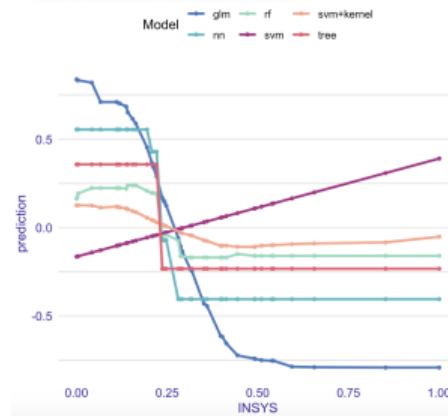
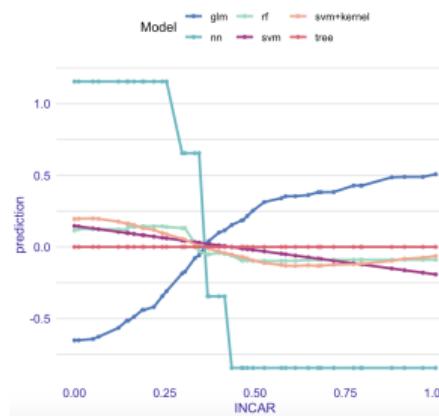
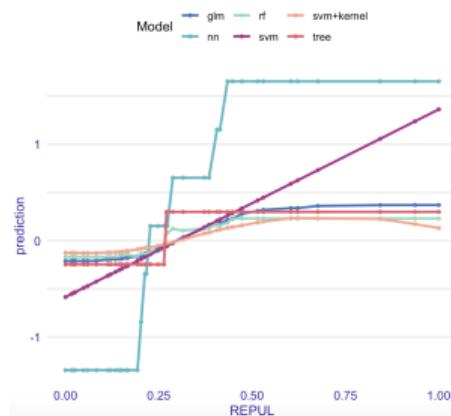


Interprétabilité VIII

$$p(\mathbf{x}_s) = \mathbb{E}[m(\mathbf{x}_s, \mathbf{S}_c)] \text{ et } \hat{p}(\mathbf{x}_s) = \frac{1}{n} \sum_{i=1}^n \hat{m}(\mathbf{x}_s, \mathbf{x}_{i,c})$$

Ici, on considère

$$a(\mathbf{x}_s) = \int_{-\infty}^{\mathbf{x}_s} \mathbb{E} \left[\frac{\partial m(\mathbf{z}_s, \mathbf{S}_c)}{\partial \mathbf{x}_s} \right] d\mathbf{z}_s$$



Paradoxe de Simpson I

La sous-identification correspond au cas où le vrai modèle serait

$y_i = \beta_0 + \mathbf{x}_1^\top \boldsymbol{\beta}_1 + \mathbf{x}_2^\top \boldsymbol{\beta}_2 + \varepsilon_i$, mais le modèle estimé est $y_i = b_0 + \mathbf{x}_1^\top \mathbf{b}_1 + \eta_i$ (autrement dit, les variables \mathbf{x}_2 ne sont pas utilisées dans la régression). L'estimateur du maximum de vraisemblance de \mathbf{b}_1 est (avec l'écriture matricielle classique en économétrie, comme [Davidson et al. \(2004\)](#) ou [Charpentier et al. \(2018\)](#))

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

de telle sorte que $\mathbb{E}[\hat{\mathbf{b}}_1] = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{12}$, le biais (ce que nous avons noté $\boldsymbol{\beta}_{12}$) étant nul uniquement dans le cas où $\mathbf{X}_1^\top \mathbf{X}_2 = \mathbf{0}$ (c'est à dire $\mathbf{X}_1 \perp \mathbf{X}_2$). Si on simplifie un peu,

Paradoxe de Simpson II

supposons que le véritable modèle sous-jacent des données

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

où x_1 et x_2 désignent des variables explicatives, y est la variable cible, et ε est un bruit aléatoire. Le modèle estimé en enlevant x_2 donne $\hat{y} = \hat{b}_0 + \hat{b}_1 x_1$. On peut penser à une variable importante manquante x_2 , ou au cas où x_2 est une variable protégée. Les estimations des coefficients de régression obtenus par moindre carrés sont (généralement) biaisées, dans le sens où

$$\hat{b}_1 = \frac{\widehat{\text{cov}}[x_1, y]}{\widehat{\text{Var}}[x_1]} = \frac{\widehat{\text{cov}}[x_1, \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon]}{\widehat{\text{Var}}[x_1]}$$

soit

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

Paradoxe de Simpson III

Admissions d'étudiant(e) gradué(e)s à U.C. Berkeley, [Bickel et al. \(1975\)](#),

	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%

Paradoxe de Simpson IV

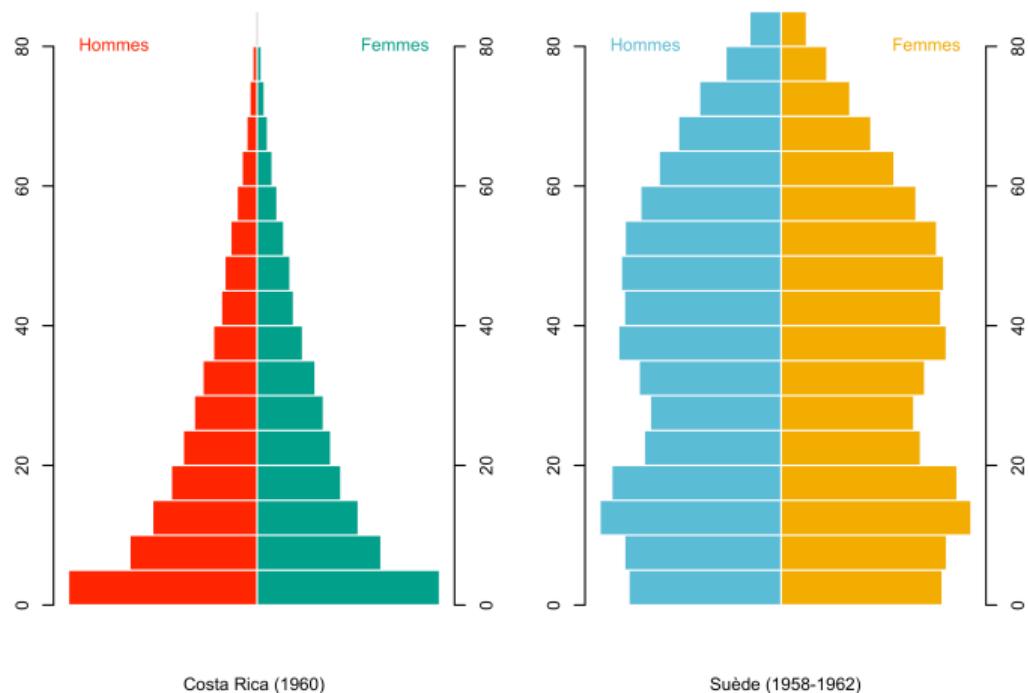
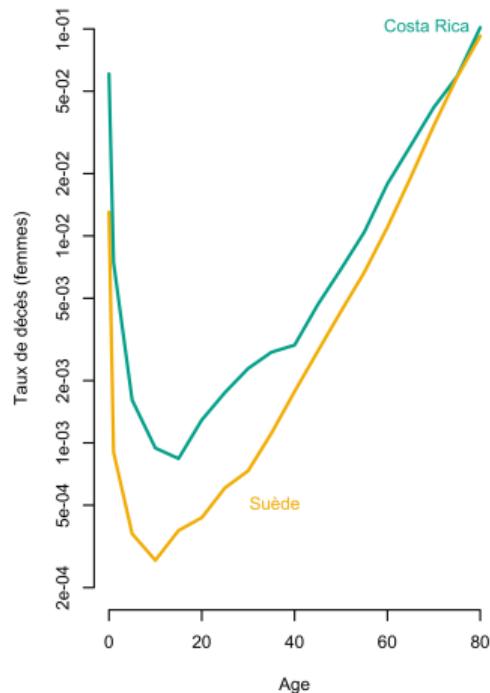
On peut aussi considérer la survie parmi les passagers du Titanic

	Total	Femmes	Hommes
troisième classe membre d'équipage	181/709 ~ 25.5% 211/890 ~ 23.7%	106/216 ~ 49.1% 20/ 23 ~ 86.9%	75/493 ~ 15.2% 191/867 ~ 22.0%

Mathématiquement, il n'y a pas vraiment de paradoxe, au sens où

$$\frac{a_1}{c_1} < \frac{a_2}{c_2} \text{ et } \frac{b_1}{d_1} < \frac{b_2}{d_2} \Leftrightarrow \frac{a_1 + b_1}{c_1 + d_1} < \frac{a_2 + b_2}{c_2 + d_2}$$

Paradoxe de Simpson V



Paradoxe de Simpson VI

Extension du paradoxe de Simpson, via Morris (2021)

	Total	jeunes	personnes âgées
vaccinés	301/5634634~ 0.053‰	11/3501118 ~ 0.003‰	290/2133516 ~ 0.136‰
non vaccinés	214/1302912~ 0.164‰	43/1116834 ~ 0.039‰	171/186078 ~ 0.919‰
efficacité	3.1×	13.0×	6.7×

Corriger une discrimination I

Lindholm et al. (2020) Statistiques de sinistralité basées sur deux variables, une variable protégée (notée p), le genre, et une variable autorisée (notée x) indiquant si la personne est fumeuse, ou pas.

nombre n	femmes	hommes	total	exposition e	femmes	hommes	total
fumeur	32	4	36	fumeur	133	24	157
non-fumeur	28	48	76	non-fumeur	131	301	432
total	60	52	112	total	264	325	589

La fréquence annuelle de sinistre Y , par année

$$\mathbb{E}[Y] = \frac{n}{e} = \frac{112}{589} \approx 19.0\%,$$

Corriger une discrimination II

et si on segmente suivant le tabagisme

$$\mathbb{E}[Y|X = \text{fumeur}] = \frac{36}{157} \approx 22.9\% \text{ et } \mathbb{E}[Y|X = \text{non-fumeur}] = \frac{76}{432} \approx 17.6\%$$

Notons que l'on retrouve la formule des espérances totales évoquée auparavant,

$$\mathbb{E}[Y] = \mathbb{E}[Y|X = \text{fumeur}] \cdot \mathbb{P}[X = \text{fumeur}] + \mathbb{E}[Y|X = \text{non-fumeur}] \cdot \mathbb{P}[X = \text{non-fumeur}].$$

On le voit sur cet exemple, le modèle tarifaire basé sur le tabagisme (et pas le genre) n'est pas indépendant du genre pour autant. En effet, si la prime est proportionnelle à $\mathbb{E}[Y|X = x]$, la prime moyenne des hommes et des femmes sont proportionnelles à

$$\begin{cases} \text{femmes} & : \frac{133}{264} \cdot \mathbb{E}[Y|X = \text{fumeur}] + \frac{131}{264} \cdot \mathbb{E}[Y|X = \text{non-fumeur}] \approx 20.3\% \\ \text{hommes} & : \frac{24}{325} \cdot \mathbb{E}[Y|X = \text{fumeur}] + \frac{301}{325} \cdot \mathbb{E}[Y|X = \text{non-fumeur}] \approx 18.0\%. \end{cases}$$

Corriger une discrimination III

Autrement dit, la prime des hommes est, en moyenne, plus élevée que celle des femmes. [Lindholm et al. \(2020\)](#) propose une méthode relativement simple pour proposer une prime dite “*discrimination-free*”, en proposant d’utiliser

$$\text{prime}(x) = \sum_p \mathbb{E}[Y|X = x, P = p] \cdot \mathbb{P}[P = p]$$

soit ici

$$\begin{cases} \text{prime(fumeur)} = \frac{32}{133} \cdot \frac{264}{584} + \frac{4}{24} \cdot \frac{325}{584} \approx 20.0\% & (< \mathbb{E}[Y|X = \text{fumeur}] \approx 22.9\%) \\ \text{prime(non-fumeur)} = \frac{28}{131} \cdot \frac{264}{584} + \frac{48}{301} \cdot \frac{325}{584} \approx 18.4\% & (> \mathbb{E}[Y|X = \text{non-fumeur}] \approx 17.6\%) \end{cases}$$

Cette technique est à rapprocher des notions de “*partial dependence plot*”, introduites par [Friedman \(2001\)](#), et utilisées pour interpréter et expliquer des modèles “boîte

Corriger une discrimination IV

noire", en apprentissage automatique. Notons que dans ce cas, on va globalement sous-tarifer

$$\left\{ \begin{array}{l} \text{total : } \frac{157}{589} \cdot \text{prime(fumeur)} + \frac{432}{589} \cdot \text{prime(non-fumeur)} \approx 18.8\% \quad (< \mathbb{E}[Y] \approx 19.0\%) \\ \text{femmes : } \frac{133}{264} \cdot \text{prime(fumeur)} + \frac{131}{264} \cdot \text{prime(non-fumeur)} \approx 19.2\% \\ \text{hommes : } \frac{24}{325} \cdot \text{prime(fumeur)} + \frac{301}{325} \cdot \text{prime(non-fumeur)} \approx 18.4\%. \end{array} \right.$$

Corriger une discrimination V

Plus généralement $\mu(\mathbf{x}, p) = \mathbb{E}[Y | \mathbf{X} = \mathbf{x}, P = p]$ est le meilleur estimateur

$\pi_u(\mathbf{x}) = \mathbb{E}[Y | \mathbf{X} = \mathbf{x}]$ is the **prix par ignorance** au sens de Kusner et al. (2017)

$$\pi_u(\mathbf{x}) = \int \mu(\mathbf{x}), p d\mathbb{P}(p|\mathbf{x})$$

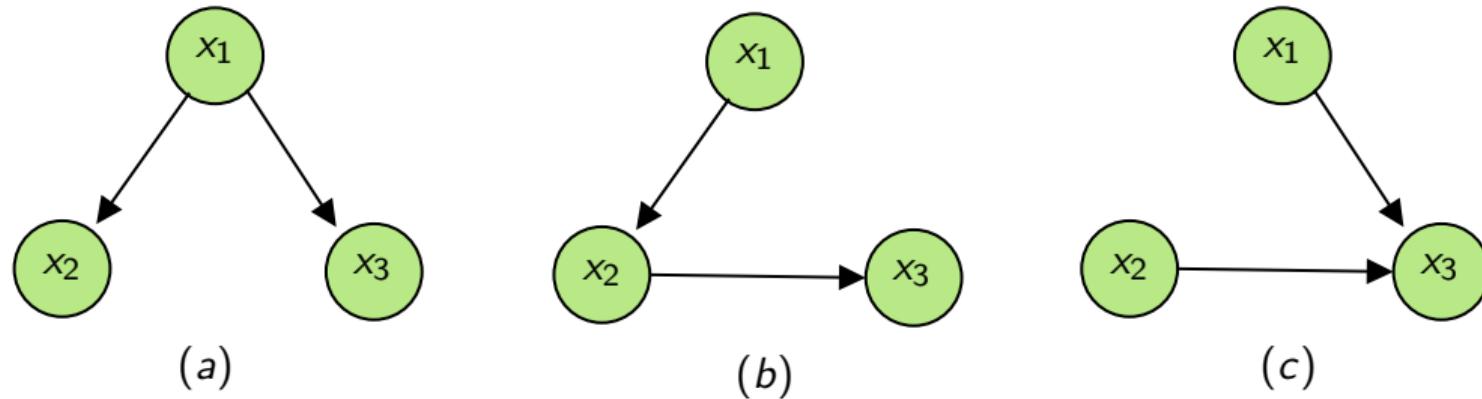
Notons que si $P \perp\!\!\!\perp \mathbf{X}$, $d\mathbb{P}(p|\mathbf{x}) = d\mathbb{P}(p)$

$\pi^*(\mathbf{x}) = \int \mu(\mathbf{x}, p) d\mathbb{P}^*(p)$ un prix *discrimination-free* (selon Lindholm et al. (2020))

"The discrimination-free price is obtained by averaging best-estimate prices over discriminatory covariates, using a (potentially arbitrary) marginal distribution"

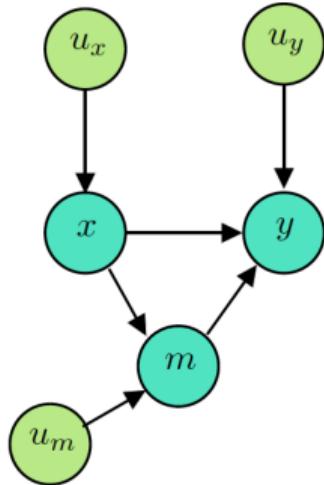
Notons que $\mathbb{E}[Y] = \mathbb{E}[\mu(\mathbf{X}, P)] = \mathbb{E}[\pi_u(\mathbf{X})] \neq \mathbb{E}[\pi^*(\mathbf{X})]$: le prix *discrimination-free* est biaisé.

Causalité et narration I

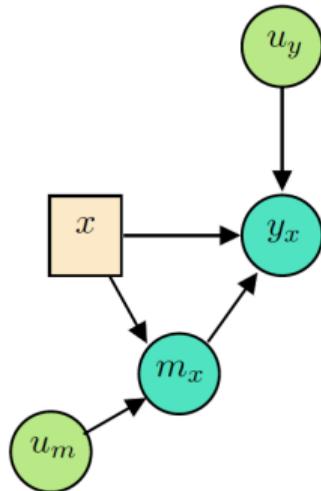


Quelques exemples de graphes dirigés, avec 3 nœuds, et 2 connections. (a) correspond au cas où x_1 est un **facteur de confusion** pour x_2 et x_3 , correspondant à un choc commun ou de dépendance mutuelle, (b) correspond au cas où x_2 est un **médiateur** pour x_1 et x_3 , et (c) correspond au cas où x_3 est un **colusionneur** pour x_1 et x_2 , correspondant à un cas de cause mutuelle.

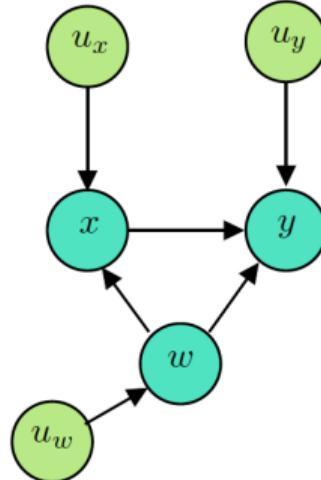
Causalité et narration II



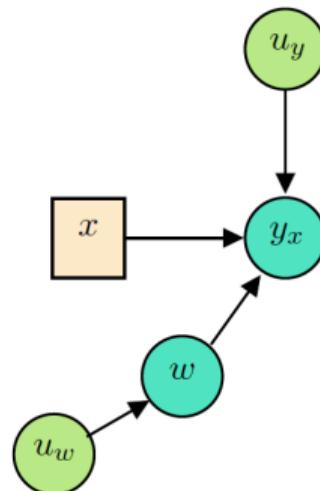
(a)



(b)



(c)



(d)

Diagrammes causaux, $x \rightarrow y$, avec un médiateur m à gauche, et avec un facteur de confusion w à droite.

Causalité et narration III

en présence d'un médiateur (m)

monde réel	avec intervention ($do(x)$)
$\begin{cases} X = f_x(U_x) \\ M = f_m(X, U_m) \\ Y = f_y(X, M, U_y) \end{cases}$	$\begin{cases} X = x \\ M_x = f_m(x, U_m) \\ Y_x = f_y(x, M_x, U_y) \end{cases}$

en présence d'un facteur de confusion (w)

monde réel	avec intervention ($do(x)$)
$\begin{cases} X = f_x(W, U_x) \\ W = f_w(U_w) \\ Y = f_y(X, W, U_y) \end{cases}$	$\begin{cases} X = x \\ W = f_w(U_w) \\ Y_x = f_y(x, W, U_y) \end{cases}$

$$\begin{cases} \text{médiateur : } & \mathbb{P}[Y_x = 1] = \mathbb{P}[Y = 1 | do(X = x)] = \mathbb{P}[Y = 1 | X = x] \\ \text{confusion : } & \mathbb{P}[Y_x = 1] = \mathbb{P}[Y = 1 | do(X = x)] \neq \mathbb{P}[Y = 1 | X = x] \end{cases}$$

Causalité et narration IV

On estime ainsi $\mathbb{E}[Y_{T \leftarrow 1}^*] - \mathbb{E}[Y_{T \leftarrow 0}^*]$ par (matching)

$$ACE = \bar{y}_{T \leftarrow 1}^* - \bar{y}_{T \leftarrow 0}^* = \frac{1}{n_1} \sum_{i:t_i=1} y_i - y_{j_i^*}, \text{ où } j_i^* = \operatorname{argmin}_{j:t_j=0} \{d(\mathbf{x}_i, \mathbf{x}_j)\},$$

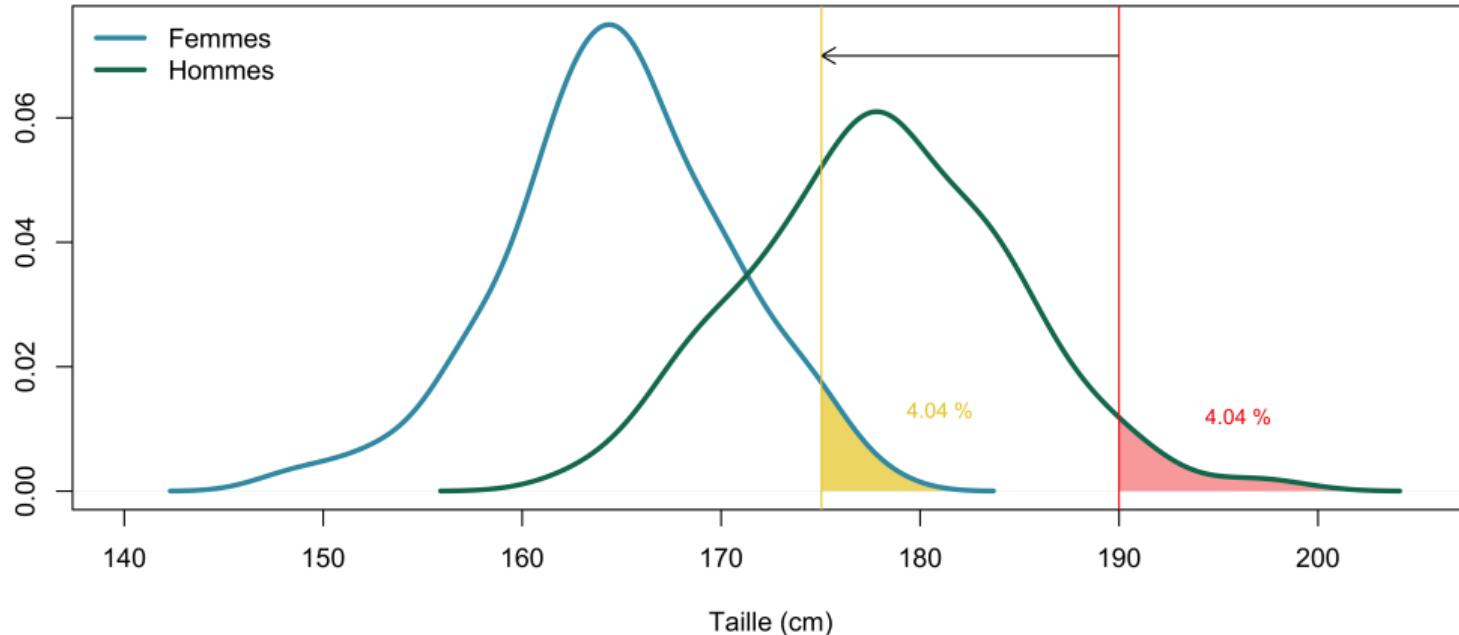
et où n_1 est le nombre d'observations traitées.

Ayant observé (\mathbf{x}, p, y) on espère créer un contrefactuel en considérant (\mathbf{x}, p^*, y^*) , où $p^* = 1 - p$

Mais si p change, il y a des chances pour que \mathbf{x} change aussi, comme le rappellent [Gordaliza et al. \(2019\)](#), [Black et al. \(2020\)](#), [Torous et al. \(2021\)](#) ou [de Lara et al. \(2021\)](#). Formellement, on dit simplement que les distributions de \mathbf{x} , conditionnellement à $p = 0$ ou $p = 1$, ne sont pas identiques, ce qui arrivera forcément en présence d'un proxy de p parmi les variables explicatives \mathbf{x} .

Cf \mathbf{x} soit la taille d'une personne et p son genre. Si on observe un homme de 190cm, le contrefactuel n'est a priori pas une femme de 190cm.

Causalité et narration V



Causalité et narration VI

L'extension à un ensemble de variables $\mathbf{x} = (x_1, \dots, x_k)$ se fait en utilisant un algorithme de transport optimal (voir [Villani \(2009\)](#) ou [Galichon \(2016\)](#) par exemple).

L'idée est de noter que dans les deux groupes, $p = 0$ et $p = 1$, les caractéristiques \mathbf{x} ont (potentiellement) deux distributions différentes, \mathbb{P}_0 (ou \mathbb{P}) et \mathbb{P}_1 (ou \mathbb{Q}).

Formellement, si \mathbb{P} est une distribution sur \mathbb{R}^k , étant donné $T : \mathbb{R}^k \rightarrow \mathbb{R}^k$ et on définit la mesure “*push-forward*” associée, \mathbb{Q} , définie par

$$\mathbb{Q}(A) = T_{\#}\mathbb{P}(A) = \mathbb{P}(T^{-1}(A)), \quad \forall A \subset \mathbb{R}^k.$$

Un transport optimal T^* (au sens de Brenier) de \mathbb{P} vers \mathbb{Q} sera solution du problème de Monge

$$T^* \in \operatorname{arginf}_{T : T_{\#}\mathbb{P} = \mathbb{Q}} \int_{\mathbb{R}^k} \|\mathbf{x} - T(\mathbf{x})\|^2 d\mathbb{P}(\mathbf{x}).$$

Il est possible de montrer (voir [Villani \(2009\)](#) ou [Galichon \(2016\)](#)) que $T^* = \nabla\psi$ où ψ est une fonction convexe. Si $k = 1$ (comme dans notre exemple précédent avec la

Causalité et narration VII

taille) T est alors une fonction croissante (en tant que dérivée d'une fonction convexe), et si $F_0(x) = \mathbb{P}_0[X \leq x]$ et $F_1(x) = \mathbb{P}_1[X \leq x]$, alors $T^*(x) = F_1^{-1} \circ F_0(x)$ vérifie $T_\#^*\mathbb{P}_0 = \mathbb{P}_1$ (car $F_1(x) = F_0(T^{*-1}(x))$) et T^* sera optimal au sens de Brenier. m sera non-équitable ($m(\mathbf{x}, p) = \mathbf{1}(s(\mathbf{x}, p) > \text{seuil})$) si $m(\mathbf{x}, 0) \neq m(T^*(\mathbf{x}), 1)$. Black et al. (2020) définit ainsi l'ensemble “*FlipSet*” comme

$$\mathcal{X}_F(m, T^*) = \{\mathbf{x} \in \mathcal{X} : m(\mathbf{x}, 0) \neq m(T^*(\mathbf{x}), 1)\}.$$

$$\begin{cases} \mathcal{X}_F^+(m, T^*) = \{\mathbf{x} \in \mathcal{X} : m(\mathbf{x}, 0) > m(T^*(\mathbf{x}), 1)\}. \\ \mathcal{X}_F^-(m, T^*) = \{\mathbf{x} \in \mathcal{X} : m(\mathbf{x}, 0) < m(T^*(\mathbf{x}), 1)\}. \end{cases}$$

et si $T_\#^*\mathbb{P}_0 = \mathbb{P}_1$, alors l'effet causal moyen vaut

$$ACE = \mathbb{E}[Y_{T \leftarrow 1}^*] - \mathbb{E}[Y_{T \leftarrow 0}^*] = \mathbb{P}_0[\mathcal{X}_F^-(m, T^*)] - \mathbb{P}_0[\mathcal{X}_F^+(m, T^*)]$$

Données et modèles (L'enfer, c'est les autres...) I

Deux sous-populations

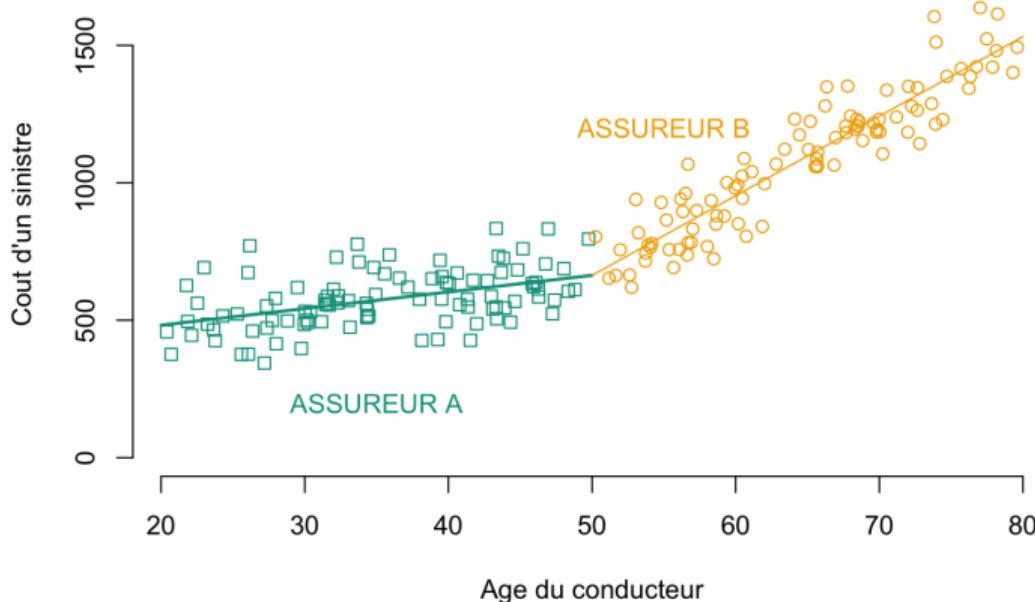
{□, ○}

Deux assureurs

{□ → assureur A
○ → assureur B}

Modèles linéaires

{ $\pi_A(x) = \alpha_0 + \alpha_1 x$
 $\pi_B(x) = \beta_0 + \beta_1 x$

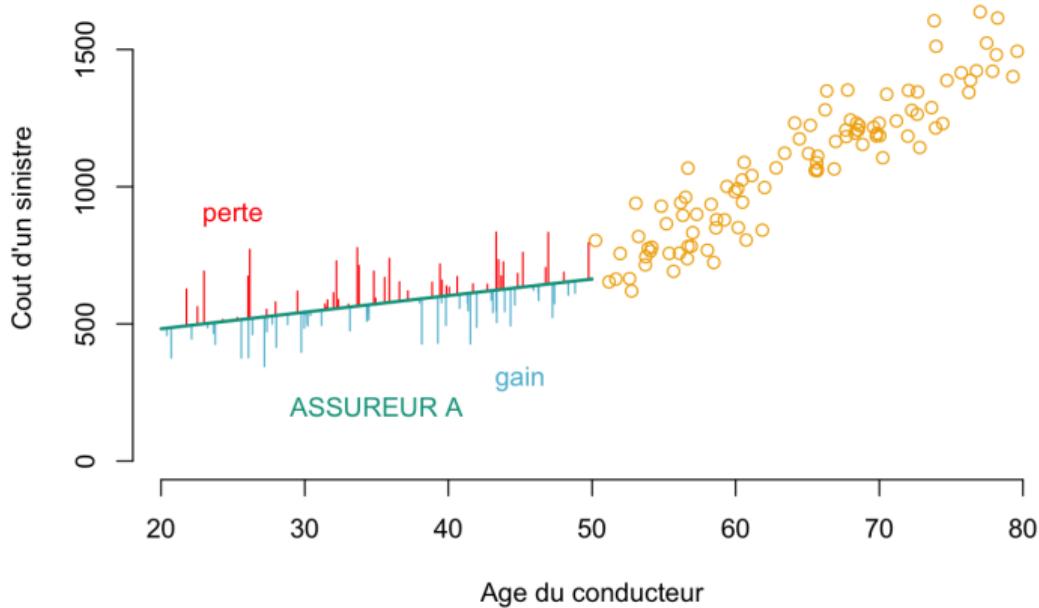


Données et modèles (L'enfer, c'est les autres...) II

Chaque assureur est à l'équilibre, en moyenne

$$\sum_{i: \square} y_i \approx \sum_{i: \square} \pi_A(x_i)$$

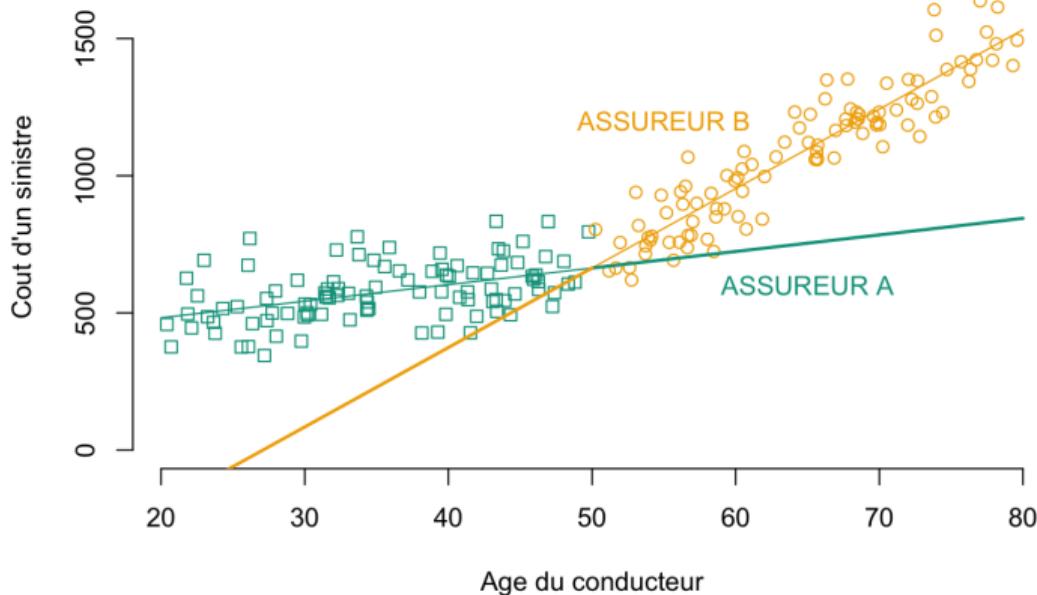
$$\sum_{i: \circ} y_i \approx \sum_{i: \circ} \pi_B(x_i)$$



Données et modèles (L'enfer, c'est les autres...) III

Que se passe-t-il si on met les assureurs en concurrence ?

- { \square : $\pi_A(x_i) > \pi_B(x_i)$
- { \circ : $\pi_A(x_i) < \pi_B(x_i)$
- { \square → assureur B
- { \circ → assureur A

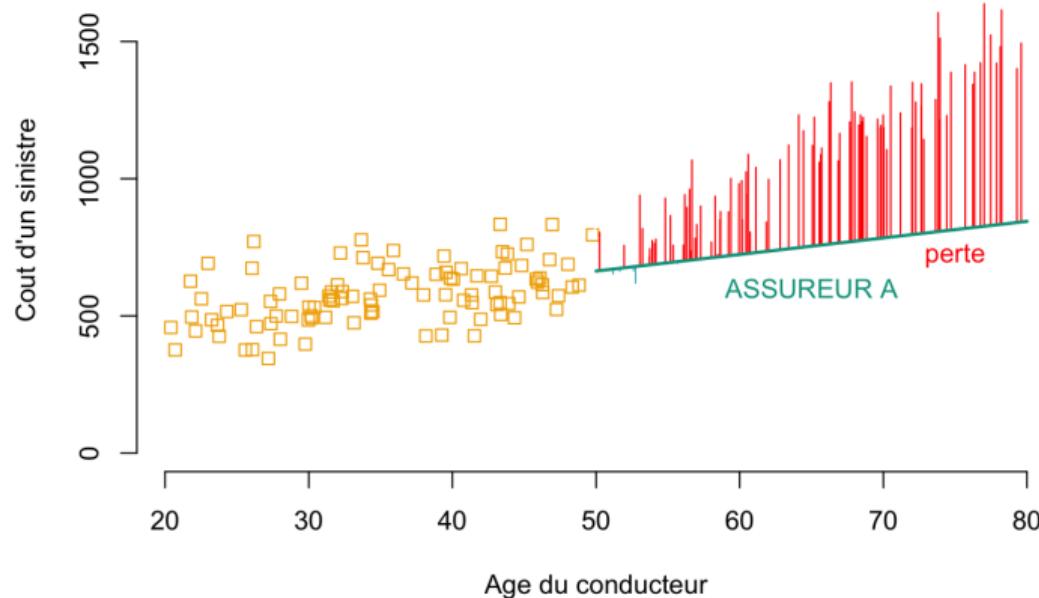


Données et modèles (L'enfer, c'est les autres...) IV

Chaque assureur
perd de l'argent

$$\sum_{i: \square} y_i > \sum_{i: \square} \pi_B(x_i)$$

$$\sum_{i: \circ} y_i > \sum_{i: \circ} \pi_A(x_i)$$

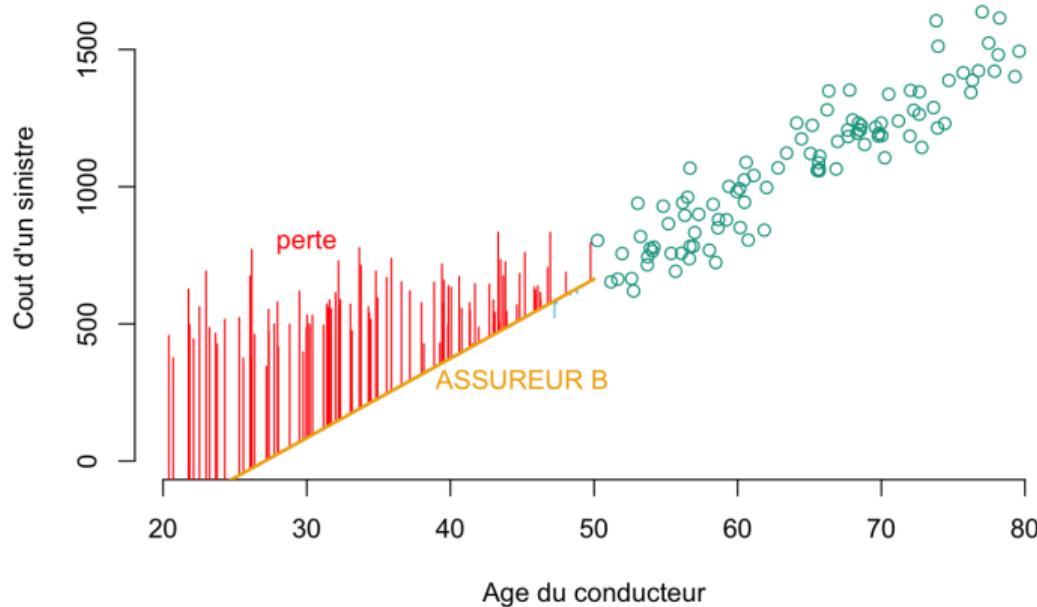


Données et modèles (L'enfer, c'est les autres...) V

Chaque assureur
perd de l'argent

$$\sum_{i: \square} y_i > \sum_{i: \square} \pi_B(x_i)$$

$$\sum_{i: \circ} y_i > \sum_{i: \circ} \pi_A(x_i)$$



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.
- Barry, L. and Charpentier, A. (2020). Personalization as a promise: Can big data change the practice of insurance? *Big Data & Society*, 7(1):2053951720935143.
- 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.
- Bigot, R. and Charpentier, A. (2019). Repenser la responsabilité, et la causalité. *Risques*, 120:123–128.
- Bigot, R. and Charpentier, A. (2020). Quelle responsabilité pour les algorithmes? *Risques*, 121.
- Black, E., Yeom, S., and Fredrikson, M. (2020). Fliptest: Fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 111121.
- Charpentier, A. (2022). *Assurance: biais, discrimination et équité*. Institut Louis Bachelier.
- Charpentier, A., Flachaire, E., and Ly, A. (2018). Econometrics and machine learning. *Economie et Statistique*, 505(1):147–169.

References II

- Davidson, R., MacKinnon, J. G., et al. (2004). *Econometric theory and methods*, volume 5. Oxford University Press New York.
- de Lara, L., González-Sanz, A., Asher, N., and Loubes, J.-M. (2021). Transport-based counterfactual models. *arXiv*, 2108.13025.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.
- Galichon, A. (2016). *Optimal transport methods in economics*. Princeton University Press.
- Gordaliza, P., Del Barrio, E., Fabrice, G., and Loubes, J.-M. (2019). Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365. PMLR.
- Hume, D. (1739). *A Treatise of Human Nature*. Cambridge University Press Archive.
- Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. NIPS.
- Lindholm, M., Richman, R., Tsanakas, A., and Wuthrich, M. V. (2020). Discrimination-free insurance pricing. *Risk Management & Analysis in Financial Institutions eJournal*.
- Lions, P.-L. (2020). *Dans la tête d'un mathématicien*. HumenSciences.

References III

- Morris, J. (2021). Israeli data: How can efficacy vs. severe disease be strong when 60% of hospitalized are vaccinated? *Covid-19 Data Science*, 08/22.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann.
- Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284.
- Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.
- 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.
- Torous, W., Gunsilius, F., and Rigollet, P. (2021). An optimal transport approach to causal inference. *arXiv*, 2108.05858.
- Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.