Apprentissage statistique pour l'échantillonnage en population finie

Mehdi Dagdoug[a]

(a) McGill University, Department of Mathematics and Statistics, Montréal, Canada

Séminaire en ligne
Groupe Enquête de la SFDS

18 Avril 2024

McGill

Plan

1) Une introduction à l'apprentissage statistique.

2) Utilisation des algorithmes d'apprentissage statistique en sondage.

3) Quelques défis liés à l'utilisation de modèles prédictifs.

4) Remarques finales.

Introduction à l'apprentissage statistique: régression

- $(\mathbf{x}, y)^\top$: Vecteur aléatoire dans $\mathbb{R}^p \times \mathbb{R}$ avec distribution $\mathbb{P}_m$, où :
  - y représente une **variable d'intérêt**.
  - $\mathbf{x}$ représente de **l'information auxiliaire**.
- Supposons être intéressés pour **prédire** y à partir de $\mathbf{x}$. Comment procéder?
- Pour toute fonction $f : \mathbb{R}^p \to \mathbb{R}$, on a

$$\mathbb{E}_m \left[(y - f(\mathbf{x}))^2 \right] \geq \mathbb{E}_m \left[(y - m(\mathbf{x}))^2 \right],$$

  où $m(\mathbf{x}) := \mathbb{E}_m [y|\mathbf{x}]$.
- La fonction m est appelée la **fonction de régression**.

Une première décomposition

- Il est possible de décomposer y de la manière suivante:

$$y = m(\mathbf{x}) + \epsilon, \tag{1}$$

où ϵ satisfait $\mathbb{E}[\epsilon|\mathbf{x}] = 0$.

- La relation (1) n'est pas une hypothèse, mais une **décomposition**:

$$y = m(\mathbf{x}) + \underbrace{y - m(\mathbf{x})}_{:=\epsilon} = m(\mathbf{x}) + \epsilon,$$

et

$$\mathbb{E}_m[\epsilon|\mathbf{x}] = \mathbb{E}_m[y - m(\mathbf{x})|\mathbf{x}] = m(\mathbf{x}) - m(\mathbf{x}) = 0.$$

Problème: la fonction de régression est inconnue

- Idéalement, pour prédire de y à partir de $\mathbf{x}$, nous utiliserions la fonction de régression m.
- Malheureusement, m est inconnue.
- En revanche, nous observons généralement N i.i.d. observations

$$\mathcal{D}_N := \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)\},$$

de loi $\mathbb{P}_m$.

- Un des objectifs de l'apprentissage statistique est "d'apprendre" (i.e., estimer) la fonction m inconnue à l'aide de $\mathcal{D}_N$.

Estimation de la fonction de régression

- Un **estimateur de la fonction de régression** est une fonction $\widehat{m}$ de la forme suivante:

$$\widehat{m}(\,\cdot\,, \mathcal{D}_N) : \begin{cases} \mathbb{R}^p \to \mathbb{R}, \\ \mathbf{x} \mapsto \widehat{m}(\mathbf{x}, \mathcal{D}_N)\,. \end{cases}$$

- Il existe de nombreuses manières d'estimer m, notamment:
  - Régression linéaire, pénalisée ou non (p. ex., Lasso, Ridge).
  - Méthodes non-paramétriques "classiques" telles que les polynomes locaux, les splines, les noyaux, etc.
  - Méthodes non-paramétriques, souvent plus algorithmiques, par exemple les k-plus proches voisins, arbres de régression, forêts aléatoires, boosting, réseaux de neurones, etc.

Illustration: modèle non-linéaire univarié (1)

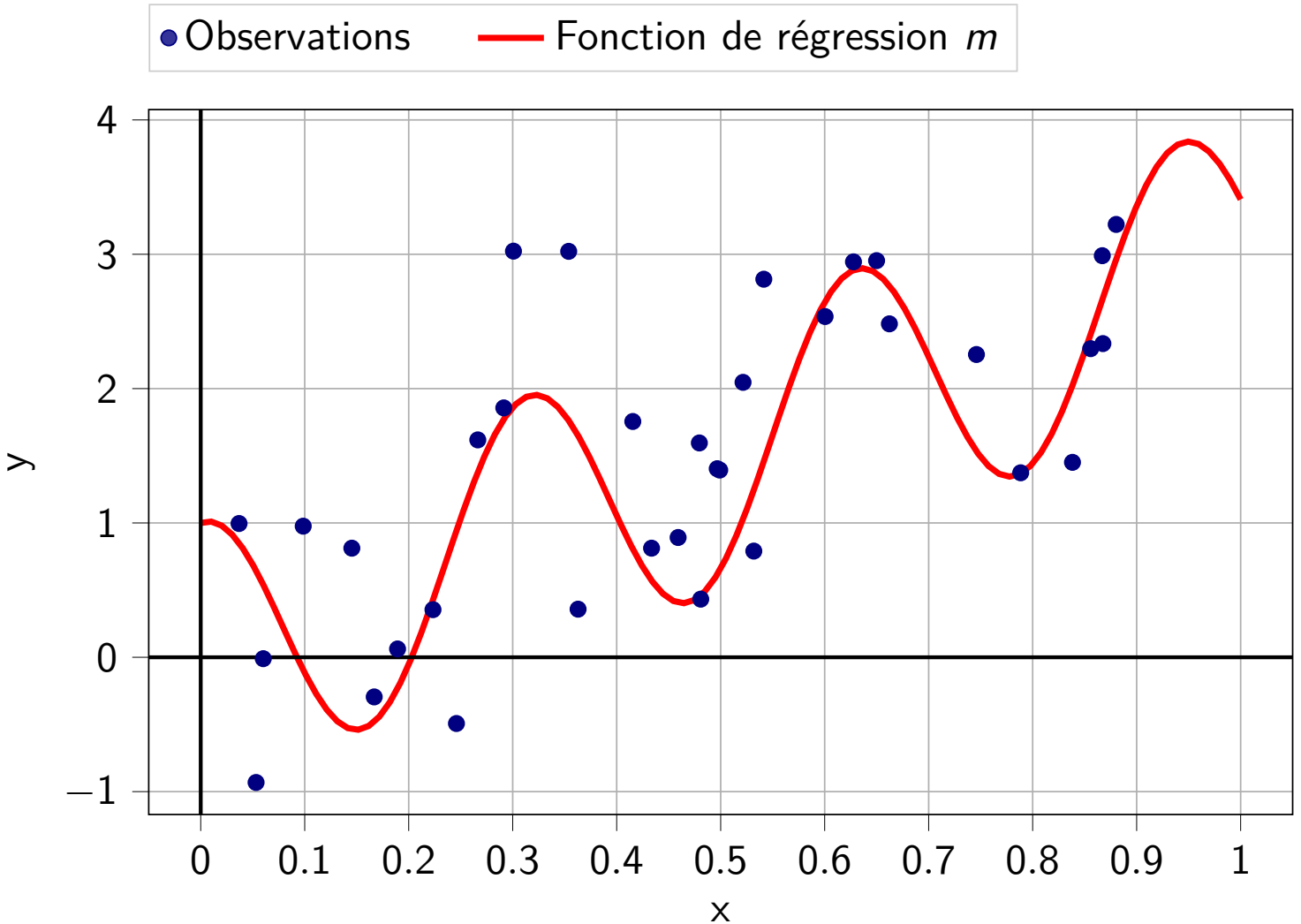


Figure: Graphe de $y = 3x + \cos(20x) + \mathcal{N}(0, 0.64)$ avec $x \sim \mathcal{U}([0; 1])$.

Illustration: modèle non-linéaire univarié (2)

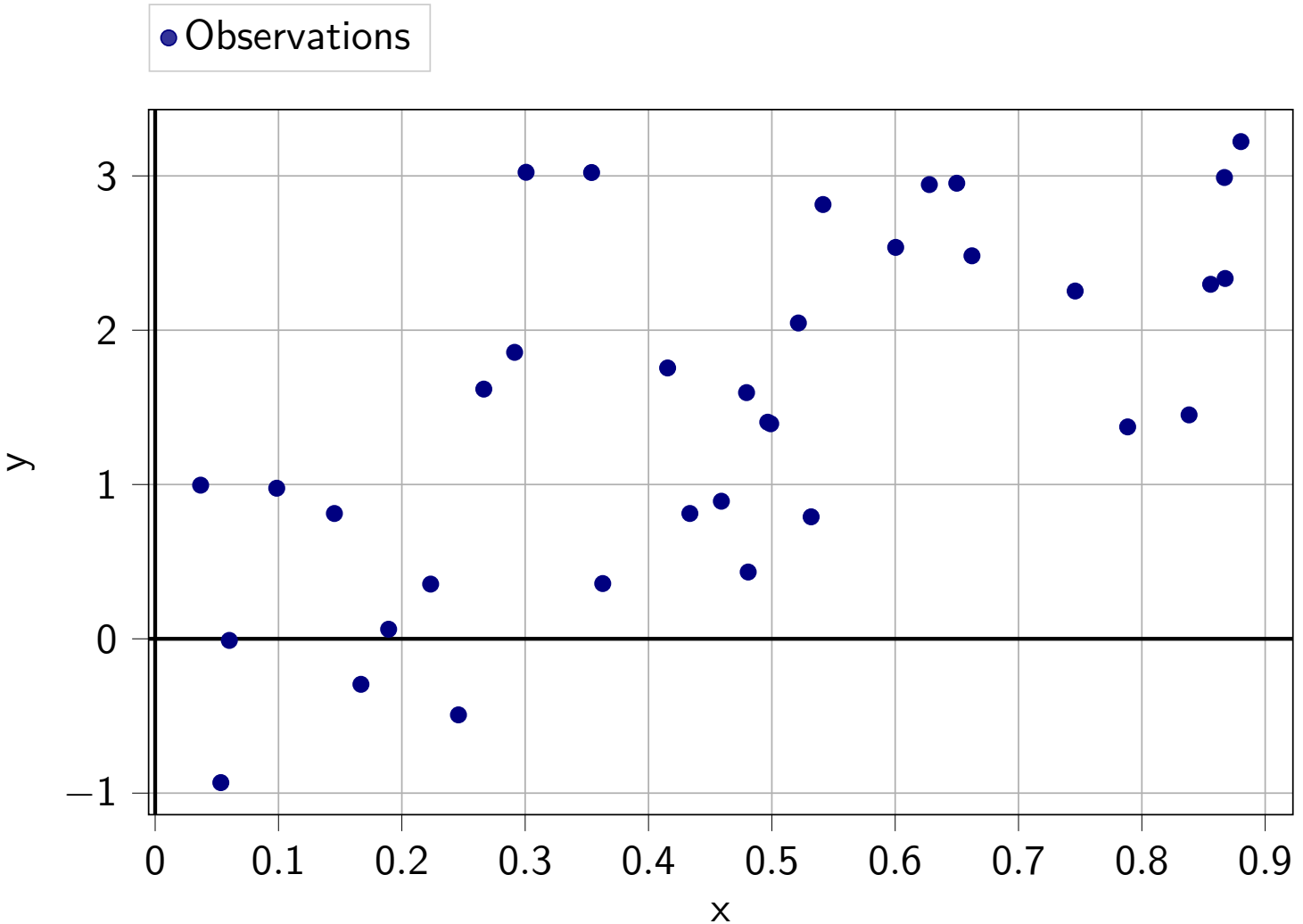


Figure: Graphe de $y = 3x + \cos(20x) + \mathcal{N}(0, 0.64)$ avec $x \sim \mathcal{U}([0; 1])$.

Arbres de régression

Objectif: Un arbre de régression $\widehat{m}_{tree}$ est un estimateur de m.

Définition. (Arbre de régression)

Étape 1: Choisir:

- Un **critère de découpe** (p. ex., CART).
- Un **critère d'arrêt** (p. ex., un nombre minimal d'éléments n_0 dans les feuilles).

Étape 2: Découper récursivement $\mathbb{R}^p$ pour en créer une **partition**

$$\mathcal{P} = \{\mathcal{A}_1, ..., \mathcal{A}_T\}.$$

Les éléments de $\mathcal{P}$ sont appelés les **feuilles de l'arbre**.

Étape 3: La prédiction à un point $\mathbf{x} \in \mathcal{A}_j$ est donnée par

$$\widehat{m}_{tree}(\mathbf{x}) := \sum_{k \in S_r} \frac{\mathbb{1}_{\mathbf{x}_k \in \mathcal{A}_j}}{\sum_{l \in S_r} \mathbb{1}_{\mathbf{x}_l \in \mathcal{A}_j}} y_k.$$

Illustration: Construction d'un arbre de régression

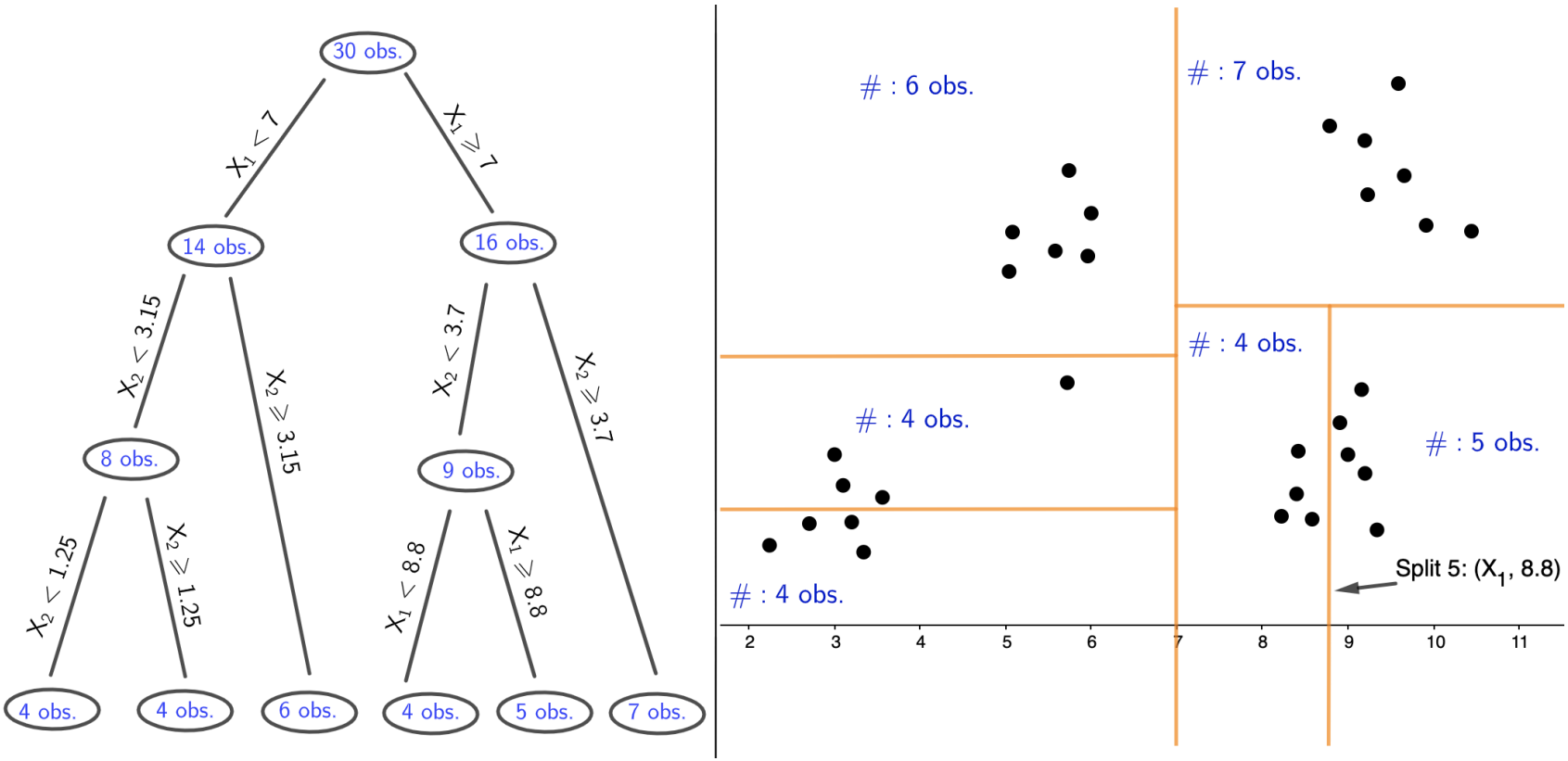


Figure: Un arbre de régression (gauche) et sa partition (droite).

Illustration: Prédictions d'un arbre (1)

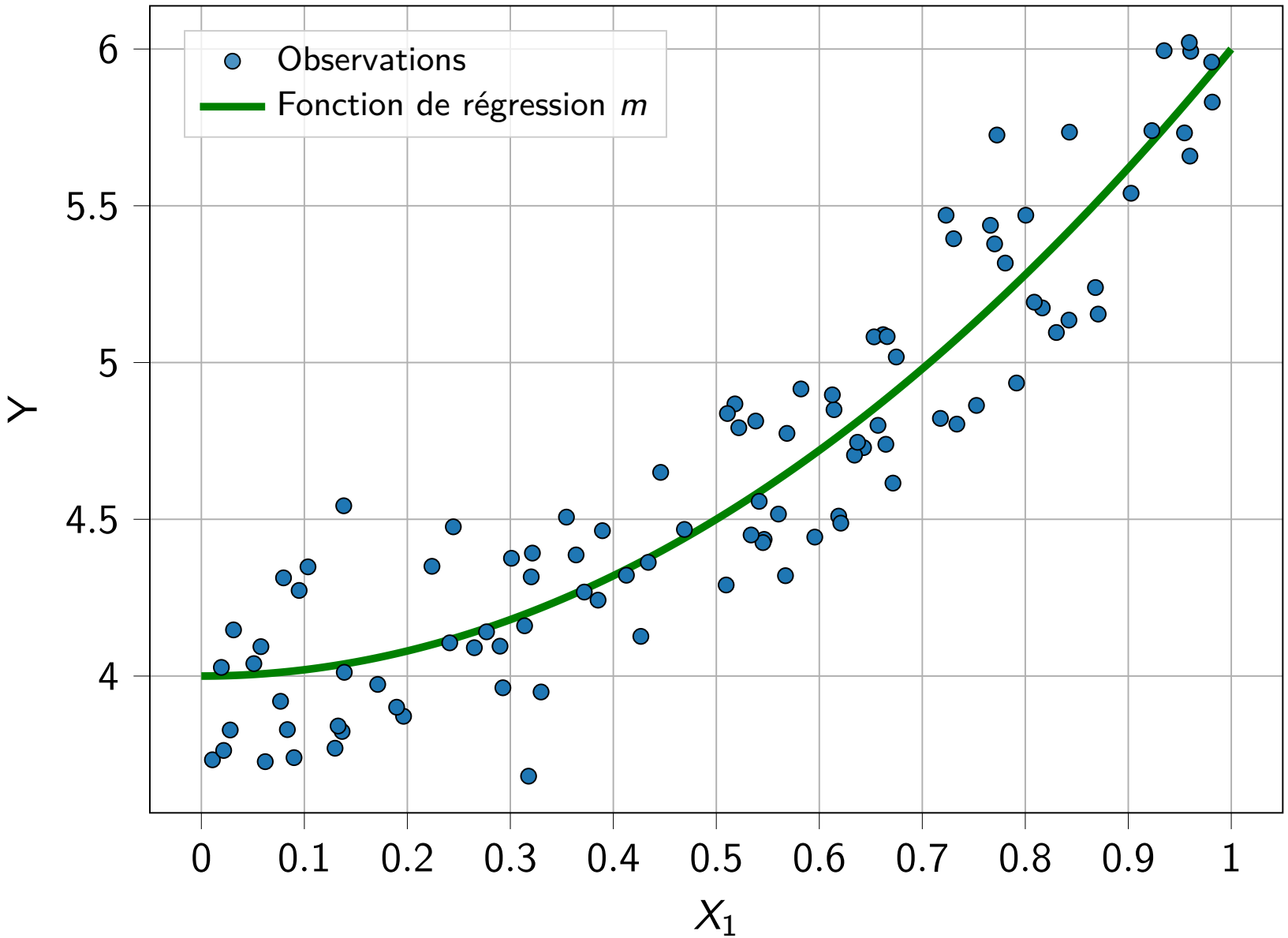
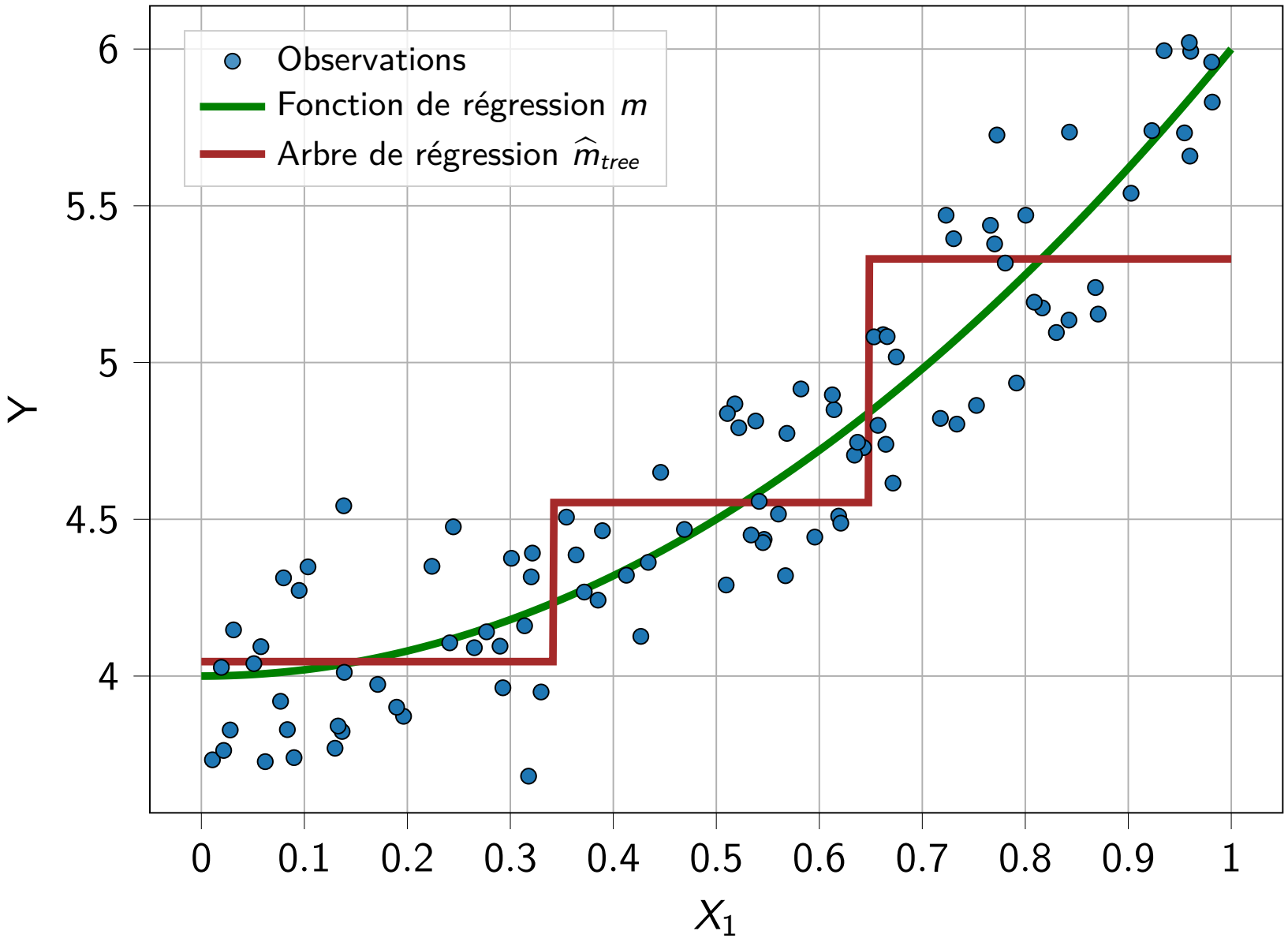


Illustration: Prédictions d'un arbre (2)



Modèles linéaires: qu'en est-il?

- Faire l'hypothèse d'un modèle linéaire revient à supposer qu'il existe $\boldsymbol{\beta} \in \mathbb{R}^p$ tel que
$$m(\mathbf{x}) := \mathbb{E}[y|\mathbf{x}] = \mathbf{x}^\top \boldsymbol{\beta}.$$
- Un estimateur classique $\widehat{m}_{lr}$ de m, est défini par
$$\widehat{m}_{lr}(\mathbf{x}) := \mathbf{x}^\top \widehat{\boldsymbol{\beta}},$$
où
$$\widehat{\boldsymbol{\beta}} := \underset{\mathbf{u}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{k=1}^{N} \left(y_k - \mathbf{x}_k^\top \mathbf{u}\right)^2 = \left(\sum_{k=1}^{N} \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \sum_{k=1}^{N} \mathbf{x}_k y_k,$$
si $\sum_{k=1}^{N} \mathbf{x}_k \mathbf{x}_k^\top$ est inversible.
- On utilise quelques fois des estimateurs pénalisés:
$$\widehat{\boldsymbol{\beta}}_{pen} := \underset{\mathbf{u}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{k=1}^{N} \left(y_k - \mathbf{x}_k^\top \mathbf{u}\right)^2 + \operatorname{pen}_\lambda(\mathbf{u}),$$
où $\operatorname{pen}_\lambda : \mathbb{R}^p \to \mathbb{R}$ est une fonction induisant la "parcimonie".

Plan

1) Une introduction à l'apprentissage statistique.

2) Utilisation des algorithmes d'apprentissage statistique en sondage.

3) Quelques défis liés à l'utilisation de modèles prédictifs.

4) Remarques finales.

Échantillonnage pour l'estimation d'une moyenne

- U: population finie de taille N.
- Y: variable d'intérêt, avec mesures $\{y_k\}_{k \in U}$.
- **Objectif**: Estimer la moyenne μ définie par :

$$\mu := \frac{1}{N} \sum_{k \in U} y_k.$$

- S: échantillon aléatoire de U.
- Probabilités d'inclusions:

$$\pi_k := \mathbb{P}\{k \in S\}, \qquad \pi_{kl} := \mathbb{P}\{k, l \in S\}, \qquad k, l \in U.$$

On fait l'hypothèse que $\pi_k > 0$ et $\pi_{kl} > 0$, pour tout $k, l \in U$.

Estimation sans information auxiliaire

- S'il n'y a pas de non-réponse, les données observées sont

$$\mathcal{D}_y := \{y_k \ ; \ k \in S\}.$$

- On peut alors utiliser l'estimateur d'Horvitz-Thompson (HT):

$$\widehat{\mu}_\pi := \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_k}.$$

- L'estimateur HT est sans biais dès lors que $\pi_k > 0, \ \forall \ k \in U$.
- Sa variance est donnée par

$$\mathbb{V}_p\left(\widehat{\mu}_\pi\right) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l},$$

avec $\Delta_{kl} := \pi_{kl} - \pi_k \pi_l$, pour tout $k, l \in U$.

Estimation avec information auxiliaire

- S'il n'y a pas de non-réponse, les données observées sont

$$\mathcal{D}_{ma} := \{(\mathbf{x}_k, y_k) \ ; \ k \in S\} \bigcup \{\mathbf{x}_k \ ; \ k \in U \backslash S\} .$$

- On peut utiliser cette information supplémentaire pour "améliorer" notre estimation de μ.

- Pour toute fonction $f : \mathbb{R}^p \to \mathbb{R}$, il est possible de définir,

$$\widehat{\mu}_{ma}(f) := \frac{1}{N} \left(\sum_{k \in U} f(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - f(\mathbf{x}_k)}{\pi_k} \right) .$$

On appellera $\widehat{\mu}_{ma}(f)$ **l'estimateur assisté par le modèle** f.

Estimateur assisté par m

Quel estimateur $\widehat{\mu}_{ma}(f)$ choisir? $\iff$ Quelle fonction f choisir?

- Pour toute fonction $f : \mathbb{R}^p \to \mathbb{R}$, indépendantes de $\mathcal{D}_{ma}$, on a

$$\mathbb{E}_{mp}\left[(\widehat{\mu}_{ma}(f) - \mu)^2\right] \geqslant \mathbb{E}_{mp}\left[(\widehat{\mu}_{ma}(m) - \mu)^2\right].$$

  L'estimateur $\widehat{\mu}_{ma}(m)$ est sans biais, avec pour variance la borne de Godambe-Joshi.

- Malheureusement, l'estimateur $\widehat{\mu}_{ma}(m)$, n'est pas utilisable: il dépend de m, qui est inconnue.

- On peut alors utiliser un estimateur $\widehat{m}$ de la fonction de régression m pour définir $\widehat{\mu}_{ma}(\widehat{m})$:

$$\widehat{\mu}_{ma}\left(\widehat{m}\right) = \frac{1}{N}\left(\sum_{k\in U}\widehat{m}\left(\mathbf{x}_k\right) + \sum_{k\in S}\frac{y_k - \widehat{m}\left(\mathbf{x}_k\right)}{\pi_k}\right).$$

Estimateurs assistés par un modèle: quelques avantages

- Considérons $\widehat{m}$, un estimateur de m, construit au niveau de l'échantillon S.
- Sous des conditions de régularité supplémentaires, il est souvent possible de montrer que

$$\frac{\widehat{\mu}_{ma}(\widehat{m}) - \mu}{\sqrt{\mathbb{AV}\left(\widehat{\mu}_{ma}(\widehat{m})\right)}} \xrightarrow[n,N\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

où

$$\mathbb{AV}\left(\widehat{\mu}_{ma}(\widehat{m})\right) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_k - \widetilde{m}(\mathbf{x}_k)}{\pi_k}\frac{y_l - \widetilde{m}(\mathbf{x}_l)}{\pi_l},$$

et $\widetilde{m}$ est l'estimateur de m, "correspondant" à $\widehat{m}$, construit au niveau de la population U.

- L'estimateur $\widehat{\mu}_{ma}(\widehat{m})$ est donc souvent plus efficace que $\widehat{\mu}_{\pi}$ si $\widetilde{m}$ "fait de bonnes prédictions", asymptotiquement.

Non-réponse: notations

- Bien souvent, Y n'est observé que partiellement: certains individus refusent de répondre.
- On décompose $S = S_r \cup S_m$, où:
  - S_r dénote l'échantillon des répondants,
  - S_m dénote l'échantillon des non-répondants.
- On fait l'hypothèse que des variables auxiliaires $X_1, ..., X_p$ sont observées pour tous les éléments de S.
- Les données observées sont:

	$\boldsymbol{X}$	Y
S_r	✓	✓
S_m	✓	✗

Missing At Random hypothèse

- r_k: indicatrice de réponse de l'élément k; i.e.,

$$r_k := \begin{cases} 1, & \text{si } y_k \text{ est observé,} \\ 0, & \text{si } y_k \text{ est manquant.} \end{cases}$$

- On suppose l'hypothèse **Missing At Random (MAR)**, c'est-à-dire,

$$\mathbb{P}\left\{r_k = 1 | y_k, \mathbf{x}_k\right\} = \mathbb{P}\left\{r_k = 1 | \mathbf{x}_k\right\} := p(\mathbf{x}_k).$$

  On fait de plus l'hypothèse que $p(\mathbf{x}) > 0$, presque sûrement.

- L'hypothèse MAR implique que

$$\mathbb{E}\left[y_k | \mathbf{x}_k, r_k = 1\right] = \mathbb{E}\left[y_k | \mathbf{x}_k, r_k = 0\right] = m\left(\mathbf{x}_k\right).$$

Un estimateur imputé

- Considérons la décomposition suivante:

$$\widehat{\mu}_{\pi} = \frac{1}{N}\left(\sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{y_k}{\pi_k}\right).$$

- L'**estimateur imputé**, basé sur la fonction de régression,

$$\widehat{\mu}_{imp}(m) = \frac{1}{N}\left(\sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{m(\mathbf{x}_k)}{\pi_k}\right)$$

  est sans biais.

- Malheureusement, $\widehat{\mu}_{imp}(m)$ est inutilisable, car m est inconnue.

- On peut utiliser un estimateur $\widehat{m}$ de la fonction de régression m pour définir

$$\widehat{\mu}_{imp}(\widehat{m}) = \frac{1}{N}\left(\sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{\widehat{m}(\mathbf{x}_k)}{\pi_k}\right).$$

Un estimateur repondéré

- Si beaucoup de variables d'intérêts sont concernées par la non-réponse, l'imputation peut être difficile.
- L'estimateur **Propensity Score Adjusted (PSA)**, basé sur p,

$$\widehat{\mu}_{psa}(p) := \frac{1}{N} \sum_{k \in S_r} \frac{y_k}{\pi_k p(\mathbf{x}_k)},$$

  est aussi sans biais.
- Malheureusement, $\widehat{\mu}_{psa}(p)$ est inutilisable, car p est inconnue.
- On peut alors utiliser un estimateur $\widehat{p}$ de p, pour définir

$$\widehat{\mu}_{psa}(\widehat{p}) = \frac{1}{N} \sum_{k \in S_r} \frac{y_k}{\widehat{p}(\mathbf{x}_k)\pi_k}.$$

Imputer et repondérer: une nécessité

- Les principaux effets de la non-réponse sont les suivants:
  - Une augmentation de la variance.
  - Un biais de non-réponse, potentiellement observable dès lors que
$$\mathbb{P}\left\{r_k = 1 | y_k\right\} \neq \mathbb{P}\left\{r_k = 1\right\},$$
    auquel cas
$$f\left(y_k | r_k = 1\right) \neq f\left(y_k | r_k = 0\right).$$
- Les estimateurs imputés et repondérés ont pour objectif de réduire les effets néfastes de la non-réponse.
- Ces estimateurs pourront réduire, voir retirer, le biais de non-réponse et minimiser l'augmentation de variance.

Plan

1) Une introduction à l'apprentissage statistique.

2) Utilisation des algorithmes d'apprentissage statistique en sondage.

3) Quelques défis liés à l'utilisation de modèles prédictifs.

4) Remarques finales.

Défi 1: Une dépendance à la dimension

- Le problème initial, estimer

$$\mu = \frac{1}{N} \sum_{k \in U} y_k,$$

  est indépendant de la dimension p.

- L'estimateur HT $\widehat{\mu}_\pi$ est lui aussi indépendant de la dimension p.

- En revanche, les estimateurs:
  - Assistés par un modèle $\widehat{\mu}_{ma}(\widehat{m})$,
  - Imputés $\widehat{\mu}_{imp}(\widehat{m})$,
  - Repondérés $\widehat{\mu}_{psa}(\widehat{p})$,

  dépendent tous les trois de l'estimation d'une fonction $f : \mathbb{R}^p \to \mathbb{R}$.

- Nous avons introduit une dépendance à la dimension:
  - Optionnelle dans le cas assisté par un modèle.
  - Nécessaire dans le cas de la non-réponse.

Simulation 1: influence de la dimension sur les estimateurs

Objectif: Estimer μ avec des estimateurs assistés par un modèle.

Set-up:

- $X_1, X_2, ..., X_{200}$: Générées i.i.d. avec distribution $\mathcal{N}(5, 1)$.
- Une variable d'intérêt Y a été créée

$$Y = 2 + 2X_1 + 3X_2 + 4X_5 + \mathcal{N}(0, 4).$$

- Sondage aléatoire simple sans remise à été utilisé.

Méthodologie:

- Estimateurs utilisés: Estimateurs assistés par régression linéaire, Ridge et Lasso.
- Critère: efficacité relative à HT, défini par

$$RE(\widehat{\mu}_{ma}) = 100 \times \frac{\sum_{r=1}^{R}(\widehat{\mu}_{ma}^{(r)} - \mu)^2}{\sum_{r=1}^{R}(\widehat{\mu}_{\pi}^{(r)} - \mu)^2}.$$

- **Nous avons donné aux modèles un nombre croissant de variables pour observer les effets de la dimension.**

Simulation 1: résultats

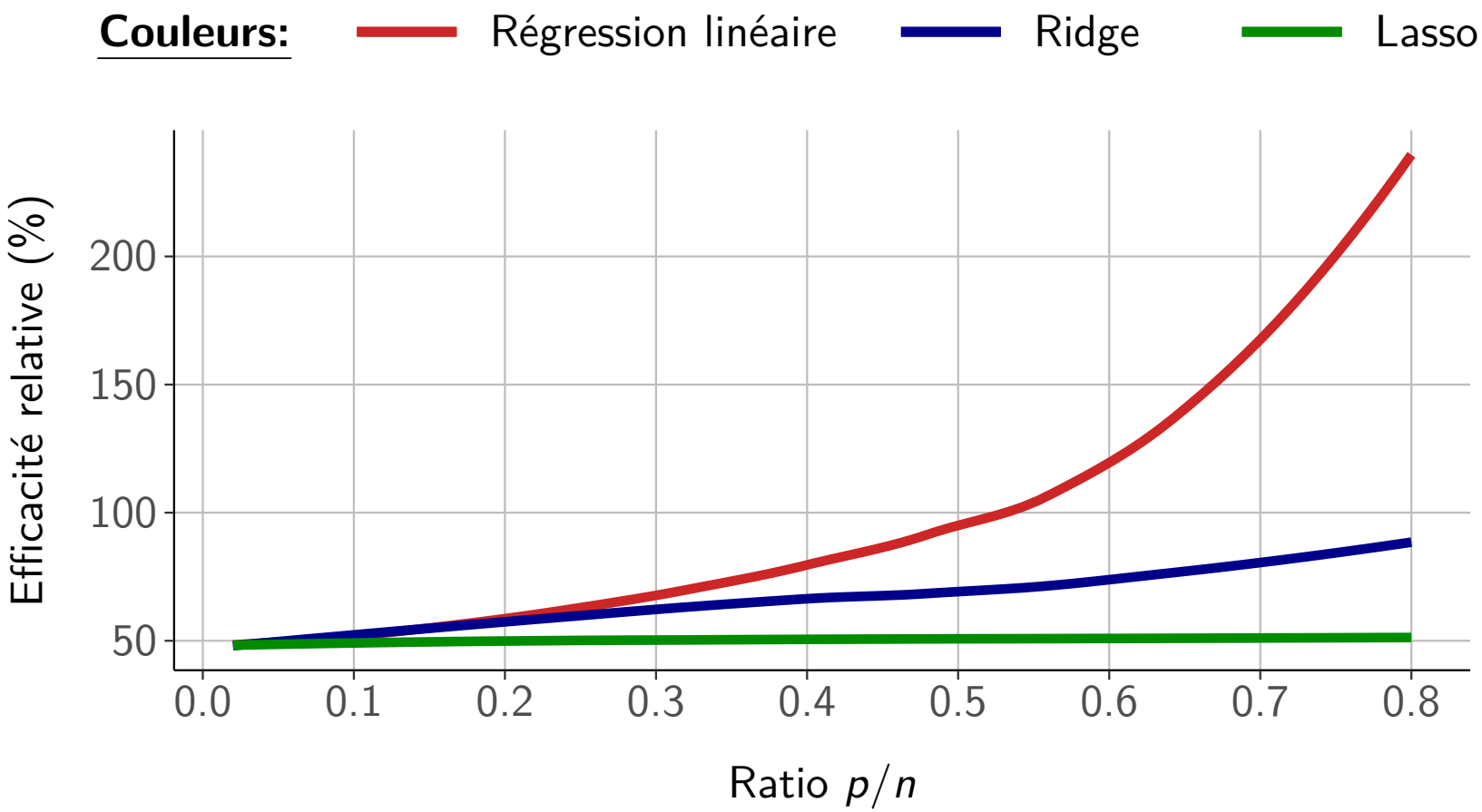


Figure: Évolution de l'efficacité relative en fonction de p/n.

Simulation 2: Estimation de la variance

- Les estimateurs de variance dépendent eux aussi de p. **À quels points sont-ils affectés?**
- Considérons les estimateurs de variance de $\widehat{\mu}_{ma}\left(\widehat{m}_{lr}\right) := \widehat{\mu}_{greg}$ suivants:

$$\widehat{V}_{taylor} := \frac{1}{N^2} \sum_{k \in S} \sum_{l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k - \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}}{\pi_k} \frac{y_l - \mathbf{x}_l^\top \widehat{\boldsymbol{\beta}}}{\pi_l},$$

$$\widehat{V}_g := \frac{1}{N^2} \sum_{k \in S} \sum_{l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_k \left(y_k - \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}\right)}{\pi_k} \frac{g_l \left(y_l - \mathbf{x}_l^\top \widehat{\boldsymbol{\beta}}\right)}{\pi_l},$$

$$\widehat{V}_{jk} := \frac{1}{N^2} \left(1 - \frac{n}{N}\right) \frac{n-1}{n} \sum_{k \in S} \left(\widehat{\mu}_{greg}^{(-k)} - \widehat{\mu}_{greg}\right)^2,$$

où $\widehat{\mu}_{greg}^{(-k)}$ dénote l'estimateur GREG, sans l'élément k.

- Critère: biais relatif, défini par

$$RB\left(\widehat{V}\right) := \frac{100}{R} \sum_{r=1}^{R} \frac{\widehat{V}^{(r)} - \mathbb{V}_{MC}(\widehat{\mu}_{greg})}{\mathbb{V}_{MC}(\widehat{\mu}_{greg})}.$$

Simulation 2: résultats

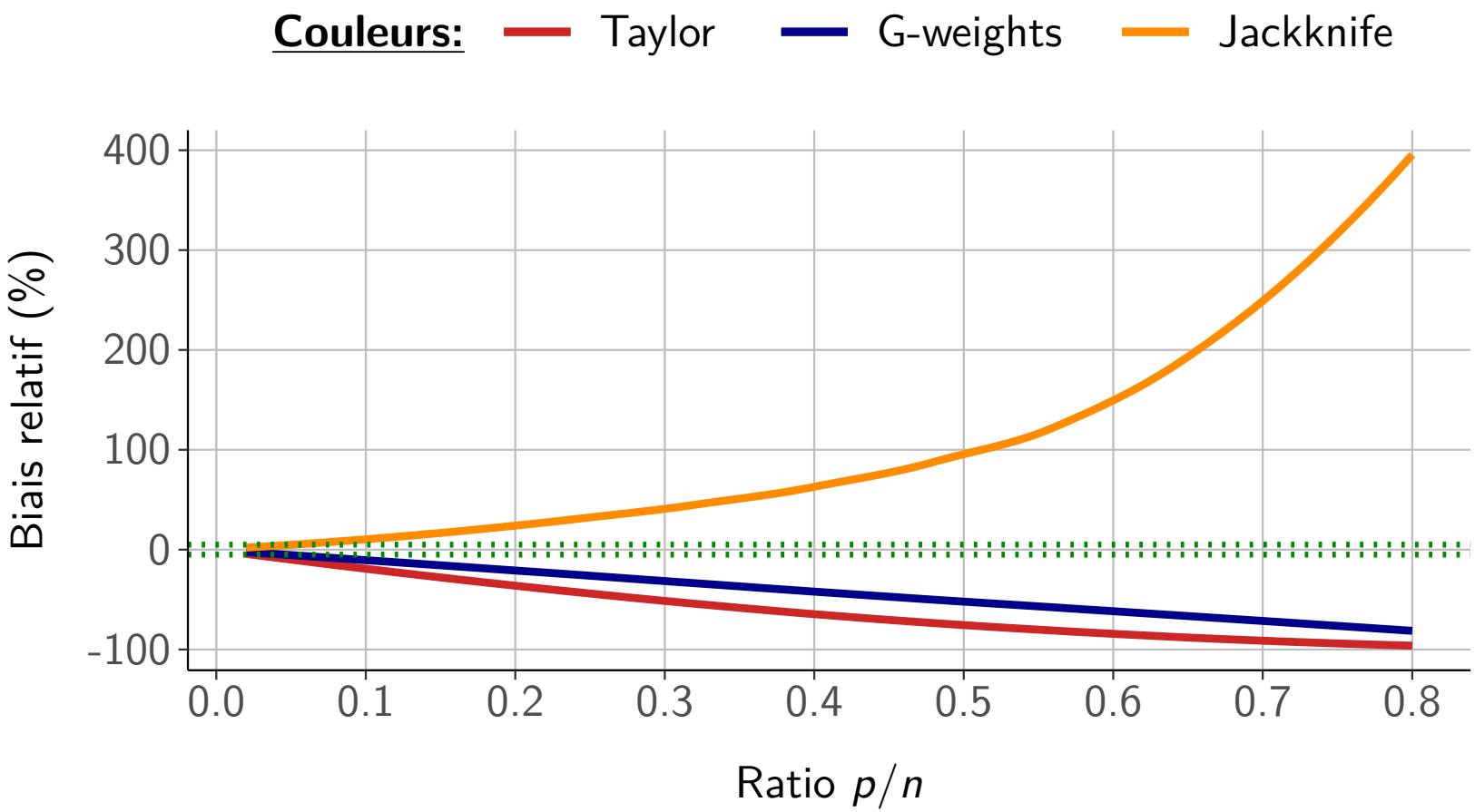


Figure: Évolution du biais relatif en fonction de p/n.

Défi 2: Estimation de la variance

- Il n'y a pas qu'en grande dimension que l'estimation de la variance est délicate...
- Lorsque les estimateurs $\widehat{m}$ de m sont très "flexibles", cela peut donner lieu à des instabilités.
- Considérons le comportement de l'estimateur de variance usuel de l'estimateur $\widehat{\mu}_{ma}(\widehat{m}_{tree})$:

$$\widehat{V}_{taylor} := \frac{1}{N^2} \sum_{k \in S} \sum_{l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k - \widehat{m}_{tree}(\mathbf{x}_k)}{\pi_k} \frac{y_l - \widehat{m}_{tree}(\mathbf{x}_l)}{\pi_l}.$$

- Dans le cas des arbres, cet estimateur est fonction de n_0: le nombre minimal d'éléments dans les feuilles.

Objectif de simulation: Construire des intervalles de confiances à 95% à l'aide de $\widehat{\mu}_{ma}(\widehat{m}_{tree})$.

Simulation 3: Taux de couverture

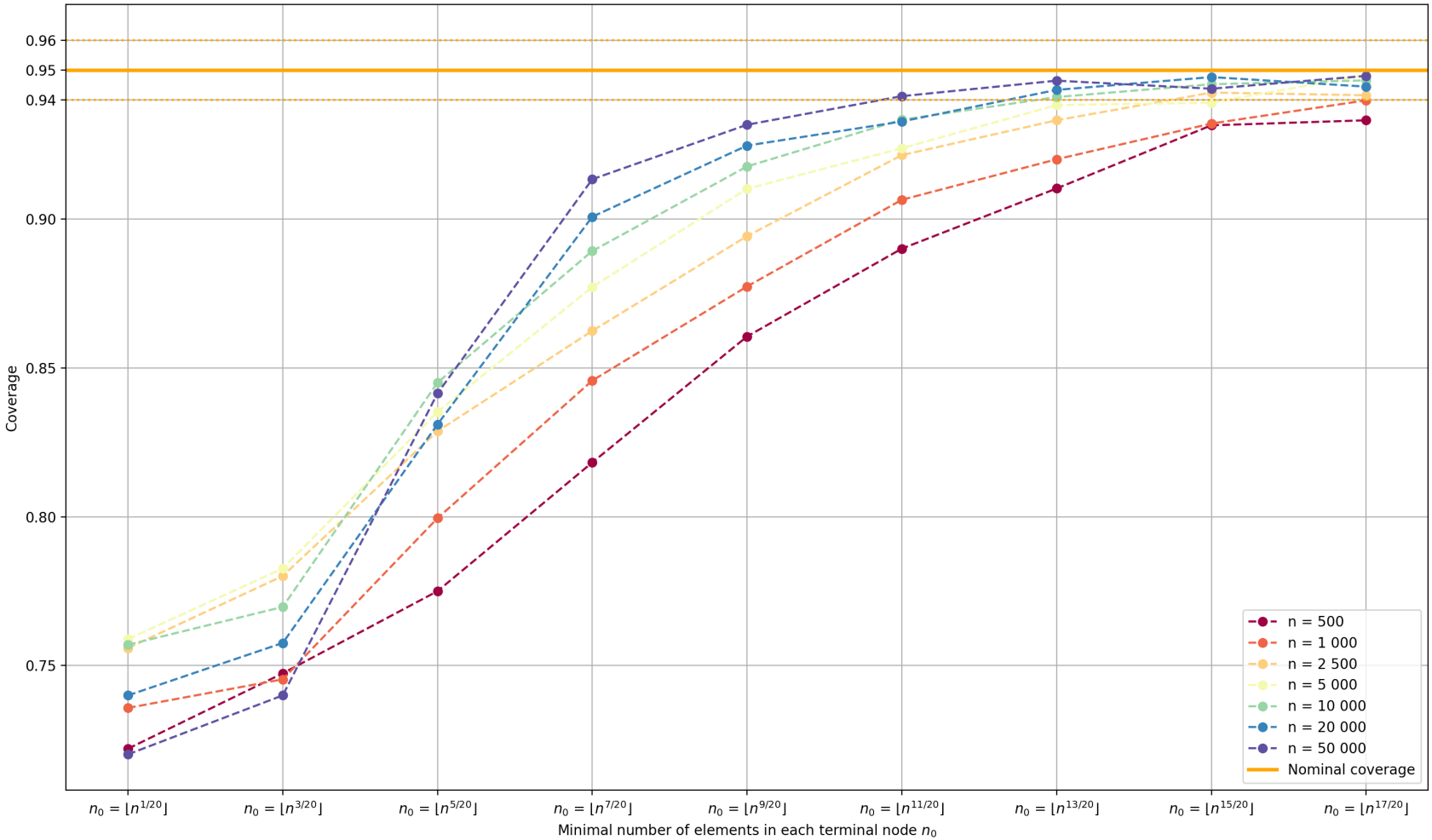


Figure: Couverture Monte-Carlo pour l'estimation de μ; tiré de Dagdoug, Goga, Haziza (2023).

Défi 3: Sélection de modèles

- En pratique, il existe énormément d'estimateurs de m ou de p:
  - Sans non-réponse: À chaque estimateur $\widehat{m}$ de m, on peut définir un estimateur $\widehat{\mu}_{ma}(\widehat{m})$ de μ assisté par $\widehat{m}$.
  - Avec non-réponse: À chaque estimateur $\widehat{m}$ de m, on peut définir un estimateur imputé $\widehat{\mu}_{imp}(\widehat{m})$ de μ.
  - Avec non-réponse: À chaque estimateur $\widehat{p}$ de p, on peut définir un estimateur repondéré $\widehat{\mu}_{psa}(\widehat{p})$ de μ.
- Dans une même famille d'estimateurs de m, (p. ex., arbres) il existe souvent une multitude de paramètres à choisir (p. ex., n_0).
- Plus précisément, étant donné une liste d'estimateurs

$$\Gamma := \{\widehat{\mu}_{ma}(\widehat{m}_1), \widehat{\mu}_{ma}(\widehat{m}_2), ..., \widehat{\mu}_{ma}(\widehat{m}_T)\},$$

quel estimateur devrions-nous choisir?

Prédictions vs estimations

- En apprentissage statistique, différentes méthodes ont été développées pour choisir le meilleur estimateur $\widehat{m}^*$ de m.
- De meilleures prédictions impliquent elles une meilleure estimation?

Méthodologie de simulation:

- Considérons un scénario avec $p = 6$ covariables et une variable d'intérêt

$$Y = 2 - 2X_1 + 4X_2 + \epsilon_k.$$

- Le modèle de non-réponse satisfait l'hypothèse MAR:

$$p(\mathbf{x}) = 0.05 + 0.95 \times \text{sigmoid}\left(\mathbf{x}^\top \boldsymbol{\beta}\right),$$

avec $\beta_j > 0$, pour tout j.

$\hookrightarrow$ m dépend de $X_1 - X_2$, et p dépend de $X_1 - X_6$.

Simulation 4: set-up

Objectif: Estimer μ en présence de non-réponse à l'aide d'estimateurs PSAs.

- L'estimateur de p considéré ici est basé sur la **méthode des scores** $\widehat{p}_{score}$.
- Méthode similaire aux arbres de régression: on partitionne les prédictions d'une régression logistique $\{\widehat{p}_{log}(\mathbf{x}_k)\}_{k \in S}$.
- Nous avons considéré 6 estimateurs de μ,

$$\Gamma := \{\widehat{\mu}_{psa}(\widehat{m}_{score1}),\ \widehat{\mu}_{psa}(\widehat{m}_{score2}),\ \ldots,\ \widehat{\mu}_{psa}(\widehat{p}_{score6})\}\,,$$

  où $\widehat{p}_{scoreJ}$ est l'estimateur de p, basé sur la méthode des scores, entraîné sur les variables $X_1, \ldots, X_J$.

- Critère additionnel:

$$\mathrm{MSE}_{MC}(\widehat{p}) = \frac{100}{B} \sum_{b=1}^{B} \frac{1}{n_r} \sum_{k \in S_r} \left(\widehat{p}_{(b)}(\mathbf{x}_k) - p(\mathbf{x}_k)\right)^2.$$

Simulation 4: résultats

Estimator	$\widehat{\mu}_{naive}$	$\widehat{\mu}_{psa}$ x_1	$\widehat{\mu}_{psa}$ x_1-x_2	$\widehat{\mu}_{psa}$ x_1-x_3	$\widehat{\mu}_{psa}$ x_1-x_4	$\widehat{\mu}_{psa}$ x_1-x_5	$\widehat{\mu}_{psa}$ x_1-x_6
RB	-13.4	-12.2	-0.2	-0.8	-0.3	-1.0	-0.4
RE	623	561	134	141	142	161	206
MSE($\widehat{p}$)	4.7	5.0	4.9	4.6	4.1	1.3	0.4

Table: Comparaison de l'efficacité des estimations versus efficacité des prédictions; tiré de Larbi, Tsang, Haziza et Dagdoug (2024+).

Conséquences:

- De meilleures prédictions n'impliquent pas nécessairement de meilleures estimations.
- L'utilisation des critères utilisés en apprentissage statistique (p. ex., validation croisée) n'est pas nécessairement adaptée.

Plan

1) Une introduction à l'apprentissage statistique.

2) Utilisation des algorithmes d'apprentissage statistique en sondage.

3) Quelques défis liés à l'utilisation de modèles prédictifs.

4) Remarques finales.

Quelques solutions: grande dimension

- Le problème de la grande dimension n'est important que si $p/n \not\approx 0$.
- Si le modèle est supposé linéaire et sparse, il est préférable d'utiliser la pénalisation (p. ex., Ridge, Lasso).
- Si le modèle n'est pas linéaire, certains algorithmes d'apprentissage statistique s'adaptent bien à la sparsité; p. ex., les forêts aléatoires.
- L'estimation de la variance en grande dimension reste toutefois délicate, mais il est possible d'obtenir des estimateurs "débiaisés" (Eustache, Dagdoug et Haziza, 2024+).

Quelques solutions: estimation de la variance

- Le problème d'estimation de variance dans le cas d'algorithmes complexes vient d'une sous-estimation des résidus $\{y_k - \widetilde{m}(\mathbf{x}_k)\}$.
- Dans le cas d'estimation assistée par un modèle, une solution revient à remplacer les résidus de l'échantillon par des résidus obtenu par validation croisée.
- Cette méthode a été proposée par Opsomer et Miller (2005) et Dagdoug, Goga et Haziza (2023) dans des contextes différents.
- Cette technique reste valable pour estimer la variance des estimateurs imputés, mais est plus délicate à bien définir.

Quelques solutions: sélection d'estimateurs

- Choisir un estimateur parmi une liste d'estimateurs

$$\Gamma := \{\widehat{\mu}_1, \widehat{\mu}_2, ..., \widehat{\mu}_T\},$$

en sondage nécessite le développement d'un nouveau critère $\mathcal{C} : \Gamma \to \mathbb{R}_+$.

- Deux possibilités: **sélectionner** ou **agréger**.
  - La sélection d'un estimateur selon $\mathcal{C}$:

$$\widehat{\mu}^* := \underset{\widehat{\mu} \in \Gamma}{\operatorname{argmin}} \mathcal{C}(\widehat{\mu}).$$

  - L'agrégation d'estimateurs:

$$\widehat{\mu}_{agg} := \sum_{j=1}^{T} w_j \widehat{\mu}_j,$$

où les poids $\{w_j\}_{j=1}^T$ peuvent dépendre des données.