# Utilisation des méthodes d'apprentissage automatique pour le traitement de la non-réponse totale dans les enquêtes

David Haziza

Département de mathematiques et de statistique
University of Ottawa

En collaboration avec

Khaled Larbi (ENSAE)

et

Mehdi Dagdoug (Université de Bourgogne Franche-Comté)

Séminaire du groupe Enquêtes, Modèles et Applications

20 Octobre 2022

# Effets de la non-réponse

- Principal problème: biais introduit quand les répondants sont différents des non-répondants en ce qui a trait aux variables d'intérêt.
- Composante de variance additionnelle: due à la taille d'échantillon observée, $n_r$, qui est plus petite que la taille d'échantillon planifiée, $n$.
- Solution: utiliser des méthodes de pondération qui mettent à profit une information auxiliaire disponible à la fois pour les répondants et les non-répondants.

# Estimation en l'absence de non-réponse

- Soit $U = \{1, 2, ..., N\}$ une population de taille $N$.
- $Y$: Variable d'intérêt
- Objectif: estimer le total

$$t_y = \sum_{k \in U} y_k.$$

- On sélectionne un échantillon $S \subset U$, avec $\pi_k = \mathbb{P}(k \in S) > 0$ et $\pi_{k\ell} = \mathbb{P}(k, \ell \in S) > 0$, $k, \ell \in U$.
- Estimateur prototype de $t_y$ (Estimateur de Horvitz-Thompson):

$$\widehat{t}_{y,\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in S} d_k y_k.$$

- Sans biais par rapport au plan de sondage: $\mathbb{E}_p(\widehat{t}_{y,\pi}) = t_y$ quelque soit la variable d'intérêt $y$.

# Mécanisme de non-réponse

- Soit $r_k$ l'indicateur de réponse associé à l'unité $k$ tel que $r_k = 1$ si l'unité $k$ est répondante et $r_k = 0$, sinon.
- L'ensemble des répondants, $S_r$, est le sous-ensemble de $S$ qui contient toutes les unités $k \in S$ telles que $r_k = 1$.
- Nous supposons que le vrai mécanisme de non-réponse inconnu dépend d'un certain vecteur de variables $\mathrm{v}_k$, $k \in S$.
- La probabilité de réponse pour l'unité $k$ est $p_k = P(r_k = 1|S, \mathrm{v}_k)$.
- Nous supposons aussi que les unités répondent indépendamment les unes des autres.
- Nous supposons que $0 < p_k \leqslant 1$ pour tout $k$.
- Mécanisme de non-réponse:

$$r_k \sim B(p_k), \quad k = 1, \ldots, n.$$

# Erreur totale

- Soit $\widehat{t}_{y,NR}$ un estimateur de $t_y$ après traitement.
- Erreur totale de $\widehat{t}_{y,NR}$:

$$\widehat{t}_{y,NR} - t_y = \left(\widehat{t}_{y,\pi} - t_y\right) + \left(\widehat{t}_{y,NR} - \widehat{t}_{y,\pi}\right).$$

- Le terme $\widehat{t}_{y,\pi} - t_y$ désigne l'erreur due à l'échantillonnage.
- Le terme $\widehat{t}_{y,NR} - \widehat{t}_{y,\pi}$ désigne l'erreur due à la non-réponse.
- Objectif du traitement: réduire le biais de non-réponse et contrôler la variance de non-réponse.

# Estimateur non-ajusté

- Estimateur non-ajusté de $t_y$:

$$\widehat{t}_{y,naif} = N\widehat{\overline{Y}}_r \quad \text{avec } \widehat{\overline{Y}}_r = \frac{\sum_{k\in S_r} d_k y_k}{\sum_{k\in S_r} d_k}$$

- Erreur de non-réponse de $\widehat{t}_{y,naif}$:

$$\widehat{t}_{y,naif} - \widehat{t}_{y,\pi} = N\left\{\frac{\widehat{N}_m}{\widehat{N}_\pi}\left(\widehat{\overline{Y}}_r - \widehat{\overline{Y}}_m\right)\right\},$$

- L'erreur de non-réponse de $\widehat{t}_{y,naif}$ tend à être grande si:
  - le taux de non-réponse est important;

    et/ou
  - $\widehat{\overline{Y}}_r$ (moyenne des répondants) est loin de $\widehat{\overline{Y}}_m$ (moyenne des non-répondants).

# La repondération par l'inverse de la probabilité de réponse estimée

- Si les $p_k$ sont connues, un estimateur sans biais de $t_y$ est l'estimateur par double dilatation :

$$\widetilde{t}_{y,DE} = \sum_{k \in S_r} \frac{d_k}{p_k} y_k = \sum_{k \in S_r} \frac{1}{\pi_k} \frac{1}{p_k} y_k.$$

- En pratique, les $p_k$ sont inconnues $\Rightarrow$ elles doivent être estimées.
- Ce problème est souvent résolu en choisissant un modèle pour les indicateurs de la réponse $r_k$, appelé modèle de non-réponse, et en obtenant ensuite les probabilités estimées $\widehat{p}_k$; par exemple, Särndal et Swensson (1987) et Ekholm et Laaksonen (1991).

# Adjusted estimators

- Système de pondération ajusté pour la non-réponse:

$$\{w_k^* = d_k/\widehat{p}_k = 1/(\pi_k \widehat{p}_k); k \in S_r\}.$$

- Estimateurs ajustés:

$$\widehat{t}_{y,PSA} = \sum_{k \in S_r} w_k^* y_k \quad \text{et} \quad \widehat{t}_{y,HA} = \frac{N}{\widehat{N}_{PSA}} \widehat{t}_{y,PSA},$$

où $\widehat{N}_{PSA} = \sum_{k \in S_r} w_k^*$.

- Deux étapes dans la modélisation:
  - Sélection de variables $v_k$ qui sont prédictrices de $r_k$
  - Choix d'un modèle afin de décrire la relation entre $r_k$ et $v_k$

# Comment choisir les prédicteurs?

- L'utilisation de variables explicatives $v_k$ qui sont fortement prédictives de la réponse tend à produire:
  - certaines probabilités de réponse estimées petites et donc de grands ajustements aux poids $\widehat{p}_k^{-1}$
  - estimateurs instables (avec une grande variance)
- Recommandation: le vecteur $v_k$ devrait être lié à la fois à $r_k$ et aux variables d'intérêt; par exemple, Little et Vartivarian (2005), Beaumont (2005) et Kim et al. (2019)
- Des prédicteurs qui sont liés à $r_k$ mais pas aux variables d'intérêt devraient être exclus:
  - Ne contribueront pas à réduire le biais de non-réponse ;
  - risque d'augmenter la variance de non-réponse de manière significative.

# Estimation paramétrique de $p_k$

- Nous supposons que les $v_k, k \in S$ ne contiennent aucune valeur manquante.
- Sous cette hypothèse, les données sont dites Missing At Random (MAR).
- Nous commençons par une estimation paramétrique des $p_k$. Un modèle de non-réponse paramétrique général peut s'écrire comme suit:

$$p_k = f(v_k, \gamma),$$

pour une certaine fonction prédéterminée $f(\cdot, \gamma)$, où $\gamma$ est un vecteur de paramètres.

- La probabilité de réponse estimée est : $\widehat{p}_k = f(v_k, \widehat{\gamma})$, où $\widehat{\gamma}$ est un estimateur de $\gamma$.
- L'estimateur PSA de $t_y$ qui en résulte est convergent pour $t_y$ si le modèle de non-réponse est correctement spécifié.

# Estimation paramétrique de $p_k$

- Il existe de nombreuses fonctions possibles $f(\cdot)$.
- Par exemple, avec la régression logistique, la probabilité de réponse est modélisée comme :

$$p_k = f(\mathbf{v}_k, \boldsymbol{\gamma}) = \frac{e^{\mathbf{v}_k^\top \boldsymbol{\gamma}}}{1 + e^{\mathbf{v}_k^\top \boldsymbol{\gamma}}}.$$

- Il existe de nombreuses méthodes pour estimer $\boldsymbol{\gamma}$.
- Méthode du maximum de vraisemblance (ML) : $\widehat{\boldsymbol{\gamma}}$ est solution de l'équation :

$$\sum_{k \in S} [r_k - f(\mathbf{v}_k, \widehat{\boldsymbol{\gamma}})] \mathbf{v}_k = 0.$$

- Pseudo ML (pondéré par les poids de sondage) :

$$\sum_{k \in S} d_k [r_k - f(\mathbf{v}_k, \widehat{\boldsymbol{\gamma}})] \mathbf{v}_k = 0.$$

# Estimation paramétrique des probabilités de réponse

- Défis liés à l'utilisation d'un modèle paramétrique : non-robuste à une mauvaise spécification du modèle
  - La fonction $f(v_k; .)$ peut ne pas être appropriée pour décrire la relation entre l'indicateur de réponse et les variables explicatives.
  - Il peut manquer des interactions dans le modèle, qui n'ont pas été détectées lors de la sélection du modèle.
  - Il peut manquer des prédicteurs qui prennent en compte la courbure (termes quadratiques, cubiques, etc.).
  - Les modèles paramétriques comme la régression logistiques peuvent produire de très petites probabilités de réponse estimées, $\widehat{p}_k$, $\longrightarrow$ très grands ajustements de poids $\widehat{p}_k^{-1}$ $\longrightarrow$ estimateurs potentiellement instables.

# Estimation non-paramétrique des probabilités de réponse

- Estimation non-paramétrique : nous ne sommes pas prêts à faire des hypothèses sur la forme de la fonction $f(v_k; .)$:
  - Méthode par noyau: Giommi (1984) et Da Silva et Opsomer (2006)
  - Polynômes locaux : Da Silva et Opsomer (2008).
  - Classes de pondération formées sur les probabilités de réponse estimées: Little (1986) ; Eltinge et Yansaneh (1997) et Haziza et Beaumont (2007) $\longrightarrow$ la méthode du score.
  - Détection automatique des interaction Khi-carré (CHi square Automatic Interaction Detection, ou CHAID): Kass (1980).
  - Arbres de régression: par exemple, Phipps et Toth (2012).
- Les méthodes non-paramétriques fournissent une certaine robustesse si la forme de $f(v_k; .)$ est mal spécifiée et protéger (dans une certaine mesure) contre la non-inclusion de prédicteurs prenant en compte la courbure ou les interactions.

# Estimation non-paramétrique: La méthode du score

- Les étapes pour former les classes sont les suivantes :
  - Étape 1 : Obtenir une estimation préliminaire des probabilités de réponse, $\widehat{p}_k^{LR}$, $k \in S$, à partir d'une régression logistique.
  - Étape 2 : Former les classes sur la base des probabilités de réponse estimées, $\widehat{p}_k^{LR}$, en utilisant l'une des méthodes suivantes:
    - méthode des quantiles égaux : consiste à ranger l'échantillon de la probabilité de réponse estimée en ordre croissant et à diviser l'échantillon ordonné en $C$ classes de taille approximativement égale.
    - Algorithme de classification basé sur les $\widehat{p}_k^{LR}$ pour former les classes.
  - Étape 3 : Ajuster le poids des répondants au sein de chaque classe (c'est-à-dire diviser le poids de base $d_k$ par le taux de réponse observé au sein de cette même classe).

# Estimation vs. prédiction: Illustration

- Nous avons généré une population de taille $N = 10,000$ avec 7 variables: une variable d'intérêt $Y$ et 6 prédicteurs $v_1$-$v_6$.
- Les variables $v_1$-$v_6$ ont été générées à partir de différentes distributions Gamma.
- Étant donné $v_1$-$v_6$, nous avons généré la variable $Y$ selon le modèle

$$y_k = 2 - 2v_{1k} + 4v_{2k} + \epsilon_k$$

- De la population, nous avons sélectionné $B = 10,000$ échantillons, de taille $n = 1000$, selon un plan aléatoire simple sans remise.

# Estimation vs. prédiction: Illustration

- Dans chaque échantillon, les unités se sont vues assignées une probabilité de réponse $p_k$ selon la fonction logistique:

$$p_k = \{1 + \exp(-0.05v_{1k} + 0.05v_{2k} - 0.05v_{3k} + 0.05v_{4k} - 0.05v_{5k} + 0.02v_{6k})\}^{-1}.$$

- Taux de réponse global: approximativement égal à 50% dans chaque échantillon.
- Dans chaque échantillon, nous avons généré les indicatrices $r_k$ à partir d'une loi de Bernoulli avec probabilité $p_k$.
- But: estimer $t_y = \sum_{k \in U} y_k$.
- Les $v_1$-$v_6$ sont observées pour toutes les unités échantillonnées (répondants et non-répondants)

# Utilisation de variables superflues: Illustration

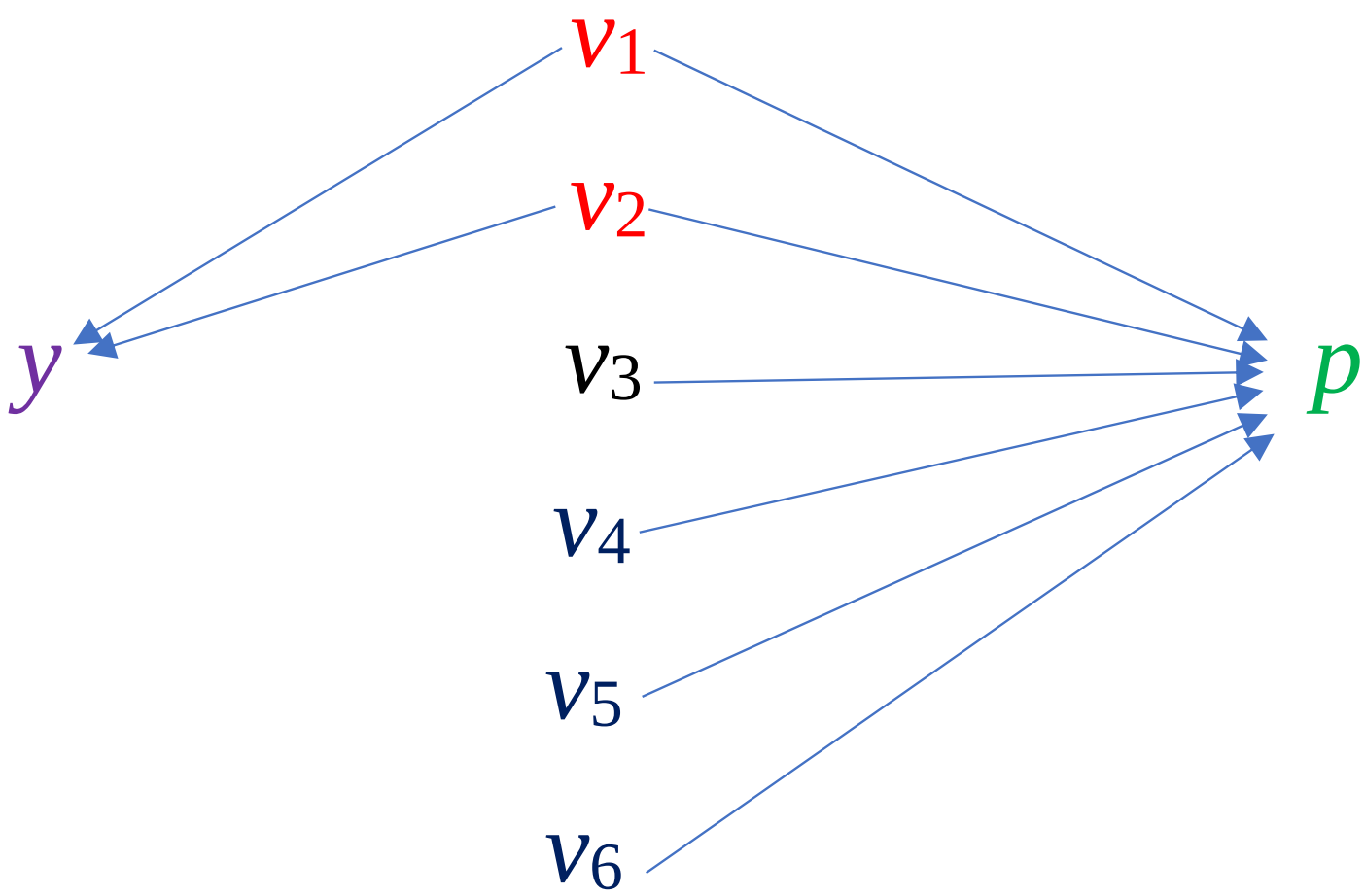



Figure 1: Relations entre les variables

# Estimation vs. prédiction: Illustration

- Nous avons calculé deux estimateurs de $t_y$:
  - L'estimateur non-ajusté $\widehat{t}_{y,naif} = N\widehat{\overline{Y}}_r$;
  - L'estimateur $\widehat{t}_{y,PSA} = \sum_{k \in S_r} \frac{d_k}{\widehat{p}_k} y_k$, où $\widehat{p}_k$ a été obtenu au moyen de la méthode des scores (avec 20 classes) et de différents sous-ensembles de $v_1$-$v_6$.
- Nous avons calculé les mesures Monte Carlo suivantes:
  - Biais relatif Monte Carlo (en %):

$$\text{BR}_{MC}(\widehat{t}) = \frac{1}{10,000} \sum_{b=1}^{10,000} \frac{(\widehat{t}_{(b)} - t_y)}{t_y} \times 100.$$

  - Erreur quadratique moyenne Monte Carlo:

$$\text{EQM}_{MC}(\widehat{t}) = \frac{1}{10,000} \sum_{b=1}^{10,000} \left(\widehat{t}_{(b)} - t_y\right)^2.$$

# Estimation vs. prédiction: Illustration

- Nous avons également calculé le coefficient de variation (en %) des poids ajustés $w_k^* = d_k/\widehat{p}_k$:

$$\mathrm{CV}_{MC}(w_k^*) = 100 \times \frac{1}{10,000} \sum_{b=1}^{10,000} \frac{s_{w^*(b)}}{\overline{w}^*_{(b)}},$$

où

$$s^2_{w^*} = \frac{1}{n_r - 1} \sum_{k \in S_r} (w_k^* - \overline{w}^*)^2$$

avec $\overline{w}^* = n_r^{-1} \sum_{k \in S_r} w_k^*$.

- Nous avons calculé l'erreur quadratique moyenne Monte Carlo des prédictions:

$$EQM_{MC}(\widehat{p}) = 100 \times \frac{1}{10,000} \sum_{b=1}^{10,000} \frac{1}{n_r} \sum_{k \in S_r} \left(\widehat{p}_{k(b)} - p_k\right)^2.$$

# Estimation vs. prédiction: Illustration

| Estimateur | $\widehat{t}_{y,naif}$ | $\widehat{t}_{y,PSA}$ $v_1$ | $\widehat{t}_{y,PSA}$ $v_1$-$v_2$ | $\widehat{t}_{y,PSA}$ $v_1$-$v_3$ | $\widehat{t}_{y,PSA}$ $v_1$-$v_4$ | $\widehat{t}_{y,PSA}$ $v_1$-$v_5$ | $\widehat{t}_{y,PSA}$ $v_1$-$v_6$ |
|---|---|---|---|---|---|---|---|
| $BR_{MC}(\widehat{t})$ (%) | -14.1 | -13.0 | -1.7 | -1.8 | -0.7 | -1.1 | -0.8 |
| $ER_{MC}(\widehat{t})$ | 540 | 480 | 112 | 118 | 117 | 149 | 218 |
| $CV_{MC}(w*)$ (%) | 0 | 17.4 | 19.6 | 21.7 | 30.1 | 46.7 | 64.2 |
| $EQM_{MC}(\widehat{p})$ | 4.8 | 5.4 | 5.3 | 5.1 | 4.6 | 1.7 | 0.9 |

Table 1: Biais relatif Monte Carlo et efficacité relative pour plusieurs estimateurs de $t_y$ basés sur la méthode des scores avec 20 classes

Note: $ER_{MC}(\widehat{t}) = 100 \times \frac{EQM_{MC}(\widehat{t})}{EQM_{MC}(\widehat{t}_{y,\pi})}$

# Même expérience avec les arbres de régression

- Nous avons répété les mêmes simulations mais, cette fois, avec des arbres de régression. Nous avons calculé :
  - L'estimateur non ajusté $\widehat{t}_{y,naif} = N\widehat{\overline{Y}}_r$;
  - L'estimateur ajusté du score de propension $\widehat{t}_{y,PSA} = \sum_{k \in S_r} \frac{d_k}{\widehat{p}_k} y_k$, où $\widehat{p}_k$ a été obtenu en utilisant un arbre de régression basé sur différents sous-ensembles de $v_1$-$v_6$ comme prédicteurs.
- Nous avons fait varier différents paramètres :
  - La taille de l'échantillon $n$ ;
  - $n_0$ : nombre minimal de répondants dans chaque nœud terminal ;
  - $c$ : seuil du paramètre de complexité.
- Remarque : Une valeur de $c = 1$ se traduira toujours par un arbre sans split. Si un split n'augmente pas le $R^2$ global du modèle d'au moins $c$, alors le split ne vaut pas la peine d'être fait. Valeur par défaut : $c = 0.01$.

# Même expérience avec les arbres de régression

| | $BR_{MC}(\widehat{t})$ (%) | $ER_{MC}(\widehat{t})$ | $EQM_{MC}(\widehat{p})$ | $CV_{MC}(w*)$ |
|---|---|---|---|---|
| $\widehat{t}_{y,naif}$ | -14.4 | 865 | 3.0 | 0 |
| $\widehat{t}_{y,PSA}(v_1)$ | -11.1 | 572 | 4.0 | 29 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_2)$ | -0.6 | 116 | 4.3 | 36 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_3)$ | -1.6 | 140 | 3.9 | 43 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_4)$ | -2.6 | 162 | 3.8 | 48 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_5)$ | -4.0 | 206 | 3.4 | 53 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_6)$ | -6.5 | 319 | 2.9 | 62 |

Table 2: Arbres de régression avec $n = 1000$, $c = 0$ et $n_0 = 10$

Note: Average number of nodes between 53-61

# Même expérience avec les arbres de régression

| | $BR_{MC}(\widehat{t})$ (%) | $ER_{MC}(\widehat{t})$ | $EQM_{MC}(\widehat{p})$ | $CV_{MC}(w*)$ |
|---|---|---|---|---|
| $\widehat{t}_{y,naif}$ | -14.4 | 866 | 3.0 | 0 |
| $\widehat{t}_{y,PSA}(v_1)$ | -11.2 | 577 | 3.9 | 29 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_2)$ | -0.7 | 117 | 4.2 | 36 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_3)$ | -1.8 | 142 | 3.8 | 43 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_4)$ | -2.8 | 164 | 3.7 | 48 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_5)$ | -4.1 | 209 | 3.3 | 53 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_6)$ | -6.6 | 321 | 2.9 | 62 |

Table 3: Arbres de régression avec $n = 1000$, $c = 0.001$ et $n_0 = 10$

Note: Nombre moyen de noeuds terminaux 50-57

# Même expérience avec les arbres de régression

| | $BR_{MC}(\widehat{t})$ (%) | $ER_{MC}(\widehat{t})$ | $EQM_{MC}(\widehat{p})$ | $CV_{MC}(w*)$ |
|---|---|---|---|---|
| $\widehat{t}_{y,naif}$ | -14.4 | 866 | 3.1 | 0 |
| $\widehat{t}_{y,PSA}(v_1)$ | -13.7 | 803 | 3.0 | 4.7 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_2)$ | -8.0 | 414 | 3.0 | 14 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_3)$ | -7.3 | 360 | 2.9 | 23 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_4)$ | -7.3 | 341 | 2.8 | 33 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_5)$ | -7.8 | 364 | 2.6 | 39 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_6)$ | -10.0 | 519 | 2.4 | 49 |

Table 4: Arbres de régression avec $n = 1000$, $c = 0.01$ et $n_0 = 10$

Note: Nombre moyen de noeuds terminaux 2-22

# Même expérience avec les arbres de régression

| | $BR_{MC}(\widehat{t})$ (in %) | $ER_{MC}(\widehat{t})$ | $EQM_{MC}(\widehat{p})$ | $CV_{MC}(w*)$ |
|---|---|---|---|---|
| $\widehat{t}_{y,naif}$ | -14.4 | 866 | 3.0 | 0 |
| $\widehat{t}_{y,PSA}(v_1)$ | -13.4 | 779 | 3.0 | 4 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_2)$ | -8.4 | 434 | 3.0 | 10 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_3)$ | -7.6 | 352 | 2.7 | 18 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_4)$ | -9.6 | 469 | 2.6 | 21 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_5)$ | -10.4 | 534 | 2.4 | 25 |
| $\widehat{t}_{y,PSA}(v_1\text{-}v_6)$ | -12.8 | 733 | 2.3 | 30 |

Table 5: Arbres de régression avec $n = 1000$, $c = 0$ et $n_0 = 50$

Note: Nombre moyen de noeuds terminaux 2-10

# Méthodes d'aggrégation

- Les méthodes d'aggrégation consistent à :
  - Obtenir des probabilités de réponse estimées en utilisant plusieurs procédures (e.g., apprentissage automatique) ;
  - Combiner ces probabilités de manière à obtenir un ensemble de poids ajustés $w_k^* = d_k/\widehat{p}_k$ ajusté pour la non-réponse ;
- Pourquoi utiliser une méthode d'aggrégation ?
  - Il est très probable qu'aucune procédure d'apprentissage automatique ne surpasse toutes les autres concurrentes dans tous les scénarios ;
  - Une procédure d'apprentissage automatique peut donner de bons résultats pour un scénario particulier mais de moins bons résultats pour un autre scénario ;
  - Difficile de prédire à l'avance quelle procédure aura une bonne performance.
  - Une méthode d'aggrégation, qui combine plusieurs procédures d'apprentissage automatique, peut être plus performante qu'une seule procédure.

# Méthodes d'aggrégation

- Trois méthodes d'aggrégation:
  (1) Calage ;
  (2) Réajustement par régression linéaire ;
  (3) Réajustement par régression linéaire suivi d'un calage.
- Supposons que nous utilisions $M$ procédures d'apprentissage automatique ;
- Soit $\widehat{\mathrm{p}}_k = (\widehat{p}_k^{(1)}, \ldots, \widehat{p}_k^{(M)})$ un vecteur $M$ de probabilités de réponse estimées associées à l'unité $k$.
- La composante $\widehat{p}_k^{(m)}$ dans $\widehat{\mathrm{p}}_k$ correspond à une probabilité de réponse estimée sur la base de la $m$ième procédure d'apprentissage automatique, $m = 1, \ldots, M$.

# Méthodes d'aggrégation: Calage

- Nous cherchons un poids de calage $w_k$ tel que

$$\sum_{k \in S_r} G(w_k, d_k)$$

est minimum sujet aux $M + 1$ contraintes de calage suivantes:

$$\sum_{k \in S_r} w_k = \sum_{k \in S} d_k \quad \text{et} \quad \sum_{k \in S_r} w_k \mathcal{L}(\widehat{\mathrm{p}}_k) = \sum_{k \in S} d_k \mathcal{L}(\widehat{\mathrm{p}}_k)$$

où $\mathcal{L}(\cdot)$ dépend de la méthode de calage utilisée.

- Les poids $w_k$ peuvent être considérés comme un résumé scalaire de l'information contenue dans le vecteur $\widehat{\mathrm{p}}_k$.

- Estimateur de $t_y$ :

$$\widehat{t}_{y,cal} = \sum_{k \in \mathcal{S}_r} w_k y_k.$$

# Méthodes d'ensemble : Réajustement par régression linéaire

- Consiste à comprimer les informations contenues dans $\widehat{\mathrm{p}}_k$ en ajustant un modèle de régression linéaire avec l'indicateur de réponse $r_k$ comme variable dépendante et $\widehat{\mathrm{p}}_k$ comme vecteur de variables explicatives:

$$r_k = \beta^{(1)}\widehat{p}_k^{(1)} + \ldots + \beta^{(M)}\widehat{p}_k^{(M)} + \varepsilon_k.$$

- Estimateur des moindres carrés régression de $\boldsymbol{\beta} = (\beta^{(1)}, \ldots, \beta^{(M)})^\top$: $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}^{(1)}, \ldots, \widehat{\beta}^{(M)})^\top$.
- Score comprimé (ou résumé scalaire) pour l'unité $k$ :

$$\widehat{\mathrm{p}}_k^{\mathrm{comp}} = \widehat{\mathrm{p}}_k^\top \widehat{\boldsymbol{\beta}}.$$

- Estimateur de $t_y$ :

$$\widehat{t}_{y,\mathrm{comp}} = \sum_{k \in \mathcal{S}_r} \frac{d_k}{\widehat{p}_k^{\mathrm{comp}}} y_k.$$

# Méthodes d'ensemble : Réajustement par régression linéaire + calage

- Nous commençons par obtenir les scores comprimés $\widehat{p}_k^{\text{com}}$ comme décrit précédemment. Ensuite, nous cherchons des poids de calage $w_k$ tels que

$$\sum_{k \in S_r} G(w_k, d_k)$$

est minimum sujet aux deux contraintes de calage suivantes :

$$\sum_{k \in S_r} w_k = \sum_{k \in S} d_k$$

et

$$\sum_{k \in S_r} w_k \mathcal{L}(\widehat{p}_k^{\text{comp}}) = \sum_{k \in S} d_k \mathcal{L}(\widehat{p}_k^{\text{comp}}).$$

- Estimateur de $t_y$ :

$$\widehat{t}_{y,\text{comp-cal}} = \sum_{k \in \mathcal{S}_r} w_k y_k.$$

# Étude par simulation: Générer les données

- Autres études par simulation: Lohr et al. (2015), Gelein (2017) et Kern et al. (2019).
- Nous avons effectué une étude par simulation afin d'évaluer les performances de plusieurs procédures d'apprentissage automatique en termes de biais et d'efficacité.
- Nous avons généré plusieurs populations finies de taille $N = 50000$.
- Chaque population est constituée d'une variable d'intérêt $Y$ et de 7 variables auxiliaires (prédicteurs): 4 continues + 3 discrètes.
- Deux scénarios :
  - les prédicteurs ont été générés indépendamment ;
  - Corrélation entre les prédicteurs au moyen de copules gaussiennes.

# Étude par simulation: Générer les données

- Étant donné les valeurs des prédicteurs, nous avons généré plusieurs variables $Y$ selon les modèles suivants :

$$y_k = \gamma_0 + \gamma_1^{(s)} X_{1k}^{(s)} + \gamma_1^{(c)} X_{1k}^{(c)} + \gamma_2^{(c)} X_{2k}^{(c)} + \gamma_3^{(c)} X_{3k}^{(c)} + \sum_{j=2}^{5} \gamma_{1j}^{(d)} (1_{\{X_{1k}^{(d)}=j\}})$$

$$+ \gamma_2^{(d)} X_{2k}^{(d)} + \sum_{k=2}^{5} \gamma_{3j}^{(d)} (1_{\{X_{3k}^{(d)}=j\}}) + \varepsilon_k$$

et

$$y_k = \delta_1 X_{2k}^{(c)} + \delta_2 (X_{2k}^{(c)})^2 (1 - 1_{\{X_{3k}^{(d)}=2\} \cup \{X_{3k}^{(d)}=3\}}) + \log(1 + \delta_3 X_{2k}^{(c)}) (1_{\{X_{3k}^{(d)}=2\} \cup \{X_{3k}^{(d)}=3\}}) + \varepsilon_k,$$

où $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

- Deux types de modèles : linéaires et non linéaires.

# Étude par simulation: Plans de sondage

- Chaque population a été divisée en dix strates sur la base de la variable auxiliaire $X^{(s)}$ en utilisant la méthode de déciles égaux.
- Dans chaque population, nous avons sélectionné $B = 1000$ échantillons selon un échantillonnage aléatoire simple stratifié sans remise de taille $n = 1000$ basé sur la répartition de Neyman.
- Deux plans de sondage:
  - Non informatif: aucune corrélation entre les poids d'échantillonnage $n_h/N_h$ et la variable de l'enquête ;
  - Informatif: corrélation entre les poids d'échantillonnage $n_h/N_h$ et la variable d'enquête fixée à 0.3 environ.
- Cela a conduit à 7 variables d'enquête différentes.

# Étude par simulation: Mécanismes de non-réponse

Six mécanismes de non-réponse:

NR1 : $p_k^{(1)} = \text{logit}^{-1}\{-0.8 - 0.05X_{1k}^{(s)} + 0.2X_{1k}^{(c)} + 0.5X_{2k}^{(c)} - 0.05X_{3k}^{(c)} + \sum_{k=2}^{5} 0.2(1_{\{X_{1k}^{(c)}=k\}}) + 0.2X_{2k}^{(d)} + \sum_{k=2}^{5} 0.3(1_{\{X_{3k}^{(d)}=k\}})\}.$

NR2 : $p_k^{(2)} = 0.1 + 0.9\,\text{logit}^{-1}(0.5 + 0.3X_{1k}^{(s)} - 1.1X_{1k}^{(c)} - 1.1X_{2k}^{(c)} - 1.1X_{3k}^{(c)} + \sum_{k=2}^{5} 0.8(1_{\{X_{1k}^{(c)}=k\}}) + 0.8X_{2k}^{(d)} + \sum_{k=2}^{5} 0.8(1_{\{X_{3k}^{(d)}=k\}})).$

NR3 : $p_k^{(3)} =$

$$0.1 + 0.9\,\text{logit}^{-1}\left\{-1 + \text{sgn}\,(X_{1k}^{c})\,(X_{1k}^{c})^2 + 3 \times 1_{\left\{X_{1k}^{(d)}<4\right\}\cap\left\{X_{2k}^{(d)}=1\right\}}\right\}.$$

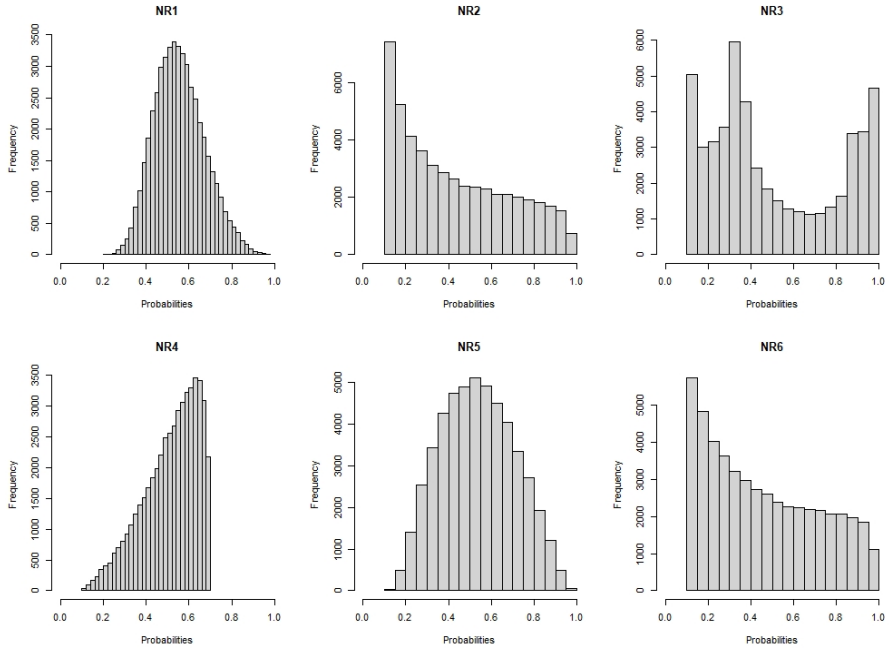
NR4 : $p_k^{(6)} = 0.1 + 0.6\,\text{logit}^{-1}(0.85X_{1k}^{(s)} + 0.85X_{2k}^{(c)} - 0.85X_{3k}^{(c)} - \sum_{k=2}^{5} 0.2(1_{\{X_{1k}^{(c)}=k\}}) + 0.2X_{2k}^{(d)} - \sum_{k=2}^{5} 0.3(1_{\{X_{3k}^{(d)}=k\}})).$

NR5 : $p_k^{(4)} = 0.55 + 0.45\tanh(0.05y_k - 0.5).$

NR6 : $p_k^{(5)} = 0.1 + 0.9\,\text{logit}^{-1}(0.2y_k - 1.2).$

# Étude par simulation: Mécanismes de non-réponse

- Les paramètres de chaque modèle de non-réponse ont été fixés de manière à obtenir un taux de réponse approximativement égal à 50 %.
- Les indicateurs de réponse $r_k^{(j)}$ ont été générés à partir d'une distribution de Bernoulli avec une probabilité $p_k^{(j)}$, $j = 1, \ldots, 6$.
- Les mécanisme de non-réponse (1)-(4) sont ignorables, tandis que les mécanismes de non-réponse (5) et (6) sont non-ignorables.



# Étude par simulation: Méthodes d'apprentissage automatique

(a) logit: Régression logistique;

(b) `logit_lasso`: Régression logistique avec sélection de variables au moyen d'un LASSO (facteur de pénalisation $\lambda$ obtenu au moyen d'une validation croisée 10-fold).

(c) Arbres de régression et de classification:

- `cart1` : Arbres élagués, au moins 10 observations dans chaque feuille.
- `cart2` : Arbres élagués, au moins 20 observations dans chaque feuille.
- `cart3` : Arbres élagués, au moins 30 observations dans chaque feuille.
- `cart4` : Arbres non-élagués, au moins 20 observations dans chaque feuille.

# Étude par simulation: Méthodes d'apprentissage automatique

(d) Forêts aléatoires:

- `rf1` : au moins 10 observations dans chaque feuille, 100 arbres.
- `rf2` : au moins 10 observations dans chaque feuille, 500 arbres.
- `rf3` : au moins 30 observations dans chaque feuille, 100 arbres.
- `rf4` : au moins 30 observations dans chaque feuille, 500 arbres.
- `rf5` : au moins 30 observations dans chaque feuille, 500 arbres, variables du plan toujours sélectionnées.

(e) $k$-plus proche voisin:

- knn : $k$ obtenu par validation croisée (10-fold) avec $k \in \{3, 12\}$;
- knn_reg : $k$ obtenu par validation croisée (10-fold) avec $k \in \{3, 30\}$.

# Étude par simulation: Méthodes d'apprentissage automatique

## (f) Bayesian additive regression trees (BART):

- `bart` Bart: classification method with parameters described in the original paper for all priors.
- `bart_reg` : Bart as a regression method with parameters described in the original paper for all priors.

## (g) Extreme Gradient Boosting (XGBoost).

- `xb1` : 500 arbres, learning rate: 0.5, max depth : 2.
- `xgb2` : 2000 arbres, learning rate: 0.5, max depth : 2.
- `xgb3` : 1000 arbres, learning rate: 0.01, max depth : 1.
- `xgb4` : 500 arbres, learning rate: 0.05, max depth : 3.

# Étude par simulation: Méthodes d'apprentissage automatique

(h) Support vector machine:

- `svm1` : $\nu-$SVM avec noyau Gaussien.
- `svm2` : $\nu-$SVM avec noyau linéaire.

(i) Algorithme Cubist:

- `cb1` : Unbiased, with extrapolation, 10 committees.
- `cb2` : Unbiased, without extrapolation, 10 committees.
- `cb3` : Biased, with extrapolation, 10 committees.
- `cb4` : Unbiased, with extrapolation, 50 committees.
- `cb5` : Unbiased, with extrapolation, 100 committees.

(j) Model-based recursive partitioning

(k) `CAL`: Méthode d'aggrégation-Calage;

(l) `COMPRESS`: Méthode d'aggrégation-Réajustement linéaire;

(m) `COMPRESS-CAL`: Méthode d'aggrégation-Réajustement linéaire/calage.

# Étude par simulation: Estimateurs ponctuels

- Dans chaque échantillon, nous avons calculé deux estimateurs :

  (i) L'estimateur ajusté (PSA): $\widehat{t}_{y,PSA} = \sum_{k \in \mathcal{S}_r} \frac{d_k}{\widehat{p}_k} y_k$;

  (ii) L'estimateur de Hajek : $\widehat{t}_{y,HA} = \frac{N}{\widehat{N}_{PSA}} \widehat{t}_{y,PSA}$.

- Biais relatif Monte Carlo en pourcentage :

$$\mathrm{RB}_{MC}(\widehat{t}_y) = \frac{100}{B} \sum_{k=1}^{B} \frac{(\widehat{t}_{y,k} - t_y)}{t_y}.$$

- Efficacité relative Monte Carlo, en utilisant l'estimateur de données complètes $\widehat{t}_{y,\pi}$ comme référence :

$$\mathrm{RE}_{MC}(\widehat{t}_y) = 100 \times \frac{\mathrm{EQM}_{MC}(\widehat{t}_y)}{\mathrm{EQM}_{MC}(\widehat{t}_{y,\pi})}$$

# Étude par simulation: Résultats

| Méthode d'apprentissage | Min | Q1 | Med | Q3 | Max | Mean |
|---|---|---|---|---|---|---|
| xgb1 | 155 | 225 | 324 | 1 124 | 12 551 | 1 677 |
| COMPRESS_CAL | 139 | 208 | 328 | 798 | 7 772 | 908 |
| xgb4 | 148 | 221 | 330 | 1 139 | 12 111 | 1 589 |
| xgb3 | 143 | 239 | 344 | 928 | 11 581 | 1 394 |
| cart3 | 175 | 259 | 345 | 1 506 | 9 627 | 1 393 |
| cart2 | 175 | 256 | 348 | 1 464 | 9 472 | 1 376 |
| COMPRESS | 137 | 199 | 348 | 906 | 10 382 | 1 317 |
| CART_reg | 162 | 269 | 350 | 1 367 | 9 522 | 1 293 |
| cart1 | 172 | 259 | 351 | 1 448 | 9 373 | 1 370 |
| xgb2 | 148 | 215 | 368 | 1 016 | 11 479 | 1 405 |
| cart4 | 145 | 262 | 369 | 1 382 | 8 881 | 1 231 |
| bart | 129 | 199 | 384 | 852 | 10 595 | 1 314 |
| knn | 172 | 282 | 392 | 921 | 11 513 | 1 621 |
| Score method | 134 | 216 | 392 | 1 252 | 9 998 | 1 359 |
| svm1 | 129 | 280 | 407 | 780 | 12 482 | 1 639 |

Table 6: Efficacité relative Monte Carlo sur les 42 scénarios pour les estimateurs PSA : les 15 meilleures méthodes (sur 33)

# Étude par simulation: Résultats

| Méthode d'apprentissage | Min | Q1 | Med | Q3 | Max | Mean |
|---|---|---|---|---|---|---|
| xgb1 | 171 | 220 | 295 | 1 751 | 12 305 | 1 864 |
| COMPRESS | 158 | 196 | 296 | 1 470 | 10 144 | 1 443 |
| xgb4 | 170 | 219 | 296 | 1 741 | 11 783 | 1 778 |
| bart | 159 | 202 | 306 | 1 417 | 10 201 | 1 457 |
| xgb3 | 147 | 201 | 307 | 1 508 | 10 815 | 1 560 |
| Score method | 135 | 217 | 308 | 1 267 | 9 984 | 1 377 |
| xgb2 | 148 | 206 | 315 | 1 520 | 10 817 | 1 567 |
| COMPRESS_CAL | 139 | 208 | 328 | 798 | 7 772 | 908 |
| CART_reg | 163 | 252 | 344 | 1 733 | 9 515 | 1 382 |
| cb4 | 165 | 224 | 345 | 1 389 | 12 223 | 1 675 |
| cb5 | 163 | 224 | 346 | 1 398 | 12 255 | 1 680 |
| cart4 | 145 | 229 | 362 | 1 413 | 8 879 | 1 255 |
| cb1 | 182 | 228 | 363 | 1 365 | 12 281 | 1 680 |
| cb2 | 138 | 211 | 419 | 1 291 | 10 922 | 1 367 |
| cart1 | 173 | 248 | 421 | 1 807 | 9 369 | 1 485 |
| cart2 | 174 | 240 | 422 | 1 807 | 9 472 | 1 487 |

Table 7: Efficacité relative Monte Carlo sur les 42 scénarios pour les estimateurs de Hajek : les 15 meilleures méthodes (sur 33)

# Étude par simulation: Résultats

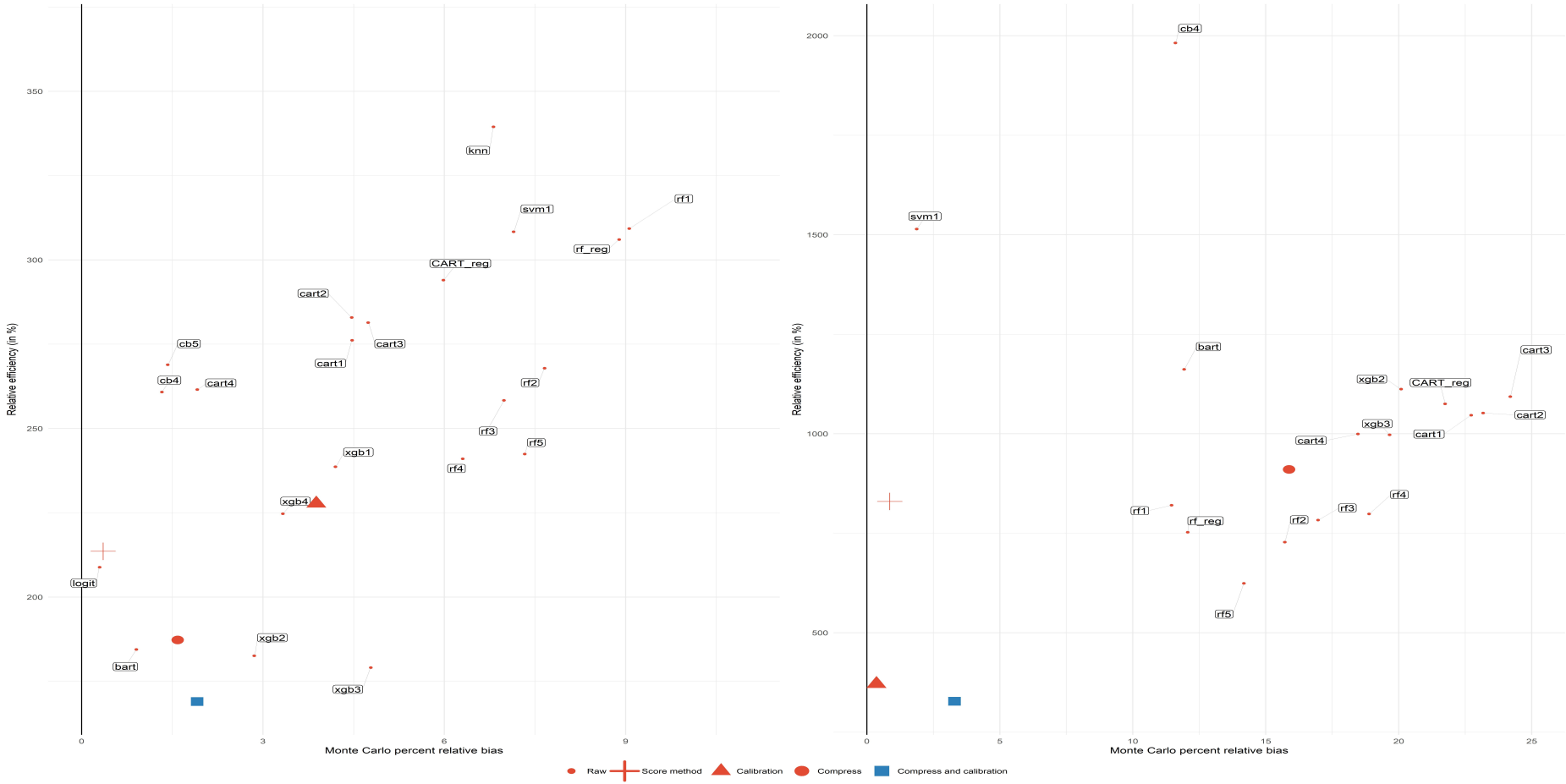



Figure 3: x (independant), y (linéaire), non-informatif, NR1 and NR2, estimateurs PSA

# Étude par simulation: Résultats

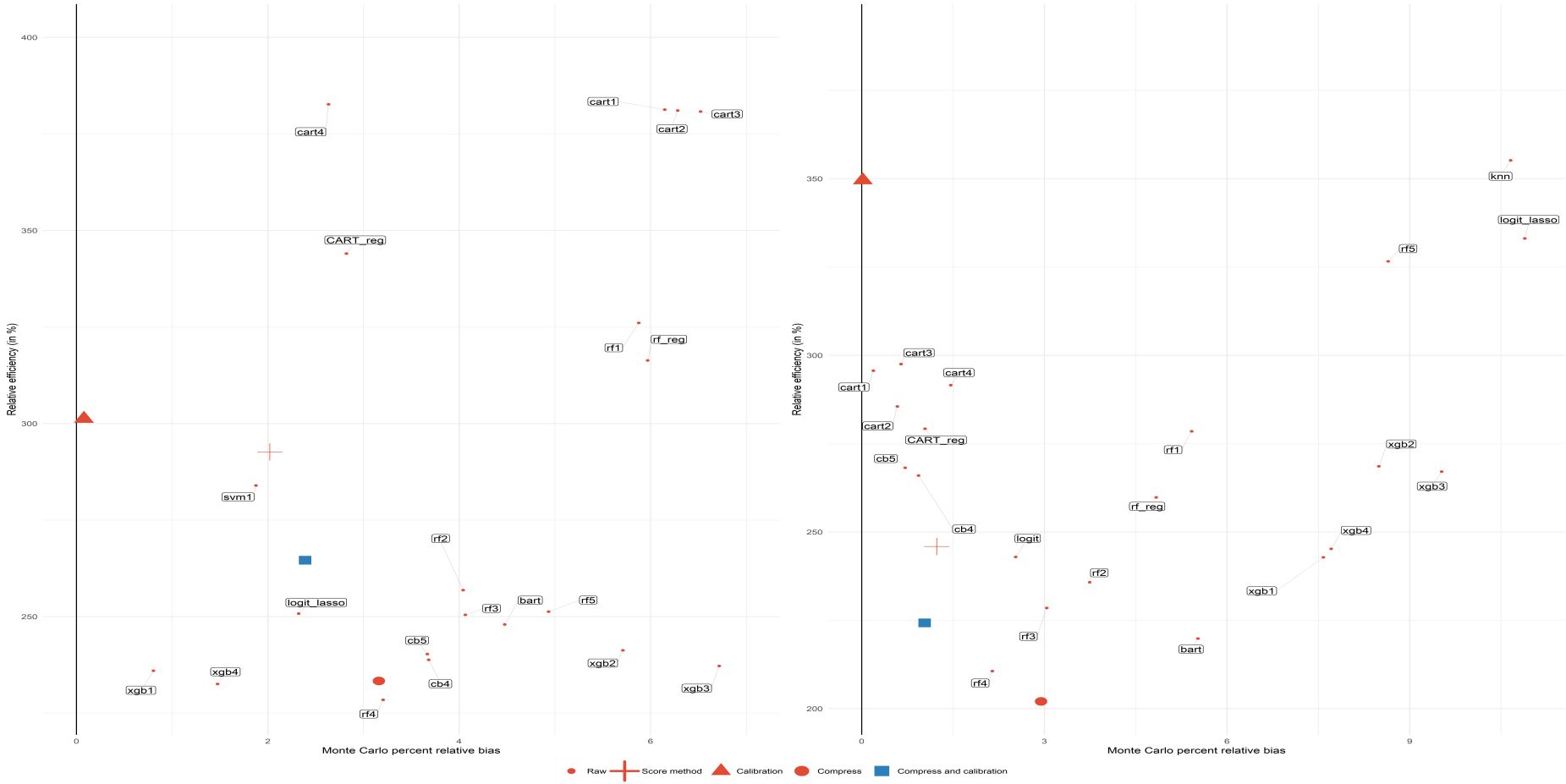



Figure 4: x (independant), y (linéaire), non-informatif, NR3 and NR4, estimateurs PSA

# Étude par simulation: Résultats

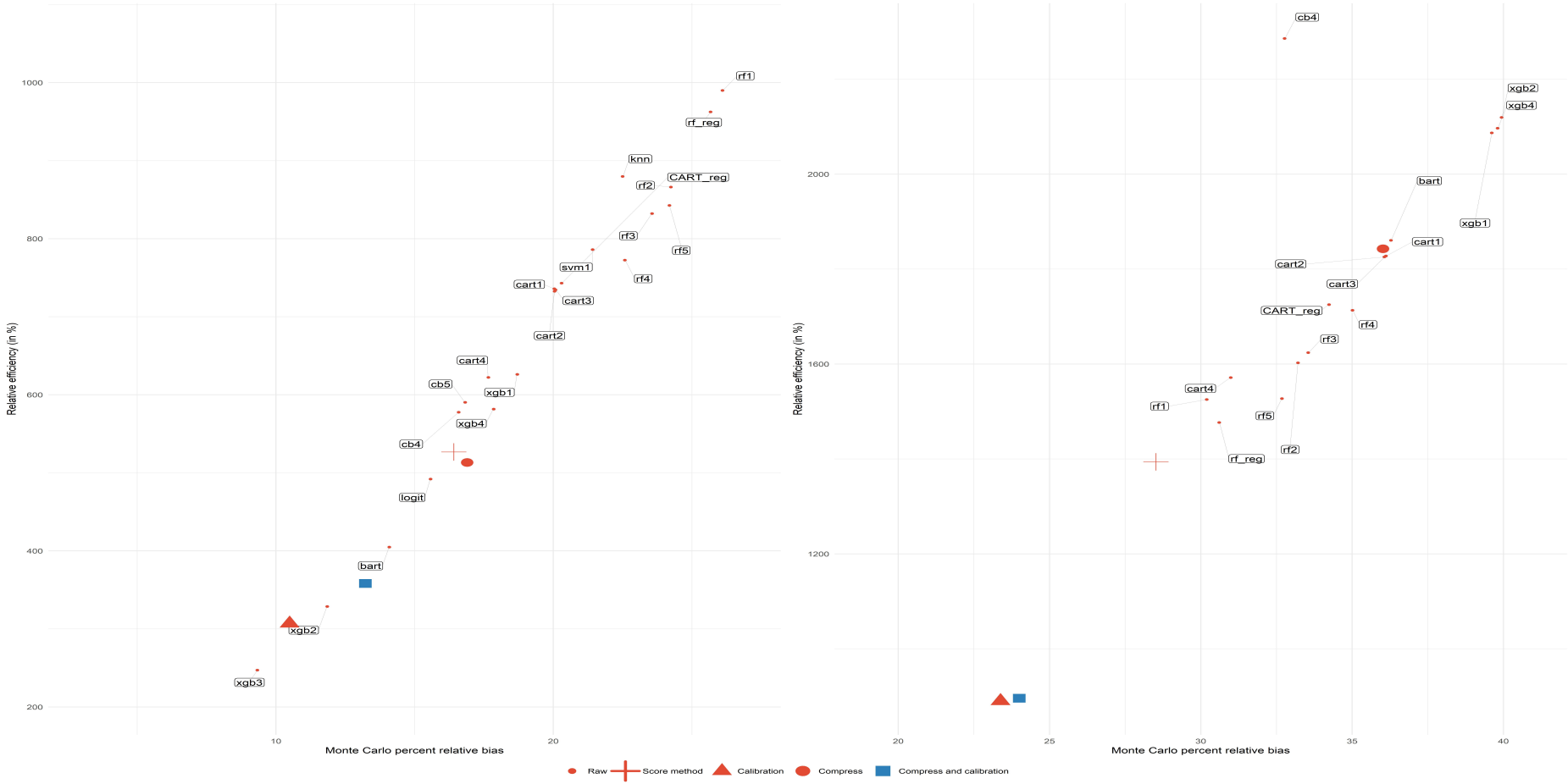



Figure 5: x (independant), y (linéaire), non-informatif, NR5 and NR6, estimateurs PSA

# Étude par simulation: Résultats

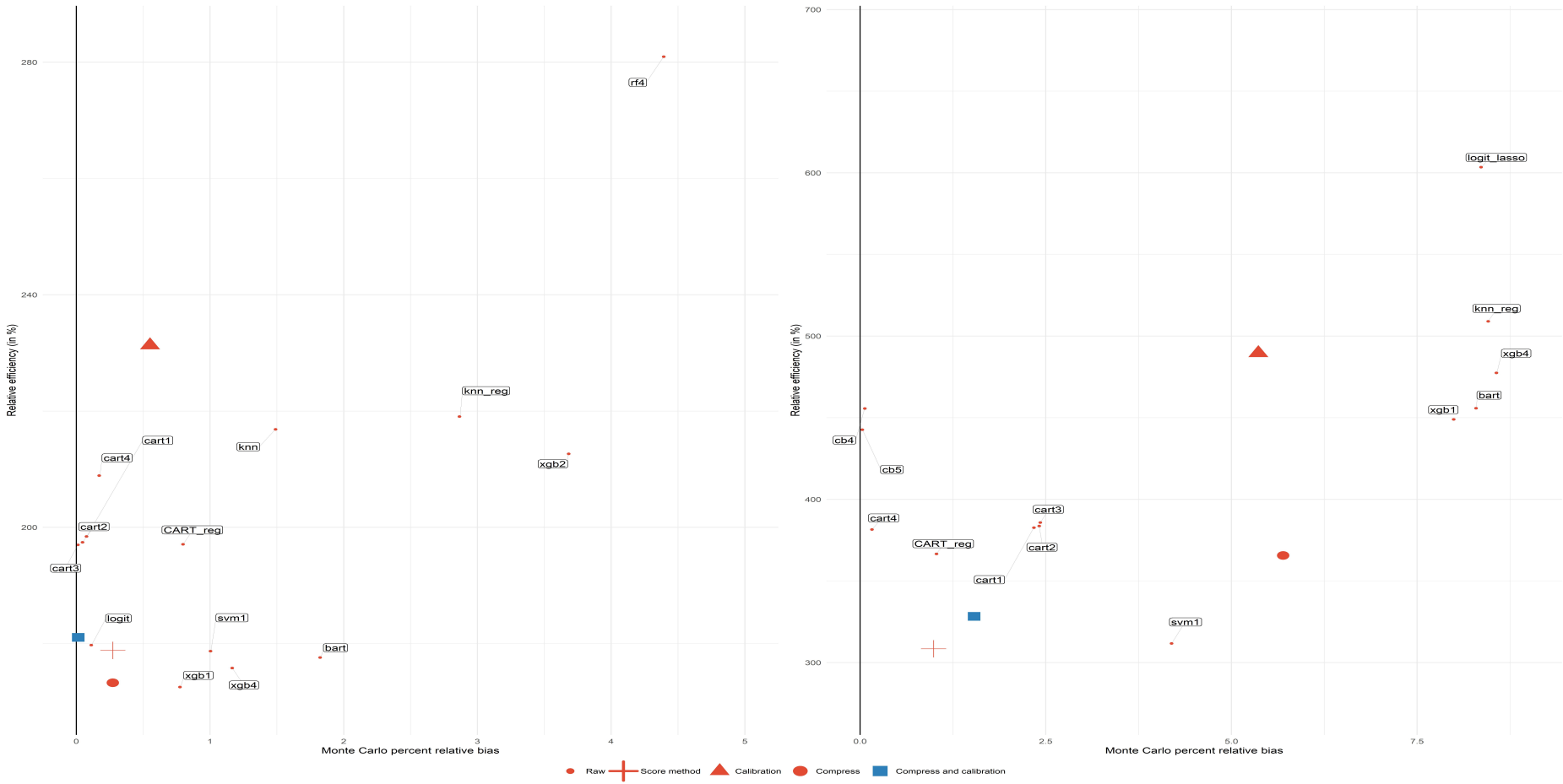



Figure 6: x (dépendant), y (non-linéaire), non-informatif, NR1 and NR2, estimateurs PSA

# Étude par simulation: Résultats

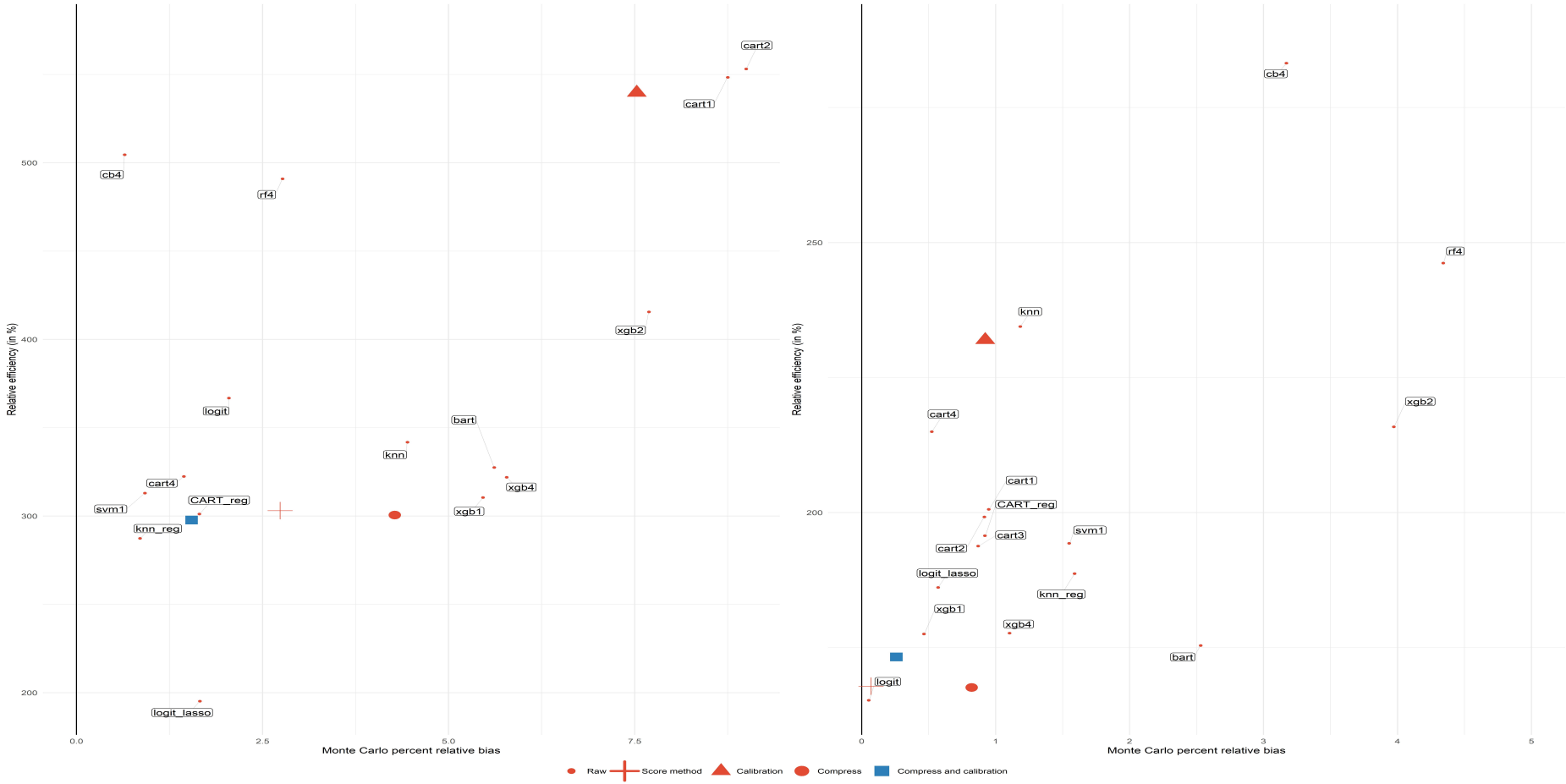



Figure 7: x (dépendant), y (non-linéaire), non-informatif, NR3 and NR4, estimateurs PSA

# Étude par simulation: Résultats

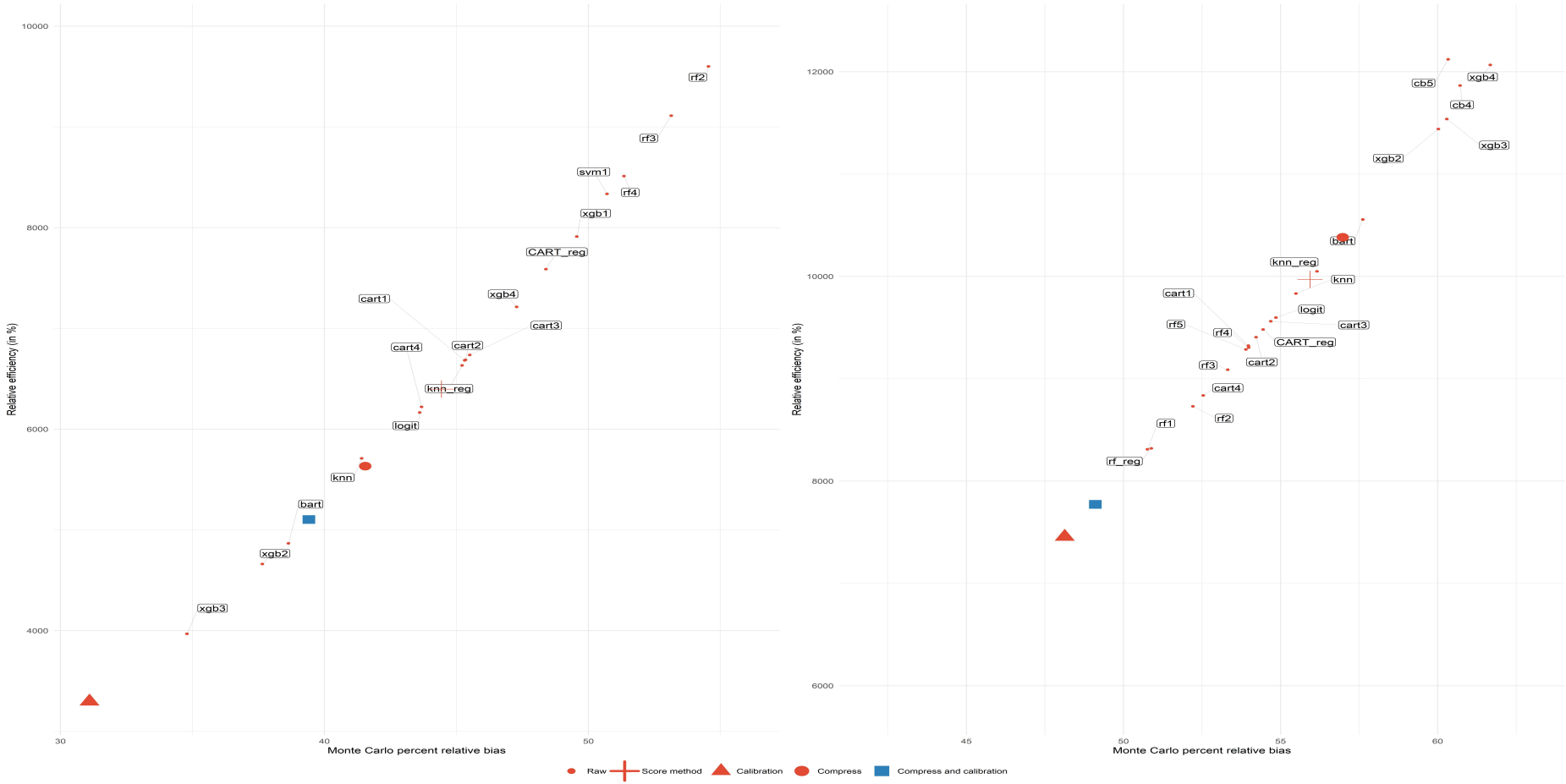



Figure 8: x (dépendant), y (non-linéaire), non-informatif, NR5 and NR6, estimateurs PSA

# Conclusions

- L'utilisation de la méthode la plus prédictive ne conduit pas nécessairement au meilleur estimateur (le plus efficace) d'un total de population.
- Les méthodes d'aggrégation se sont bien comportées dans nos expériences. Des travaux additionnels sont nécessaires.
- Méthodes d'ensemble liées à des procédures d'estimation à la multi-robustesse (par exemple, Han et Wang, 2013 ; Chen et Haziza, 2017) et à l'algorithme Superlearner (van der laan et al., 2007)
- Les résultats théoriques sur la converence des estimateurs PSA feront l'objet de travaux futurs.
- Des critères différents pour, par exemple, les arbres de régression, sont actuellement sous étude.