

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Inférence statistique pour une moyenne de population finie avec imputation fondée sur l'apprentissage automatique pour données d'enquête manquantes

par Mehdi Dagdoug et David Haziza

Date de diffusion : le 8 septembre 2025



Statistique
Canada

Statistics
Canada

Canada

Inférence statistique pour une moyenne de population finie avec imputation fondée sur l'apprentissage automatique pour données d'enquête manquantes¹

Mehdi Dagdoug et David Haziza²

Résumé

Les offices nationaux de statistique ont de plus en plus adopté l'apprentissage automatique (AA) en raison de son potentiel à améliorer les estimations issues des enquêtes. Les techniques d'AA offrent des avantages considérables, notamment la capacité de traiter des données de grande dimension et de capter des relations complexes et non linéaires, améliorant ainsi la qualité globale des statistiques d'enquête. Dans cet article, suivant l'approche de Chernozhukov et al. (2018), nous décrivons un cadre d'apprentissage automatique doublement débiaisé permettant de réaliser une inférence statistique valide lorsque les estimateurs imputés proviennent de procédures d'AA. Les résultats de simulations indiquent que le cadre proposé fonctionne bien dans un large éventail de scénarios.

Mots-clés : Ajustement croisé; apprentissage automatique débiaisé; estimation doublement robuste; non-réponse partielle; estimation de la variance.

1. Introduction

L'apprentissage automatique (AA) a suscité un vif intérêt au sein des offices nationaux de statistique en raison de son potentiel à améliorer la qualité des estimations issues des enquêtes. Les techniques d'AA offrent plusieurs avantages, notamment la capacité de traiter des données de grande dimension et une robustesse face à des relations complexes et non linéaires. En particulier, les procédures d'AA peuvent s'avérer utiles pour le traitement des données manquantes. C'est l'objet de cet article. Malgré ces atouts, la littérature sur l'inférence statistique lorsque les valeurs imputées proviennent de procédures d'AA demeure relativement limitée. Une exception notable est l'article de Dagdoug et al. (2025), qui ont examiné le problème de l'imputation au moyen d'arbres de régression et de forêts aléatoires.

Dans cet article, nous nous intéressons à l'estimation de la moyenne d'une population finie pour une variable d'intérêt Y , sujette à des valeurs manquantes. Si une méthode paramétrique (par ex., la régression linéaire) est utilisée pour produire un ensemble de valeurs imputées, l'estimateur imputé de la moyenne est convergent à la vitesse $1/\sqrt{n}$, à condition que le premier moment du modèle d'imputation soit correctement spécifié. En présence d'une mauvaise spécification du premier moment, l'estimateur imputé peut présenter un biais important, et ce, même asymptotiquement.

En ce qui concerne les méthodes d'AA, bien qu'elles soient robustes à une mauvaise spécification du modèle, les estimateurs imputés obtenus à partir de ces approches n'atteignent généralement pas la vitesse de convergence paramétrique $1/\sqrt{n}$. De nombreuses méthodes d'AA, y compris celles basées sur les arbres de régression, les réseaux de neurones ou les méthodes à noyau, convergent à des vitesses plus lentes que $1/\sqrt{n}$, souvent à des vitesses qui dépendent de la régularité et de la dimension de la fonction à estimer. Cette lente vitesse est le prix à payer pour avoir formulé des hypothèses plus faibles. Par conséquent, le développement d'outils d'inférence statistique peut s'avérer plus complexe. En effet, les estimateurs ponctuels issus

¹Ce travail a été financé par des subventions du Conseil de recherches en sciences naturelles et en génie du Canada.

²Mehdi Dagdoug, McGill University, Montréal, Québec; David Haziza, Université d'Ottawa, Ottawa, Ontario

de procédures d'AA peuvent souffrir d'un biais de petit échantillon en présence de surapprentissage. De plus, le développement d'estimateurs de variance qui tiennent compte à la fois du plan de sondage et de la non-réponse, pour des procédures d'AA quelconques, peut se révéler très complexe. Enfin, puisque les estimateurs imputés convergent à des vitesses plus lentes, ils ne satisfont pas naturellement aux conditions asymptotiques requises par le théorème central limite. Cela rend difficile la justification d'intervalles de confiance normaux à $100(1 - \alpha)\%$ comme c'est le cas pour les estimateurs issus de méthodes paramétriques.

Dans un article fondateur, Chernozhukov et al. (2018) ont montré que les défis associés à l'AA peuvent être surmontés lorsque ces méthodes sont combinées à des techniques telles que l'apprentissage automatique double. C'est là que les estimateurs doublement robustes (DR) jouent un rôle clé. Les estimateurs DR offrent deux possibilités de réduire le biais potentiel lié à la non-réponse en s'appuyant sur deux modèles distincts : (i) le modèle d'imputation, qui prédit la variable d'intérêt à partir d'un ensemble de covariables entièrement observées, et (ii) le modèle de score de propension, qui estime la probabilité de réponse étant donné des covariables. Un avantage important des estimateurs DR est qu'ils restent convergents si au moins l'un de ces deux modèles est correctement spécifié (voir, par ex., Robins et al., 1994; Haziza et Rao, 2006; Kang et Schafer, 2008; Cao et al., 2009; Kim et Haziza, 2014). De plus, lorsqu'ils sont combinés à des procédures d'AA, les estimateurs DR présentent une autre caractéristique intéressante: ils peuvent atteindre la vitesse de convergence paramétrique, malgré les vitesses plus lentes généralement observées pour les estimateurs issus de l'AA. Cette propriété découle de l'orthogonalité de Neyman — un ingrédient clé mis en évidence par Chernozhukov et al. (2018) dans le contexte de l'estimation d'effets causaux — qui garantit que les erreurs d'estimation des fonctions auxiliaires issues de l'AA (à savoir, le modèle d'imputation et le modèle de score de propension) n'ont qu'un effet de second ordre sur l'estimateur ponctuel d'un paramètre de population finie. En d'autres termes, tant que les modèles d'AA convergent à une vitesse suffisamment rapide, l'estimateur DR conserve la convergence en $1/\sqrt{n}$ nécessaire à une inférence valide. La combinaison de l'estimation DR et de la validation croisée en blocs (cross-fitting) réduit encore le biais de surapprentissage, permettant ainsi une inférence asymptotiquement normale.

2. Inférence statistique basée sur un modèle d'imputation unique

Considérons une population finie $U = \{1, \dots, k, \dots, N\}$ de taille N . Nous nous intéressons à l'estimation de la moyenne de population finie d'une variable d'intérêt Y , $\mu = N^{-1} \sum_{k \in U} y_k$. Nous sélectionnons un échantillon S , de taille n , selon un plan de sondage $\mathcal{F}(\mathbf{I} | \mathbf{Z})$, où $\mathbf{Z}$ désigne la matrice des variables du plan disponibles avant le sondage pour toutes les unités de la population, et $\mathbf{I} = (I_1, \dots, I_k, \dots, I_N)^\top$ désigne le N -vecteur des indicatrices de sélection dans l'échantillon, tel que $I_k = 1$ si $k \in S$ et $I_k = 0$ sinon. Les probabilités d'inclusion du premier et du deuxième ordre sont définies respectivement par $\pi_k = \mathbb{E}(I_k | \mathbf{Z})$ et $\pi_{k\ell} = \mathbb{E}(I_k I_\ell | \mathbf{Z})$. L'estimateur sur l'échantillon complet de μ est l'estimateur de Horvitz-Thompson donné par $\hat{\mu}_F = N^{-1} \sum_{k \in S} w_k y_k$, où $w_k := \pi_k^{-1}$ désigne le poids de sondage associé à l'unité k .

En pratique, la variable Y est sujette à des valeurs manquantes. Soit r_k une indicatrice de réponse pour l'unité $k \in U$ telle que $r_k = 1$ si y_k est observé et $r_k = 0$ si y_k est manquant. Soit $\mathbf{r}$ le N -vecteur contenant r_k en sa k -ième composante. Nous définissons l'ensemble des répondants à la variable Y comme $S_r := \{k \in S ; r_k = 1\}$ et l'ensemble des non-répondants comme $S_m := \{k \in S ; r_k = 0\}$.

Soit $\mathbf{x}_k$ un vecteur de variables entièrement observées associé à $k \in U$. Nous limitons notre attention aux plans de sondage non informatifs ; c'est-à-dire, $\mathbb{P}[I_k = 1 | \mathbf{x}_k, y_k] = \mathbb{P}[I_k = 1 | \mathbf{x}_k]$, voir par ex. Pfeffermann et Sverchkov (2009). Pour rendre le plan de sondage non informatif, il suffit d'inclure les variables de plan $\mathbf{Z}$ dans le modèle d'imputation, si approprié. Les vecteurs $[r_k, y_k, \mathbf{x}_k^\top]^\top$ $k \in U$ sont supposés indépendants et identiquement distribués (i.i.d.). Nous supposons en outre que le modèle de non-réponse satisfait : (i) L'hypothèse de positivité, c.-à-d., il existe $\delta > 0$ tel que $\mathbb{P}[r_k = 1 | \mathbf{x}_k] > \delta$, presque sûrement ; (ii) L'hypothèse de données manquantes au hasard (MAR, Rubin, 1976), c.-à-d., $\mathbb{P}[r_k = 1 | y_k, \mathbf{x}_k] = \mathbb{P}[r_k = 1 | \mathbf{x}_k] := p(\mathbf{x}_k)$ où p désigne la fonction de score de propension ; (iii) La propriété forte d'invariance (Beaumont et Haziza, 2016), qui implique que $\mathbb{P}[r_k = 1 | \mathbf{x}_k, \mathbf{I}] = \mathbb{P}[r_k = 1 | \mathbf{x}_k]$.

La relation entre la variable d'intérêt et les variables $\mathbf{x}$ peut être décrite comme

$$y_k = m(\mathbf{x}_k) + \epsilon_k, \quad (1)$$

où $m(\mathbf{x}) := \mathbb{E}[Y|\mathbf{x}]$ désigne la fonction de régression inconnue. Les erreurs ϵ_k dans (1) sont supposées être des variables aléatoires i.i.d., satisfaisant $\mathbb{E}[\epsilon_k|\mathbf{x}_k] = 0$ et $\mathbb{V}(\epsilon_k|\mathbf{x}_k) = \sigma_k^2$. Dans les Sections 2.1-2.3, nous revenons sur le problème de l'inférence statistique pour deux procédures d'imputation couramment rencontrées : l'imputation par régression linéaire (LRI) et l'imputation par arbre de régression (RTI).

2.1. Estimation ponctuelle

Un estimateur de μ après imputation est défini par

$$\hat{\mu}_I = \frac{1}{N} \sum_{k \in S} w_k \tilde{y}_k, \quad (2)$$

où $\tilde{y}_k = r_k y_k + (1 - r_k) \hat{m}(\mathbf{x}_k)$, $\hat{m}(\mathbf{x}_k)$ désignant la valeur imputée utilisée pour remplacer le y_k manquant. La forme (2) est pratique pour les utilisateurs de données car $\hat{\mu}_I$ s'exprime comme une moyenne pondérée des valeurs observées et imputées, ce qui la rend facilement calculable en pratique.

Pour la LRI, les valeurs imputées $\hat{m}(\mathbf{x}_k)$ sont données par $\hat{m}_{LR}(\mathbf{x}_k) = \mathbf{x}_k^\top \hat{\beta}_r$, pour $k \in S_m$, où

$$\hat{\beta}_r = \left(\sum_{k \in S} w_k r_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S} w_k r_k \mathbf{x}_k y_k$$

est l'estimateur des moindres carrés ordinaires de β basé sur les unités répondantes. Sous des conditions de régularité faibles, l'estimateur imputé (2) basé sur la LRI est convergent à la vitesse $1/\sqrt{n}$ pour μ , à condition que le premier moment du modèle soit correctement spécifié.

En ce qui concerne la RTI, la prédiction en un point $\mathbf{x}$, notée $\hat{m}_{tree}(\mathbf{x})$, est obtenue en partitionnant l'espace des prédicteurs — engendré par les variables $\mathbf{x}$ observées sur les unités répondantes — en nœuds terminaux disjoints selon un critère donné. La valeur imputée pour $k \in S_m$ appartenant au nœud terminal A est alors calculée comme la moyenne des valeurs y des unités répondantes appartenant au même nœud :

$$\hat{m}_{tree}(\mathbf{x}) := \frac{1}{n_{rA}} \sum_{k \in S_r: \mathbf{x}_k \in A} y_k, \quad \mathbf{x} \in A,$$

où n_{rA} désigne le nombre de répondants appartenant au nœud terminal A . Dagdoug et al. (2025) ont montré que l'estimateur $\hat{\mu}_I$, basé soit sur la RTI, soit sur les forêts aléatoires (RF), est convergent en moyenne quadratique pour μ . Cependant, bien qu'ils n'aient pas fourni de vitesse de convergence explicite, celle-ci est généralement plus lente que la vitesse paramétrique $1/\sqrt{n}$. Il en va de même pour la plupart des méthodes d'AA.

2.2. Vitesse de convergence: Une étude empirique

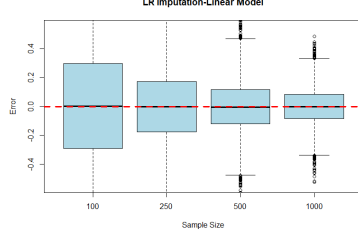
Pour illustrer la lente vitesse de convergence de l'imputation non paramétrique, nous avons mené une étude de simulation limitée. Nous avons généré 20 prédicteurs indépendants $X_1, \dots, X_{20}$, chacun selon une loi Gamma avec un paramètre de forme égal à 1 et un paramètre d'échelle égal à 10. Nous avons ensuite répété $R = 10,000$ itérations du processus suivant :

- 1) Une population finie U de taille $N = 20,000$ a été générée, comprenant 2 variables d'enquête, Y_1 et Y_2 . Ces variables ont été générées selon deux modèles : un modèle linéaire et un modèle non linéaire.

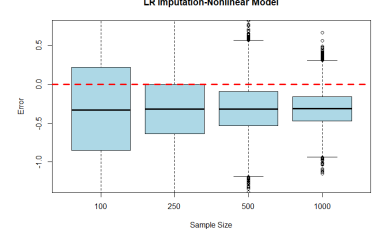
Figure 2.1-1

Distribution de l'erreur de l'estimateur imputé sur 10 000 itérations

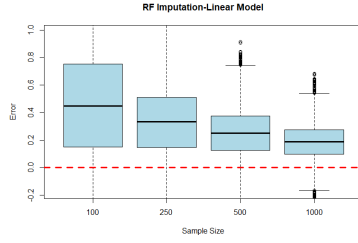
(b)
Imputation par la régression linéaire:
Modèle linéaire



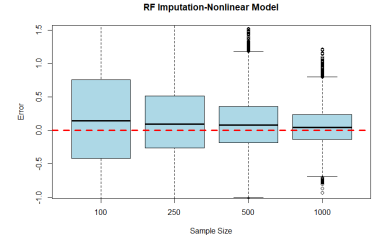
Imputation par la régression linéaire:
Modèle non-linéaire



(d)
Imputation par forêts aléatoires:
Modèle linéaire



Imputation par forêts aléatoires:
Modèle non-linéaire



Plus précisément,

$$Y_1 = 0.15 \{(x_1 - 10) + (x_2 - 10) + (x_3 - 10)\} + \mathcal{N}(0, 4); \quad (\text{Modèle linéaire})$$

$$Y_2 = 5, \mathbb{I}(x_1 > 10) + 3 \sin(5(x_2 - 10)) + 0.02((x_1 - 10)^2 - 100) + 2, \mathbb{I}(x_2 > 10) \cos(7(x_3 - 10)) + \mathcal{N}(0, 4); \quad (\text{Modèle non linéaire})$$

Dans les deux modèles, le coefficient de détermination R^2 a été fixé à 0.5.

- 2) Un échantillon S de taille $n \in 100; 250; 500; 1000$ a été sélectionné à partir de U selon un sondage aléatoire simple sans remise.
- 3) Dans chaque échantillon, les indicatrices de réponse r_k ont été générées indépendamment selon une loi de Bernoulli de probabilité

$$p(\mathbf{x}_k) = \frac{1}{2} \left\{ 1 + \frac{0.48 + \sin\left(\frac{x_1 - 10}{10}\right) + \sin\left(\frac{x_2 - 10}{10}\right) + \sin\left(\frac{x_3 - 10}{10}\right)}{3} \right\}, \quad k \in U. \quad (3)$$

Les paramètres de (3) ont été choisis de façon à obtenir une proportion de valeurs manquantes d'environ 50% dans chaque échantillon.

- 4) Dans chaque échantillon, nous avons calculé l'estimateur imputé $\hat{\mu}_I$ défini en (2) à partir de la LRI et des forêts aléatoires (RF). Pour RF, le nombre minimal d'observations dans chaque nœud terminal et le nombre d'arbres ont été fixés à $n_0 = 10$ et $B = 250$, respectivement. L'ensemble complet des prédictors $X_1, \dots, X_{20}$ a été inclus dans la procédure d'ajustement des modèles.
- 5) Dans chaque échantillon, nous avons calculé l'erreur $\hat{\mu}_I - \mu$ pour chaque valeur de n sous LR et RF. La distribution de $\hat{\mu}_I - \mu$ pour les différents scénarios est illustrée à la Figure 2.1-1.

D'après la Figure 2.1-1, lorsque le vrai modèle est linéaire, l'imputation par LR produit des estimateurs presque sans biais, avec des erreurs centrées autour de zéro et une variance qui diminue lorsque la taille de l'échantillon augmente. En revanche, l'imputation par RF présente un biais sous le modèle linéaire, particulièrement pour les petites tailles d'échantillon. Tant le biais que la variance diminuent avec la taille de l'échantillon, ce qui suggère que l'estimateur imputé sous RF est convergent. Toutefois, nous notons que la vitesse de convergence semble relativement lente. Lorsque le vrai modèle est non linéaire, l'imputation par RF donne de bons résultats, produisant des estimations presque sans biais et avec une variance décroissante à mesure que n augmente, alors que l'imputation par LR sous-estime systématiquement, avec un biais qui reste à peu près constant et une variance qui se réduit avec des échantillons plus grands. Cela n'est pas surprenant, puisque l'estimateur imputé par LR n'est pas convergent dans le cas d'un modèle non linéaire en raison de la mauvaise spécification du premier moment.

2.3. Estimation de la variance

Plusieurs approches ont été proposées dans la littérature pour estimer la variance de $\hat{\mu}_I$; voir Haziza et Vallée (2020) pour un aperçu des procédures d'estimation de la variance pour les données d'enquête imputées. Dans cet article, nous considérons l'approche dite « renversée » proposée initialement par Fay (1991) et Shao et Steel (1999). La variance totale de $\hat{\mu}_I$ peut s'exprimer comme

$$\mathbb{V}_{tot}(\hat{\mu}_I) = \mathbb{E}\mathbb{V}(\hat{\mu}_I \mid \mathbf{X}, \mathbf{r}, \mathbf{y}) + \mathbb{E}\mathbb{V}\{\mathbb{E}(\hat{\mu}_I - \mu \mid \mathbf{X}, \mathbf{r}, \mathbf{y}) \mid \mathbf{X}, \mathbf{r}\}, \quad (4)$$

où $\mathbf{y}$ est le N -vecteur dont le k -ième élément est y_k , et $\mathbf{X}$ est la matrice dont la k -ième ligne est $\mathbf{x}_k^\top$. Sous des conditions de régularité faibles, la contribution à la variance totale du second terme du côté droit de (4) est négligeable si la fraction d'échantillonnage n/N est négligeable (Shao et Steel, 1999). Dans cette section, nous supposons que n/N est négligeable de sorte que nous omettons le calcul du second terme du côté droit de (4). En utilisant un développement de Taylor au premier ordre, un estimateur de $\mathbb{E}\mathbb{V}(\hat{\mu}_I \mid \mathbf{X}, \mathbf{r}, \mathbf{y})$, noté $\hat{\mathbb{V}}_1(\hat{\mu}_I)$, est donné par

$$\hat{\mathbb{V}}_1(\hat{\mu}_I) := \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\hat{\phi}_k}{\pi_k} \frac{\hat{\phi}_\ell}{\pi_\ell}, \quad (5)$$

où $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$. La pseudo-valeur $\hat{\phi}_k, k \in S$, dans (5) est donnée par

$$\hat{\phi}_k := \hat{m}(\mathbf{x}_k) + r_k \hat{C}_k \{y_k - \hat{m}(\mathbf{x}_k)\}, \quad k \in S, \quad (6)$$

où $\hat{C}_k$ est égal à

$$\hat{C}_k = 1 + \left(\sum_{\ell \in S} w_\ell (1 - r_\ell) \mathbf{x}_\ell \right)^\top \left(\sum_{\ell \in S} w_\ell r_\ell \mathbf{x}_\ell \mathbf{x}_\ell^\top \right)^{-1} \mathbf{x}_k, \quad k \in S,$$

pour la LRI, et égal à

$$\hat{C}_k = \frac{\hat{N}(\mathbf{x}_k, S)}{\hat{N}(\mathbf{x}_k, S_r)}, \quad k \in S,$$

pour la RTI, où $\hat{N}(\mathbf{x}, S) := \sum_{\ell \in S} \mathbb{1}_{\{\mathbf{x}_\ell \in A(\mathbf{x})\}} \pi_\ell^{-1}$ et $\hat{N}(\mathbf{x}_k, S_r)$ est défini de manière similaire.

Nous faisons les remarques suivantes : (i) Pour la RTI, la partition de l'espace des prédicteurs (qui varie d'un échantillon à l'autre) a été traitée comme fixe dans le développement de Taylor au premier ordre. (ii) L'estimateur de variance (5) correspond à l'estimateur de variance d'Horvitz–Thompson usuel appliqué aux pseudo-valeurs $\hat{\phi}_k$. (iii) Le terme $\hat{C}_k$ dans (6) peut être interprété comme une estimation implicite de la propension à répondre $p(\mathbf{x}_k)$, puisque, sous la LRI et la RTI, l'estimateur imputé $\hat{\mu}_I$ donné par (2) peut s'écrire sous la forme d'un estimateur ajusté par score de propension, c'est-à-dire $\hat{\mu}_I = \sum_{k \in S_r} (w_k / \hat{C}_k) y_k$.

Comme illustré dans Dagdoug et al. (2025) dans le contexte des arbres de régression et des forêts aléatoires, l'estimateur $\hat{\mathbb{V}}_1(\hat{\mu}_I)$ peut sous-estimer substantiellement la vraie variance, $\mathbb{E}\mathbb{V}(\hat{\mu}_I \mid \mathbf{X}, \mathbf{r}, \mathbf{y})$, en présence de surapprentissage. Le surapprentissage peut produire des résidus artificiellement petits, $y_k - \hat{m}(\mathbf{x}_k)$, menant

à des estimateurs de variance biaisés vers le bas. Dans le cas de la LRI, le surapprentissage peut survenir lorsque le nombre de prédicteurs est grand relativement à la taille de l'échantillon. Dans le cas de la RTI, le surapprentissage peut survenir lorsque l'arbre est trop profond ou effectue trop de divisions par rapport aux données disponibles. Pour pallier ce problème, les résidus $y_k - \hat{m}(\mathbf{x}_k)$ sont remplacés par des « résidus honnêtes » calculés sur un jeu de test distinct. Dagdoug et al. (2025) ont proposé de construire ces résidus honnêtes en utilisant une procédure de validation croisée en L volets, décrite comme suit :

1. Diviser l'échantillon traité S en L sous-échantillons mutuellement exclusifs, $\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_L$. Chaque sous-échantillon $\mathcal{F}_j$ servira une fois de jeu de test, $j = 1, \dots, L$.
2. Pour chaque sous-échantillon $j = 1, \dots, L$, définir le jeu d'entraînement comme $S_{(-j)} = S \setminus \mathcal{F}_j$. Utiliser les données de $S_{(-j)}$ pour ajuster le modèle d'imputation
3. Pour chaque unité k dans le jeu de test $\mathcal{F}_j$, obtenir la prédiction $\hat{m}_{(-j)}(\mathbf{x}_k)$.
4. Calculer les résidus honnêtes pour les unités du jeu de test comme $e_{k,cv} = y_k - \hat{m}_{(-j)}(\mathbf{x}_k)$, pour $k \in \mathcal{F}_j$.

Dagdoug et al. (2025) ont montré empiriquement que, dans le cas d'une imputation par arbres de régression, l'estimateur $\hat{V}_1(\hat{\mu}_I)$ basé sur les pseudo-valeurs

$$\hat{\phi}_{k,cv} := \hat{m}(\mathbf{x}_k) + r_k \hat{C}_k e_{k,cv}, \quad k \in S, \quad (7)$$

a conduit à une amélioration substantielle en termes de biais et a atteint des taux de couverture d'intervalles de confiance basés sur la loi normale proches du niveau nominal. Un inconvénient de l'approche de Dagdoug et al. (2025) est qu'il semble difficile de la généraliser à des méthodes d'apprentissage automatique arbitraires. La procédure de validation croisée décrite ci-dessus est étroitement liée à l'ajustement croisé (par exemple, Chernozhukov et al., 2018; Wager et Athey, 2018); voir Section 3.3.

3. Un cadre de travail doublement débiaisé

Dans cette section, nous étendons le cadre du double apprentissage automatique débiaisé (« double debiased machine learning ») de Chernozhukov et al. (2018) au cas de populations finies.

3.1. Estimation oracle doublement robuste

Nous commençons par considérer un estimateur « oracle » doublement robuste (ou estimateur AIPW) de μ basé sur les quantités vraies $m(\mathbf{x}_k)$ et $p(\mathbf{x}_k)$. Il est défini comme

$$\begin{aligned} \hat{\mu}_{aipw}(m, p) &= \frac{1}{N} \left\{ \sum_{k \in S} w_k m(\mathbf{x}_k) + \sum_{k \in S_r} w_k \frac{y_k - m(\mathbf{x}_k)}{p(\mathbf{x}_k)} \right\}, \\ &= \frac{1}{N} \sum_{k \in S} w_k \eta_k, \end{aligned} \quad (8)$$

où

$$\eta_k = m(\mathbf{x}_k) + \frac{r_k}{p(\mathbf{x}_k)} \{y_k - m(\mathbf{x}_k)\}, \quad \text{pour } k \in S. \quad (9)$$

Il est facile de voir que $\hat{\mu}_{aipw}(m, p)$ est sans biais pour μ , au sens où $\mathbb{E}\{\hat{\mu}_{aipw}(m, p) - \mu\} = 0$, où $\mathbb{E}(\cdot)$ désigne l'espérance par rapport à la distribution jointe induite par le modèle d'imputation, le plan d'échantillonnage et le modèle de score de propension. De plus, sous des conditions de régularité faibles, $\hat{\mu}_{aipw}(m, p)$ est convergent à la vitesse $1/\sqrt{n}$ pour μ .

En utilisant la décomposition de variance (4), la variance totale de $\hat{\mu}_{aipw}(m, p)$ peut s'exprimer comme

$$\begin{aligned} \mathbb{V}_{tot}(\hat{\mu}_{aipw}(m, p)) &:= \mathbb{E} \left\{ \frac{1}{N^2} \sum_{k \in U} \sum_{\ell \in U} \Delta_{k\ell} \frac{\eta_k}{\pi_k} \frac{\eta_\ell}{\pi_\ell} \right\} + \frac{1}{N^2} \sum_{k \in U} \frac{1 - p(\mathbf{x}_k)}{p(\mathbf{x}_k)} \sigma_k^2 \\ &\equiv \mathbb{V}_1(\hat{\mu}_{aipw}(m, p)) + \mathbb{V}_2(\hat{\mu}_{aipw}(m, p)). \end{aligned} \quad (10)$$

Sous des conditions de régularité faibles, le premier terme du côté droit de (10) est d'ordre $O(1/n)$, tandis que le second est d'ordre $O(1/N)$. Par conséquent, la contribution du second terme à la variance totale est d'ordre $O(n/N)$, ce qui est négligeable lorsque la fraction d'échantillonnage n/N est négligeable.

En supposant que $\mathbb{V}(\epsilon_k | \mathbf{x}_k) = \sigma_k^2$ est connue, un estimateur sans biais de $\mathbb{V}_{tot}(\hat{\mu}_{aipw}(m, p))$ peut être obtenu en estimant séparément chaque terme de (10) sans biais, ce qui conduit à

$$\hat{\mathbb{V}}_{tot}(\hat{\mu}_{aipw}(m, p)) := \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\eta_k}{\pi_k} \frac{\eta_\ell}{\pi_\ell} + \frac{1}{N^2} \sum_{k \in S_r} w_k \frac{1 - p(\mathbf{x}_k)}{\{p(\mathbf{x}_k)\}^2} \sigma_k^2. \quad (11)$$

Bien que (8) et (11) soient tous deux irréalisables en pratique, ils servent de base aux développements des Sections 3.2 et 3.3. Enfin, sous certaines conditions de régularité (par exemple, Chen et Rao, 2007), on peut montrer que

$$\left\{ \hat{\mathbb{V}}_{tot}(\hat{\mu}_{aipw}(m, p)) \right\}^{-1/2} (\hat{\mu}_{aipw}(m, p) - \mu) \rightarrow \mathcal{N}(0, 1).$$

3.2. Estimation doublement robuste basée sur des fonctions de nuisance estimées

Comme estimateur de μ , nous considérons un estimateur DR de la forme

$$\begin{aligned} \hat{\mu}_{aipw}(\hat{m}, \hat{p}) &= \frac{1}{N} \left\{ \sum_{k \in S} w_k \hat{m}(\mathbf{x}_k) + \sum_{k \in S_r} w_k \frac{y_k - \hat{m}(\mathbf{x}_k)}{\hat{p}(\mathbf{x}_k)} \right\}, \\ &= \frac{1}{N} \sum_{k \in S} w_k \hat{\eta}_k, \end{aligned} \quad (12)$$

où

$$\hat{\eta}_k = \hat{m}(\mathbf{x}_k) + \frac{r_k}{\hat{p}(\mathbf{x}_k)} \{y_k - \hat{m}(\mathbf{x}_k)\}, \quad \text{pour } k \in S, \quad (13)$$

avec $\hat{m}(\mathbf{x}_k)$ représentant la valeur prédite de y_k manquant, $k \in S$, obtenue à partir d'une méthode d'apprentissage automatique, et $\hat{p}(\mathbf{x}_k)$ la probabilité de réponse estimée pour $k \in S_r$ à partir d'une autre méthode d'apprentissage automatique. Comme ni $\hat{m}(\mathbf{x})$ ni $\hat{p}(\mathbf{x})$ ne sont d'intérêt principal, elles peuvent être considérées comme des fonctions de nuisance, servant d'étapes intermédiaires vers l'objectif final qui est d'estimer μ .

En supposant que la fraction d'échantillonnage n/N est négligeable, un estimateur de $\mathbb{V}_{tot}(\hat{\mu}_{aipw}(\hat{m}, \hat{p}))$, noté $\hat{\mathbb{V}}_1(\hat{\mu}_{aipw}(\hat{m}, \hat{p}))$, est donné par le premier terme du côté droit de (11) avec η_k remplacé par $\hat{\eta}_k$ donné par (13). Cependant, $\hat{\mu}_{aipw}(\hat{m}, \hat{p})$ dans (12) et son estimateur de variance associé, $\hat{\mathbb{V}}_1(\hat{\mu}_{aipw}(\hat{m}, \hat{p}))$, présentent un inconvénient majeur : ils impliquent tous deux les résidus $y_k - \hat{m}(\mathbf{x}_k)$, $k \in S_r$, ce qui les rend vulnérables à un biais en présence de surapprentissage. Pour remédier à ce problème, il est recommandé de combiner l'estimation DR avec l'ajustement croisé à la fois pour l'estimation ponctuelle et pour l'estimation de la variance. Ceci est discuté dans la section suivante.

3.3. Estimation doublement robuste basée sur l'ajustement croisé

L'ajustement croisé se déroule selon les étapes suivantes :

1. Diviser l'échantillon S en L sous-échantillons mutuellement exclusifs, $\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_L$. Chaque sous-échantillon $\mathcal{F}_j$ servira une fois d'ensemble de test, $j = 1, \dots, L$.
2. Pour chaque sous-échantillon $j = 1, \dots, L$, définir l'ensemble d'entraînement comme $S_{(-j)} = S \setminus \mathcal{F}_j$. Utiliser les données de $S_{(-j)}$ pour estimer le modèle d'imputation m avec $\hat{m}_{(-j)}$ et le modèle de probabilité de réponse p avec $\hat{p}_{(-j)}$.
3. Pour chaque unité k de l'ensemble de test $\mathcal{F}_j$, obtenir les prédictions obtenues par ajustement croisé $\hat{m}_{(-j)}(\mathbf{x}_k)$ et $\hat{p}_{(-j)}(\mathbf{x}_k)$.

Une version obtenue par ajustement croisé de $\hat{\mu}_{aipw}(\hat{m}, \hat{p})$ est donnée par

$$\begin{aligned}\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}) &= \frac{1}{N} \left\{ \sum_{k \in S} w_k \hat{m}_{(-j)}(\mathbf{x}_k) + \sum_{k \in S_r} w_k \frac{y_k - \hat{m}_{(-j)}(\mathbf{x}_k)}{\hat{p}_{(-j)}(\mathbf{x}_k)} \right\} \\ &= \frac{1}{N} \sum_{k \in S} w_k \hat{\eta}_{k,cf},\end{aligned}\tag{14}$$

où

$$\hat{\eta}_{k,cf} = \hat{m}_{(-j)}(\mathbf{x}_k) + \frac{r_k}{\hat{p}_{(-j)}(\mathbf{x}_k)} \{y_k - \hat{m}_{(-j)}(\mathbf{x}_k)\}, \quad \text{pour } k \in S.\tag{15}$$

Contrairement à l'estimateur (2) obtenu au moyen d'une seule méthode d'apprentissage automatique, l'estimateur doublement robuste $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$ est convergent à la vitesse $1/\sqrt{n}$, même si $\hat{m}(\mathbf{x})$ et $\hat{p}(\mathbf{x})$ convergent plus lentement. Par exemple, la convergence $1/\sqrt{n}$ est obtenue si $\hat{m}(\mathbf{x})$ et $\hat{p}(\mathbf{x})$ convergent toutes deux vers leurs cibles à une vitesse plus rapide que $n^{-1/4}$. Le théorème 1 ci-dessous montre que l'estimateur obtenu par ajustement croisé (14) est asymptotiquement équivalent à l'estimateur oracle (8), pourvu que les fonctions de nuisance estimées $\hat{m}(\mathbf{x}_k)$ et $\hat{p}(\mathbf{x}_k)$ convergent vers leurs cibles « suffisamment rapidement ». Par conséquent, l'estimateur obtenu par ajustement croisé $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$ hérite des propriétés de sa version oracle. En particulier, $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$ est asymptotiquement normal avec une variance asymptotiquement égale à $V_{tot}(\hat{\mu}_{aipw}(m, p))$; voir l'Expression (10).

Pour établir les propriétés théoriques de $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$, nous considérons le cadre asymptotique d'Isaki et Fuller (1982). Nous considérons une suite croissante de populations finies $\{U_v\}_{v \in \mathbb{N}}$, de tailles $\{N_v\}_{v \in \mathbb{N}}$, telle que $U_v \subset U_{v+1}$, pour tout $v \in \mathbb{N}$. À partir de U_v , un échantillon S_v est sélectionné selon le plan de sondage $\mathcal{P}_v$ pour estimer μ_v . Les probabilités d'inclusion d'ordre un et deux de $\mathcal{P}_v$ sont notées respectivement $\{\pi_{k,v}\}_{k \in U_v}$ et $\{\pi_{kl,v}\}_{k \neq l \in U_v}$. Pour alléger la notation, l'indice v sera omis autant que possible. Nous considérons les hypothèses suivantes sur le plan de sondage.

(H1) La fraction de sondage vérifie $\lim_{v \rightarrow \infty} n_v/N_v := \pi_* > 0$.

(H2) Il existe $\lambda > 0$ tel que, pour tout $v \in \mathbb{N}$, $\min_{k \in U_v} \pi_{k,v} > \lambda$.

(H3) Les covariances d'échantillonnage vérifient

$$\limsup_{v \rightarrow \infty} n_v \max_{\substack{k, l \in U_v \\ k \neq l}} |\pi_{kl,v} - \pi_{k,v} \pi_{l,v}| < \infty.$$

Théorème 1. *Considérons les suites $\{\hat{\mu}_{aipw,cf}(\hat{m}_v, \hat{p}_v)\}_{v \in \mathbb{N}}$ et $\{\hat{\mu}_{aipw,v}(m, p)\}_{v \in \mathbb{N}}$ des estimateurs DR obtenus par ajustement croisé et oracles, respectivement. Nous supposons les conditions suivantes.*

(H1) *Les suites d'estimateurs $\{\hat{m}_v\}_{v \in \mathbb{N}}$ et $\{\hat{p}_v\}_{v \in \mathbb{N}}$ satisfont la condition de vitesse suivante :*

$$\lim_{v \rightarrow \infty} \sqrt{n_v} \times \mathbb{E} \left[\frac{1}{|\mathcal{F}_j|} \sum_{k \in \mathcal{F}_j} (\hat{m}_{(-j)}(\mathbf{x}_k) - m(\mathbf{x}_k))^2 \right]^{1/2} \mathbb{E} \left[\frac{1}{|\mathcal{F}_j|} \sum_{k \in \mathcal{F}_j} (\hat{p}_{(-j)}(\mathbf{x}_k) - p(\mathbf{x}_k))^2 \right]^{1/2} = 0,\tag{16}$$

pour tout $j = 1, \dots, L$.

(H2) Les probabilités prédites sont bornées loin de 0, c'est-à-dire qu'il existe $\rho > 0$ tel que, pour tout $v \in \mathbb{N}$, $\hat{p}_v(\mathbf{x}_k) > \rho$.

Alors,

$$\sqrt{n_v} \{ \hat{\mu}_{aipw,cf,v}(\hat{m}_v, \hat{p}_v) - \hat{\mu}_{aipw,v}(m, p) \} = o_p(1).$$

Le théorème 1 suggère que la variabilité associée aux fonctions de nuisance estimées $\hat{m}(\mathbf{x}_k)$ et $\hat{p}(\mathbf{x}_k)$ peut être ignorée sans risque. Par conséquent, un estimateur de $\mathbb{V}_1(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}))$ dans (10) est obtenu à partir du premier terme du côté droit de (11) en remplaçant η_k par $\hat{\eta}_{k,cf}$ donné par (15). Nous notons cet estimateur de variance $\hat{\mathbb{V}}_1(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}))$. Remarquons que $\hat{\eta}_{k,cf}$ implique l'inverse des probabilités de réponse estimées par ajustement croisé, $\hat{p}_{(-j)}(\mathbf{x}_k)$. Des valeurs très petites de ces probabilités peuvent affecter la performance de $\hat{\mathbb{V}}_1(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}))$. Dans de tels cas, une méthode de troncature pourrait être appliquée afin d'atténuer l'impact des valeurs extrêmement faibles de $\hat{p}_{(-j)}(\mathbf{x}_k)$. Cependant, des recherches supplémentaires sont nécessaires pour déterminer un seuil approprié. On pourrait aussi utiliser la stratification selon le score de propension — aussi appelée méthode du score (par ex., Little, 1986; Lunceford et Davidian, 2004; Haziza et Beaumont, 2007) — pour construire C cellules homogènes par rapport à $\hat{p}_{(-j)}(\mathbf{x}_k)$ et utiliser le taux de réponse dans chaque cellule à la place de $\hat{p}_{(-j)}(\mathbf{x}_k)$ dans (15). Là encore, cela nécessite des investigations supplémentaires.

Pour obtenir un estimateur de $\mathbb{V}_2(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}))$ dans (10), nous procédons comme suit :

- (1) Partir des résidus obtenus par ajustement croisé, $y_k - \hat{m}_{(-j)}(\mathbf{x}_k) \equiv e_{k,cf}$ dans (15).
- (2) Ajuster la procédure d'apprentissage automatique utilisée pour le modèle d'imputation en prenant $\log(e_{k,cf}^2)$ comme variable dépendante et $\mathbf{x}_k$ comme vecteur de prédicteurs. Obtenir les valeurs prédites.
- (3) Exponentier les valeurs ajustées obtenues à l'étape 2; désigner ces valeurs ajustées par $\tilde{e}_{k,cf}^2$.

Un estimateur de $\mathbb{V}_2(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}))$ dans (10) est donné par

$$\hat{\mathbb{V}}_2(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})) = \frac{1}{N^2} \sum_{k \in S_r} w_k \frac{1 - \hat{p}_{(-j)}(\mathbf{x}_k)}{\{\hat{p}_{(-j)}(\mathbf{x}_k)\}^2} \tilde{e}_{k,cf}^2. \quad (17)$$

Enfin, un estimateur de la variance totale, $\mathbb{V}_{tot}(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}))$, est donné par

$$\hat{\mathbb{V}}_{tot}(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})) = \hat{\mathbb{V}}_1(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})) + \hat{\mathbb{V}}_2(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})).$$

Remarque 1. Si la propension de réponse $p(\mathbf{x}_k)$ était connue, on pourrait construire un estimateur DR obtenu par ajustement croisé de la forme

$$\hat{\mu}_{aipw,cf}(\hat{m}, p) = \frac{1}{N} \left\{ \sum_{k \in S} w_k \hat{m}_{(-j)}(\mathbf{x}_k) + \sum_{k \in S_r} w_k \frac{y_k - \hat{m}_{(-j)}(\mathbf{x}_k)}{p(\mathbf{x}_k)} \right\}, \quad (18)$$

où $\hat{m}(\mathbf{x}_k)$ est obtenu à l'aide de n'importe quelle méthode d'apprentissage automatique. Dans ce cas, il est facile de voir que $\mathbb{E}(\hat{\mu}_{aipw}(\hat{m}, p) - \mu) = 0$, quelle que soit la méthode utilisée pour obtenir les prédictions $\hat{m}(\mathbf{x}_k)$. De plus, la variance de $\hat{\mu}_{aipw,cf}(\hat{m}, p)$ est asymptotiquement égale à (10), tant que $\hat{m}(\mathbf{x}_k)$ est consistant pour $m(\mathbf{x}_k)$. En d'autres termes, aucune vitesse de convergence n'est requise.

3.4. Mise en œuvre

L'estimateur $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$ ne s'exprime pas sous la forme pratique (2). Pour résoudre ce problème, nous proposons la mise en œuvre suivante. Pour les unités répondantes (c'est-à-dire les unités pour lesquelles $r_k = 1$), nous reportons les valeurs observées y_k dans le fichier de données, tandis que pour les unités non répondantes (c'est-à-dire les unités pour lesquelles $r_k = 0$), nous reportons les valeurs imputées modifiées

$$\hat{m}^*(\mathbf{x}_k) = \hat{m}_{(-j)}(\mathbf{x}_k) + \frac{1}{\sum_{\ell \in S_m} w_\ell} \sum_{\ell \in S_r} w_\ell \left\{ \frac{1 - \hat{p}_{(-j)}(\mathbf{x}_\ell)}{\hat{p}_{(-j)}(\mathbf{x}_\ell)} \right\} \{y_\ell - \hat{m}_{(-j)}(\mathbf{x}_\ell)\} \quad (19)$$

Avec les valeurs imputées modifiées (19), il est facile de vérifier que

$$\frac{1}{N} \left[\sum_{k \in S} w_k \{r_k y_k + (1 - r_k) \hat{m}^*(\mathbf{x}_k)\} \right] = \hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}). \quad (20)$$

En conséquence, les utilisateurs de données peuvent facilement obtenir l'estimateur $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$ en utilisant les procédures habituelles d'estimation ponctuelle.

3.5. Résumé de l'approche proposée

Pour appliquer l'approche proposée, nous procédons comme suit :

- **Étape 1** : Sélectionner une procédure d'apprentissage automatique pour le modèle d'issue et une autre pour le score de propension, puis optimiser les hyperparamètres de chaque méthode (en utilisant, par exemple, la validation croisée). L'optimisation des hyperparamètres est une étape importante pour s'assurer que la Condition (H_1) donnée en (16) est satisfaite.
- **Étape 2** : Calculer

$$\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}) = \frac{1}{N} \left[\sum_{k \in S} w_k \hat{m}_{(-j)}(\mathbf{x}_k) + \sum_{k \in S_r} w_k \frac{y_k - \hat{m}_{(-j)}(\mathbf{x}_k)}{\hat{p}_{(-j)}(\mathbf{x}_k)} \right].$$

- **Étape 3** : Obtenir les valeurs imputées modifiées $\hat{m}^*(\mathbf{x}_k)$.
- **Étape 4** : Estimer la variance de $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$ par

$$\begin{aligned} \hat{V}_{tot}(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})) &= \hat{V}_1(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})) + \hat{V}_2(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})) \\ &= \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\hat{\eta}_{k,cf}}{\pi_k} \frac{\hat{\eta}_{\ell,cf}}{\pi_\ell} + \frac{1}{N^2} \sum_{k \in S_r} w_k \frac{1 - \hat{p}_{(-j)}(\mathbf{x}_k)}{\{\hat{p}_{(-j)}(\mathbf{x}_k)\}^2} \tilde{e}_{k,cf}^2. \end{aligned}$$

- **Étape 5** : Calculer un intervalle de confiance de niveau $1 - \alpha\%$ pour μ :

$$\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}) \pm z_{\alpha/2} \sqrt{\hat{V}_{tot}(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}))},$$

où $z_{\alpha/2}$ est le quantile $1 - \alpha/2$ de la loi normale standard.

4. Étude de simulation

Afin d'évaluer la performance des méthodes proposées en termes de biais, d'efficacité et de taux de couverture, nous avons réalisé une étude de simulation. Tout d'abord, nous avons généré 5 prédicteurs mutuellement indépendants $X_1, \dots, X_5$, selon une loi normale standard. Nous avons ensuite répété $R = 10,000$ itérations du processus suivant :

- 1) Étant donnés les prédicteurs $X_1, \dots, X_5$, nous avons généré deux populations finies avec différentes valeurs de N : 2000 et 16000. Chaque population était constituée de deux variables d'enquête Y_1 et Y_2 générées selon

$$\begin{aligned} Y_1 &= 1 + 2x_1 - 3x_2^2 + 1.5x_3^3 + 0.5x_4^2 - x_5^3 + \mathcal{N}(0, 1); \\ Y_2 &= 20 \times \mathbf{1}_{x_1 > 0} + 10 \sin(x_2) + 5x_3^2 + \mathcal{N}(0, 1); \end{aligned}$$

- 2) Un échantillon S a été sélectionné dans chaque population selon un plan de sondage aléatoire simple sans remise, avec une fraction d'échantillonnage $n/N = 10\%$, ce qui a conduit aux valeurs de n suivantes: 200 et 1600.

- 3) Dans chaque échantillon, les valeurs manquantes de la variable d'intérêt Y_1 ont été générées indépendamment selon une loi de Bernoulli de probabilité

$$p(\mathbf{x}_k) = 0.05 + 0.95 \frac{\exp \{(x_1 - 0.5) + (x_2 - 0.5) + (x_3 - 0.5) + (x_4 - 0.5) + (x_5 - 0.5)\}}{1 + \exp \{(x_1 - 0.5) + (x_2 - 0.5) + (x_3 - 0.5) + (x_4 - 0.5) + (x_5 - 0.5)\}}, \quad k \in S. \quad (21)$$

Les valeurs manquantes de la variable d'intérêt Y_2 ont été générées indépendamment selon une loi de Bernoulli de probabilité

$$p(\mathbf{x}_k) = \frac{1}{6} \{0.5 + \sin(x_1) + \sin(x_2) + \sin(x_3)\}, \quad k \in S. \quad (22)$$

Dans (21) et (22), les paramètres ont été fixés de manière à obtenir une proportion de valeurs manquantes d'environ 50% dans chaque échantillon.

- 4) Dans chaque échantillon, nous avons calculé les estimateurs doublement robustes $\hat{\mu}_{aipw}(\hat{m}, \hat{p})$ et $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$ donnés par (12) et (14), respectivement, pour différentes combinaisons de méthodes d'apprentissage automatique. Pour le modèle de régression de l'issue, nous avons utilisé : RF (Breiman, 2001) avec 15 observations par nœud terminal et $B = 250$ arbres ; Cubist (Quinlan, 1992) ; et XG-boost (Chen et Guestrin, 2016) avec un taux d'apprentissage fixé à 1, une profondeur de 5, et $B = 50$ arbres. Pour le modèle du score de propension, nous avons utilisé soit des arbres de régression avec 40 observations par nœud terminal, soit RF avec 20 observations par nœud terminal et $B = 250$ arbres ; voir les tableaux 4-1 et 4-2 pour les différentes combinaisons.

- 5) Dans chaque échantillon, nous avons calculé des intervalles de confiance à 95% de la forme $\hat{\mu} \pm 1.96\sqrt{\hat{V}_{tot}(\hat{\mu})}$, où le couple $[\hat{\mu}, \hat{V}_{tot}(\hat{\mu})]$ correspondait soit à $[\hat{\mu}_{aipw}(\hat{m}, \hat{p}), \hat{V}_{tot}(\hat{\mu}_{aipw}(\hat{m}, \hat{p}))]$, soit à $[\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}), \hat{V}_{tot}(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}))]$.

Comme mesure du biais d'un estimateur ponctuel générique $\hat{\mu}$, nous avons calculé son biais relatif en pourcentage de Monte Carlo, défini par $RB(\hat{\mu}) = 100 \times E_{MC}(\hat{\mu} - \mu)/\mu$, où $E_{MC}(\cdot)$ désigne la moyenne de Monte Carlo sur les R itérations. Comme mesure d'efficacité relative de $\hat{\mu}$, nous avons calculé $RE(\hat{\mu}) = 100 \times \{MSE_{MC}(\hat{\mu})/MSE_{MC}(\hat{\mu}_F)\}$, où $\hat{\mu}_F$ désigne l'estimateur basé sur l'échantillon complet (voir Section 2). Comme mesure du biais d'un estimateur de variance générique $\hat{V}(\hat{\mu})$, nous avons calculé son biais relatif en pourcentage de Monte Carlo défini par

$$RB(\hat{V}(\hat{\mu})) = 100 \times \frac{E_{MC}(\hat{V}(\hat{\mu})) - V_{MC}(\hat{\mu})}{V_{MC}(\hat{\mu})},$$

où $V_{MC}(\cdot)$ désigne la variance de Monte Carlo. Enfin, nous avons calculé le taux de couverture de Monte Carlo des intervalles de confiance normaux à 95%. Les résultats sont présentés dans les tableaux 4-1 et 4-2. Dans les deux tableaux, pour une taille d'échantillon n et un estimateur ponctuel donnés, la première ligne indique le RB de Monte Carlo, alors que l'efficacité relative (RE) de Monte Carlo est indiquée entre parenthèses sur la deuxième ligne. De même, pour une taille d'échantillon n et un estimateur de variance donnés, la première ligne indique le RB de Monte Carlo, alors que le taux de couverture de Monte Carlo est indiqué entre parenthèses sur la deuxième ligne.

Nous commençons par discuter la performance des estimateurs ponctuels $\hat{\mu}_{aipw}(\hat{m}, \hat{p})$ et de sa version avec l'ajustement croisé $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$. Les deux ont présenté un biais négligeable dans tous les scénarios sauf un. Plus précisément, lorsque RF était utilisé pour estimer $m(\mathbf{x}_k)$ et $p(\mathbf{x}_k)$, $\hat{\mu}_{aipw}(\hat{m}, \hat{p})$ et $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$ ont montré un léger biais pour Y_2 , avec des valeurs de RB égales à -2.7% et 2.6%, respectivement. En termes d'efficacité relative (RE), l'estimateur $\hat{\mu}_{aipw}(\hat{m}, \hat{p})$ était généralement légèrement plus efficace que sa version avec l'ajustement croisé, bien que les différences tendent à disparaître lorsque la taille de l'échantillon augmente.

En ce qui concerne la performance des estimateurs de variance, nous notons que $\hat{V}_{tot}(\hat{\mu}_{aipw}(\hat{m}, \hat{p}))$ a présenté

Tableau 4-1
Statistiques de Monte Carlo pour l'estimateur ponctuel et l'estimateur de variance de la variable d'intérêt Y_1

	$\hat{\mu}_{aipw}$	$\hat{V}(\hat{\mu}_{aipw})$	$\hat{\mu}_{aipw,cf}$	$\hat{V}(\hat{\mu}_{aipw,cf})$
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{Cubist}; \text{Trees}]$				
$n = 200$	0.4 (162)	-4.0 (94.4)	0.4 (173)	4.0 (95.1)
$n = 1,600$	0.0 (158)	-3.7 (95.0)	0.0 (165)	-2.1 (95.8)
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{Cubist}; \text{RF}]$				
$n = 200$	0.3 (161)	-14.0 (92.8)	0.4 (166)	-2.0 (94.4)
$n = 1,600$	0.1 (148)	-12.1 (93.9)	0.1 (148)	-3.0 (94.9)
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{RF}; \text{RF}]$				
$n = 200$	0.1 (156)	-17.8 (93.2)	0.0 (160)	2.5 (95.5)
$n = 1,600$	0.1 (152)	-23.0 (91.8)	0.1 (152)	-6.4 (94.9)
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{XGboost}; \text{Trees}]$				
$n = 200$	-0.4 (203)	-53.0 (82.6)	0.2 (217)	-0.4 (95.1)
$n = 1,600$	-0.1 (195)	-50.4 (83.9)	0.0 (218)	0.4 (95.0)

un biais négatif dans tous les scénarios. Dans plusieurs cas, le biais était important, avec des valeurs absolues de RB dépassant 15%. Dans ces scénarios, le taux de couverture des intervalles de confiance basés sur la loi normale était inférieur au taux nominal. Cette mauvaise performance est très probablement due au problème de surapprentissage. En revanche, l'estimateur de variance $\hat{V}_{tot}(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}))$ a bien fonctionné dans tous les scénarios, produisant des taux de couverture proches du taux nominal.

5. Remarques finales

Dans cet article, nous avons utilisé une seule méthode d'apprentissage automatique (ML) pour modéliser la variable d'intérêt Y et une autre méthode ML pour le modèle de score de propension. Une approche alternative consisterait à utiliser des méthodes d'agrégation, qui combinent plusieurs algorithmes ML afin d'améliorer la performance prédictive et la robustesse. Une de ces méthodes est l'algorithme *Super Learner*, qui construit un ensemble de modèles candidats et les pondère de manière optimale à l'aide de la validation croisée (van der Laan et al., 2007). Cette approche peut être appliquée à la fois au modèle de l'issue et au modèle de score de propension, améliorant potentiellement l'efficacité et réduisant le biais. Sous des conditions modérées, le *Super Learner* atteint un taux de convergence asymptotique au moins aussi rapide que celui du meilleur apprenant candidat dans la bibliothèque de modèles. L'étude des propriétés des méthodes d'agrégation dans le contexte de l'imputation pour données manquantes dans les enquêtes demeure un sujet de recherche à explorer.

Tableau 4-2
Statistiques de Monte Carlo pour l'estimateur ponctuel et l'estimateur de variance de la variable d'intérêt Y_2

	$\hat{\mu}_{aipw}$	$\hat{V}(\hat{\mu}_{aipw})$	$\hat{\mu}_{aipw,cf}$	$\hat{V}(\hat{\mu}_{aipw,cf})$
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{Cubist}; \text{Trees}]$				
$n = 200$	0.1 (131)	-18.5 (92.1)	0.6 (144)	-0.9 (94.5)
$n = 1,600$	-0.1 (106)	-6.7 (94.0)	0.2 (110)	-3.6 (94.3)
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{Cubist}; \text{RF}]$				
$n = 200$	0.1 (113)	-25.9 (90.9)	0.4 (147)	-11.0 (93.7)
$n = 1,600$	-0.1 (106)	-4.8 (94.3)	0.1 (107)	-1.8 (94.6)
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{RF}; \text{RF}]$				
$n = 200$	2.7 (151)	-14.0 (91.1)	2.6 (156)	1.7 (93.6)
$n = 1,600$	-0.1 (112)	-8.8 (93.8)	-0.1 (113)	-0.8 (94.6)
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{XGboost}; \text{Trees}]$				
$n = 200$	-0.4 (121)	-20.2 (91.8)	0.2 (135)	-1.1 (94.7)
$n = 1,600$	-0.1 (116)	-11.3 (93.7)	0.0 (120)	0.0 (95.0)

Bibliographie

- Beaumont, J.-F. et Haziza, D. (2016), A note on the concept of invariance in two-phase sampling designs. *Survey Methodology*, **42**, p. 319–323.
- Breiman, L. (2001), Random forests. *Machine Learning*, **45**, p. 5–32.
- Cao, W., Tsiatis, A. A. et Davidian, M. (2009), Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, **96**, p. 723–734.
- Chen, T. et Guestrin, C. (2016), XGBoost. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16*. ACM Press.
- Chen, J. et Rao, J.N.K. (2007), Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, **17**, p. 1047–1064.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., et Newey, W. (2018), Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, **21**, p. C1–C68.
- Dagdoug, M., Goga, C., et Haziza, D. (2023b), Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison. *Journal of Survey Statistics and Methodology*, **11**, p. 141–188.
- Dagdoug, M., Goga, C., et Haziza, D. (2025), Statistical Inference in the presence of imputed survey data through random forests. To appear in *Scandinavian Journal of Statistics*.
- Haziza, D. et Beaumont, J. F. (2007), On the construction of imputation classes in surveys. *International Statistical Review*, **75**, p. 25–43.

- Haziza, D. et Rao, J. N. K. (2006), A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology*, **32**, p. 53–64.
- Haziza, D. et Vallée, A.-A. (2020). Variance estimation in the presence of singly imputed data: A critical review. *Japanese Journal of Statistics and Data Science*, **3**, p. 583–623.
- Isaki, C.-T. et Fuller, W.-A. (1982), Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, **77**, p. 49–61.
- Kang, J.D.Y. et Schafer, J.L. (2008). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, **22**, p. 523–539.
- Kim, J. K. et Haziza, D. (2014), Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, **24**, p. 375–394.
- van der Laan, M. J., Polley, E. C., et Hubbard, A. E. (2007), Super learner. *Statistical Applications in Genetics and Molecular Biology*, **6**, p. 1–23.
- Little, R. J. (1986), Survey nonresponse adjustments for estimates of means. *International Statistical Review*, **54**, p. 139–157.
- Lunceford, J. K. et Davidian, M. (2004), Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, **23**, p. 2937–2960.
- Pfeffermann, D. et Sverchkov, M. (2009), Inference under informative sampling. *In Handbook of Statistics*, vol. 29, p. 455–487. Elsevier.
- Robins, J.M., Rotnitzky, A. et Zhao, L.P. (1994), Estimation of regression coefficient when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, p. 846–866.
- Rubin, D. B. (1976), Inference and missing data. *Biometrika*, **63**, p. 581–592.
- Shao, J. et Steel, P. (1999), Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, **94**, p. 254–265.
- Wager, S., et Athey, S. (2018), Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, **113**, p. 1228–1242.