

Techniques d'enquête

Estimation par la régression généralisée en cas d'erreur de spécification du plan de sondage

par Joseph D. Engmark et Jean D. Opsomer

Date de diffusion : le 23 décembre 2025



Statistique
Canada

Statistics
Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- | | |
|---|----------------|
| • Service de renseignements statistiques | 1-800-263-1136 |
| • Service national d'appareils de télécommunications pour les malentendants | 1-800-363-7629 |
| • Télécopieur | 1-514-283-9350 |

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « [Normes de service à la clientèle](#) ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Publication autorisée par la ministre responsable de Statistique Canada

© Sa Majesté le Roi du chef du Canada, représenté par la ministre de l'Industrie, 2025

L'utilisation de la présente publication est assujettie aux modalités de l'[entente de licence ouverte](#) de Statistique Canada.

Une [version HTML](#) est aussi disponible.

This publication is also available in English.

Estimation par la régression généralisée en cas d'erreur de spécification du plan de sondage

Joseph D. Engmark et Jean D. Opsomer¹

Résumé

Une estimation à partir de données d'enquête classique fondée sur le plan repose sur un plan de sondage adéquatement défini afin d'obtenir une inférence valable. Nous tenons compte des propriétés de l'estimation par régression en cas d'erreur de spécification du plan de sondage, dans lequel les probabilités d'inclusion nominale et réelle ne sont pas nécessairement les mêmes. Cette situation générale d'erreur de spécification d'un plan de sondage comporte beaucoup de défis dans le contexte moderne dans lequel sont réalisées les enquêtes. Dans ce contexte, une analyse asymptotique de l'estimateur par la régression, une expression du biais et une expression de la variance sont présentées. En outre, un estimateur de la variance convergent est obtenu et une expression qui estime le biais en partie ou en totalité est présentée. Cette dernière expression peut être utilisée en tant qu'indicateur de la présence d'un biais causé par l'erreur de spécification d'un spécialiste. Une étude par simulation est menée afin d'étayer la théorie exposée.

Mots-clés : Assisté par un modèle; échantillonnage; échantillonnage probabiliste; fondé sur le plan de sondage; propriétés asymptotiques de l'enquête.

1. Introduction

Le cadre d'échantillonnage fondé sur le plan de sondage est un cadre robuste et général qui permet de faire des inférences sur les caractéristiques d'une population finie comme des totaux ou des moyennes. Les estimateurs fondés sur le plan de sondage dépendent d'un plan de sondage adéquatement spécifié, dans lequel chaque élément de la population a une probabilité de sélection connue et autre que zéro afin d'obtenir une inférence valable. Dans l'environnement moderne où les enquêtes sont menées, cependant, il ne s'agit souvent plus d'une attente raisonnable, d'où la raison d'être du présent article.

Ce problème a toujours été examiné dans le contexte des cas de non-réponses et réglé en apportant des ajustements pour la non-réponse par pondération (Bethlehem, 1988; Särndal et Lundström, 2005). Cette approche intègre habituellement un mécanisme de réponse à ses estimateurs, ce qui donne lieu à une probabilité de réponse inconnue ou à un modèle supposé pour la réponse. Le contexte examiné ici est plus général : le plan réel est mal spécifié et inconnu. Contrairement aux situations de non-réponse où l'inférence est examinée dans le cadre d'un modèle mixte de plan de sondage et de réponse, nous supposons ici que l'on continue d'obtenir l'échantillon par l'intermédiaire d'un mécanisme de sélection unique.

Le type d'erreur de spécification que nous examinons survient lorsque la probabilité qu'un élément apparaisse dans l'échantillon diffère de celle attribuée par le plan de sondage nominal. L'une des principales sources est lorsque les processus de sélection ne sont pas mis en œuvre comme prévu, notamment lors de défaillances dans la base de sondage, de mesures erronées ou désuètes de la taille utilisée dans l'échantillonnage avec probabilité proportionnelle à la taille (PPT), d'une stratification ou de variables du plan de sondage inexactes, du non-respect des méthodes de sélection indiquées sur le terrain, des limites

1. Joseph D. Engmark, Department of Mathematics, University of Maryland, College Park, MD 20742, États-Unis. Courriel : engmark@umd.edu; Jean D. Opsomer, Department of Mathematics, University of Maryland, College Park, MD 20742, États-Unis. Courriel : jopsomer@umd.edu.

erronées de l'unité d'échantillonnage primaire ou de l'utilisation d'un échantillonnage par marche aléatoire. L'erreur de spécification peut également être attribuable aux effets du répondant sur les processus de sélection, comme la non constance des probabilités de sélection qui n'est pas prise en compte et même les cas de non-réponse (par exemple certaines personnes sont plus susceptibles de répondre, certains animaux sont plus susceptibles d'être attrapés), lorsqu'ils ne sont pas traités comme une phase d'échantillonnage supplémentaire. On peut également l'appliquer au contexte général de l'échantillonnage non probabiliste, lorsqu'un pseudo-plan de sondage (nominal) qui reflète la façon dont les éléments ont été sélectionnés afin de se retrouver dans l'échantillon est utilisé. Voir Elliott et Valliant (2017) pour une analyse approfondie sur l'inférence pour les échantillons non probabilistes.

L'estimateur par la régression généralisée (ERG) est un estimateur basé sur un modèle bien connu pour le total d'une population. Lorsque le plan de sondage est bien spécifié, l'estimateur par la régression est asymptotiquement non biaisé et à moyenne carrée convergente, peu importe la qualité du modèle de régression. Qui est plus, une expression de la variance asymptotique de l'estimateur est fournie et l'expression peut être estimée. Ces résultats sont présentés aux pages 225 à 238 des chapitres 6.4 à 6.6, de Särndal, Swensson et Wretman (1992) [SSW ci-après], un article exhaustif sur les estimateurs basés sur un modèle historiques, et récemment aux sections 1 à 6 de Breidt et Opsomer (2017) [BO17 ci-après], un article qui expose un cadre asymptotique pour les estimateurs basés sur un modèle où l'estimateur ERG est une technique de prédiction précise.

L'estimateur ERG a été étudié en présence de cas de non-réponse. L'une des principales difficultés réside dans l'obtention d'expressions et d'estimateurs pour le biais et la variance à utiliser afin d'établir une inférence. Bethlehem (1988) présente une expression du biais et de la variance, tandis que Särndal et Lundström (2005) fournissent un estimateur pour la variance. Dans notre contexte plus général, nous produisons une expression pour le biais et la variance, un estimateur pour la variance et un estimateur pour une partie du biais lorsque des données de regroupement sont disponibles. Le biais est la composante la plus difficile à examiner. La littérature actuelle sur l'échantillonnage a traité de la relation entre le biais et la corrélation entre le processus de sélection et la variable réponse d'intérêt, ainsi que de la nécessité d'avoir de l'information auxiliaire sur le processus de sélection afin de réduire le biais (Meng, 2022; Chambers, Ranjbar, Salvati et Pacini, 2022). L'expression pour le biais de l'estimateur ERG présentée dans le présent article affirme cette relation dans notre contexte et des simulations illustrent l'utilité de l'information auxiliaire pour réduire la corrélation entre le processus de sélection et le résidu, et ainsi le biais de l'estimateur ERG.

Le reste de l'article est organisé comme suit. La section 2 présente la notation et le cadre asymptotique des estimateurs fondés sur le plan; la section 3 examine brièvement les propriétés bien établies de l'estimateur ERG dans un plan adéquatement spécifié et présente les propriétés en cas d'erreur de spécification du plan. Cette section se termine par l'obtention d'un estimateur de la variance pour l'estimateur ERG réalisable en cas d'erreur de spécification du plan et une expression qui estime le biais en partie ou en totalité lorsque des variables auxiliaires catégoriques sont disponibles. En pratique, cette expression peut servir d'indicateur de biais ou d'erreur de spécification du plan. Dans la section 4, on confirme la théorie et les résultats au moyen d'une simulation. Enfin, nous concluons l'exposé à la section 5.

2. Notation et cadre asymptotique

2.1 Notation

Le cadre du présent article est le cadre classique fondé sur le plan, qui évite les hypothèses de modèle quant aux variables recueillies par l'enquête, comme indiqué dans la section 1 de BO17 et au chapitre 2 de SSW. Désignons y_k comme la valeur d'une variable d'intérêt pour l'élément k dans la population $U = \{1, 2, \dots, N\}$ et $t_y = \sum_{k \in U} y_k$ comme le total de la population, ici traité comme fixe, mais inconnu avant l'échantillonnage.

Le *plan de sondage*, que nous désignons comme $p(s)$, correspond à une distribution probabiliste de tous les sous-ensembles possibles 2^N de U . Autrement dit, $p(s)$ correspond à la probabilité de sélectionner l'échantillon particulier s . Définissons l'indicateur d'appartenance à l'échantillon $I_k = 1$ si $k \in s$ et $I_k = 0$ autrement. Pour $k, l \in U$, désignons les *probabilités d'inclusion du premier degré* du plan comme $\pi_k = \mathbb{E}[I_k] = P[k \in s] = \sum_{s \subset U: k \in s} p(s)$, et les *probabilités d'inclusion du deuxième degré* du plan comme $\pi_{kl} = \mathbb{E}[I_k I_l] = P[k, l \in s] = \sum_{s \subset U: k, l \in s} p(s)$. Le plan $p(s)$ correspond à un *plan de sondage probabiliste* si $\pi_k > 0$ pour tout $k \in U$ et on dit qu'il est *mesurable* si $\pi_{kl} > 0$ pour tout $k, l \in U$.

L'estimateur Horvitz-Thompson (HT) pour un total est obtenu comme suit :

$$\hat{t}_{y\pi} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k.$$

L'estimateur HT est reconnu comme sans biais par rapport au plan pour t_y , avec une variance

$$\text{Var}(\hat{t}_{y\pi}) = \sum_{k, l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l},$$

où $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. Si le plan est mesurable, un estimateur sans biais de la variance est obtenu comme suit :

$$\widehat{\text{Var}}(\hat{t}_{y\pi}) = \sum_{k, l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

2.2 Cadre asymptotique

Les estimateurs fondés sur le plan sont élaborés à partir d'un échantillon probabiliste sélectionné dans la population. Il est naturel de commencer les arguments asymptotiques à partir d'une séquence de populations finies U_N de taille N , avec $N \rightarrow \infty$. Une séquence de plans de sondage $p_N(\cdot)$ est associée à cette séquence. Pour chaque N de la séquence, un échantillon $s_N \subset U_N$ est prélevé selon le plan $p_N(\cdot)$, avec une taille d'échantillon n_N . Les probabilités d'inclusion π_{kN}, π_{klN} sont associées au plan $p_N(\cdot)$. On laissera tomber l'indice N dans la mesure du possible, comme le veut la pratique courante.

Des conditions de régularité sont requises pour la population, U_N , et le plan, $p_N(\cdot)$, afin de garantir que les résultats asymptotiques sont bien définis parce que les deux changent, étant donné N . Le premier exposé

porte uniquement sur l'estimateur HT simple. La section 2 de BO17 présente l'estimateur par la différence, qui peut être considéré comme un estimateur HT décalé, suivi de l'estimateur ERG, qui peut quant à lui être considéré comme un estimateur par la différence qui utilise des coefficients estimés et un terme du reste. Par conséquent, les résultats asymptotiques de l'estimateur HT sont directement transférés à l'estimateur par la différence, qui est à son tour transféré à l'estimateur ERG, pourvu que le terme du reste soit négligeable par rapport aux totaux estimés. Les conditions de régularité suivantes sont supposées, pour un plan d'une taille non aléatoire n :

- D1. Comme $N \rightarrow \infty, \frac{n}{N} \rightarrow \theta \in (0, 1)$. Pour toute N , $\min_{k \in U} \pi_k \geq \lambda_1 > 0$ et $\limsup_{N \rightarrow \infty} n \max_{k, l \in U: k \neq l} |\Delta_{kl}| < \infty$.
- D2. Pour toute N , $\min_{k \in U} \pi_{kl} \geq \lambda_2 > 0$.
- D3. Les variables de l'étude $\{y_k\}_{k \in U}$ satisfont à $\limsup_{N \rightarrow \infty} \frac{\sum_{k \in U} y_k^4}{N} < \infty$.
- D4. Comme $N \rightarrow \infty, \frac{\hat{t}_{y\pi} - t_y}{\sqrt{\text{Var}(\hat{t}_{y\pi})}} \xrightarrow{d} N(0, 1)$.

Les conditions D1 et D2 sont imposées au plan de sondage $p_N(\cdot)$, ce qui garantit que les probabilités de sélection sont délimitées aussi loin de zéro que nécessaire pour que le plan de sondage soit un plan de sondage probabiliste. La condition D3 est relative à la séquence de populations finies U_N . Les conditions D1 et D3 sont utilisées afin de montrer que $\hat{t}_{y\pi}$ est constant pour t_y . Les conditions D2 et D3 sont utilisées afin de produire une estimation constante de la variance de l'estimateur. Ces conditions, combinées à la condition D4, soit la normalité asymptotique supposée de l'estimateur HT, permettent de créer des intervalles de confiance avec une couverture adéquate aux fins de l'estimateur HT. Dans les plans de sondage dont la taille de l'échantillon est aléatoire, le n ci-dessus indique la taille d'échantillon prévue, $\mathbb{E}(\sum_{k \in U} I_k)$. On trouvera une analyse plus détaillée sur le cadre asymptotique dans BO17, à la section 2.

3. Estimation par régression et propriétés

3.1 L'estimateur par la régression généralisée (ERG)

Nous définissons l'estimateur ERG et ses composantes. Se reporter au chapitre 6 de SSW et à la section 6 de BO17 pour une analyse complète dans le cas d'un plan adéquatement spécifié. Désignons $\mathbf{x}_k$ comme un vecteur de variables auxiliaires. Supposons que les totaux d'une population $\mathbf{t}_x = \sum_{k \in U} \mathbf{x}_k$ sont connus et que des vecteurs d'échantillon $\{\mathbf{x}_k\}_{k \in S}$ sont constatés. L'estimateur par la régression par calage (ERG) reconnu se définit comme suit :

$$\begin{aligned} \hat{t}_{yr} &= \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \hat{\mathbf{B}} \\ &= \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \hat{\mathbf{T}}^{-1} \hat{\mathbf{t}}, \end{aligned} \quad (3.1)$$

où $\hat{\mathbf{t}} = \sum_{k \in S} \mathbf{x}_k y_k / (\pi_k \sigma_k^2)$ et $\hat{\mathbf{T}} = \sum_{k \in S} \mathbf{x}_k \mathbf{x}_k' / (\pi_k \sigma_k^2)$ sont des estimateurs HT pour $\mathbf{t} = \sum_{k \in U} \mathbf{x}_k y_k / \sigma_k^2$ et $\mathbf{T} = \sum_{k \in U} \mathbf{x}_k \mathbf{x}_k' / \sigma_k^2$; $\hat{\mathbf{B}} = \hat{\mathbf{T}}^{-1} \hat{\mathbf{t}}$ est un estimateur pour $\mathbf{B} = \mathbf{T}^{-1} \mathbf{t}$; et on suppose que $\hat{\mathbf{T}}$ est non singulier.

L'estimateur ERG peut aussi être exprimé en tant que $\hat{t}_{yr} = \sum_{k \in U} g_{ks} y_k / \pi_k$, où $g_{ks} = 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \hat{\mathbf{T}}^{-1} \mathbf{x}_k / \sigma_k^2$. Pour des raisons de simplicité, nous établissons que $\sigma_k^2 = 1$ pour tout $k \in U$, ce qui limite la portée à la régression multiple homoscédastique et omet le paramètre de la variance pour le reste de l'article.

L'estimateur ERG est une fonction non linéaire d'estimateurs HT, que l'on peut approximer au moyen d'une *linéarisation en série de Taylor* étant donné $\hat{t}_{yr,0} = \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \mathbf{B}$ (voir la page 236 du chapitre 6.6 de SSW). L'approximation linéarisée a une variance

$$\text{Var}(\hat{t}_{yr,0}) = \sum_{k,l \in U} \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l},$$

où $E_k = y_k - \mathbf{x}_k' \mathbf{B}$, avec un estimateur

$$\widehat{\text{Var}}(\hat{t}_{yr,0}) = \sum_{s,k,l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_{ks}}{\pi_k} \frac{e_{ls}}{\pi_l},$$

où $e_{ks} = y_k - \mathbf{x}_k' \hat{\mathbf{B}}$. Selon les conditions D1 à D4, l'estimateur ERG est asymptotiquement non biaisé et à moyenne carrée convergente pour t_y , et l'estimateur de la variance approximative peut être utilisé aux fins d'inférence asymptotiquement correcte.

3.2 Estimateur par la régression en cas d'erreur de spécification du plan de sondage

Les propriétés ci-dessus demeurent, en supposant que le plan de sondage est bien spécifié, c'est-à-dire $\mathbb{E}(I_k) = \pi_k$, $\mathbb{E}(I_{kl}) = \pi_{kl}$ et $\text{Cov}(I_k, I_l) = \Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. Dans un cas d'erreur de spécification du plan de sondage, les probabilités d'inclusion de premier degré véritables sont $\mathbb{E}(I_k) = \pi_k^* \neq \pi_k$, où π_k correspond à la probabilité d'inclusion définie par le plan de sondage nominal, tout comme pour la probabilité d'inclusion de deuxième degré $\mathbb{E}(I_{kl}) = \pi_{kl}^* \neq \pi_{kl}$; et la covariance est de $\text{Cov}(I_k, I_l) = \Delta_{kl}^* = \pi_{kl}^* - \pi_k^* \pi_l^* \neq \Delta_{kl}$. Dans ce contexte, les estimateurs HT qui utilisent le plan nominal ne sont plus des estimateurs sans biais pour leurs quantités respectives au niveau de la population. En particulier,

$$\mathbb{E}(\hat{t}_{y\pi}) = \mathbb{E}\left(\sum_{k \in U} I_k \frac{y_k}{\pi_k}\right) = \sum_{k \in U} \frac{\pi_k^*}{\pi_k} y_k \equiv t_y^* \neq t_y.$$

Les estimateurs HT restants des composantes de l'estimateur ERG estiment maintenant des totaux « * » :

$$\begin{aligned} \mathbb{E}(\hat{\mathbf{t}}_{x\pi}) &= \sum_{k \in U} \frac{\pi_k^*}{\pi_k} \mathbf{x}_k \equiv \mathbf{t}_x^* \neq \mathbf{t}_x; & \mathbb{E}(\hat{\mathbf{t}}) &= \sum_{k \in U} \frac{\pi_k^*}{\pi_k} \mathbf{x}_k y_k \equiv \mathbf{t}^* \neq \mathbf{t}; \\ \mathbb{E}(\hat{\mathbf{T}}) &= \sum_{k \in U} \frac{\pi_k^*}{\pi_k} \mathbf{x}_k \mathbf{x}_k' \equiv \mathbf{T}^* \neq \mathbf{T}; & \mathbf{T}^{*-1} \mathbf{t}^* &\equiv \mathbf{B}^* \neq \mathbf{B}. \end{aligned}$$

Tout au long de notre exposé, « * » indique une quantité inconnue étant donné la présence de π_k^* dans leur élaboration. Afin d'étayer les déclarations asymptotiques sur les estimateurs HT en cas d'erreur de spécification du plan de sondage, nous devons avoir une condition de régularité supplémentaire pour le vrai plan inconnu :

$$\text{D5. Pour toute } N, \min_{k \in U} \pi_k^* \geq \lambda_1 > 0 \text{ et } \limsup_{N \rightarrow \infty} n \max_{k,l \in U: k \neq l} |\Delta_{kl}^*| < \infty,$$

ce qui garantit que les vraies probabilités de sélection sont délimitées aussi loin de zéro que nécessaire pour que le plan de sondage soit probabiliste. La normalité asymptotique de l'estimateur HT (D4) est maintenant supposée selon le vrai plan de sondage. Dans le cas d'une taille d'échantillon aléatoire, n correspondrait maintenant à la taille d'échantillon attendue selon le vrai plan de sondage.

Au moyen d'une application directe d'une linéarisation en série de Taylor (voir la page 236 du chapitre 6.6 de SSW), évaluée selon les nouveaux totaux « * » des estimateurs HT, l'estimateur ERG (3.1) peut être exprimé comme

$$\hat{t}_{yr} = \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \mathbf{B}^* + (\mathbf{t}_x - \mathbf{t}_x^*)' \mathbf{T}^{*-1} (\hat{\mathbf{t}} - \hat{\mathbf{T}} \mathbf{B}^*) + O_p \left(\frac{N}{n} \right).$$

Quand $n \rightarrow \infty$, le terme du reste est asymptotiquement négligeable par rapport aux totaux estimés, et l'estimateur ERG continue de se comporter comme une combinaison linéaire d'estimateurs HT. Les estimateurs HT sont désormais sans biais et à moyenne carrée convergente pour leurs quantités « * » respectives. Si l'on omet le terme du reste, l'expression suivante constitue une approximation linéarisée de l'estimateur ERG en cas d'erreur de spécification du plan de sondage :

$$\tilde{t}_{yr,0} = \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \mathbf{B}^* + (\mathbf{t}_x - \mathbf{t}_x^*)' \mathbf{T}^{*-1} (\hat{\mathbf{t}} - \hat{\mathbf{T}} \mathbf{B}^*). \quad (3.2)$$

L'expression (3.2), qui correspond à l'*estimateur linéarisé*, est utilisée afin d'explorer les propriétés asymptotiques de l'estimateur ERG en cas d'erreur de spécification du plan de sondage :

3.3 Biais de l'estimateur par la régression en cas d'erreur de spécification du plan de sondage

Le biais approximatif de l'estimateur ERG en cas d'erreur de spécification du plan de sondage découle directement de l'espérance de l'expression (3.2) en ce qui concerne le plan de sondage :

$$\mathbb{E}(\tilde{t}_{yr,0}) = t_y^* + (\mathbf{t}_x - \mathbf{t}_x^*)' \mathbf{B}^*,$$

et donc

$$\text{Biais}(\tilde{t}_{yr,0}) = \mathbb{E}(\tilde{t}_{yr,0}) - t_y = \sum_{k \in U} \left(\frac{\pi_k^*}{\pi_k} - 1 \right) E_k^*, \quad (3.3)$$

étant donné $E_k^* = y_k - \mathbf{x}_k' \mathbf{B}^*$.

L'expression (3.3) sera définie comme le biais linéarisé. Elle sous-entend que si π_k^* / π_k n'est pas corrélé avec E_k^* , le biais sera asymptotiquement négligeable. On peut obtenir d'autres représentations du biais, mais elles ne seront pas abordées ici. Nous évitons ici de formuler des hypothèses de modélisation explicites sur les données et mettons plutôt l'accent sur la corrélation de la population finie en tant que facteur à l'origine du biais. Cette corrélation dépend du pouvoir de prédiction des variables auxiliaires dans la variable du sondage y_k et du ratio des probabilités d'inclusion réelles à nominales. Nous l'examinerons de façon empirique à la section 4.

3.4 Variance de l'estimateur par la régression en cas d'erreur de spécification du plan de sondage

Il est facile d'obtenir des expressions de la variance en cas d'erreur de spécification au moyen des techniques de SSW (voir la page 237, au chapitre 6.6). L'estimateur linéarisé en cas d'erreur de spécification du plan de sondage (3.2) est

$$\begin{aligned}\tilde{t}_{yr,0} &= \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \mathbf{B}^* + (\mathbf{t}_x - \mathbf{t}_x^*)' \mathbf{T}^{*-1} (\hat{\mathbf{t}} - \hat{\mathbf{T}} \mathbf{B}^*) \\ &= \mathbf{t}_x' \mathbf{B}^* + \sum_{k \in s} \frac{G_k^* E_k^*}{\pi_k},\end{aligned}$$

où $G_k^* = 1 + (\mathbf{t}_x - \mathbf{t}_x^*)' \mathbf{T}^{*-1} \mathbf{x}_k$ peut être considéré comme un niveau de population analogue des poids- g $g_{ks} = 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_x)' \hat{\mathbf{T}}^{-1} \mathbf{x}_k$ dépendants de l'échantillon. Dans le contexte de l'erreur de spécification, $g_{ks} \xrightarrow{p} G_k^*$ et $e_{ks} \xrightarrow{p} E_k^*$ en ce qui concerne une série croissante de plans de sondage. Ainsi, l'estimateur linéarisé peut une fois de plus être exprimé en tant qu'ajustement en population ($\mathbf{t}_x' \mathbf{B}^*$, une constante) auquel on ajoute la somme d'un résidu pondéré ($\sum_{k \in s} G_k^* E_k^* / \pi_k$, un estimateur). En tant qu'estimateur par la différence, la variance et un estimateur de la variance sont immédiatement offerts :

$$\text{Var}(\tilde{t}_{yr,0}) = \sum_{k,l \in U} \Delta_{kl}^* \left(\frac{G_k^* E_k^*}{\pi_k} \right) \left(\frac{G_l^* E_l^*}{\pi_l} \right), \quad (3.4)$$

$$\widetilde{\text{Var}}(\tilde{t}_{yr,0}) = \sum_{k,l \in s} \frac{\Delta_{kl}^*}{\pi_{kl}} \left(\frac{g_{ks} e_{ks}}{\pi_k} \right) \left(\frac{g_{ls} e_{ls}}{\pi_l} \right). \quad (3.5)$$

Il n'est toutefois pas possible d'obtenir l'estimateur de la variance (3.5) étant donné la covariance inconnue, Δ_{kl}^* . Au lieu de tenter d'étudier le cas d'un inconnu général Δ_{kl}^* , nous nous concentrerons uniquement sur le contexte, courant dans les enquêtes à grande échelle, où il est raisonnable de traiter le plan de sondage comme un plan avec remise.

Il est raisonnable de substituer les résultats d'échantillonnage avec remise par des résultats d'échantillonnage sans remise pour les grandes populations, où la probabilité de sélectionner un élément plus d'une fois est extrêmement faible. Par ailleurs, on sait qu'un plan de sondage avec remise accroît généralement l'estimation de la variance par rapport à un échantillonnage sans remise, étant donné que l'on omet la correction de la population finie, tandis que la variance linéarisée estimée tend à sous-estimer en pratique. Ensemble, ces effets opposés donnent lieu à un estimateur de la variance aux propriétés raisonnables en pratique.

Le calcul de l'estimateur de la variance avec remise découle du résultat 2.9.1 de SSW (page 51), où un estimateur de la variance sans biais pour l'estimateur « avec remise dilaté par la probabilité p_k » est présenté. Suivant cette approche, nous définissons $p_k = \pi_k / n$ pour $k = 1, \dots, N$, où $p_k = \mathbb{P}$ (en sélectionnant l'élément k dans un tirage donné selon le plan de sondage nominal), en ignorant la possibilité qu'un élément puisse apparaître plus d'une fois dans l'échantillon. Définissons $Z_i^*, i = 1, \dots, n$ comme des variables aléatoires, de sorte que $Z_i^* = G_k^* E_k^* / p_k$ si l'élément k est sélectionné au tirage i . La distribution de Z_i^* est $\mathbb{P}(Z_i^* = G_k^* E_k^* / p_k) = p_k^*$, où p_k^* est inconnu à cause de l'erreur de spécification et défini par $\pi_k^* = n p_k^*$. Avec

un léger abus de notation, nous supprimerons l'indice du tirage $i = 1, \dots, n$, et inscrirons Z^* afin d'indiquer la variable aléatoire et $k \in s$ pour indiquer les éléments contenus dans l'échantillon. Ensuite,

$$\text{Var}\left(\sum_{k \in s} \frac{G_k^* E_k^*}{\pi_k}\right) = \text{Var}(\bar{Z}^*) \quad (3.6)$$

étant donné $\bar{Z}^* = \sum_{k \in s} Z_k^* / n$. Nous faisons remarquer que, même si les distributions (3.6) et (3.4) ont la même forme, elles sont évaluées selon différentes distributions d'échantillonnage. Selon l'échantillonnage avec remise, les hypothèses sur les éléments de covariance des conditions D1 et D4 sont satisfaites de façon insignifiante parce que les éléments hors diagonale dans la matrice de covariance équivalent à zéro. De même, la condition de régularité D2 n'est plus nécessaire étant donné $\pi_{kl} = n^2 \pi_k \pi_l > 0$ par les conditions sur π_k dans la condition D1.

Les Z^* sont tirées indépendamment sur les mêmes ensembles de probabilités $p_1^*, \dots, p_N^*$ dans chaque tirage. Étant donné que les Z^* sont indépendantes et identiquement distribuées (i.i.d.), $\text{Var}(\bar{Z}^*) = n^{-1} \text{Var}(Z^*)$ où, $\text{Var}(Z^*) = \sum_{k \in U} (Z_k^* - \mathbb{E}(Z^*))^2 p_k^*$. Un estimateur de la variance sans biais de $\text{Var}(Z^*)$ est $\widehat{\text{Var}}(Z^*) = (n-1)^{-1} \sum_{k \in s} (Z_k^* - \bar{Z}^*)^2$. Cela contient des quantités inconnues « * », qui motivent l'estimateur réalisable $\widehat{\text{Var}}(Z^*) = (n-1)^{-1} \sum_{k \in s} (z_{ks} - \bar{z}_s)^2$, où $z_{ks} = g_{ks} e_{ks} / p_k$ et $\bar{z}_s = n^{-1} \sum_{k \in s} z_{ks}$. Lorsque l'on réunit ces composantes, l'estimation de la variance réalisable pour l'estimateur ERG en cas d'erreur de spécification correspond à

$$\widehat{\text{Var}}(\tilde{t}_{yr,0}) = \frac{1}{n} \widehat{\text{Var}}(Z^*) = \frac{1}{n(n-1)} \sum_{k \in s} (z_{ks} - \bar{z}_s)^2. \quad (3.7)$$

Dans la section S.1 du matériel supplémentaire (voir Engmark et Opsomer, 2025), on montre que l'estimateur de la variance réalisable (3.7) est convergent pour la variance de l'estimateur linéarisé (3.4), selon la distribution d'échantillonnage avec remise. Par conséquent, en pratique, une erreur de spécification du plan de sondage n'a aucune incidence sur la capacité d'estimer de façon cohérente la variance de l'estimateur par la régression dans un échantillonnage avec remise.

Nous avançons que ce résultat demeure valable tant que la structure de corrélation du plan nominal et du plan comportant une erreur de spécification est la même, soit $\pi_{kl} / \pi_k \pi_l = \pi_{kl}^* / \pi_k^* \pi_l^*$ (ce qui est particulièrement vrai dans le cas d'un échantillonnage avec remise). En pratique, cela voudrait dire que la stratification et la corrélation intra-grappe du plan demeurent correctement spécifiées. Nous ne poursuivons pas cette généralisation officiellement dans le cas présent et mettons plutôt l'accent sur l'utilisation de la distribution (3.7) pour l'estimation de la variance.

3.5 Contrôler le biais de l'estimateur par la régression en cas d'erreur de spécification du plan de sondage

Il est impossible d'estimer directement l'expression du biais approximative (3.3) en raison de l'inconnu π_k^* . Nous proposons ici une mesure qui peut être utile pour détecter le biais dans un contexte commun

d'enquêtes sociales, où la plupart, voire la totalité des variables auxiliaires sont catégoriques. Dans ce cas, l'incidence possible des variables auxiliaires sur le biais surviendrait uniquement par l'intermédiaire de la structure de groupe dans les probabilités de sélection inconnues déterminées par l'ensemble de valeurs de ces variables. On ignore toutefois quelles sont les variables qui déterminent cette structure. Nous examinons un scénario où nous tentons de déterminer si une variable auxiliaire donnée éclaire les probabilités de sélection, en ce sens où son omission des prédicteurs dans l'estimateur ERG causerait un biais. S'il est effectivement déterminé que cette variable auxiliaire est informative, elle peut être ajoutée aux prédicteurs de l'estimateur ERG afin de réduire le biais causé par l'erreur de spécification.

Les valeurs des variables catégoriques examinées séparent la population en un nombre fixe de groupes, $g = 1, \dots, G$, où nous avons besoin de $G \leq |s| = n$. Définissons θ_g comme la probabilité de sélection moyenne pour le groupe g dans la population, $\theta_g = N_g^{-1} \sum_{k \in U_g} \pi_k^*$. Le biais linéarisé (3.3) peut être exprimé comme

$$\begin{aligned} \text{Biais}(\tilde{t}_{y^r,0}) &= \sum_{g=1}^G \sum_{k \in U_g} \left(\frac{\pi_k^*}{\pi_k} - 1 + \frac{\theta_g}{\pi_k} - \frac{\theta_g}{\pi_k} \right) E_k^* \\ &= \sum_{g=1}^G \sum_{k \in U_g} \frac{1}{\pi_k} (\theta_g - \pi_k) E_k^* + \sum_{g=1}^G \sum_{k \in U_g} \frac{1}{\pi_k} (\pi_k^* - \theta_g) E_k^* \\ &= B_1 + B_2. \end{aligned}$$

Au sens large, B_1 peut être considéré comme un « biais entre groupes » et B_2 comme un « biais au sein du même groupe » (3.2).

Nous examinons d'abord le cas où les probabilités de sélection correspondent effectivement à la structure du groupe, c'est-à-dire $\pi_k^* = \theta_g$ pour tout $k \in U_g$, $g = 1, \dots, G$. Dans ce cas, la composante de biais B_2 sera de 0. Étant donné que la quantité $N_g^0 = \sum_{k \in U_g} 1/\pi_k$ est connue pour le plan de sondage nominal et que $\hat{N}_g = \sum_{k \in s_g} 1/\pi_k$ correspond à l'estimateur HT de

$$N_g^* = \sum_{k \in U_g} \frac{\pi_k^*}{\pi_k} = \theta_g N_g^0,$$

θ_g peut être estimé par $\hat{\theta}_g = \hat{N}_g / N_g^0 = \left(\sum_{k \in s_g} 1/\pi_k \right) / \left(\sum_{k \in U_g} 1/\pi_k \right)$. Ainsi, nous pouvons définir un estimateur par substitution de B_1 par

$$\hat{B}_1 = \sum_{g=1}^G \sum_{k \in s_g} \frac{1}{\hat{\theta}_g} \frac{1}{\pi_k} (\hat{\theta}_g - \pi_k) e_{ks}. \quad (3.8)$$

Par conséquent, si cette structure de propension du groupe tient, B_2 sera exactement zéro et $\hat{B}_1$ sera approximativement sans biais pour B_1 et estimera le biais linéarisé en entier.

Si la structure de groupe ne tient pas dans les probabilités de sélection, B_2 ne correspondra pas à zéro et $\hat{B}_1$ sera un estimateur biaisé de B_1 . Une grande valeur de $\hat{B}_1$ peut néanmoins servir d'indicateur de biais possible en raison d'une erreur de spécification du plan. La section S.2 du matériel supplémentaire montre

le comportement de $\hat{B}_1$ à partir du contexte général où l'on suppose que les probabilités de sélection ne sont pas constantes dans les groupes.

4. Simulation

L'étude par simulation vise à évaluer le biais de l'estimateur ERG en cas d'erreur de spécification du plan, ainsi que la capacité de l'estimation de la variance de rétablir la variance sous le plan et de la mesure du biais d'indiquer la nécessité d'ajouter une variable de calage.

4.1 Configuration de la simulation

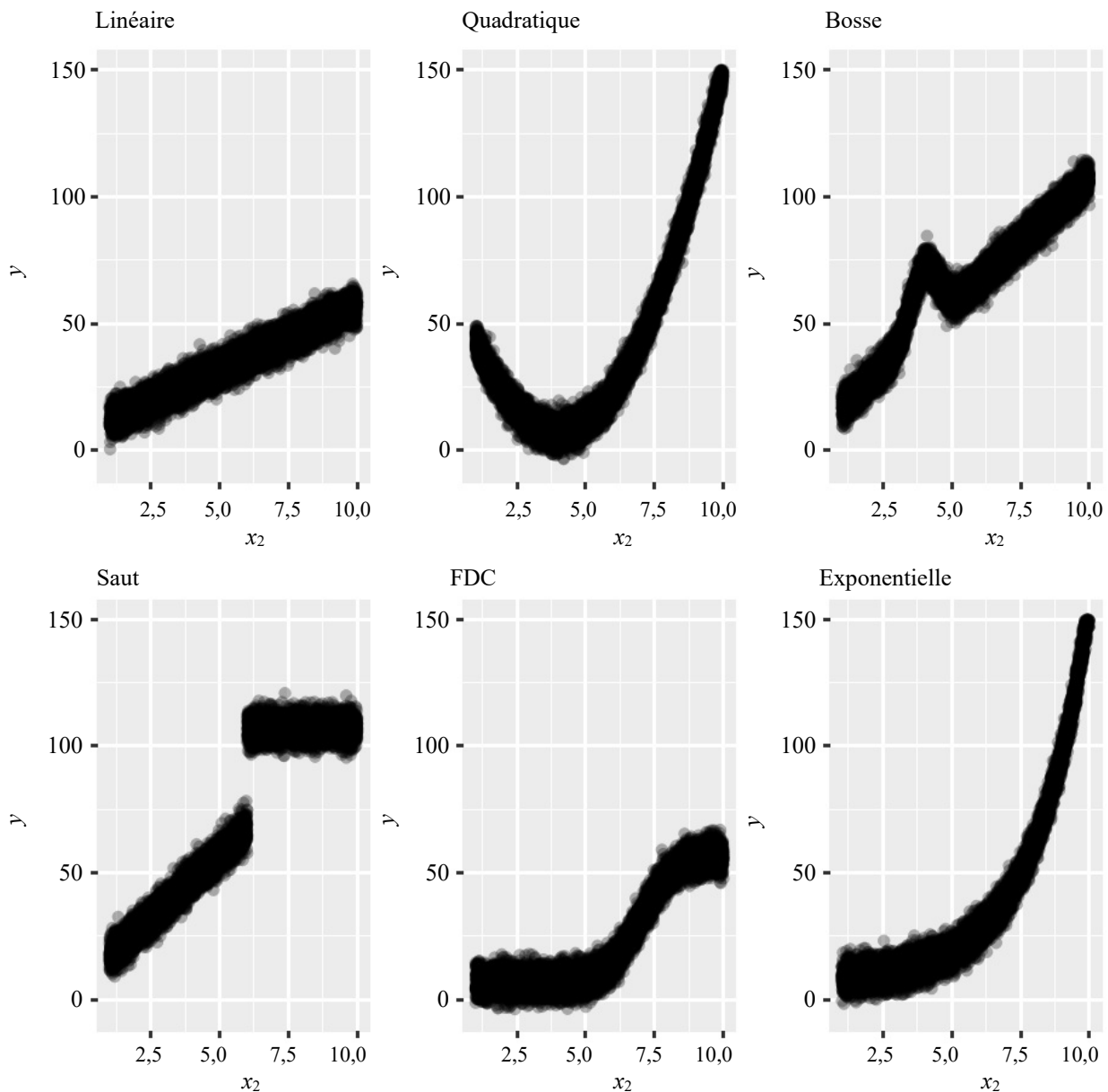
4.1.1 Génération de données

La population simulée est composée de $\{y_k, x_{1k}, x_{2k}, w_k, z_k, \pi_k, \pi_k^*\}$ pour $k = 1, \dots, N$. Les variables d'intérêt dans les populations sont créées au moyen des fonctions moyennes $\mu_i(\mathbf{x}), i = 1, \dots, 6$ données ci-dessous. Les fonctions sont tirées de la section 3 de Breidt et Opsomer (2000).

linéaire :	$\mu_1(x_1, x_2) = 1 + 3x_1 + 5x_2 + \varepsilon,$
quadratique :	$\mu_2(x_1, x_2) = 1 + 3x_1 + 4(x_2 - 4)^2 + \varepsilon,$
bosse :	$\mu_3(x_1, x_2) = 1 + 3x_1 + 10x_2 + 25\exp(-2(x_2 - 4)^2) + \varepsilon,$
saut :	$\mu_4(x_1, x_2) = 1 + 3x_1 + 10x_2 I_{\{x_2 \leq 6\}} + 100 I_{\{x_2 > 6\}} + \varepsilon,$
fonction de distribution cumulative :	$\mu_5(x_1, x_2) = 1 + 3x_1 + 50\Phi(x_2 - 7) + \varepsilon,$ avec Φ en tant que fonction de distribution cumulative normale standard,
exponentielle :	$\mu_6(x_1, x_2) = 1 + 3x_1 + \exp((1/2)x_2) + \varepsilon.$

Les fonctions moyennes représentent un éventail de relations différentes entre y et (x_1, x_2) , comme illustré à la figure 4.1. Elles sont aussi utilisées afin de démontrer le comportement de l'estimateur ERG dans un contexte typique d'enquête, où les mêmes poids sont appliqués à toutes les variables d'enquête. Pour μ_1 , les données sont linéaires dans x_2 , et l'estimateur ERG devrait bien fonctionner. Pour les autres fonctions moyennes, il est intéressant de voir le rendement de l'estimateur ERG à mesure que les données s'éloignent de la fonction linéaire : pour μ_2 , les données sont quadratiques dans x_2 ; μ_3 , les données quittent localement la fonction linéaire dans x_2 sous la forme d'une « bosse »; μ_4 , les données affichent une discontinuité x_2 ; μ_5 , les données sont sigmoïdales dans x_2 ; μ_6 , les données sont exponentielles dans x_2 .

Figure 4.1 Représentation graphique des fonctions moyennes : linéaire, quadratique, bosse, saut, distribution cumulative et exponentielle



Note : FDC = fonction de distribution cumulative.

Les variables auxiliaires de la population x_{1k} et x_{2k} sont générées en tant que variables aléatoires i.i.d. uniformes (1, 3) et que variables aléatoires uniformes (1, 10), respectivement, et elles sont indépendantes les unes des autres. Le plan réel dans la simulation dépend de x_2 . Les variables d'intérêt de la population, y_k , sont générées à partir des fonctions moyennes en ajoutant des erreurs i.i.d. $N(0, \sigma^2)$, étant donné $\sigma = 3$. La taille de la population est de $N = 10\,000$ et celle de l'échantillon, de $n = 99$.

4.1.2 Sélection de l'échantillon

Le but est de simuler des contextes où l'estimateur de régression est biaisé. L'échantillon est sélectionné de manière que le plan soit informatif, comme le contrôle de la variable auxiliaire x_2 , c'est-à-dire que la variable d'intérêt y_k et les probabilités de sélection réelles dépendent de x_2 . Le plan nominal ne dépend pas de x_2 . Quatre différents contextes sont examinés et ils porteront les noms de contexte I à IV.

Dans le contexte I, le plan de sondage nominal est un échantillonnage aléatoire simple sans remise (SI), tandis que le plan réel est un échantillonnage avec probabilité proportionnelle à la taille (PPT) en ce qui concerne le x_2 .

Dans le contexte II, nous simulons des probabilités d'inclusion non constantes dans le plan nominal et le plan réel. Le contexte II est une version stratifiée du contexte I, le plan nominal est un échantillonnage aléatoire simple stratifié sans remise (STSI), tandis que le plan réel est un échantillonnage stratifié avec probabilité proportionnelle à la taille (STPPT) pour x_2 . Les strates sont déterminées par une variable w_k , qui est corrélée avec y_k et créée en ajoutant $N(0,1)$ erreurs à x_{1k} . La population est triée par w_k , et trois strates sont formées avec des tailles de $N_1 = 5\,000$, $N_2 = 3\,000$ et $N_3 = 2\,000$ respectivement, selon les valeurs de w_k . Des échantillons d'une taille de $n_1 = n_2 = n_3 = 33$ sont tirés pour chaque strate.

Dans les contextes III et IV, nous examinons le comportement de l'estimateur de la composante du biais lorsque les probabilités de sélection sont constantes ou non constantes au sein des groupes. Dans ces contextes, les éléments de population sont divisés en groupes et en strates. Les groupes sont déterminés par une variable z_k , qui est corrélée avec y_k et créée en ajoutant $N(0,1)$ erreurs à x_{2k} . La population est triée par z_k , et deux groupes de taille égale sont formés en fonction des valeurs de z_k et sans égard aux strates. Les strates sont déterminées comme ci-dessus. Dans le contexte III, le plan nominal est STSI, tandis que le plan réel est STSI avec des probabilités d'inclusion de groupe constantes dans les strates. Dans le contexte IV, le plan nominal est STSI, tandis que le plan réel est STPPT en ce qui concerne x_2 , avec des probabilités d'inclusion de groupe non constantes dans les strates. On trouve une description plus détaillée du processus d'échantillonnage à la section S.3 du matériel supplémentaire.

4.1.3 Estimateurs

L'estimateur Horvitz-Thompson utilisé dans le plan nominal, appelé HT(π), sert d'estimateur naïf en n'utilisant que le plan nominal dans sa création. Les estimateurs ERG dans le plan nominal, calibrés selon x_1 uniquement, x_2 uniquement et x_1 et x_2 , appelés $\text{ERG}(x_1, \pi)$, $\text{ERG}(x_2, \pi)$, $\text{ERG}(x_1 x_2, \pi)$ respectivement, servent d'estimateurs pratiques en cas d'erreur de spécification du plan et lorsque de l'information auxiliaire est disponible. Pour obtenir une comparaison complète, les estimateurs mentionnés ci-dessus sont aussi calculés en fonction du plan réel inconnu et appelés $\text{HT}(\pi^*)$, $\text{ERG}(x_1, \pi^*)$, $\text{ERG}(x_2, \pi^*)$, $\text{ERG}(x_1 x_2, \pi^*)$. Ils agissent en tant qu'estimateurs idéaux.

Pour obtenir une comparaison plus directe de l'estimation de la variance, la variance de l'estimateur HT(π) est aussi estimée par l'estimateur de la variance « avec remise dilaté par la probabilité p_k » (équation 2.9.9 à la

page 52 de SSW), en fonction de l'hypothèse selon laquelle chaque élément de l'échantillon n'apparaît qu'une fois, comme celui utilisé pour obtenir l'expression (3.7) :

$$\widehat{\text{Var}}(\hat{t}_{y\pi}) = \frac{1}{n(n-1)} \sum_{k \in S} \left(n \frac{y_k}{\pi_k} - \hat{t}_{y\pi} \right)^2. \quad (4.1)$$

Lorsque le plan nominal est un plan STSI, on tient compte de la stratification dans l'estimation de la variance de l'estimateur HT (π) en estimant la variance dans chaque strate de l'échantillon au moyen de l'expression (4.1) et on additionne toutes les strates.

Les variances pour l'estimateur ERG selon le plan nominal sont estimées au moyen de l'expression (3.7). Lorsque le plan nominal est un plan STSI, on tient compte de la stratification dans l'estimation de la variance de l'estimateur ERG en estimant la variance dans chaque strate de l'échantillon au moyen de l'expression (3.7) et on additionne toutes les strates.

$$\widehat{\text{Var}}_{\text{ST}}(\hat{t}_{y\pi}) = \sum_{h=1}^H \frac{1}{n_h(n_h-1)} \sum_{k \in S_h} (z_{hks} - \bar{z}_{hs})^2,$$

où $z_{hks} = g_{ks} e_{ks} / p_{hk}$, $\bar{z}_{hs} = n_h^{-1} \sum_{k \in S_h} z_{hks}$, $p_{hk} = \pi_k / n_h$, avec e_{ks} et g_{ks} comme défini aux sections 3.1 et 3.4 respectivement.

L'estimation de la variance de l'estimateur HT (π^*) et de l'estimateur ERG selon le plan réel est obtenue en remplaçant π_k par π_k^* dans les estimateurs précédemment décrits. Encore une fois, lorsque le plan réel est un plan STPPT, on tient compte de la stratification lorsqu'on estime la variance de l'estimateur HT (π^*) et de l'ERG.

4.2 Résultats de simulation

La population demeure fixe pendant 10 000 échantillons répétés, ce qui permet d'évaluer le rendement moyen des estimateurs en fonction du plan. Pour chaque échantillon de la simulation, la variance et le total estimés de chaque estimateur, appelés de façon générale dans le présent article $\hat{t}$ et $\widehat{\text{Var}}(\hat{t})$, sont calculés. Enfin, un intervalle de confiance (IC), $\hat{t} \pm z_{1-\alpha/2} \widehat{\text{Var}}(\hat{t})^{1/2}$, est élaboré avec $z_{1-\alpha/2}$ pour $\alpha = 0,05$.

Ces valeurs, combinées au total connu de la population, t_y , nous permettent de calculer les éléments suivants présentés dans les tableaux ci-dessous : i) le biais relatif des estimateurs, $(\mathbb{E}(\hat{t}) - t_y) / t_y$; ii) le ratio de biais des estimateurs, $\text{RB}(\hat{t}) = (\mathbb{E}(\hat{t}) - t_y) / \text{Var}(\hat{t})^{1/2}$; iii) la largeur moyenne des intervalles de confiance, normalisée en fonction de la largeur moyenne des intervalles de confiance obtenus avec l'estimateur HT (π^*); iv) la probabilité de couverture des intervalles de confiance; v) le biais relatif des estimations de la variance, $(\mathbb{E}(\widehat{\text{Var}}(\hat{t})) - \text{Var}(\hat{t})) / \text{Var}(\hat{t})$.

Nous présentons aussi la corrélation entre les résidus de niveau de population et le ratio π_k^* / π_k pour $k = 1, \dots, N$, $\text{Corr}(E_k^*, \pi_k^* / \pi_k)$. Cette valeur n'est pas fondée sur une simulation; il s'agit plutôt d'une caractéristique de la population et du facteur à l'origine des biais que nous constatons dans les estimateurs.

Pour l'estimateur HT (π), nous indiquons $\text{Corr}(y_k, \pi_k^* / \pi_k)$. Pour les estimateurs utilisés dans le plan réel, nous omettons les corrélations, qui sont équivalentes à zéro et donc sans importance. Les résultats de simulation pour les trois premières variables de l'enquête sont présentés dans les tableaux 4.1 et 4.2. Les résultats complets, y compris les variables d'enquête restantes, sont présentés aux tableaux S.1 et S.2 du matériel supplémentaire.

Les tableaux 4.1 et 4.2 montrent que l'estimateur ERG calé en fonction de la variable non informative seulement, $\mathbf{x}_1$, se comporte de la même façon que l'estimateur HT naïf, HT (π). Les deux affichent un biais important déterminé par la corrélation, des IC dont la largeur se ressemble, une très mauvaise couverture et un biais négligeable dans l'estimateur de la variance. Lorsque l'estimateur ERG est calé en fonction de la variable informative, $\mathbf{x}_2$ ou $\mathbf{x}_1\mathbf{x}_2$, le biais et la corrélation sont réduits, les IC sont plus petits que ceux de l'estimateur HT naïf, la couverture est meilleure que celle de l'estimateur HT naïf, et l'estimateur de la variance a un biais négligeable. Dans de nombreux cas, le biais de l'estimateur de la variance utilisé pour le plan nominal est plus petit que celui de l'estimateur utilisé pour le plan réel, mais il n'y a aucune différence majeure dans l'ampleur. La mauvaise couverture est attribuable à la présence d'un biais dans les estimateurs, et pas aux estimateurs de la variance.

Tableau 4.1

Résultats de simulation dans le contexte I – Plan nominal - SI, plan réel - PPT en ce qui concerne $\mathbf{x}_2$

Population	Estimateur	Biais relatif	$\text{Corr}\left(E_k^*, \frac{\pi_k^*}{\pi_k}\right)$	Largeur normalisée de l'IC	Couverture	Biais relatif de l'estimateur de la variance
linéaire	HT (π)	0,18	0,97	2,05	0,00	-0,01
	ERG ($\mathbf{x}_1, \pi$)	0,18	0,97	2,03	0,00	-0,02
	ERG ($\mathbf{x}_2, \pi$)	0,00	0,01	0,67	0,94	-0,02
	ERG ($\mathbf{x}_1\mathbf{x}_2, \pi$)	0,00	0,02	0,58	0,94	-0,02
	HT (π^*)	0,00	—	—	0,93	0,01
	ERG ($\mathbf{x}_1, \pi^*$)	0,01	—	2,93	0,92	-0,08
	ERG ($\mathbf{x}_2, \pi^*$)	0,00	—	0,68	0,93	-0,06
	ERG ($\mathbf{x}_1\mathbf{x}_2, \pi^*$)	0,00	—	0,58	0,93	-0,07
quadratique	HT (π)	0,35	0,79	1,63	0,08	-0,01
	ERG ($\mathbf{x}_1, \pi$)	0,35	0,79	1,63	0,08	-0,02
	ERG ($\mathbf{x}_2, \pi$)	-0,15	-0,47	1,16	0,49	-0,07
	ERG ($\mathbf{x}_1\mathbf{x}_2, \pi$)	-0,15	-0,47	1,15	0,48	-0,09
	HT (π^*)	0,00	—	—	0,94	-0,01
	ERG ($\mathbf{x}_1, \pi^*$)	0,01	—	1,36	0,95	-0,02
	ERG ($\mathbf{x}_2, \pi^*$)	-0,01	—	1,16	0,91	-0,15
	ERG ($\mathbf{x}_1\mathbf{x}_2, \pi^*$)	-0,02	—	1,13	0,90	-0,19

Note : IC = intervalle de confiance; ERG = estimateur par la régression; HT = Horvitz-Thompson; PPT = échantillonnage avec probabilité proportionnelle à la taille; SI = échantillonnage aléatoire simple sans remise.

Tableau 4.1 (suite)

Résultats de simulation dans le contexte I – Plan nominal - SI, plan réel - PPT en ce qui concerne x_2

Population	Estimateur	Biais relatif	$\text{Corr}\left(E_k^*, \frac{\pi_k^*}{\pi_k}\right)$	Largeur normalisée de l'IC	Couverture	Biais relatif de l'estimateur de la variance
bosse	HT (π)	0,18	0,95	1,89	0,00	-0,02
	ERG ($\mathbf{x}_1, \pi$)	0,18	0,96	1,88	0,00	-0,03
	ERG ($\mathbf{x}_2, \pi$)	0,01	0,18	0,88	0,91	-0,04
	ERG ($\mathbf{x}_1\mathbf{x}_2, \pi$)	0,01	0,18	0,86	0,90	-0,05
	HT (π^*)	0,00	—	—	0,94	0,01
	ERG ($\mathbf{x}_1, \pi^*$)	0,01	—	3,07	0,91	-0,10
	ERG ($\mathbf{x}_2, \pi^*$)	0,00	—	0,86	0,93	-0,07
	ERG ($\mathbf{x}_1\mathbf{x}_2, \pi^*$)	0,00	—	0,84	0,93	-0,09

Note : IC = intervalle de confiance; ERG = estimateur par la régression; HT = Horvitz-Thompson; PPT = échantillonnage avec probabilité proportionnelle à la taille; SI = échantillonnage aléatoire simple sans remise.

L'inférence en vertu du plan nominal est considérablement meilleure pour tous les types de données de la population lorsque l'on effectue un calage en fonction de la variable informative. Si les données sont linéaires ou qu'elles s'éloignent légèrement de la fonction linéaire (bosse), la couverture et l'inférence sont très bonnes. Si les données sont non linéaires (quadratique), la couverture et l'inférence sont considérablement meilleures que celles de l'estimateur HT naïf.

Tableau 4.2

Résultats de simulation dans le contexte II – Plan nominal - STSI, plan réel - STPPT en ce qui concerne x_2

Population	Estimateur	Biais relatif	$\text{Corr}\left(E_k^*, \frac{\pi_k^*}{\pi_k}\right)$	Largeur normalisée de l'IC	Couverture	Biais relatif de l'estimateur de la variance
linéaire	HT (π)	0,18	0,97	2,11	0,00	0,00
	ERG ($\mathbf{x}_1, \pi$)	0,18	0,97	2,08	0,00	-0,01
	ERG ($\mathbf{x}_2, \pi$)	0,00	0,01	0,68	0,94	-0,01
	ERG ($\mathbf{x}_1\mathbf{x}_2, \pi$)	0,00	0,02	0,59	0,94	-0,05
	HT (π^*)	0,00	—	—	0,93	0,02
	ERG ($\mathbf{x}_1, \pi^*$)	0,01	—	2,97	0,92	-0,07
	ERG ($\mathbf{x}_2, \pi^*$)	0,00	—	0,68	0,93	-0,05
	ERG ($\mathbf{x}_1\mathbf{x}_2, \pi^*$)	0,00	—	0,59	0,93	-0,10
quadratique	HT (π)	0,34	0,79	1,63	0,12	0,00
	ERG ($\mathbf{x}_1, \pi$)	0,34	0,79	1,62	0,11	-0,01
	ERG ($\mathbf{x}_2, \pi$)	-0,15	-0,47	1,15	0,51	-0,01
	ERG ($\mathbf{x}_1\mathbf{x}_2, \pi$)	-0,15	-0,47	1,14	0,51	-0,04
	HT (π^*)	0,00	—	—	0,95	0,03
	ERG ($\mathbf{x}_1, \pi^*$)	0,01	—	1,36	0,95	-0,01
	ERG ($\mathbf{x}_2, \pi^*$)	-0,02	—	1,13	0,91	-0,11
	ERG ($\mathbf{x}_1\mathbf{x}_2, \pi^*$)	-0,02	—	1,10	0,90	-0,16

Note : IC = intervalle de confiance; ERG = estimateur par la régression; HT = Horvitz-Thompson; STPPT = échantillonnage stratifié avec probabilité proportionnelle à la taille; STSI = échantillonnage aléatoire simple stratifié sans remise.

Tableau 4.2 (suite)**Résultats de simulation dans le contexte II – Plan nominal - STSI, plan réel - STPPT en ce qui concerne x_2**

Population	Estimateur	Biais relatif	$\text{Corr}\left(E_k^*, \frac{\pi_k^*}{\pi_k}\right)$	Largeur normalisée de l'IC	Couverture	Biais relatif de l'estimateur de la variance
bosse	HT (π)	0,18	0,95	1,90	0,00	0,00
	ERG (x_1, π)	0,18	0,96	1,88	0,00	-0,01
	ERG (x_2, π)	0,01	0,18	0,89	0,91	-0,04
	ERG ($x_1 x_2, \pi$)	0,01	0,18	0,86	0,90	-0,06
	HT (π^*)	0,00		–	0,94	0,00
	ERG (x_1, π^*)	0,01		3,03	0,91	-0,07
	ERG (x_2, π^*)	0,00		0,86	0,93	-0,08
	ERG ($x_1 x_2, \pi^*$)	0,00		0,83	0,92	-0,11

Note : IC = intervalle de confiance; ERG = estimateur par la régression; HT = Horvitz-Thompson; STPPT = échantillonnage stratifié avec probabilité proportionnelle à la taille; STSI = échantillonnage aléatoire simple stratifié sans remise.

Examinons maintenant les estimateurs selon le plan réel. Tous les estimateurs sont sans biais, les IC sont plus petits que ceux obtenus avec l'estimateur HT idéal, HT (π^*), lorsque les données sont linéaires ou qu'elles s'éloignent légèrement de la fonction linéaire (bosse) et l'estimateur est calé en fonction d'une variable informative. Lorsque les données sont non linéaires (quadratique), nous constatons que les IC de l'estimateur ERG ne sont pas plus petits que ceux de l'estimateur HT idéal, peu importe les variables de calage. Tous les estimateurs ont une bonne couverture pour établir une inférence généralement valable. L'estimateur HT idéal a un biais négligeable dans l'estimateur de la variance. Dans tous les calages, l'estimateur ERG a un biais légèrement négatif dans l'estimateur de la variance. Il en est ainsi parce que l'on utilise l'estimateur de la variance linéarisé, qui a tendance à faire des sous-estimations.

La largeur de l'IC, la couverture et le biais de l'estimateur de la variance ont un comportement un peu moins bon que prévu. Lorsque la simulation est répétée avec une population plus grande et une taille d'échantillon de $N = 100\,000$ et $n = 999$, nous constatons que la largeur de l'IC est meilleure que celle de l'estimateur HT lorsqu'on le cale en fonction de la variable informative, que la couverture s'approche de 95 % et que le biais relatif de l'estimateur de la variance correspond à zéro.

Examinons ensuite les contextes de groupe. Dans les contextes III et IV, pour chaque échantillon répété, nous calculons deux estimateurs supplémentaires pour l'estimateur ERG selon le plan nominal : l'estimateur de la composante de biais $\hat{B}_1$ et un estimateur de biais relatif réalisable $\hat{B}_1 / \hat{t}$. Ces estimateurs nous permettent de calculer vi) la valeur attendue de l'estimateur de biais relatif réalisable, $\mathbb{E}(\hat{B}_1 / \hat{t})$; vii) la valeur attendue de l'estimation de la composante de biais mise à l'échelle en fonction du biais réel, $\mathbb{E}(\hat{B}_1) / \text{biais}$.

En outre, les biais relatifs des estimateurs i) sont présentés aux fins de comparaison avec la valeur estimée de vi). La valeur vii) indique la partie moyenne du biais général qui est saisi en moyenne par l'estimateur de la composante de biais. Enfin, nous présentons la caractéristique de population $B_1 / (B_1 + B_2)$, qui indique

la proportion du biais linéarisé qui correspond à un biais entre groupes, B_1 . Les résultats sont présentés dans les tableaux 4.3 et 4.4 ci-dessous.

Le tableau 4.3 résume le contexte III et la condition idéale de la propension de groupe constante. Ici, le biais est entièrement un biais au sein du même groupe. L'estimateur de la composante de biais, $\hat{B}_1$, estime le biais entier en moyenne, tandis que l'estimateur de biais relatif pratique, $\hat{B}_1 / \hat{t}$, estime le biais relatif réel bien dans la moyenne. Cette estimation est meilleure lorsque les données sont linéaires ou qu'elles s'éloignent légèrement de la fonction linéaire, mais aussi raisonnable lorsque les données sont non linéaires.

Le tableau 4.4 résume le contexte IV et les conditions pratiques de la propension de groupe non constante. Ici, seulement une partie du biais est un biais au sein du même groupe. L'estimateur de la composante de biais, $\hat{B}_1$, estime une partie du biais en moyenne, tandis que l'estimateur de biais relatif pratique, $\hat{B}_1 / \hat{t}$, n'estime plus avec exactitude le biais relatif réel bien dans la moyenne. Cependant, l'estimateur du biais relatif pratique saisit bel et bien une partie et la direction du biais relatif réel, ce qui lui confère une utilité en tant qu'indicateur de biais.

Tableau 4.3

Résultats de simulation de l'estimation du biais dans le contexte III – Plan nominal - STSI, plan réel - STSI en ce qui concerne les strates et les groupes

Population	Estimateur	Biais relatif	$\mathbb{E}\left(\frac{\hat{B}_1}{\hat{t}}\right)$	$\frac{\mathbb{E}(\hat{B}_1)}{\text{Biais}}$	$\frac{B_1}{B_1 + B_2}$
linéaire	ERG ($\mathbf{x}_1, \pi$)	0,10	0,09	0,99	1,00
	ERG ($\mathbf{x}_2, \pi$)	0,00	0,00	-0,99	1,00
	ERG ($\mathbf{x}_1 \mathbf{x}_2, \pi$)	0,00	0,00	0,07	1,00
quadratique	ERG ($\mathbf{x}_1, \pi$)	0,20	0,16	0,99	1,00
	ERG ($\mathbf{x}_2, \pi$)	-0,04	-0,04	0,93	1,00
	ERG ($\mathbf{x}_1 \mathbf{x}_2, \pi$)	-0,04	-0,04	0,92	1,00
bosse	ERG ($\mathbf{x}_1, \pi$)	0,10	0,09	0,99	1,00
	ERG ($\mathbf{x}_2, \pi$)	0,00	0,00	1,03	1,00
	ERG ($\mathbf{x}_1 \mathbf{x}_2, \pi$)	0,00	0,00	1,02	1,00

Note : ERG = estimateur par la régression; STSI = échantillonnage aléatoire simple stratifié sans remise.

Tableau 4.4

Résultats de simulation de l'estimation du biais dans le contexte IV – Plan nominal - STSI, plan réel - STPPT en ce qui concerne les strates et les groupes

Population	Estimateur	Biais relatif	$\mathbb{E}\left(\frac{\hat{B}_1}{\hat{t}}\right)$	$\frac{\mathbb{E}(\hat{B}_1)}{\text{Biais}}$	$\frac{B_1}{B_1 + B_2}$
linéaire	ERG ($\mathbf{x}_1, \pi$)	0,16	0,08	0,57	0,63
	ERG ($\mathbf{x}_2, \pi$)	0,00	0,00	-0,11	0,28
	ERG ($\mathbf{x}_1 \mathbf{x}_2, \pi$)	0,00	0,00	-0,16	0,33

Note : ERG = estimateurs par la régression; STPPT = échantillonnage stratifié avec probabilité proportionnelle à la taille; STSI = échantillonnage aléatoire simple stratifié sans remise.

Tableau 4.4 (suite)

Résultats de simulation de l'estimation du biais dans le contexte IV – *Plan nominal - STSI, plan réel - STPPT en ce qui concerne les strates et les groupes*

Population	Estimateur	Biais relatif	$\mathbb{E}\left(\frac{\widehat{B}_1}{\widehat{t}}\right)$	$\frac{\mathbb{E}(\widehat{B}_1)}{\text{Biais}}$	$\frac{B_1}{B_1 + B_2}$
quadratique	ERG ($\mathbf{x}_1, \pi$)	0,31	0,18	0,76	0,65
	ERG ($\mathbf{x}_2, \pi$)	-0,14	-0,03	0,15	0,60
	ERG ($\mathbf{x}_1 \mathbf{x}_2, \pi$)	-0,14	-0,03	0,15	0,60
bosse	ERG ($\mathbf{x}_1, \pi$)	0,16	0,07	0,50	0,59
	ERG ($\mathbf{x}_2, \pi$)	0,01	-0,01	-0,45	-0,02
	ERG ($\mathbf{x}_1 \mathbf{x}_2, \pi$)	0,01	0,00	-0,45	-0,02

Note : ERG = estimateurs par la régression; STPPT = échantillonnage stratifié avec probabilité proportionnelle à la taille; STSI = échantillonnage aléatoire simple stratifié sans remise.

5. Conclusion

Il est reconnu qu'une erreur de spécification crée un biais dans les estimateurs d'enquête classique. Pour l'estimateur HT, le biais est déterminé par la corrélation entre la variable d'intérêt et la disparité entre la probabilité d'inclusion nominale et réelle. Pour l'estimateur ERG, le biais est déterminé par la corrélation entre les résidus et la même disparité.

L'estimateur ERG peut atténuer cette corrélation et corriger le biais lorsque la régression convient aux données et que la corrélation entre le mécanisme de réponse et les résidus est faible. Autrement dit, l'estimateur ERG fonctionne mieux lorsque les données sont presque linéaires et que les variables de régression décrivent les probabilités d'inclusion réelles. Lorsque les données sont non linéaires, l'estimateur ERG peut réduire la corrélation et améliorer le biais causé par l'erreur de spécification, pourvu que les variables de régression soient informatives. Cette correction améliore considérablement l'inférence et montre l'utilité de l'estimateur ERG, peu importe la forme des données. Ainsi, l'estimateur ERG résiste à une erreur de spécification du processus de sélection ou au modèle de régression sous-entendu. Si le modèle de régression motivant est bon, les probabilités de sélection erronées ne causent donc aucun biais. Si les probabilités de sélection sont bonnes, un modèle de régression motivant erroné ne cause donc aucun biais. En ce sens, l'estimateur ERG est « doublement robuste ».

Cette conclusion enrichit la littérature sur l'application de l'estimateur ERG des écarts par rapport à un plan donné, comme les cas de non-réponse, à un cadre général avec erreur de spécification. En outre, en vertu de ce cadre général, nous avons proposé un estimateur de la variance qui est raisonnable dans tous les contextes de données et de plans, peu importe les variables de régression utilisées.

Enfin, nous supposons que nous avons une variable auxiliaire à partir de laquelle nous pouvons créer des groupes et qui serait liée aux mécanismes de réponse. S'il y a une probabilité de sélection constante au sein des groupes, nous pouvons estimer le biais de l'estimateur ERG. S'il y a une propension de groupe informative, mais non constante, nous pouvons estimer une partie du biais de l'estimateur ERG. Cette estimation peut être utilisée en tant qu'indicateur de biais ou d'erreur de spécification par les spécialistes.

Bibliographie

- Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4(3), 251-260.
- Breidt, F.J. et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4), 1026-1053.
- Breidt, F.J. et Opsomer, J.D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2), 190-205.
- Chambers, R., Ranjbar, S., Salvati, N. et Pacini, B. (2022). Weighting, informativeness and causal inference, with an application to rainfall enhancement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185.
- Elliott, M.R. et Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.
- Engmark, J.D. et Opsomer, J.D. (2025). Supplementary material: Generalized regression estimation under misspecified sample design. Disponible à l'adresse <https://doi.org/10.13140/RG.2.2.29212.19848>.
- Meng, X.-L. (2022). [Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste » : La miniaturisation de la corrélation due à un défaut des données : une stratégie polyvalente de traitement des échantillons non probabilistes](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2022002/article/00006-fra.pdf). *Techniques d'enquête*, 48(2), 369-392. Article accessible à l'adresse <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2022002/article/00006-fra.pdf>.
- Särndal, C.-E. et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.
- Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.