Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Intégration des données d’enquête pour l’analyse de régression au moyen du calage assisté par un modèle

par Zhonglei Wang, Hang J. Kim et Jae Kwang Kim

Date de diffusion : le 30 juin 2023

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 le ministre responsable de Statistique Canada

© Sa Majesté le Roi du chef du Canada, représenté par le ministre de l'Industrie 2023

Tous droits réservés. 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.

Techniques d'enquête, juin 2023
Vol. 49, nº 1, pp. 97-125
Statistique Canada, nº 12-001-X au catalogue

Intégration des données d'enquête pour l'analyse de régression au moyen du calage assisté par un modèle

Zhonglei Wang, Hang J. Kim et Jae Kwang Kim[1]

Résumé

Nous envisageons ici l'analyse de régression dans le contexte de l'intégration de données. Pour combiner des renseignements partiels de sources externes, nous utilisons l'idée de calage assisté par un modèle qui introduit un modèle « de travail » réduit fondé sur les covariables observées. Ce modèle de travail réduit n'est pas nécessairement spécifié correctement, mais il peut être un outil utile pour intégrer les renseignements partiels provenant de données externes. La mise en œuvre en tant que telle est fondée sur une application nouvelle de la projection d'information et de la pondération par calage du modèle. La méthode proposée est particulièrement intéressante pour combiner des renseignements de plusieurs sources présentant différentes tendances en matière de données manquantes. La méthode est appliquée à un exemple de données réelles combinant les données d'enquête de l'enquête KNHANES (enquête nationale coréenne sur la santé et la nutrition) et les mégadonnées du NHISS (service national coréen de partage de l'assurance maladie).

Mots-clés : Mégadonnées; vraisemblance empirique; projection d'information; modèles d'erreur de mesure; covariables manquantes.

1. Introduction

L'intégration de données est un nouveau champ de recherche dans le domaine de l'échantillonnage d'enquête. En intégrant des renseignements partiels d'échantillons externes, nous pouvons améliorer l'efficacité de l'estimateur obtenu et augmenter la fiabilité de l'analyse. Lohr et Raghunathan (2017), Yang et Kim (2020) et Rao (2020) présentent des examens de méthodes statistiques d'intégration des données aux fins d'inférence d'une population finie. De nombreuses méthodes (par exemple Hidiroglou, 2001; Merkouris, 2010; Zubizarreta, 2015) visent principalement à estimer des moyennes ou des totaux de population, alors que la combinaison de renseignements à des fins d'inférence analytique, comme l'analyse de régression, n'est pas pleinement étudiée selon les ouvrages publiés.

Dans la présente étude, nous envisageons l'analyse de régression dans le contexte de l'intégration de données. Lorsque nous combinons les sources de données pour effectuer une analyse de régression combinée, nous pouvons nous heurter à certains problèmes, comme le fait que les covariables ne sont peut-être pas entièrement observées ou qu'elles sont sujettes à des erreurs de mesure. On peut donc considérer le problème comme un problème de régression où des covariables étaient manquantes. Robins, Rotnitzky et Zhao (1994) et Wang, Wang, Zhao et Ou (1997) ont traité de l'estimation semiparamétrique dans une analyse de régression où il manquait des données de covariables selon l'hypothèse de répartition au hasard des données de covariables manquantes. Dans notre scénario, la source de données externe pour laquelle il manque des covariables peut être un recensement ou des mégadonnées.

1. Zhonglei Wang, professeur adjoint, Wang Yanan Institute for Studies in Economics (WISE) and School of Economics, Xiamen University, Xiamen, (Fujian) 361005, Chine. Courriel : wangzl@xmu.edu.cn; Hang J. Kim, professeur agrégé, Division of Statistics and Data Science, University of Cincinnati, Cincinnati, (Ohio) 45221, États-Unis. Courriel : kim3h4@ucmail.uc.edu; Jae Kwang Kim, professeur, Department of Statistics, Iowa State University, Ames, (Iowa) 50011, États-Unis. Courriel : jkim@iastate.edu.

Dans le cadre de ce scénario, Chatterjee, Chen, Maas et Carroll (2016) ont élaboré une méthodologie d'intégration des données fondée sur le maximum de vraisemblance contraint, qui s'appuie sur un modèle entièrement paramétrique pour spécifier la vraisemblance et une contrainte tirée d'un modèle réduit d'intégration des données. La méthode du maximum de vraisemblance contraint est efficace quand le modèle est correctement spécifié, mais ne s'applique pas quand il est difficile ou impossible de spécifier une fonction de densité correcte. Kundu, Tang et Chatterjee (2019) ont généralisé la méthode de Chatterjee et coll. (2016) pour étudier plusieurs modèles de régression fondés sur la théorie de la méthode généralisée des moments (Hansen, 1982, GMM). Récemment, Xu et Shao (2020) ont élaboré une méthode d'intégration des données au moyen d'une méthode généralisée de la technique des moments, mais leur méthode suppose implicitement une spécification correcte du modèle réduit. Dans une étude de cas-témoins emboîtés, Shin, Pfeiffer, Graubard et Gail (2020a) ont proposé l'utilisation de l'échantillon entièrement observé à la phase 2 pour correspondre à un modèle paramétrique, et les covariables manquantes dans l'échantillon de la phase 1 sont imputées; voir également Shin, Pfeiffer, Graubard et Gail (2020b). Zhang, Deng, Wheeler, Qin et Yu (2021) ont élaboré un cadre de vraisemblance empirique rétrospective pour tenir compte du biais d'échantillonnage dans des études de cas-témoins. Sheng, Sun, Huang et Kim (2021) ont mis au point une méthode de vraisemblance empirique pénalisée pour intégrer de telles informations dans le scénario de régression logistique.

Pour combiner des renseignements partiels de sources externes, nous utilisons l'idée de calage au moyen d'un modèle (Wu et Sitter, 2001) qui introduit un modèle « de travail » réduit fondé sur les covariables observées. Dans le modèle réduit, les paramètres du modèle sont estimés à partir de sources externes, puis combinés selon une application nouvelle de la méthodologie de vraisemblance empirique (Owen, 1991; Qin et Lawless, 1994), qui peut être considérée comme une projection d'information (Csiszár et Shields, 2004). Le modèle de travail réduit n'est pas nécessairement spécifié correctement, mais un bon modèle de travail peut améliorer l'efficacité de l'analyse qu'il produit. La méthode proposée est particulièrement intéressante pour combiner des renseignements provenant de plusieurs sources de données présentant différentes tendances en matière de données manquantes. Dans ce cas, il suffit de spécifier différents modèles de travail pour différentes tendances en matière de données manquantes.

Par ailleurs, la méthode que nous proposons est fondée sur les conditions des premiers moments, comme les analyses de régression habituelles, si bien que des hypothèses faibles peuvent permettre d'élargir l'applicabilité de la méthode proposée à de nombreux problèmes pratiques. En particulier, la méthode proposée s'applique directement aux données d'échantillon d'enquête, qui est l'objet principal de notre étude. Nous examinons un scénario de régression plus général. Notre méthode de vraisemblance empirique proposé ne nécessite pas de modèle de travail réduit spécifié correctement.

Nous présenterons nos travaux dans le présent article comme suit. Premièrement, nous proposons un cadre unifié d'intégration de sources de données externes dans le contexte d'analyse de régression. La méthodologie proposée repose sur des hypothèses plus faibles que la méthodologie fondée sur des modèles paramétriques de Chatterjee et coll. (2016) et fournit donc des résultats d'estimation plus robustes.

Deuxièmement, la méthodologie proposée est largement applicable, car elle peut facilement traiter plusieurs sources de données externes, comme le montre la section 5. Elle peut également être appliquée quand la source de données externe présente un biais de sélection. Dans l'application sur des données réelles de la section 7, nous montrons que la méthodologie proposée peut s'appuyer sur des mégadonnées externes présentant des probabilités de sélection inconnues, en appliquant un ajustement de pondération du score de propension. Enfin, la méthodologie que nous proposons est facile à mettre en œuvre et pleinement justifiée en théorie. Le calcul est simple, puisqu'il s'agit d'une application directe de la méthode type de vraisemblance empirique et peut être mis en oeuvre au moyen de logiciels existants.

La présente étude est structurée de la manière suivante. La section 2 présente un scénario de base ainsi que les méthodologies existantes. La section 3 présente l'approche proposée et la section 4, ses propriétés asymptotiques. À la section 5, nous exposons une application à une intégration de données multiples. La section 6 présente des études par simulations, suivies de l'application de la méthodologie proposée à des données réelles à la section 7. Des conclusions sont proposées à la section 8.

2. Scénario de base

Supposons une population finie $\mathcal{U}=\{1,\ldots,N\}$ de taille N. Associé à l'unité i^{e}, supposons que y_i désigne la variable d'intérêt à l'étude et $\mathbf{x}_i=(\mathbf{x}_{i1},\mathbf{x}_{i2})$, le vecteur auxiliaire correspondant de longueur p. Nous souhaitons estimer un paramètre de population $\boldsymbol{\beta}_0$, qui résout $\mathbf{U}_1(\boldsymbol{\beta})=\sum_{i\in\mathcal{U}}\mathbf{U}_1(\boldsymbol{\beta};\mathbf{x}_i,y_i)=\mathbf{0}$, où $\mathbf{U}_1(\boldsymbol{\beta};\mathbf{x},y)$ est une fonction d'estimation prédéfinie pour $\boldsymbol{\beta}$. Un exemple de la fonction d'estimation est $\mathbf{U}_1(\boldsymbol{\beta};\mathbf{x}_i,y_i)=\{y_i-m_1(\mathbf{x}_i;\boldsymbol{\beta})\}\mathbf{h}_1(\mathbf{x}_i;\boldsymbol{\beta})$, qui est implicitement fondé sur un modèle de régression $E(Y_i|\,\mathbf{x}_i)=m_1(\mathbf{x}_i;\boldsymbol{\beta})$ au niveau de la superpopulation pour certains $\mathbf{h}_1(\mathbf{x}_i;\boldsymbol{\beta})$ satisfaisant à certaines conditions d'identification (par exemple, Kim et Rao, 2009). Dans la population finie, un échantillon probabiliste $S_1\subset\mathcal{U}$ est sélectionné et un estimateur de Z, $\hat{\boldsymbol{\beta}}$, peut être obtenu en résolvant l'équation :

$$\hat{\mathbf{U}}_1(\boldsymbol{\beta})\equiv\sum_{i\in S_1}d_{i1}\,\mathbf{U}_1(\boldsymbol{\beta};\mathbf{x}_i,y_i)=\mathbf{0}, \tag{2.1}$$

où d_i est le poids d'échantillonnage de l'unité $i\in S_1$.

En plus de S_1, supposons que nous observons $\mathbf{x}_{i1}$ et y_i dans toute la population finie et que nous souhaitons intégrer ces renseignements supplémentaires pour améliorer l'efficacité d'estimation de $\hat{\boldsymbol{\beta}}$. Avant de proposer notre méthodologie, nous présentons deux travaux connexes, notamment ceux de Chen et Chen (2000) et de Chatterjee et coll. (2016).

Chen et Chen (2000) ont d'abord examiné ce problème dans le contexte de modèles d'erreur de mesure. Pour expliquer leur idée dans notre scénario, nous considérons d'abord un modèle « de travail » réduit :

$$E(Y_i|\,\mathbf{x}_{i1})=m_2(\mathbf{x}_{i1};\boldsymbol{\alpha}) \tag{2.2}$$

pour certaines valeurs $\boldsymbol{\alpha}$. Selon le modèle de travail (2.2), nous pouvons obtenir un estimateur $\hat{\boldsymbol{\alpha}}$ à partir de l'échantillon actuel S_1 en résolvant l'équation :

$$\hat{\mathbf{U}}_2(\boldsymbol{\alpha}) \equiv \sum_{i \in \mathcal{S}_1} d_i \, \mathbf{U}_2\left(\boldsymbol{\alpha}; \mathbf{x}_{i1}, y_i\right) = \mathbf{0}, \tag{2.3}$$

où $\mathbf{U}_2\left(\boldsymbol{\alpha}; \mathbf{x}_{i1}, y_i\right) = \left\{y_i - m_2\left(\mathbf{x}_{i1}; \boldsymbol{\alpha}\right)\right\} \mathbf{h}_2\left(\mathbf{x}_{i1}; \boldsymbol{\alpha}\right)$ pour certaines valeurs $\mathbf{h}_2\left(\mathbf{x}_{i1}; \boldsymbol{\alpha}\right)$ satisfaisant aux conditions similaires à celles imposées à $\mathbf{h}_1\left(\mathbf{x}_i; \boldsymbol{\beta}\right)$. De plus, nous pouvons obtenir $\boldsymbol{\alpha}^*$ en résolvant $\sum_{i=1}^{N} \mathbf{U}_2\left(\boldsymbol{\alpha}; \mathbf{x}_{i1}, y_i\right) = \mathbf{0}$. Chen et Chen (2000) ont proposé d'utiliser :

$$\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} + \widehat{\mathrm{Cov}}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}\right) \left\{\hat{V}\left(\hat{\boldsymbol{\alpha}}\right)\right\}^{-1} \left(\boldsymbol{\alpha}^* - \hat{\boldsymbol{\alpha}}\right)$$

comme estimateur efficace de $\boldsymbol{\beta}$, où $\hat{V}(\cdot)$ et $\widehat{\mathrm{Cov}}(\cdot)$ désignent respectivement les estimateurs de variance et de covariance fondés sur le plan. Le modèle de travail de (2.2) n'est pas nécessairement spécifié correctement, mais un bon modèle de travail peut améliorer l'efficacité de l'estimateur définitif. Alors que l'estimateur de Chen et Chen (2000) est théoriquement justifié, il peut être instable numériquement, car les erreurs d'estimation de la variance et de la matrice de covariance peuvent être importantes.

Chatterjee et coll. (2016) ont envisagé une approche fondée sur la vraisemblance à l'aide d'une distribution conditionnelle de Y_i, avec $\mathbf{X}_i$, de densité $f\left(y_i \middle| \mathbf{x}_i; \boldsymbol{\beta}\right)$, et ont imposé une contrainte fondée sur des renseignements externes. En particulier, ils proposent de maximiser :

$$\prod_{i \in \mathcal{S}_1} f\left(y_i \middle| \mathbf{x}_i; \boldsymbol{\beta}\right) dF\left(\mathbf{x}_i\right) \tag{2.4}$$

en fonction de la contrainte :

$$\int\int \mathbf{U}_2\left(\boldsymbol{\alpha}^*; \mathbf{x}_1, y\right) f\left(y \middle| \mathbf{x}; \boldsymbol{\beta}\right) dy dF(\mathbf{x}) = \mathbf{0}, \tag{2.5}$$

où $F(\mathbf{x})$ est une fonction de distribution non spécifiée pour $\mathbf{x}$, $dF(\mathbf{x})$ est la dérivée de Radon-Nikodym de la fonction de répartition $F(\mathbf{x})$ pour une certaine mesure dominante et $\boldsymbol{\alpha}^*$ est le paramètre du modèle disponible d'une source externe. Suivant l'approche fondée sur la vraisemblance de Chatterjee et coll. (2016), $\mathbf{U}_2\left(\boldsymbol{\alpha}; \mathbf{x}_{i1}, y\right)$ correspond à la fonction d'estimation faisant intervenir une fonction de distribution « réduite » $g\left(y_i \middle| \mathbf{x}_{i1}; \boldsymbol{\alpha}_0\right)$ avec un paramètre de modèle $\boldsymbol{\alpha}_0$, où $g\left(y_i \middle| \mathbf{x}_{i1}; \alpha_0\right)$ peut être spécifié incorrectement. En d'autres termes, $\boldsymbol{\alpha}^*$ est l'information externe pour $\boldsymbol{\alpha}_0$. Chatterjee et coll. (2016) ont estimé $F(\mathbf{x})$ de façon non paramétrique au moyen de la vraisemblance empirique. En imposant cette contrainte dans l'estimation par la méthode du maximum de vraisemblance, on peut intégrer naturellement l'information externe $\boldsymbol{\alpha}^*$.

La méthode du maximum de vraisemblance contraint (MVC) n'est pas directement applicable à notre modèle conditionnel moyen de (2.1), car la fonction de vraisemblance pour $\boldsymbol{\beta}$ n'est pas définie dans notre scénario. De plus, la caractéristique du plan pour l'échantillon probabiliste S_1 n'est pas directement applicable dans leur méthode. Néanmoins, on peut utiliser une fonction objective, comme celle d'une méthode généralisée des moments, pour appliquer le problème d'optimisation contrainte, qui est asymptomatiquement équivalent à la méthode de vraisemblance empirique (Imbens, 2002). La mise en œuvre de la vraisemblance empirique de l'approche du MVC fait l'objet d'une discussion par Han et Lawless (2019).

3. Approche proposée

Considérons maintenant une autre approche permettant de combiner des renseignements de plusieurs sources. Pour cela, nous utilisons la mesure de divergence Kullback-Leibler (KL), afin d'appliquer la projection d'information (Csiszár et Shields, 2004) sur l'espace du modèle avec contraintes. Soit $\hat{P}$, la distribution empirique de l'échantillon :

$$\hat{P}(x, y) = \frac{1}{\sum_{i \in \mathcal{S}_1} d_i} \sum_{i \in \mathcal{S}_1} d_i \, \mathbf{I}\{(x, y) = (x_i, y_i)\}. \tag{3.1}$$

Étant donnée la distribution empirique $\hat{P}$, nous souhaitons trouver le minimiseur de :

$$D(\hat{P} \,\|\, P) = \int \log\{d\hat{P}(\mathbf{x}, y)\}\, d\hat{P}(\mathbf{x}, y) - \int \log\{dP(\mathbf{x}, y)\}\, d\hat{P}(\mathbf{x}, y) \tag{3.2}$$

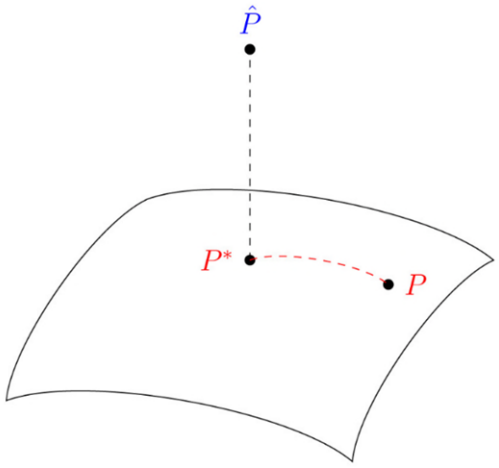
relativement à P dans l'espace du modèle. Remarquez que le premier terme est une constante et que le minimiseur de (3.2) est l'estimateur du pseudo maximum de vraisemblance de $\hat{P}$.

Nous considérons les contraintes suivantes dans notre modèle au niveau de la population finie :

$$\sum_{i=1}^{N} \mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i)\, p(\mathbf{x}_i, y_i) = 0 \qquad \text{et} \qquad \sum_{i=1}^{N} \mathbf{U}_2(\boldsymbol{\alpha}^*; \mathbf{x}_{i1}, y_i)\, p(\mathbf{x}_i, y_i) = 0, \tag{3.3}$$

où $p(\mathbf{x}_i, y_i)$ est le point matériel attribué au point $(\mathbf{x}_i, y_i)$ dans la population finie satisfaisant $\sum_{i=1}^{N} p(\mathbf{x}_i, y_i) = 1$. La figure 3.1 fournit une représentation graphique de cette projection d'information.

Figure 3.1 Projection d'information pour la répartition empirique $\hat{P}$.



Il convient de mentionner que P^* minimise $D(\hat{P} \,\|\, P)$ au sein de P satisfaisant aux contraintes de (3.3).

Lorsque nous utilisons la répartition empirique pondérée de (3.1), la mesure de la divergence KL de (3.2) se réduit à $D(\hat{P} \,\|\, P) = \text{constante} - \hat{N}^{-1} \sum_{i \in \mathcal{S}_1} d_i \log\{p(\mathbf{x}_i, y_i)\}$, où $\hat{N} = \sum_{i \in \mathcal{S}_1} d_i$. Ainsi, nous avons uniquement à maximiser $l(\mathbf{p}) = \sum_{i \in \mathcal{S}_1} d_i \log(p_i)$ en fonction de $\sum_{i=1}^{N} p_i = 1$ et des contraintes de (3.3), où

p_i simplifie $p(\mathbf{x}_i, y_i)$. Il convient de mentionner que $p_i > 0$ pour $i \notin \mathcal{S}_1$ diminuera la valeur de $l(\mathbf{p}) = \sum_{i \in \mathcal{S}_1} d_i \log(p_i)$; la solution $\hat{p}_i$ de ce problème d'optimisation devrait donner $\hat{p}_i = 0$ pour $i \notin S_1$. Nous pouvons donc définir $p_i = 0$ pour $i \notin S_1$ et exprimer le problème comme la recherche du maximiseur de :

$$Q(\mathbf{d}, \mathbf{w}) = \sum_{i \in \mathcal{S}_1} d_i \log(w_i) \tag{3.4}$$

en fonction de la contrainte :

$$\sum_{i \in \mathcal{S}_1} w_i = 1, \tag{3.5}$$

$$\sum_{i \in \mathcal{S}_1} w_i \mathbf{U}_2(\boldsymbol{\alpha}^*; \mathbf{x}_{i1}, y_i) = \mathbf{0}, \tag{3.6}$$

$$\sum_{i \in \mathcal{S}_1} w_i \mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i) = \mathbf{0}.$$

Nous utilisons w_i au lieu de p_i pour représenter les poids finaux attribués aux éléments de l'échantillon.

Remarque 1. *Maximiser la fonction objective dans (3.4) équivaut à minimiser l'entropie croisée suivante :*

$$-\sum_{i \in \mathcal{S}_1} \tilde{d}_i \log(w_i), \tag{3.7}$$

où $\tilde{d}_i = d_i / \left(\sum_{i \in \mathcal{S}_1} d_i\right)$. *La fonction objective (3.7) est également la fonction logarithme du rapport de vraisemblance pseudo empirique qu'ont étudiée Chen et Sitter (1999) ainsi que Wu et Rao (2006). Au lieu de (3.4), nous pouvons considérer d'autres fonctions objectives, notamment la vraisemblance empirique de population proposée par Chen et Kim (2014), par exemple.*

La méthodologie que nous proposons est différente de celle de Chatterjee et coll. (2016), en ce que nous utilisons une contrainte intégrale plus générale (2.5) qui ne fait pas intervenir la fonction conditionnelle de densité $f(y | \mathbf{x}; \boldsymbol{\beta})$. La contrainte (3.6) intègre toujours l'information supplémentaire dans $\boldsymbol{\alpha}^*$. L'optimisation susmentionnée peut être résolue en appliquant la méthodologie de vraisemblance empirique de profil standard ou en utilisant la méthodologie suivante d'estimation en deux étapes.

1. Déterminer des poids de calage $\hat{\mathbf{w}} = \{\hat{w}_i : i \in \mathcal{S}_1\}$ en maximisant $Q(\mathbf{d}, \mathbf{w})$ en fonction de (3.5) et (3.6).
2. Une fois la solution $\hat{\mathbf{w}}$ obtenue à partir du calage, estimer $\boldsymbol{\beta}$ en résolvant :

$$\sum_{i \in \mathcal{S}_1} \hat{w}_i \mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i) = \mathbf{0}. \tag{3.8}$$

Si la donnée repère $\boldsymbol{\alpha}^*$ n'est pas disponible à partir de la population finie, mais peut être estimée à partir d'un échantillon externe indépendant, nous pouvons utiliser l'information de l'échantillon interne d'origine et de l'échantillon externe pour obtenir l'estimation de la donnée repère. En pratique, nous n'avons pas nécessairement accès aux données brutes de l'échantillon externe, mais nous sommes souvent en mesure d'obtenir ses statistiques sommaires. Supposons que l'échantillon externe fournit un estimateur ponctuel

$\hat{\boldsymbol{\alpha}}_2$ et son estimateur de variance $\mathbf{V}_2 = \hat{V}(\hat{\boldsymbol{\alpha}}_2)$ pour le modèle de travail réduit de (2.2). Il est alors possible d'obtenir un estimateur de la donnée repère $\boldsymbol{\alpha}^*$ par :

$$\hat{\boldsymbol{\alpha}}^* = \left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1}\left(\mathbf{V}_1^{-1}\hat{\boldsymbol{\alpha}}_1 + \mathbf{V}_2^{-1}\hat{\boldsymbol{\alpha}}_2\right), \tag{3.9}$$

où $\hat{\boldsymbol{\alpha}}_1$ et $\mathbf{V}_1$ sont estimés avec l'échantillon interne $\mathcal{S}_1$. Une fois la valeur $\hat{\boldsymbol{\alpha}}^*$ obtenue à partir de (3.9), elle remplace $\boldsymbol{\alpha}^*$ dans l'équation de calage de (3.6).

De la même façon que Wu et Sitter (2001), la méthodologie proposée ne nécessite pas de modèle de travail « réel », comme il est expliqué ci-dessous. Soit $\hat{\mathbf{U}}_{\text{ext}}(\boldsymbol{\alpha}) = 0$ l'équation d'estimation pour obtenir la valeur $\boldsymbol{\alpha}^*$ calculée à partir de l'échantillon externe $\mathcal{S}_2$. Nous pouvons maintenant calculer l'approximation de la fonction d'estimation définitive pour $\boldsymbol{\beta}$ en utilisant l'expression de calage de modèle $\hat{\mathbf{U}}_{\text{cal}}(\boldsymbol{\beta}) = \sum_{i\in\mathcal{S}_1} \hat{w}_i \mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i)$ comme suit :

$$\hat{\mathbf{U}}_{\text{cal}}(\boldsymbol{\beta}) \doteq \hat{\mathbf{U}}_1(\boldsymbol{\beta}) + \mathbf{K}\left\{\hat{\mathbf{U}}_{\text{ext}}(\boldsymbol{\alpha}^*) - \hat{\mathbf{U}}_2(\boldsymbol{\alpha}^*)\right\} \tag{3.10}$$

pour certaines valeurs $\mathbf{K}$, où $\hat{\mathbf{U}}_1(\boldsymbol{\beta})$ et $\hat{\mathbf{U}}_2(\boldsymbol{\alpha})$ sont respectivement calculées à l'aide de (2.1) et (2.3), à partir de l'échantillon interne $\mathcal{S}_1$. L'approximation dans (3.10) peut être facilement dérivée à l'aide de l'équivalence asymptotique de l'estimateur de calage et de l'estimateur de régression. Ainsi, même si $E\left\{\hat{\mathbf{U}}_{\text{ext}}(\boldsymbol{\alpha}^*)\right\}$ n'est pas égal à zéro, la résolution de $\hat{\mathbf{U}}_{\text{cal}}(\boldsymbol{\beta}) = 0$ est convergente, puisque $E\left\{\hat{\mathbf{U}}_{\text{ext}}(\boldsymbol{\alpha}) - \hat{\mathbf{U}}_2(\boldsymbol{\alpha})\right\} = 0$ par définition.

Remarque 2. *Même s'il n'est pas nécessaire de spécifier correctement le modèle de travail* $E(Y_i | \mathbf{x}_{i1}) = m_2(\mathbf{x}_{i1}; \boldsymbol{\alpha})$, *nous pouvons systématiquement trouver* $\mathbf{U}_2(\boldsymbol{\alpha}; \mathbf{x}_{i1}, y_i)$ *en calculant sa construction comme un problème de covariable manquante, en nous fiant à la technique de calage de régression. Supposons, par exemple, que* $\mathbf{x}_i = (x_{i1}, x_{i2})$, *nous définissons une variable explicative* $\hat{x}_{i2} = \beta_0 + \beta_1 x_{i1}$ *et une équation d'estimation sous la forme :*

$$\mathbf{U}_1(\boldsymbol{\beta}; x_{i1}, \hat{x}_{i2}, y_i) = \left\{y_i - m_1(x_{i1}, \hat{x}_{i2}; \boldsymbol{\beta})\right\}\mathbf{h}_1(x_{i1}, \hat{x}_{i2}; \boldsymbol{\beta}) \tag{3.11}$$

pour la fonction de contrôle de la méthode de calage assisté par un modèle où $\boldsymbol{\beta} = (\beta_0, \beta_1)$. *Nous pouvons soit estimer* $\boldsymbol{\beta}$ *à partir de l'échantillon* $\mathcal{S}_1$, *soit utiliser une valeur de paramètre fixe tant que la solution à* $\sum_{i\in\mathcal{S}_1} d_i \mathbf{U}_1(\boldsymbol{\beta}; x_{i1}, \hat{x}_{i2}, y_i) = \mathbf{0}$ *est unique. Un estimateur de donnée repère de* $\boldsymbol{\beta}$ *peut être obtenu à l'aide d'échantillons externes pour appliquer la méthodologie de calage de modèle proposée. Si nous utilisons la fonction de contrôle de (3.11), nous traitons donc essentiellement une régression de* y *sur* x_1 *et* $\hat{x}_2$ *comme modèle « de travail » pour le calage. Cela est uniquement possible lorsque nous disposons d'un accès direct à un échantillon externe* $\mathcal{S}_2$ *en plus de l'échantillon interne* $\mathcal{S}_1$.

4. Propriétés théoriques

Dans la section qui suit, nous examinons les propriétés asymptotiques de l'estimateur proposé $\hat{\boldsymbol{\beta}}$ pour (3.8). Puisque les paramètres de population, notamment $\boldsymbol{\beta}_0$ et $\boldsymbol{\alpha}^*$, sont déterminés par la population finie de taille N, nous utilisons explicitement un indice N pour les paramètres de population de la présente

section, par exemple, $\boldsymbol{\beta}_{0N}$ et $\boldsymbol{\alpha}_N^*$, mais nous omettons cet indice pour $(d_i, \mathbf{x}_i, y_i)$ par souci de simplicité. Nous considérons deux scénarios : lorsque $\boldsymbol{\alpha}_N^*$ est disponible à partir de la population finie et lorsque nous ne disposons que d'un échantillon externe pour estimer $\boldsymbol{\alpha}_N^*$ par la méthode des moindres carrés généralisés de (3.9).

4.1 $\boldsymbol{\alpha}_N^*$ est disponible

Soit $\tilde{d}_i = \hat{N}^{-1} d_i$, où $\hat{N} = \sum_{i \in \mathcal{S}_1} d_i$ est l'estimateur Horvitz-Thompson de la taille de population N. En remplaçant d_i par $\tilde{d}_i$ dans (3.4), nous considérons le problème de type lagrangien qui maximise :

$$l(\mathbf{w}, \boldsymbol{\lambda}, \phi) = \sum_{i \in \mathcal{S}_1} \tilde{d}_i \log(w_i) + \boldsymbol{\lambda}^\top \sum_{i \in \mathcal{S}_1} w_i \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right) + \phi\left(\sum_{i \in \mathcal{S}_1} w_i - 1\right),$$

où $\boldsymbol{\lambda}$ et ϕ sont les multiplicateurs de Lagrange.

En définissant $\partial l(\mathbf{w}, \boldsymbol{\lambda}, \phi) / \partial \boldsymbol{\lambda} = \mathbf{0}$, $\partial l(\mathbf{w}, \boldsymbol{\lambda}, \phi) / \partial \phi = 0$ et $\partial l(\mathbf{w}, \boldsymbol{\lambda}, \phi) / \partial w_i = 0$ pour $i \in \mathcal{S}_1$, nous obtenons $\hat{\phi} = -1$ et $\hat{w}_i = \tilde{d}_i \left\{1 - \boldsymbol{\lambda}^\top \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right)\right\}^{-1}$. La méthode proposée est alors équivalente à la résolution de $g(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathbf{0}$, où :

$$g(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \begin{pmatrix} \sum_{i \in \mathcal{S}_1} \dfrac{\tilde{d}_i}{1 - \boldsymbol{\lambda}^\top \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right)} \mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i) \\ \sum_{i \in \mathcal{S}_1} \dfrac{\tilde{d}_i}{1 - \boldsymbol{\lambda}^\top \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right)} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right) \end{pmatrix}. \tag{4.1}$$

Désignons la solution de (4.1) par $\hat{\boldsymbol{\eta}} = \left(\hat{\boldsymbol{\beta}}^\top, \boldsymbol{\lambda}^\top\right)^\top$. Pour examiner les propriétés asymptotiques de $\hat{\boldsymbol{\eta}}$, nous proposons les conditions de régularité suivantes.

C1. Il existe un ensemble compact $\mathcal{A}$, de telle sorte que $Z_S = \sup_{\boldsymbol{\alpha} \in \mathcal{A}} \max_{i \in \mathcal{S}_1} \left\| \mathbf{U}_2(\boldsymbol{\alpha}; \mathbf{x}_{i1}, y_i) \right\| = o_p\left(n^{1/2}\right)$ et $\boldsymbol{\alpha}_N^* \in \mathcal{A}$ pour $N \in \mathbb{N}$, où $\|\cdot\|$ désigne la norme euclidienne et l'ordre stochastique par rapport au plan d'échantillonnage.

C2. Le plan d'échantillonnage satisfait aux résultats de convergence suivants.

a. Il existe un ensemble compact Ω, de telle sorte que $\boldsymbol{\beta}_{0N} \in \Omega$ pour $N \in \mathbb{N}$ et un point intérieur de Ω, $\boldsymbol{\beta}_p$, de telle sorte que $\lim_{N \to \infty} \boldsymbol{\beta}_{0N} = \boldsymbol{\beta}_p$.

b. Il existe une fonction continue $\mathbf{U}_0(\boldsymbol{\beta})$ sur Ω, de telle sorte que $\sup_{\boldsymbol{\beta} \in \Omega} \left\| \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i) - \mathbf{U}_0(\boldsymbol{\beta}) \right\| \to 0$ en probabilité, où $\boldsymbol{\beta}_p$ est l'unique solution de $\mathbf{U}_0(\boldsymbol{\beta}) = \mathbf{0}$.

c. $\sum_{i \in \mathcal{S}_1} \tilde{d}_i \partial \mathbf{U}_1(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i) / \partial \boldsymbol{\beta}^\top = \boldsymbol{\mathcal{I}}_{11} + o_p(1)$, où $\boldsymbol{\mathcal{I}}_{11}$ est non stochastique et inversible.

d. $\sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i) \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right)^\top = \boldsymbol{\mathcal{I}}_{12} + o_p(1)$, où $\boldsymbol{\mathcal{I}}_{12}$ est non stochastique.

e. $\sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right)^{\otimes 2} = \boldsymbol{\mathcal{I}}_{22} + o_p(1)$, où $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^\top$ pour toute matrice $\mathbf{A}$ et $\boldsymbol{\mathcal{I}}_{22}$ est non stochastique et positivement définitive.

C3. Le plan d'échantillonnage satisfait à :

$$n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\begin{pmatrix}\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i\right)\\ \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{i1},y_i\right)\end{pmatrix}\to\mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}_u\right)$$

dans une distribution où $\mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}_u\right)$ est une distribution normale de moyenne nulle et de matrice de covariance :

$$\boldsymbol{\Sigma}_u=\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}.$$

C1 est une condition technique pour obtenir l'ordre asymptotique de $\hat{\boldsymbol{\lambda}}$ et Wu et Rao (2006) supposent également une condition similaire; voir la condition C1 que ceux-ci ont établie pour obtenir de plus amples renseignements. C2 suppose plusieurs résultats de convergence pour les deux fonctions d'estimation. Plus précisément, C2a désigne l'espace de paramètre pour le paramètre $\boldsymbol{\beta}_{0N}$ de population finie et la convergence de $\boldsymbol{\beta}_{0N}$ peut être satisfaite dans des conditions de régularité. La condition C2b est nécessaire pour désigner $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_p\to 0$ en probabilité, puis $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0N}\to 0$ en probabilité, couplé avec C2a. Les conditions C2c-C2e garantissent le théorème de limite centrale pour $\hat{\boldsymbol{\eta}}$. Fait à noter, $\boldsymbol{\mathcal{I}}_{22}$ est symétrique selon C2e, mais $\boldsymbol{\mathcal{I}}_{11}$ de C2c peut être asymétrique pour une certaine fonction d'estimation $\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x},y\right)$. La condition C3 est satisfaite dans les conditions de régularité pour les plans d'échantillonnage généraux; voir Fuller (2009, section 1.3) pour obtenir de plus amples renseignements.

Théorème 1. *Supposons que les conditions C1-C3 sont vérifiées. Alors,* $n^{1/2}\left(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\right)\to\mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}_\eta\right)$ *dans une distribution où* $\boldsymbol{\Sigma}_\eta=\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Sigma}_u\left(\boldsymbol{\mathcal{I}}^{-1}\right)^{\mathrm{T}}$ *et :*

$$\boldsymbol{\mathcal{I}}=\begin{pmatrix}\boldsymbol{\mathcal{I}}_{11} & \boldsymbol{\mathcal{I}}_{12}\\ \mathbf{0} & \boldsymbol{\mathcal{I}}_{22}\end{pmatrix}.$$

La démonstration du théorème 1 est présentée en annexe A. D'après le théorème 1, nous pouvons déduire que $n^{1/2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0N}\right)\to\mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}_\beta\right)$ dans une distribution où :

$$\boldsymbol{\Sigma}_\beta=\boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\Sigma}_{11}\left(\boldsymbol{\mathcal{I}}_{11}^{-1}\right)^{\mathrm{T}}-\boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\Sigma}_{21}\left(\boldsymbol{\mathcal{I}}_{11}^{-1}\right)^{\mathrm{T}}-\boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{12}^{\mathrm{T}}\left(\boldsymbol{\mathcal{I}}_{11}^{-1}\right)^{\mathrm{T}}+\boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{12}^{\mathrm{T}}\left(\boldsymbol{\mathcal{I}}_{11}^{-1}\right)^{\mathrm{T}}$$

et $\boldsymbol{\Sigma}_{11}$ et $\boldsymbol{\Sigma}_{22}$ correspondent respectivement aux variances asymptotiques de $n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i\right)$ et $n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{i1},y_i\right)$. De plus, nous avons le résultat suivant pour l'optimalité de $\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{i1},y_i\right)$.

Corollaire 1. *Supposons que les conditions du théorème 1 se vérifient. Pour une fonction d'estimation fixe* $\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x},y\right),\hat{\boldsymbol{\beta}}$ *est optimal si* $\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_1,y\right)=E\left\{\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x},y\right)\middle|\,\mathbf{x}_1,y\right\}$ *se vérifie presque certainement pour le modèle de travail réduit, où* $\mathbf{x}=\left(\mathbf{x}_1,\mathbf{x}_2\right)$ *et l'espérance dépend du modèle de superpopulation.*

La démonstration du corollaire 1 est présentée à l'annexe B. Le corollaire 1 présente une condition suffisante pour le modèle réduit permettant de garantir un estimateur optimal $\hat{\boldsymbol{\beta}}$, si le modèle de travail est correctement spécifié; autrement dit, même si nous n'avons pas besoin d'un modèle réduit spécifié correctement par souci de cohérence, le gain d'efficacité est garanti uniquement pour la spécification

correcte du modèle. Selon le corollaire 1, un estimateur optimal de $\boldsymbol{\alpha}_N^*$ peut être obtenu en résolvant $E\left\{\mathbf{U}_1\left(\boldsymbol{\beta}_{0N}; \mathbf{x}, y\right) \middle| \mathbf{x}_1, y\right\} = \mathbf{0}$.

Dans des conditions de régularité, il est possible de démontrer que $\boldsymbol{\Sigma}_\beta = \boldsymbol{\mathcal{I}}_{11}^{-1}\left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)\left(\boldsymbol{\mathcal{I}}_{11}^{-1}\right)^\top$ pour un échantillonnage aléatoire simple avec ou sans remise. Puisque $\boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\Sigma}_{11}\left(\boldsymbol{\mathcal{I}}_{11}^{-1}\right)^\top$ est la variance asymptotique de $n^{1/2}\left(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_{0N}\right)$, où $\hat{\boldsymbol{\beta}}_m$ résout $\sum_{i\in\mathcal{S}_1} d_i \mathbf{U}_1\left(\boldsymbol{\beta}; \mathbf{x}_i, y_i\right) = 0$, l'approche proposée offre une estimation efficace pour un échantillonnage aléatoire simple; voir la section S1 de la documentation supplémentaire pour obtenir de plus amples renseignements.

4.2 Un estimateur externe $\hat{\boldsymbol{\alpha}}_2$ est disponible

Lorsque $\boldsymbol{\alpha}^*$ n'est pas disponible, mais qu'un échantillon externe l'est pour obtenir $\hat{\boldsymbol{\alpha}}^*$ de (3.9), nous considérons :

$$\tilde{g}\left(\boldsymbol{\eta}\right) = \begin{pmatrix} \sum\limits_{i\in\mathcal{S}_1} \dfrac{\tilde{d}_i}{1 - \boldsymbol{\lambda}^\top \mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*; \mathbf{x}_{i1}, y_i\right)} \mathbf{U}_1\left(\boldsymbol{\beta}; \mathbf{x}_i, y_i\right) \\ \sum\limits_{i\in\mathcal{S}_1} \dfrac{\tilde{d}_i}{1 - \boldsymbol{\lambda}^\top \mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*; \mathbf{x}_{i1}, y_i\right)} \mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*; \mathbf{x}_{i1}, y_i\right) \end{pmatrix}. \tag{4.2}$$

Désignons par $\tilde{\boldsymbol{\eta}}$ la solution de $\tilde{g}\left(\boldsymbol{\eta}\right) = \mathbf{0}$. Les hypothèses supplémentaires suivantes sont ensuite nécessaires pour obtenir les propriétés asymptotiques pour $\tilde{\boldsymbol{\eta}}$.

C4. $\sum_{i\in\mathcal{S}_1} \tilde{d}_i \partial \mathbf{U}_2\left(\boldsymbol{\alpha}; \mathbf{x}_{i1}, y_i\right) / \partial \boldsymbol{\alpha}^\top = \boldsymbol{\mathcal{I}}\left(\boldsymbol{\alpha}\right) + o_p\left(1\right)$ uniformément pour $\boldsymbol{\alpha} \in \mathcal{A}$ où $\boldsymbol{\mathcal{I}}\left(\boldsymbol{\alpha}\right)$ est non stochastique. De plus, il existe une matrice inversible $\boldsymbol{\mathcal{I}}_0$, de telle sorte que $\lim_{N\to\infty} \boldsymbol{\mathcal{I}}\left(\boldsymbol{\alpha}_N^*\right) = \boldsymbol{\mathcal{I}}_0$.

C5. Le plan d'échantillonnage et l'échantillon externe satisfont aux résultats de convergence suivants.

a. $\hat{\boldsymbol{\alpha}}_1$ et $\hat{\boldsymbol{\alpha}}_2$ sont convergents pour $\boldsymbol{\alpha}^*$.

b. $\mathbf{V}_1$ et $\mathbf{V}_2$ sont respectivement des estimateurs de variance convergents de plan de sondage de $\hat{\boldsymbol{\alpha}}_1$ et $\hat{\boldsymbol{\alpha}}_2$.

c. $\mathbf{V}_1^{-1}$, $\mathbf{V}_2^{-1}$ et $\left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1}$ existent en probabilité.

d. $\left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1} \mathbf{V}_2^{-1} = \mathbf{W} + o_p\left(1\right)$, où $\mathbf{W}$ est non stochastique.

e. Il existe une fonction de démultiplication $\gamma(n)$, de telle sorte que $\gamma(n)\left(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}^*\right) \to \mathcal{N}\left(0, \boldsymbol{\Sigma}_2\right)$ dans une distribution où $\boldsymbol{\Sigma}_2$ satisfait à $\gamma(n)^2 \mathbf{V}_2 = \boldsymbol{\Sigma}_2 + o_p\left(1\right)$.

C4 est utilisée pour obtenir l'ordre asymptotique et la variance de $\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha}_N^*$; Yuan et Jennrich (1998) ont utilisé une condition similaire. C5a et C5b supposent la convergence de $\hat{\boldsymbol{\alpha}}_2$ et $\mathbf{V}_2$ obtenue par un échantillon externe. Pour la convergence de $\hat{\boldsymbol{\alpha}}_1$, une condition suffisante est similaire pour ce qui est de C2b. La convergence par rapport au plan de l'estimateur de variance $\mathbf{V}_1$ peut être obtenue pour les plans d'échantillonnage généraux; voir Fuller (2009, chapitre 1) pour obtenir de plus amples renseignements. C5c garantit l'existence de $\hat{\boldsymbol{\alpha}}^*$ pour la méthode proposée. C5e démontre le théorème de la limite centrale relativement à la statistique sommaire $\hat{\boldsymbol{\alpha}}_2$ et est utilisée pour dériver un résultat similaire à C3, $\boldsymbol{\alpha}^*$ étant

remplacé par $\hat{\boldsymbol{\alpha}}^*$. Plus précisément, le taux de convergence de $\left(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}^*\right)$ est $\gamma(n)^{-1}$, déterminé par l'échantillon externe.

Le théorème suivant établit une distribution asymptotique similaire à celle de C3.

Théorème 2. *Supposons que les conditions C1 et C3-C5 sont vérifiées. Alors :*

$$n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\begin{pmatrix}\mathbf{U}_1\left(\boldsymbol{\beta}_0;\mathbf{x}_i,y_i\right)\\ \mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{i1},y_i\right)\end{pmatrix}\to\mathcal{N}\left(\mathbf{0},\tilde{\boldsymbol{\Sigma}}_u\right)$$

dans une distribution où :

$$\tilde{\boldsymbol{\Sigma}}_u=\begin{pmatrix}\tilde{\boldsymbol{\Sigma}}_{11} & \tilde{\boldsymbol{\Sigma}}_{12}\\ \tilde{\boldsymbol{\Sigma}}_{21} & \tilde{\boldsymbol{\Sigma}}_{22}\end{pmatrix}.$$

Cas 1. Plus précisément, s'il existe une matrice non stochastique $\boldsymbol{\Sigma}_c$, *de telle sorte que* $n\mathbf{V}_2=\boldsymbol{\Sigma}_c+o_p(1)$, *alors* $\tilde{\boldsymbol{\Sigma}}_{11}=\boldsymbol{\Sigma}_{11}$, $\tilde{\boldsymbol{\Sigma}}_{12}=\boldsymbol{\Sigma}_{12}\left(\boldsymbol{\mathcal{I}}_0^{-1}\right)^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{\mathcal{I}}_0^{\mathrm{T}}$, $\tilde{\boldsymbol{\Sigma}}_{21}=\tilde{\boldsymbol{\Sigma}}_{12}^{\mathrm{T}}$ *et* $\tilde{\boldsymbol{\Sigma}}_{22}=\boldsymbol{\mathcal{I}}_0\mathbf{W}\left\{\boldsymbol{\Sigma}_c+\boldsymbol{\mathcal{I}}_0^{-1}\boldsymbol{\Sigma}_{22}(\boldsymbol{\mathcal{I}}_0^{-1})^{\mathrm{T}}\right\}\mathbf{W}^{\mathrm{T}}\boldsymbol{\mathcal{I}}_o^{\mathrm{T}}$;
Cas 2. Si $\mathbf{W}=0$, *alors* $\tilde{\boldsymbol{\Sigma}}_{ij}=\mathbf{0}$ *pour* $(i,j)\neq(1,1)$ *et* $\tilde{\boldsymbol{\Sigma}}_{11}=\boldsymbol{\Sigma}_{11}$.

La démonstration du théorème 2 est présentée en annexe C. Pour le cas 1, si $\hat{\boldsymbol{\alpha}}_2$ estimé à partir d'un échantillon externe est bien plus efficace que $\hat{\boldsymbol{\alpha}}$, dans le sens de $\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)=o_p\left(n^{-1/2}\right)$, alors $\mathbf{W}$ est une matrice d'identité et $\tilde{\boldsymbol{\Sigma}}_{ij}=\boldsymbol{\Sigma}_{ij}$ pour $i,j=1,2$. Par conséquent, nous pouvons ignorer la variabilité de la statistique sommaire $\hat{\boldsymbol{\alpha}}_2$ de l'échantillon externe et obtenir la même distribution asymptotique que dans C3. Même si ces distributions asymptotiques sont les mêmes, C3 avec $\boldsymbol{\alpha}_N^*$ connu n'est pas un cas spécial du théorème 2, puisque $\hat{\boldsymbol{\alpha}}_2=\boldsymbol{\alpha}_N^*$ a une variance nulle, ce qui ne respecte pas C5c-C5e. Par ailleurs, si $\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)\asymp n^{-1/2}$ en probabilité, alors $\hat{\boldsymbol{\alpha}}_2$ est aussi efficace que $\hat{\boldsymbol{\alpha}}_1$. Par conséquent, $\mathbf{W}$ n'est pas une matrice d'identité ni une matrice nulle et la méthode proposée est plus efficace qu'une méthode remplaçant $\boldsymbol{\alpha}^*$ par $\hat{\boldsymbol{\alpha}}^*=\hat{\boldsymbol{\alpha}}_2$, du fait de l'information supplémentaire fournie par l'échantillon externe. Le fait de ne pas pouvoir utiliser $\hat{\boldsymbol{\alpha}}_1$ pour remplacer $\boldsymbol{\alpha}^*$ dans (3.6) est sans importance; sinon, nous obtenons $\hat{w}_i=\tilde{d}_i$ et (3.8) équivaut à l'équation d'estimation traditionnelle $\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right)=\mathbf{0}$ sans calage. Si $\hat{\boldsymbol{\alpha}}_2$ est bien moins efficace que $\hat{\boldsymbol{\alpha}}_1$ en ce qui a trait au taux de convergence, alors nous ne devrions pas utiliser un tel échantillon externe pour la méthode proposée, car $\hat{\boldsymbol{\alpha}}^*-\boldsymbol{\alpha}^*=\hat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}^*+o_p\left(n^{-1/2}\right)$ et $n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{i1},y_i\right)=o_p(1)$; voir l'annexe C pour obtenir de plus amples renseignements. Selon C5, nous pouvons obtenir les mêmes résultats de convergence dans les lemmes A1-A2 pour (4.2) dans les mêmes conditions. Ainsi, selon le théorème 2, nous obtenons la distribution asymptotique suivante pour $\tilde{\boldsymbol{\eta}}$.

Corollaire 2. *Supposons que les conditions C1-C5 sont vérifiées. Nous avons alors* $n^{1/2}\left(\tilde{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\right)\to\mathcal{N}\left(0,\tilde{\boldsymbol{\Sigma}}_\eta\right)$ *en distribution où* $\tilde{\boldsymbol{\Sigma}}_\eta=\boldsymbol{\mathcal{I}}^{-1}\tilde{\boldsymbol{\Sigma}}_u\left(\boldsymbol{\mathcal{I}}^{-1}\right)^{\mathrm{T}}$; *la forme de* $\boldsymbol{\mathcal{I}}$ *se trouve dans le théorème 1 et la forme de* $\tilde{\boldsymbol{\Sigma}}_\eta$, *dans le théorème 2.*

Remarque 3. *Il convient de souligner que, dans la présente section, lorsque nous dérivons les propriétés asymptotiques, nous ne tenons pas compte des ajustements de pondération, comme l'ajustement pour la non-réponse, l'élagage des poids et le ratissage. Toutefois, ces ajustements de pondération sont*

couramment utilisés dans l'échantillonnage. Il s'agit par conséquent d'un sujet de recherche prometteur pour généraliser la méthode proposée en incorporant ces ajustements de pondération.

5. Intégration de données multiples

Nous examinons maintenant une analyse de régression combinant des renseignements partiels provenant d'échantillons externes. Pour expliquer cette idée, le tableau 5.1 donne un exemple de structure de données comptant trois sources de données (A, B, C) où l'échantillon A contient toutes les observations, tandis que les échantillons B et C contiennent des observations partielles.

Tableau 5.1
Structure de données pour l'intégration de données d'enquête

Échantillon	Poids d'échantillonnage	z	x_1	x_2	y
A	d_a	X	X	X	X
B	d_b	X	X		X
C	d_c	X		X	X

Dans le scénario du tableau 5.1, supposons que nous souhaitions estimer les paramètres du modèle de régression $E(Y \mid x_1, x_2) = m_1(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$, où $m_1(\cdot)$ est connu, mais $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ ne l'est pas. L'équation d'estimation pour $\boldsymbol{\beta}$ au moyen de l'échantillon A peut être formulée ainsi :

$$\hat{\mathbf{U}}_a(\boldsymbol{\beta}) \equiv \sum_{i \in A} d_{a,i} \{y_i - m(x_{i1}, x_{i2}; \boldsymbol{\beta})\} \mathbf{h}(x_{i1}, x_{i2}; \boldsymbol{\beta}) = \mathbf{0} \tag{5.1}$$

pour certains $\mathbf{h}(x_{i1}, x_{i2}; \boldsymbol{\beta})$, de telle sorte que $\hat{\mathbf{U}}_a(\boldsymbol{\beta})$ est linéairement indépendant presque partout.

Nous souhaitons maintenant intégrer l'information partielle de l'échantillon B. Pour ce faire, supposons que nous disposons d'un modèle « de travail » pour $E(Y \mid x_1, z)$:

$$E(Y \mid x_1, z) = m_2(x_1, z; \boldsymbol{\alpha}) \tag{5.2}$$

pour certaines valeurs $\boldsymbol{\alpha}$. Il convient de mentionner que, puisque (z_i, x_{1i}, y_i) sont observés, nous pouvons utiliser l'échantillon B pour estimer $\boldsymbol{\alpha}$ en résolvant $\sum_{i \in B} d_{b,i} \mathbf{U}_b(\boldsymbol{\alpha}; x_{i1}, z_i, y_i) = \mathbf{0}$ pour certaines valeurs de $\mathbf{U}_b$ satisfaisant à $E\{\mathbf{U}_b(\boldsymbol{\alpha}; x_1, z, Y) \mid x_1, z\} = \mathbf{0}$ dans le modèle de travail (5.2).

De même, pour intégrer l'information partielle de l'échantillon C, nous supposons que nous disposons d'un modèle « de travail » pour $E(Y \mid x_2, z)$:

$$E(Y \mid x_2, z) = m_3(x_2, z; \boldsymbol{\gamma}) \tag{5.3}$$

pour certaines valeurs $\boldsymbol{\gamma}$. Nous pouvons également construire une équation d'estimation sans biais $\sum_{i\in C} d_{c,i}\mathbf{U}_c\left(\boldsymbol{\gamma}; x_{i2}, z_i, y_i\right)=\mathbf{0}$ pour certaines valeurs de $\mathbf{U}_c$ satisfaisant à $E\left\{\mathbf{U}_c\left(\boldsymbol{\gamma}; x_2, z, Y\right)\middle|\, x_2, z\right\}=\mathbf{0}$ dans le modèle de travail (5.3). Une fois $\hat{\boldsymbol{\alpha}}$ et $\hat{\boldsymbol{\gamma}}$ obtenus, nous pouvons utiliser cette information supplémentaire pour améliorer l'efficacité de $\hat{\boldsymbol{\beta}}$ dans (5.1). Pour intégrer l'information supplémentaire, nous pouvons la formuler comme maximisant $Q\left(\mathbf{d}_a, \mathbf{w}\right)=\sum_{i\in A} d_{a,i}\log\left(w_i\right)$ en fonction de $\sum_{i\in A} w_i = N$ et de :

$$\sum_{i\in A} w_i\left[\mathbf{U}_b\left(\hat{\boldsymbol{\alpha}}; x_{i1}, z_i, y_i\right), \mathbf{U}_c\left(\hat{\boldsymbol{\gamma}}; x_{i2}, z_i, y_i\right)\right]=\mathbf{0}, \tag{5.4}$$

où $\mathbf{d}_a$ et $\mathbf{w}$ sont des ensembles contenant les poids d'échantillonnage et les poids de calage relativement à l'échantillon A. La contrainte (5.4) intègre l'information supplémentaire. Une fois la solution $\hat{w}_i$ obtenue, nous pouvons utiliser $\sum_{i\in A}\hat{w}_i\left\{y_i - m\left(x_{i1}, x_{i2}; \boldsymbol{\beta}\right)\right\}\mathbf{h}\left(x_{i1}, x_{i2}; \boldsymbol{\beta}\right)=\mathbf{0}$ pour estimer $\boldsymbol{\beta}$. Les résultats asymptotiques peuvent être obtenus de la même façon que dans la section 4.

Remarque 4. *Dans le cadre de la présente étude, nous supposons implicitement que les populations pour l'échantillon interne et les échantillons externes sont les mêmes, mais il est possible que ces populations diffèrent dans certains scénarios. L'estimateur externe* $\hat{\boldsymbol{\alpha}}$ *peut, par exemple, être obtenu en fonction d'un échantillonnage non probabiliste, dont la base de sondage diffère de celle de l'échantillonnage probabiliste du fait du biais de couverture de nombreuses enquêtes à participation volontaire. Diverses méthodologies d'intégration de données existent incorporant l'information des populations hétérogènes. Par exemple, Taylor, Choi et Han (2022) ont proposé d'utiliser des ratios de coefficients pour intégrer l'information externe dans des conditions de régularité, même lorsque les populations des échantillons internes et externes diffèrent. Voir également Zhai et Han (2022) ainsi que Sheng, Sun, Huang et Kim (2022) pour obtenir des approches pénalisées lorsqu'on intègre l'information externe provenant de populations hétérogènes. Les méthodes existantes susmentionnées ne tiennent pas compte des propriétés d'échantillonnage complexe; il est donc prometteur d'étudier, dans un futur projet, l'intégration de données pour des populations hétérogènes échantillonnées.*

6. Étude par simulations

Pour évaluer les performances de l'échantillon fini des estimateurs proposés, nous avons réalisé une étude par simulations en supposant plusieurs scénarios. Nous avons généré une population finie de taille $N=100\,000,$ chaque enregistrement se composant des variables auxiliaires $\mathbf{x}_i=\left(x_{i1}, x_{i2}\right)^{\mathsf{T}}$ de longueur $p=2$ et d'une variable réponse y_i. Nous supposons que $\left(\mathbf{x}_i, y_i\right)$ est disponible pour l'échantillon interne $\mathcal{S}_1,$ alors que seuls $\left(x_{i1}, y_i\right)$ sont disponibles pour l'échantillon externe $\mathcal{S}_2$.

Nous évaluons le rendement de l'estimateur proposé dans un scénario de régression linéaire. Dans ce cas, nous souhaitons faire une inférence statistique pour $\boldsymbol{\beta}=\left(\beta_0, \beta_1, \beta_2\right)^{\mathsf{T}}$ qui permet de résoudre $\sum_{i=1}^{N}\left(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}\right)\left(1, x_{i1}, x_{i2}\right)^{\mathsf{T}}=\mathbf{0}.$

Premièrement, nous considérons deux scénarios pour générer des covariables pour la population finie : i) $x_{i1} \sim N(3,1)$ et $x_{i2} \sim N(11; 6{,}5^2)$, où x_{i1} et x_{i2} sont indépendantes; ii) $x_{i1} \sim N(3,1)$ et $x_{i2} = x_{i1}^2 + \epsilon_i$ avec $\epsilon_i \sim N(0,1)$. Les paramètres de simulation sont choisis de sorte que la moyenne marginale et la variance de x_{i2} soient semblables dans les scénarios de dépendance et d'indépendance. Deuxièmement, la variable réponse est générée sous la forme $Y_i = \mu_i + \varepsilon_i$, où $\mu_i = 1 + 2x_{i1} + x_{i2}$ dans les deux scénarios : i) variance homogène, où $\varepsilon_i \sim N(0,9)$ et ii) variance hétérogène sous la forme $\varepsilon_i \mid \mathbf{x}_i \sim N(0, \sigma_i^2)$, où $\sigma_i = 0.2 \mid \mu_i \mid$. Troisièmement, nous considérons deux plans d'échantillonnage, afin de générer un échantillon probabiliste $\mathcal{S}_1$ de taille (attendue) $n_1 = 1\,000$: i) échantillonnage aléatoire simple sans remise (EAS), et ii) échantillonnage de Poisson avec probabilités d'inclusion satisfaisant à $\pi_{1i} \propto \left(y_i - \min\{y_i : i = 1, \ldots, N\} + 10\right)^{1/2}$ et $\sum_{i=1}^{N} \pi_{1i} = n_1$. Pour finir, nous considérons deux plans d'échantillonnage, afin de générer un échantillon externe $\mathcal{S}_2$ de taille (attendue) $n_2 = 10\,000$: i) l'EAS et ii) l'échantillonnage de Poisson avec probabilités d'inclusion satisfaisant à $\pi_{2i} \propto \{1 + \exp(0{,}2x_{i1} + 0{,}1x_{i2} - 0{,}6)\}^{-1}$ et $\sum_{i=1}^{N} \pi_{2i} = n_2$. Il convient de mentionner que le plan d'échantillonnage pour l'échantillon interne est informatif (Pfeffermann, 1993) dans le cadre de l'échantillonnage de Poisson. Ne pas tenir compte des caractéristiques du plan peut donc donner lieu à une inférence erronée.

Pour l'estimateur proposé, nous envisageons un modèle de travail réduit, $\sum_{i \in \mathcal{S}_2} \pi_{i2}^{-1} (y_i - \alpha_0 - \alpha_1 x_{i1}) (1, x_{i1})^{\mathsf{T}} = \mathbf{0}$, dont la solution est désignée par $\hat{\boldsymbol{\alpha}}_2$. En fonction de l'échantillon externe $\mathcal{S}_2$, nous supposons qu'un estimateur ponctuel $\hat{\boldsymbol{\alpha}}_2$ et son estimateur de variance $\mathbf{V}_2 = \hat{V}(\hat{\boldsymbol{\alpha}}_2)$ sont disponibles conformément à la discussion de la section 3. Une linéarisation est adoptée pour obtenir un estimateur de variance $\mathbf{V}_2$; voir la démonstration du théorème 1 en annexe A pour obtenir de plus amples précisions.

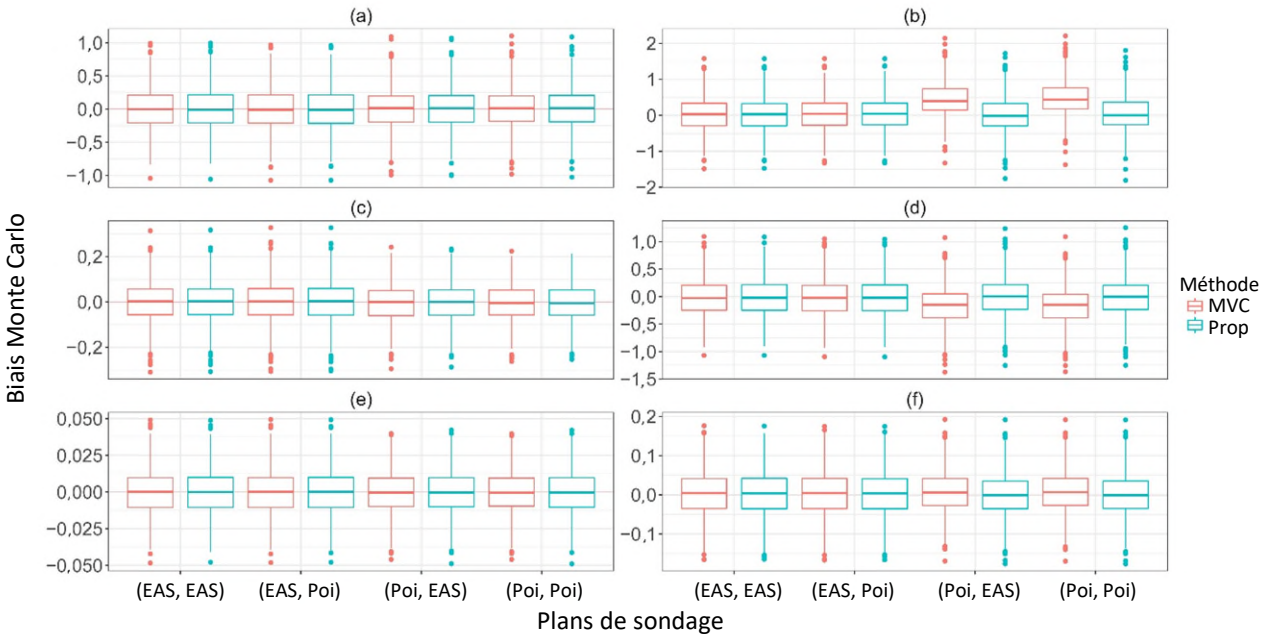
Dans l'étude par simulations, l'estimateur proposé est comparé à l'estimateur par le maximum de vraisemblance contraint (MVC) (Chatterjee et coll., 2016). Nous supposons une distribution normale pour la fonction de vraisemblance, c'est-à-dire $y_i \mid \mathbf{x}_i \sim N\{(1, \mathbf{x}_i^{\mathsf{T}})\boldsymbol{\beta}, \sigma_{\text{ent}}^2\}$. Nous supposons également qu'un analyste suppose que $y_i \mid \mathbf{x}_{i1} \sim N\{(1, x_{i1})\boldsymbol{\alpha}, \sigma_{\text{red}}^2\}$ pour le modèle de travail réduit. Voir la section S2 de la documentation supplémentaire pour connaître les détails du calcul. Nous considérons l'estimateur par le MVC dans le scénario pour lequel l'information supplémentaire de (y_i, x_{i1}) est disponible pour un échantillon externe, et non pour la population entière.

Nous menons $M = 1\,000$ simulations de Monte Carlo; les figures 6.1 et 6.2 présentent le biais de la méthode de Monte Carlo de l'estimateur proposé et de l'estimateur par le MVC respectivement pour les scénarios de variance homogène et hétérogène. Selon la figure 6.1, où la variance du terme d'erreurs est homogène et un EAS génère l'échantillon interne, l'estimateur proposé fournit un rendement relativement identique à l'estimateur par le MVC en ce qui a trait au biais Monte Carlo et à la variance. Toutefois, lorsque les auxiliaires sont corrélés et l'échantillon interne est généré par l'échantillonnage de Poisson, l'estimateur par le MVC est discutable, puisque son modèle est mal spécifié dans le plan d'échantillonnage informatif de Poisson. Par exemple, le biais Monte Carlo de l'estimateur par le MVC n'est pas négligeable lors de l'estimation de β_0 et β_1. L'estimateur proposé intégrant les caractéristiques du plan, son rendement est satisfaisant pour tous les scénarios. Comme l'illustre la figure 6.2, même lorsqu'un EAS génère

l'échantillon interne, l'estimateur par le MVC est légèrement moins efficace que l'estimateur proposé. En effet, l'estimateur par le MVC ne tient pas compte de la variance hétérogène, alors que l'estimateur proposé n'avance aucune hypothèse de distribution. Lorsqu'un plan d'échantillonnage informatif de Poisson génère l'échantillon interne, l'estimateur par le MVC ne fournit pas un rendement efficace, puisqu'il est biaisé et que sa variance est supérieure à celle de l'estimateur proposé.

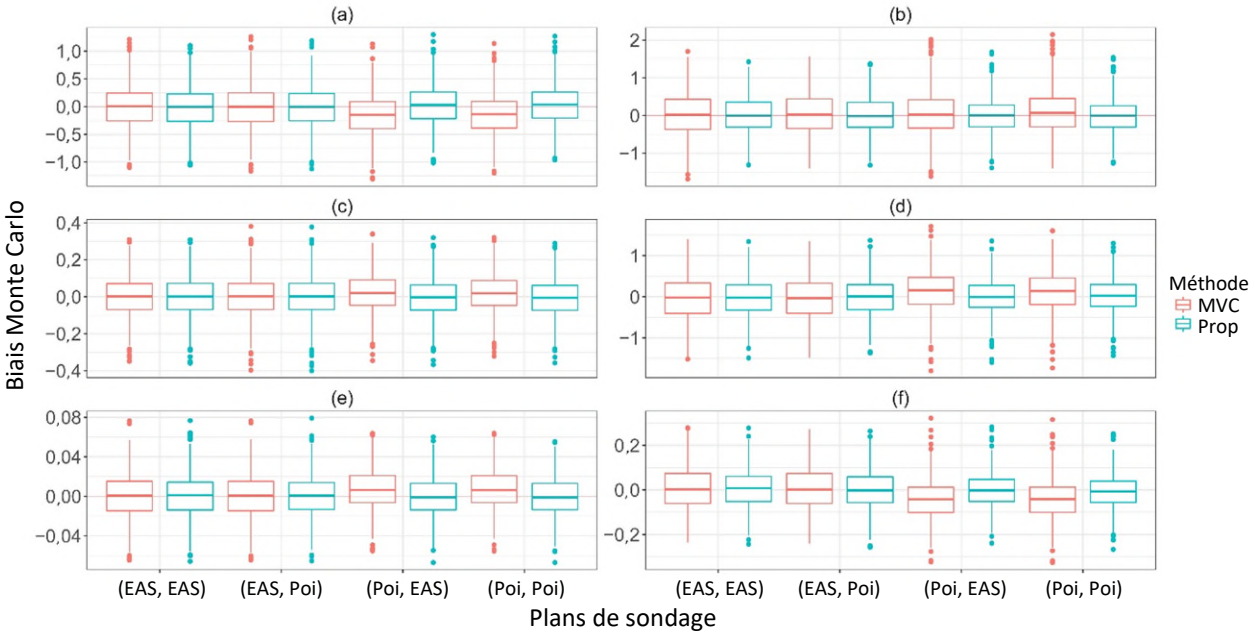
Le tableau 6.1 présente le taux de couverture d'un intervalle de confiance à 95 % pour l'estimateur proposé dans le cadre de différents scénarios. Chatterjee et coll. (2016) ont uniquement étudié les propriétés théoriques de leur estimateur lorsque l'information au niveau de la population était disponible. Par conséquent, aucun estimateur d'intervalle ne peut être fourni si seul un échantillon externe est disponible. Dans le tableau 6.1, nous concluons que les taux de couverture des intervalles de confiance sont tous proches de la définition nominale de 0,95 dans différents scénarios. Une raison possible de ce phénomène est que l'estimateur proposé est sans modèle; le modèle proposé est donc plus robuste et peut être utilisé dans le cadre de plans d'échantillonnage complexe.

Figure 6.1 Biais Monte Carlo de l'estimateur proposé et de l'estimateur par le maximum de vraisemblance contraint (MVC) selon 1 000 simulations de Monte Carlo dans le scénario de variance homogène.



Les première, deuxième et troisième lignes représentent respectivement le biais Monte Carlo pour l'estimation de β_0, β_1 et β_2. Les trois diagrammes, (a), (c) et (e), de la colonne de gauche, présentent les résultats lorsque les variables auxiliaires sont générées indépendamment et les diagrammes (b), (d) et (f) de la colonne de droite représentent le cas lorsque les auxiliaires sont dépendants. « MVC » et « Prop » désignent respectivement l'estimateur par le MVC et l'estimateur proposé. Le premier plan de sondage entre parenthèses est utilisé pour générer l'échantillon interne $\mathcal{S}_1$ et le deuxième, pour générer l'échantillon externe $\mathcal{S}_2$. « EAS » et « Poi » désignent respectivement l'échantillonnage aléatoire simple et l'échantillonnage de Poisson.

Figure 6.2 Biais Monte Carlo de l'estimateur proposé et de l'estimateur par le maximum de vraisemblance contraint (MVC) selon 1 000 simulations de Monte Carlo dans le scénario de variance hétérogène.



Les première, deuxième et troisième lignes représentent respectivement le biais Monte Carlo pour l'estimation de β_0, β_1 et β_2. Les trois diagrammes, (a), (c) et (e), de la colonne de gauche, présentent les résultats lorsque les variables auxiliaires sont générées indépendamment et les diagrammes (b), (d) et (f) de la colonne de droite représentent le cas lorsque les auxiliaires sont dépendants. « MVC » et « Prop » désignent respectivement l'estimateur par le MVC et l'estimateur proposé. Le premier plan de sondage entre parenthèses est utilisé pour générer l'échantillon interne $\mathcal{S}_1$ et le deuxième, pour générer l'échantillon externe $\mathcal{S}_2$. « EAS » et « Poi » désignent respectivement l'échantillonnage aléatoire simple et l'échantillonnage de Poisson.

Tableau 6.1
Taux de couverture d'un intervalle de confiance à 95 % par la méthode proposée selon 1 000 simulations de Monte Carlo dans des scénarios différents

	Plan $\mathcal{S}_1$	Plan $\mathcal{S}_2$	Indépendantes			Dépendantes		
			β_0	β_1	β_2	β_0	β_1	β_2
Homo	EAS	EAS	0,948	0,952	0,939	0,945	0,948	0,934
		Poi	0,945	0,951	0,938	0,946	0,946	0,934
	Poi	EAS	0,957	0,966	0,949	0,935	0,943	0,940
		Poi	0,962	0,964	0,951	0,936	0,943	0,938
Hété	EAS	EAS	0,944	0,942	0,933	0,933	0,925	0,935
		Poi	0,949	0,942	0,935	0,935	0,934	0,931
	Poi	EAS	0,959	0,955	0,935	0,948	0,950	0,941
		Poi	0,961	0,956	0,944	0,952	0,949	0,946

Note : « Homo » et « Hété » désignent respectivement les variances homogène et hétérogène du terme d'erreurs. « Plan $\mathcal{S}_1$ » et « Plan $\mathcal{S}_2$ » désignent le plan d'échantillonnage utilisé pour générer l'échantillon interne $\mathcal{S}_1$ et l'échantillon externe $\mathcal{S}_2$. « EAS » et « Poi » désignent respectivement l'échantillonnage aléatoire simple et l'échantillonnage de Poisson. « Indépendantes » et « Dépendantes » correspondent respectivement aux cas où les variables auxiliaires sont indépendantes et dépendantes.

Une simulation supplémentaire avec scénario de régression logistique est présentée à la section S3 de la documentation supplémentaire; des conclusions similaires peuvent en être tirées.

7. Étude d'application

7.1 Description des données et formulation du problème

Pour illustrer la méthode proposée, nous l'appliquons à l'analyse d'un sous-ensemble de données de l'enquête KNHANES. Cette enquête annuelle, portant sur environ 5 000 personnes chaque année, permet de recueillir des renseignements concernant les comportements liés à la santé au moyen d'interviews, l'état de santé de base au moyen d'examens physiques et d'analyses sanguines et l'apport alimentaire au moyen d'une enquête sur la nutrition. Le plan de sondage de l'enquête KNHANES constitue un échantillonnage stratifié dans lequel l'âge, le sexe et la région sont utilisés comme variables de stratification. Les poids d'échantillonnage finaux sont calculés au moyen d'un ajustement pour la non-réponse et d'une post-stratification, puis sont fournis aux utilisateurs de données avec les variables d'enquête.

Pour améliorer l'efficacité de l'analyse des données au moyen d'une enquête KNHANES de taille $n_1 = 4\,929,$ nous avons utilisé une base de données publique externe fournie par le NHISS (service national coréen de partage de l'assurance maladie). Les mégadonnées fournies par le NHISS contiennent des renseignements relatifs à la santé d'environ $n_2 = 1\,000\,000$ de personnes, dont certaines variables sont un sous-ensemble de variables de l'enquête KNHANES.

Ces structures de données, où figurent le petit $n_1,$ le grand n_2 et les mégadonnées comportant un sous-ensemble de variables dans l'échantillon interne, conviennent bien au scénario présenté à la section 2. Cependant, au moment d'appliquer la méthode proposée à la situation réelle, nous nous heurtons à une autre complication. Dans les données du NHISS, les probabilités de sélection sont inconnues, de sorte que l'estimateur convergent par rapport au plan $\hat{\boldsymbol{\alpha}}_2$ de (3.9) n'est pas disponible. Nous abordons cette question à la section 7.2 en utilisant une méthode de pondération de la propension à répondre et nous présentons les résultats de l'analyse de l'étude d'application à la section 7.3.

7.2 Pondération de la propension à répondre pour des données externes avec probabilité de sélection inconnue

Nous envisageons maintenant d'étendre la méthode proposée à une situation où l'échantillon externe $\mathcal{S}_2$ se compose de mégadonnées dont les probabilités de sélection sont inconnues. Dans ce cas, le modèle de travail pour $E\left(Y_i \mid \mathbf{x}_{i1}\right) = m\left(\boldsymbol{\alpha}^{\top}\mathbf{x}_{i1}\right)$ peut ne pas se vérifier pour l'échantillon $S_2.$ Néanmoins, nous pouvons tout de même résoudre :

$$\sum_{i \in \mathcal{S}_2} \left\{ y_i - m\left(\boldsymbol{\alpha}^{\top}\mathbf{x}_{i1}\right) \right\} \mathbf{x}_{i1} = \mathbf{0} \tag{7.1}$$

pour obtenir $\hat{\alpha}_0$ et $\hat{\alpha}_1.$ Si le mécanisme d'échantillonnage pour $\mathcal{S}_2$ est ignorable ou non informatif, la solution de (7.1) est alors sans biais; sinon, l'estimateur obtenu est biaisé.

Pour éliminer les biais de sélection dans l'estimation des mégadonnées, Kim et Wang (2019) ont proposé d'utiliser des poids de score de propension dans (7.1) pour obtenir un estimateur sans biais de $\boldsymbol{\alpha}.$ Pour construire les poids des scores de propension, nous utilisons un modèle de non-réponse non ignorable, $P\left(\delta_i = 1 \mid \mathbf{x}_{i1}, y_i\right) = \pi\left(\mathbf{x}_{i1}, y_i; \boldsymbol{\phi}\right),$ où $\delta_i = 1$ si $i \in S_2$ et nul dans les autres cas. Il convient de mentionner

que nous pouvons exprimer $\pi\left(\mathbf{x}_{i1}, y_i\right)^{-1} = 1 + \left(N_0 / N_1\right) r\left(\mathbf{x}_{i1}, y_i\right)$ où $r\left(\mathbf{x}_{i1}, y_i\right) = f\left(\mathbf{x}_{i1}, y_i \mid \delta_i = 0\right) / f\left(\mathbf{x}_{i1}, y_i \mid \delta_i = 1\right)$ est la fonction du ratio de densité avec $N_1 = \sum_{i=1}^{N} \delta_i$ et $N_0 = N - N_1$. Au moyen de la justification de Wang et Kim (2021), nous pouvons supposer un modèle de ratio de densité loglinéaire, $\log\left\{r\left(x_{i1}, y_i; \boldsymbol{\phi}\right)\right\} = \phi_0 + \phi_1 x_{i1} + \phi_2 y_i$. L'estimateur d'entropie maximale de $\boldsymbol{\phi}$ est obtenu en résolvant $\left(1/N_1\right) \sum_{i=1}^{N} \delta_i \exp\left(\phi_0 + \phi_1 x_{i1} + \phi_2 y_i\right) \left(1, x_{i1}, y_i\right) = \left(1, \hat{\bar{x}}_1, \hat{\bar{y}}\right)$ où $\left(\hat{\bar{x}}_1, \hat{\bar{y}}\right) = \left(1/\hat{N}_0\right)\left\{\sum_{i \in \mathcal{S}_1} d_i \left(x_{i1}, y_i\right) - \sum_{i=1}^{N} \delta_i \left(x_{i1}, y_i\right)\right\}$, $\hat{N}_0 = \sum_{i \in \mathcal{S}_1} d_i - N_1$, et $\mathcal{S}_1$ est l'échantillon interne. Une fois $\hat{\boldsymbol{\phi}}$ obtenu, nous pouvons construire $\hat{\pi}\left(x_{i1}, y_i\right)$ et résoudre :

$$\sum_{i \in \mathcal{S}_2} \frac{1}{\hat{\pi}\left(x_{i1}, y_i\right)} \left\{y_i - m\left(\alpha_0 + \alpha_1 x_{i1}\right)\right\} \left(1, x_{i1}\right) = \left(0, 0\right) \tag{7.2}$$

pour obtenir $\hat{\boldsymbol{\alpha}}_2 = \left(\hat{\alpha}_0, \hat{\alpha}_1\right)$.

De plus, nous pouvons utiliser l'échantillon interne $\mathcal{S}_1$ pour ajuster le même modèle de travail afin d'obtenir $\hat{\boldsymbol{\alpha}}_1$. Ensuite, nous obtenons $\hat{\boldsymbol{\alpha}}^*$ en utilisant (3.9) et nous appliquons la méthode de pondération par calage proposée pour combiner les renseignements des mégadonnées. En pratique, il est difficile de calculer $\mathbf{V}_2$ dans (3.9), mais sa taille est négligeable si la taille de l'échantillon de $\mathcal{S}_2$ est très grande. Dans ce cas, nous pouvons simplement utiliser $\hat{\boldsymbol{\alpha}}^* = \hat{\boldsymbol{\alpha}}_2$ dans le problème de calage.

7.3 Résultats de l'étude d'application : enquête KNHANES

Dans la présente étude d'application, nous utilisons $n_1 = 4\,929$ enregistrements des données de l'enquête KNHANES qui n'ont pas de valeurs manquantes pour quatre variables : le cholestérol total, l'hémoglobine, le triglycéride et le cholestérol de lipoprotéine de haute densité (LHD). Aux fins de démonstration, nous supposons qu'un analyste souhaite réaliser l'analyse de régression linéaire suivante :

$$E\left(\text{Cholestérol Total}_i \mid \mathbf{x}_i\right) = \beta_0 + \beta_1 \text{Hémoglobine}_i + \beta_2 \text{Triglycéride}_i + \beta_3 \text{LHD}_i \quad \text{pour} \quad i \in \mathcal{S}_1;$$

voir la section S4 de la documentation supplémentaire pour obtenir des précisions sur l'hypothèse linéaire. Dans nos données, la valeur absolue la plus élevée de la corrélation par paires entre covariables est de $-0{,}40$ et est observée entre le triglycéride et le cholestérol de LHD; ce qui ressemble au scénario de la section 6 où les covariables étaient fortement corrélées. Les mégadonnées externes consistent en $n_2 = 1\,000\,000$ enregistrements de données du NHISS, ayant des valeurs entièrement observées pour le cholestérol total, l'hémoglobine et le triglycéride. Le modèle de travail réduit adopté est :

$$E\left(\text{Cholestérol Total}_i \mid \mathbf{x}_{i1}\right) = \alpha_0 + \alpha_1 \text{Hémoglobine}_i + \alpha_2 \text{Triglycéride}_i \quad \text{pour} \quad i \in \mathcal{S}_1 \cup \mathcal{S}_2.$$

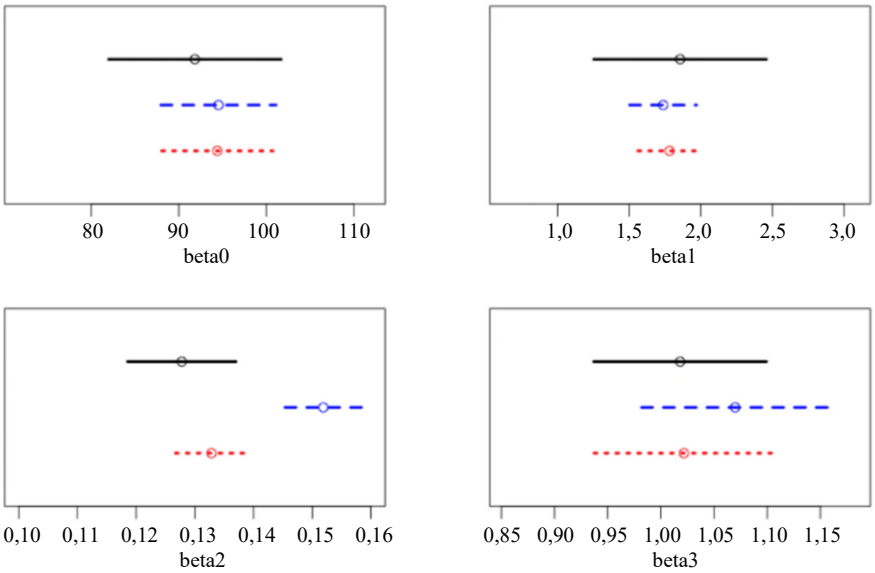
Dans la présente étude d'application, nous mettons en œuvre les méthodes proposées dans l'échantillon externe, où $\hat{\boldsymbol{\alpha}}_2$ est utilisé au lieu de $\boldsymbol{\alpha}^*$ qui est n'est pas disponible, puisque nous ne disposons pas de renseignements sur l'ensemble de la population. Les probabilités de sélection de l'échantillon externe étant inconnues, nous préparons deux versions de méthodes proposées : i) considérant $\mathcal{S}_2$ comme EAS, c'est-à-dire sans pondération de la propension à répondre; et ii) par l'ajustement de la pondération de propension à répondre présenté à la section 7.2. Pour la pondération de propension à répondre, nous ajustons le modèle de ratio de densité loglinéaire aux données externes, $\log\left\{r\left(\mathbf{x}_{i1}, y_i; \boldsymbol{\phi}\right)\right\} = \phi_0 + \phi_1 \text{Hémoglobine}_i + \phi_2 \text{Triglycéride}_i + \phi_3 \text{Cholestérol Total}_i$, nous calculons $\hat{\pi}\left(\mathbf{x}_{i1}, y_i\right)$ en fonction de $\hat{\boldsymbol{\phi}}$, puis nous résolvons

$$\sum_{i\in\mathcal{S}_2}\frac{1}{\hat{\pi}\left(x_{i1},y_i\right)}\left\{y_i-m\left(\alpha_0+\alpha_1x_{i1}\right)\right\}\left(1,x_{i1}\right)=\left(0,0\right)$$

pour obtenir $\hat{\boldsymbol{\alpha}}_2$. Le modèle de régression logistique ci-dessus est une hypothèse courante dans la documentation; voir Elliott et Valliant (2017), Chen, Li et Wu (2020), Wang et Kim (2021) ainsi que les références qui y figurent pour obtenir de plus amples renseignements. Puisque l'estimateur MVC n'intègre pas les caractéristiques du plan, il n'en est pas tenu compte dans la section d'application. Le rendement des méthodes proposées est comparé à la méthode de référence, qui s'appuie sur l'échantillon interne $\mathcal{S}_1$ uniquement pour obtenir des estimations des moindres carrés pondérées en tenant compte des poids d'échantillonnage.

La figure 7.1 présente les estimations ponctuelles et les intervalles de confiance à 95 % de $\hat{\boldsymbol{\beta}}=\left(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right)$ pour chaque méthode. Les méthodes proposées donnent lieu à des variances inférieures pour $\hat{\beta}_0$, $\hat{\beta}_1$ et $\hat{\beta}_2$ par rapport à l'utilisation de l'échantillon interne seulement. Ce résultat coïncide avec les résultats de nos études par simulations de la section précédente. Pour β_2, l'estimateur de la méthode proposée sans pondération de propension à répondre présente une différence systématique par rapport aux deux autres estimateurs. Quand l'ajustement de pondération de la propension à répondre est conjugué à la méthode proposée, son intervalle de confiance de β_2 est compris dans celui obtenu au moyen de l'échantillon interne seulement. Ce résultat signifie que l'ajustement de la pondération de propension à répondre compense le biais systématique attribuable au fait que les probabilités d'échantillonnage ne sont pas prises en compte. Aucun gain d'efficacité de l'estimation de β_3 n'était attendu, car les données externes contiennent des renseignements sur x_{i1} (hémoglobine) et x_{i2} (triglycéride), et non sur x_{i3} (LHD).

Figure 7.1 Comparaison de l'analyse de régression pour $E\left(\text{Cholestérol total}_i \mid \mathbf{x}_i\right)=\beta_0+\beta_1\,\text{Hémoglobine}_i+\beta_2\,\text{Triglycéride}_i+\beta_3\,\text{LHD}_i$ au moyen des données internes de l'enquête Korea National Health and Nutrition Examination (KNHANES) appuyées par les mégadonnées externes de la base de données du National Health Insurance Sharing Service (NHISS).



Pour chaque section, les cercles représentent des estimations ponctuelles et les lignes sont leurs intervalles de confiance à 95 % pour l'utilisation de l'échantillon interne $\mathcal{S}_1$ seulement, avec les moindres carrés pondérés (ligne pleine supérieure), la méthode proposée sans ajustement (ligne en tirets du milieu) et la méthode proposée avec ajustement de la pondération du score de propension (ligne en pointillé du bas).

8. Conclusion

L'intégration de sources de données externes à l'analyse de régression d'un échantillon interne pose un problème important dans la pratique. Nous avons étudié ce problème en ayant recours à une application nouvelle de la projection d'information (Csiszár et Shields, 2004) et de la pondération au moyen d'un calage assisté par un modèle (Wu et Sitter, 2001). La méthode proposée s'applique directement à l'échantillonnage d'enquête et peut facilement être étendue à l'intégration de données multiples. Elle est facile à mettre en œuvre et ne nécessite pas d'accès direct à des données externes. Tant que nous disposons des coefficients de régression estimés et de leurs erreurs-types pour le modèle de travail réduit, nous pouvons intégrer les renseignements supplémentaires à notre analyse.

Plusieurs pistes sont envisageables pour étendre notre recherche. Tout d'abord, une approche bayésienne peut être élaborée pour le même scénario. Il est possible d'utiliser la méthodologie de vraisemblance empirique bayésienne de Zhao, Ghosh, Rao et Wu (2020) dans ce scénario. La méthode proposée peut aussi servir à combiner les données d'essais cliniques randomisés avec des mégadonnées réelles (Yang, Zheng et Wang, 2020); ces extensions de la méthode seront présentées ailleurs. Il sera également intéressant de faire le lien entre l'approche proposée et le plan d'échantillonnage à deux degrés (double), dont le plan et l'estimation efficaces ont été activement étudiés récemment (Rivera-Rodriguez, Spiegelman et Haneuse, 2019; Rivera-Rodriguez, Haneuse, Wang et Spiegelman, 2020; Wang, Williams, Chen et Chen, 2020). La structure des données de l'échantillonnage à deux degrés, dont l'échantillon au premier degré comporte un grand n et un petit p et l'échantillon au second degré, un petit n et un grand p, convient bien au scénario supposé par l'approche proposée de calage assisté par modèle.

Remerciements

Nous remercions les examinateurs et le rédacteur associé pour leurs commentaires constructifs. Les recherches de Z. Wang ont été partiellement financées par des subventions de la fondation nationale de sciences naturelles de la Chine (numéro : 11901487, 72033002) et une subvention du centre scientifique fondamental de la fondation nationale de sciences naturelles de la Chine (numéro : 71988101). Les recherches de J.K. Kim ont été partiellement financées par la subvention de la Fondation nationale des sciences (numéro : CSSI-1931380), une entente de coopération entre le service de conservation des ressources naturelles du ministère de l'Agriculture des États-Unis et l'université d'État de l'Iowa, ainsi que la station expérimentale d'agriculture et d'économie domestique de l'Iowa (Iowa Agriculture and Home Economics Experiment Station) d'Ames (Iowa).

Documentation supplémentaire

La documentation supplémentaire figure dans le document https://arxiv.org/abs/2107.06448; elle comprend un cas spécial dans le cadre d'un échantillonnage aléatoire simple (S1), une mise en œuvre de Chatterjee et coll. (2016) (S2), une étude par simulations supplémentaire (S3) et la validation de l'hypothèse de linéarité pour l'ensemble de données de l'enquête KNHANES (S4).

Annexe

A. Démonstration du théorème 1

Lemme A1. *Supposons que les conditions C1, C2e et C3 se vérifient. Alors,* $\left\| \hat{\boldsymbol{\lambda}} \right\| = O_p\left(n^{-1/2}\right)$.

Démonstration du lemme A1. Désignons $\hat{\boldsymbol{\lambda}} = \rho\boldsymbol{\theta}$, *où* $\rho = \left\| \hat{\boldsymbol{\lambda}} \right\|$ *et* $\boldsymbol{\theta} = \rho^{-1}\hat{\boldsymbol{\lambda}}$ *est un vecteur de longueur d'unité. Nous obtenons alors :*

$$
\begin{aligned}
\mathbf{0} &= \left| \sum_{i\in\mathcal{S}_1} \frac{\tilde{d}_i}{1-\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right) \right| \\
&= \left| \boldsymbol{\theta}^{\mathrm{T}} \sum_{i\in\mathcal{S}_1} \frac{\tilde{d}_i}{1-\rho\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right) \right| \\
&= \left| \sum_{i\in\mathcal{S}_1} \tilde{d}_i\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right) + \rho\sum_{i\in\mathcal{S}_1} \frac{\tilde{d}_i\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\left\{\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right\}^{\mathrm{T}}\boldsymbol{\theta}}{1-\rho\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)} \right| \\
&\geq \frac{\rho}{1+\rho Z_S} \left| \sum_{i\in\mathcal{S}_1} \tilde{d}_i\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\left\{\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right\}^{\mathrm{T}}\boldsymbol{\theta} \right| - \left| \sum_{i\in\mathcal{S}_1} \tilde{d}_i\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right) \right|, \qquad \text{(A.1)}
\end{aligned}
$$

où la première égalité se vérifie, puisque $g\left(\hat{\boldsymbol{\eta}}\right) = \mathbf{0}$, et la dernière inégalité se vérifie par l'inégalité triangulaire.

Selon C2e et le théorème de Rayleigh-Ritz (Horn et Johnson, 2012, section 4.2), une constante $\sigma_0 > 0$ existe, de telle sorte que :

$$
\sum_{i\in\mathcal{S}_1} \tilde{d}_i\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\left\{\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right\}^{\mathrm{T}}\boldsymbol{\theta} > \sigma_0 + o_p(1). \qquad \text{(A.2)}
$$

Selon C3 et le théorème de Slutsky, nous avons :

$$
\sum_{i\in\mathcal{S}_1} \tilde{d}_i\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right) = O_p\left(n^{-1/2}\right). \qquad \text{(A.3)}
$$

Ainsi, selon C1 et (A.1)-(A.3), nous démontrons le lemme A1.

Lemme A2. *Supposons que les conditions C1, C2a-C2e et C3 se vérifient. Alors,* $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{0N} = o_p(1)$.

Démonstration du lemme A2. Selon les lemmes A1 et C1, nous concluons que :

$$
\begin{aligned}
\max_{i\in\mathcal{S}_1} \left| \hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right) \right| &\leq \max_{i\in\mathcal{S}_1} \left\| \hat{\boldsymbol{\lambda}} \right\| \left\| \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right) \right\| \\
&= \left\| \hat{\boldsymbol{\lambda}} \right\| \max_{i\in\mathcal{S}_1} \left\| \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right) \right\| = o_p(1). \qquad \text{(A.4)}
\end{aligned}
$$

Tout d'abord, nous savons que :

$$
\sum_{i\in\mathcal{S}_1} \frac{\tilde{d}_i}{1-\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)} \mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right) - \sum_{i\in\mathcal{S}_1} \tilde{d}_i\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right) \to \mathbf{0} \qquad \text{(A.5)}
$$

uniformément en probabilité pour $\boldsymbol{\beta} \in \Omega$. Selon (A.4), nous avons :

$$\begin{aligned}
&\left\| \sum_{i \in \mathcal{S}_1} \tilde{d}_i \left\{ \frac{1}{1 - \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right)} - 1 \right\} \mathbf{U}_1\left(\boldsymbol{\beta}; \mathbf{x}_i, y_i\right) \right\| \\
&\quad = \left\| \sum_{i \in \mathcal{S}_1} \tilde{d}_i \left\{ \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) + o_p\left(\hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right)\right) \right\} \mathbf{U}_1\left(\boldsymbol{\beta}; \mathbf{x}_i, y_i\right) \right\| \\
&\quad \leq \left(1 + o_p(1)\right) \max_{i \in \mathcal{S}_1} \left| \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) \right| \left\| \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1\left(\boldsymbol{\beta}; \mathbf{x}_i, y_i\right) \right\|. \qquad \text{(A.6)}
\end{aligned}$$

Selon C2a-C2b, une constante $C_{u1} > 0$ existe, de telle sorte que $\sup_{\boldsymbol{\beta} \in \Omega} \left\| \mathbf{U}_0\left(\boldsymbol{\beta}\right) \right\| < C_{u1}$. Puisque $\sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1\left(\boldsymbol{\beta}; \mathbf{x}_i, y_i\right)$ converge de façon uniforme vers $\mathbf{U}_0\left(\boldsymbol{\beta}\right)$ en probabilité, nous concluons que :

$$\left\| \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1\left(\boldsymbol{\beta}; \mathbf{x}_i, y_i\right) \right\| < C_{u1} + o_p(1) \qquad \text{(A.7)}$$

uniformément sur Ω. Selon (A.4) et (A.6)-(A.7), nous avons validé (A.5).

Selon C2b et (A.5), nous concluons que $\sum_{i \in \mathcal{S}_1} \tilde{d}_i \left\{1 - \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right)\right\}^{-1} \mathbf{U}_1\left(\boldsymbol{\beta}; \mathbf{x}_i, y_i\right)$ converge de façon uniforme vers $U_0\left(\boldsymbol{\beta}\right)$ en probabilité. Désignons $Q_0\left(\boldsymbol{\beta}\right) = -U_0\left(\boldsymbol{\beta}\right)^2$ et $Q_s\left(\boldsymbol{\beta}\right) = -\left[\sum_{i \in \mathcal{S}_1} \tilde{d}_i \{1 - \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right)\}^{-1} \mathbf{U}_1\left(\boldsymbol{\beta}; \mathbf{x}_i, y_i\right)\right]^2$. Alors, $\boldsymbol{\beta}_p$ maximise $Q_0\left(\boldsymbol{\beta}\right)$ de façon unique selon (C2b) et $\hat{\boldsymbol{\beta}}$ maximise $Q_s\left(\boldsymbol{\beta}\right)$. De plus, $Q_s\left(\boldsymbol{\beta}\right)$ converge de façon uniforme vers $Q_0\left(\boldsymbol{\beta}\right)$ en probabilité sur l'ensemble compact Ω. Ainsi, selon C2a et le théorème 2.1 d'Engle et McFadden (1994, chapitre 36), nous démontrons le lemme A2.

Démonstration du théorème 1. Selon les lemmes A1-A2, nous avons montré que :

$$\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}_0 + o_p(1), \qquad \text{(A.8)}$$

où $\hat{\boldsymbol{\eta}}^{\mathrm{T}} = \left(\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\boldsymbol{\lambda}}^{\mathrm{T}}\right)$, $\boldsymbol{\eta}_0^{\mathrm{T}} = \left(\boldsymbol{\beta}_{0N}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}\right)$ et $\mathbf{0}$ est un vecteur nul de même longueur de $\hat{\boldsymbol{\lambda}}$.

Selon (A.8) et le développement en série de Taylor, nous avons :

$$\begin{aligned}
\mathbf{0} \;=\; & g\left(\hat{\boldsymbol{\eta}}\right) = g\left(\boldsymbol{\eta}_0\right) + \frac{\partial g}{\partial \boldsymbol{\eta}^{\mathrm{T}}}\left(\boldsymbol{\eta}_0\right)\left(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\right) + o_p\left(\left\| \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 \right\|\right) \\
= & \begin{pmatrix} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1\left(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i\right) \\ \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) \end{pmatrix} \\
& + \begin{pmatrix} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \dfrac{\partial \mathbf{U}_1\left(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i\right)}{\partial \boldsymbol{\beta}^{\mathrm{T}}} & \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1\left(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i\right)\left\{\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right)\right\}^{\mathrm{T}} \\ \mathbf{0} & \sum_{i \in \mathcal{S}_1} \tilde{d}_i \left\{\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right)\right\}^{\otimes 2} \end{pmatrix} \left(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\right) \\
& + o_p\left(\left\| \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 \right\|\right). \qquad \text{(A.9)}
\end{aligned}$$

Selon (C3), nous avons :

$$n^{1/2}\begin{pmatrix}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i\right)\\ \sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\end{pmatrix}\to\mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}_u\right) \tag{A.10}$$

en distribution. Selon (C2e)-(C2c), nous concluons que :

$$\begin{pmatrix}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\dfrac{\partial\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}} & \sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i\right)\left\{\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right\}^{\mathrm{T}}\\ \mathbf{0} & \sum_{i\in\mathcal{S}_1}\tilde{d}_i\left\{\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right\}^{\otimes 2}\end{pmatrix}\to\boldsymbol{\mathcal{I}} \tag{A.11}$$

en probabilité, où :

$$\boldsymbol{\mathcal{I}}=\begin{pmatrix}\boldsymbol{\mathcal{I}}_{11} & \boldsymbol{\mathcal{I}}_{12}\\ \mathbf{0} & \boldsymbol{\mathcal{I}}_{22}\end{pmatrix}.$$

Selon (A.9)-(A.11), nous concluons que :

$$n^{1/2}\left(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\right)\to\mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}_\eta\right) \tag{A.12}$$

en distribution, où $\boldsymbol{\Sigma}_\eta=\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Sigma}_u\left(\boldsymbol{\mathcal{I}}^{-1}\right)^{\mathrm{T}}$.

B. Démonstration du corollaire 1

Puisque $\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x},y\right)$ est fourni, il est suffisant de considérer :

$$\boldsymbol{\Sigma}_{11}-\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\Sigma}_{21}-\boldsymbol{\Sigma}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{12}^{\mathrm{T}}+\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{12}^{\mathrm{T}},$$

la variance asymptotique de $\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N})-\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\tilde{\mathbf{U}}_2\left(\boldsymbol{\alpha}_N^*\right)$, où $\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N})=n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i\right)$ et $\tilde{\mathbf{U}}_2\left(\boldsymbol{\alpha}_N^*\right)=n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{i1},y_i\right)$.

Considérons ce qui suit :

$$\begin{aligned}&\mathrm{Var}\left\{\tilde{\mathbf{U}}_1\left(\boldsymbol{\beta}_{0N}\right)-\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\tilde{\mathbf{U}}_2\left(\boldsymbol{\alpha}_N^*\right)\right\}\\ &\quad=E\left[\mathrm{Var}\left\{\tilde{\mathbf{U}}_1\left(\boldsymbol{\beta}_{0N}\right)-\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\tilde{\mathbf{U}}_2\left(\boldsymbol{\alpha}_N^*\right)\middle|\,\mathcal{A}_N\right\}\right]+\mathrm{Var}\left[E\left\{\tilde{\mathbf{U}}_1\left(\boldsymbol{\beta}_{0N}\right)-\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\tilde{\mathbf{U}}_2\left(\boldsymbol{\alpha}_N^*\right)\middle|\,\mathcal{A}_N\right\}\right]\\ &\quad\succeq E\left[\mathrm{Var}\left\{\tilde{\mathbf{U}}_1\left(\boldsymbol{\beta}_{0N}\right)\right\}\right],\end{aligned}$$

où $\mathcal{A}_N=\left\{\left(\mathbf{x}_{i1},y\right):i\in\mathcal{S}_1\right\}$, $A\succeq B$ est équivalent de sorte que $A-B$ est non négativement définitif pour les deux matrices A et B de mêmes dimensions, et la dernière inégalité se vérifie, puisque $\tilde{\mathbf{U}}_2\left(\boldsymbol{\alpha}_N^*\right)$ est non stochastiquement conditionnel à $\mathcal{A}_N$. Ainsi, $\mathrm{Var}\left\{\tilde{\mathbf{U}}_1\left(\boldsymbol{\beta}_{0N}\right)-\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\tilde{\mathbf{U}}_2\left(\boldsymbol{\alpha}_N^*\right)\right\}$ atteint un minimum si $\mathrm{Var}\left[E\left\{\tilde{\mathbf{U}}_1\left(\boldsymbol{\beta}_{0N}\right)-\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\tilde{\mathbf{U}}_2\left(\boldsymbol{\alpha}_N^*\right)\middle|\,\mathcal{A}_N\right\}\right]=0$, qui est induit par la condition $\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_1,y\right)=E\left\{\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x},y\right)\middle|\,\mathbf{x}_1,y\right\}$.

C. Démonstration du théorème 2

Avant de démontrer le théorème 2, nous avons besoin du résultat suivant.

Lemme A3. *Supposons que les conditions C1, C3-C5 sont vérifiées. Nous obtenons alors :*

$$\boldsymbol{\alpha}^* - \boldsymbol{\alpha}_N^* = O_p\left(n^{-1/2}\right).$$

Démonstration du lemme A3. Puisque $\hat{\boldsymbol{\alpha}}_2$ est obtenu par une enquête externe indépendante, nous concluons que $\left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1}$ peut permettre d'estimer la variance de $\hat{\boldsymbol{\alpha}}^*$. Ainsi, l'ordre de la variance de $\hat{\boldsymbol{\alpha}}^*$ est déterminé par l'estimateur le moins efficace entre $\hat{\boldsymbol{\alpha}}_1$ et $\hat{\boldsymbol{\alpha}}_2$. Si nous démontrions :

$$\hat{\boldsymbol{\alpha}}_1 - \hat{\boldsymbol{\alpha}}_N^* = O_p\left(n^{-1/2}\right), \tag{C.1}$$

nous pourrions obtenir $\mathbf{V}_1 = O_p\left(n^{-1}\right)$ selon (C5b). Puisque $\hat{\boldsymbol{\alpha}}^*$ est au moins aussi efficace que $\hat{\boldsymbol{\alpha}}_1$, nous avons démontré le lemme A3.

Ainsi, il reste à démontrer (C.1). Suivant C4 et C5, nous obtenons :

$$\begin{aligned}
\mathbf{0} &= \sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}_1; \mathbf{x}_{1i}, y_i\right) \\
&= \sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) + \left\{\frac{\partial}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} \sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\tilde{\boldsymbol{\alpha}}; \mathbf{x}_{1i}, y_i\right)\right\}\left(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_N^*\right),
\end{aligned} \tag{C.2}$$

où $\tilde{\boldsymbol{\alpha}}$ réside dans le segment reliant $\hat{\boldsymbol{\alpha}}_1$ et $\hat{\boldsymbol{\alpha}}_N^*$. Selon C3-C5 et (C.2), nous concluons que :

$$\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_N^* = -\boldsymbol{\mathcal{I}}_0^{-1} \sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) + o_p\left(n^{-1/2}\right). \tag{C.3}$$

Ainsi, selon C3 et (C.3), nous avons démontré (C.1).

Démonstration du théorème 2. Considérons :

$$\begin{aligned}
&\sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*; \mathbf{x}_{1i}, y_i\right) \\
&\quad= \sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) + \left\{\frac{\partial}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} \sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\tilde{\boldsymbol{\alpha}}; x_{1i}, y_i\right)\right\}\left(\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha}_N^*\right) \\
&\quad= \sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) + \boldsymbol{\mathcal{I}}_0\left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1} \mathbf{V}_2^{-1}\left(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^*\right) \\
&\qquad + \boldsymbol{\mathcal{I}}_0\left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1} \mathbf{V}_1^{-1}\left(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_N^*\right) + o_p\left(n^{-1/2}\right), \\
&\quad= \sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) + \boldsymbol{\mathcal{I}}_0\left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1} \mathbf{V}_2^{-1}\left(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^*\right) \\
&\qquad - \boldsymbol{\mathcal{I}}_0\left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1} \mathbf{V}_1^{-1} \boldsymbol{\mathcal{I}}_0^{-1} \sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) + o_p\left(n^{-1/2}\right), \\
&\quad= \boldsymbol{\mathcal{I}}_0\left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1} \mathbf{V}_2^{-1}\left(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^*\right) + \boldsymbol{\mathcal{I}}_0\left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1} \mathbf{V}_2^{-1} \boldsymbol{\mathcal{I}}_0^{-1} \sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) \\
&\qquad + o_p\left(n^{-1/2}\right) \\
&\quad= \boldsymbol{\mathcal{I}}_0 \mathbf{W}\left(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^*\right) + \boldsymbol{\mathcal{I}}_0 \mathbf{W} \boldsymbol{\mathcal{I}}_0^{-1} \sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i\right) + o_p\left(\kappa(n)\right),
\end{aligned} \tag{C.4}$$

où $\tilde{\boldsymbol{\alpha}}$ réside dans le segment reliant $\hat{\boldsymbol{\alpha}}^*$ et $\boldsymbol{\alpha}_N^*$, la deuxième égalité se vérifie selon C4 et le lemme A3, la troisième égalité se vérifie selon (C.3), la dernière égalité se vérifie selon C5d, $\kappa(n)=\gamma(n)$ si $\gamma(n)\,n^{1/2}\to\infty$, et $\kappa(n)=n^{-1/2}$ autrement, et $\gamma(n)$ est l'ordre de convergence de $\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)$ dans (C5e).

S'il existe une matrice non stochastique $\boldsymbol{\Sigma}_c$, de telle sorte que $n\mathbf{V}_2=\boldsymbol{\Sigma}_c+o_p(1)$, alors $\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)=O_p\left(n^{1/2}\right)$ et $\mathbf{W}$ n'est pas une matrice nulle. Selon (C.4), nous avons alors :

$$n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\begin{pmatrix}\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i\right)\\ \mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{1i},y_i\right)\end{pmatrix}$$
$$=\begin{pmatrix}\mathbf{0}\\ n^{1/2}\boldsymbol{\mathcal{I}}_0\mathbf{W}\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)\end{pmatrix}+n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\begin{pmatrix}\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i\right)\\ \boldsymbol{\mathcal{I}}_0\mathbf{W}\boldsymbol{\mathcal{I}}_0^{-1}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\end{pmatrix}+o_p(1). \qquad \text{(C.5)}$$

Puisque l'échantillon externe est indépendant de l'échantillon interne et que $\boldsymbol{\Sigma}_c$ est la variance asymptotique de $n^{1/2}\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)$, selon (C3), (C5e) et (C.5), nous concluons que :

$$n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\begin{pmatrix}\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i\right)\\ \mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{1i},y_i\right)\end{pmatrix}\to\mathcal{N}\left(0,\tilde{\boldsymbol{\Sigma}}_u\right),$$

où $\tilde{\boldsymbol{\Sigma}}_{11}=\boldsymbol{\Sigma}_{11}$, $\tilde{\boldsymbol{\Sigma}}_{12}=\boldsymbol{\Sigma}_{12}\left(\boldsymbol{\mathcal{I}}_0^{-1}\right)^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{\mathcal{I}}_0^{\mathrm{T}}$, $\tilde{\boldsymbol{\Sigma}}_{21}=\tilde{\boldsymbol{\Sigma}}_{12}^{\mathrm{T}}$ et $\tilde{\boldsymbol{\Sigma}}_{22}=\boldsymbol{\mathcal{I}}_0\mathbf{W}\left\{\boldsymbol{\Sigma}_c+\boldsymbol{\mathcal{I}}_0^{-1}\boldsymbol{\Sigma}_{22}\left(\boldsymbol{\mathcal{I}}_0^{-1}\right)^{\mathrm{T}}\right\}\mathbf{W}^{\mathrm{T}}\boldsymbol{\mathcal{I}}_o^{\mathrm{T}}$. Ainsi, nous avons démontré le premier cas du théorème 2.

Si $\mathbf{W}=0$, alors $\gamma(n)\,n^{1/2}\to\infty$ et le taux de $\kappa(n)$ est inférieur à $n^{-1/2}$ dans (C.4). Ainsi, le reste du terme de (C.4) n'est plus $o_p\left(n^{-1/2}\right)$ pour $\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{1i},y_i\right)$. Au lieu de cela, nous étudions d'abord dans ce cas l'ordre asymptotique de $(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1})^{-1}\mathbf{V}_2^{-1}$. Selon C3, C5b et (C.3), nous avons :

$$\mathbf{V}_1\asymp n^{-1}\quad\text{et}\quad\mathbf{V}_2\asymp\gamma(n)^{-2} \qquad \text{(C.6)}$$

en probabilité. Ainsi, (C.6) donne :

$$\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1}\asymp n^{-1}\gamma(n)^2 \qquad \text{(C.7)}$$

en probabilité par le fait que $\gamma(n)\,n^{1/2}\to\infty$. Ainsi, selon (C5e) et (C.7), nous obtenons :

$$\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1}\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)\asymp n^{-1}\gamma(n) \qquad \text{(C.8)}$$

en probabilité. Selon $\gamma(n)\,n^{1/2}\to\infty$, (C.7) et (C.8), nous avons démontré :

$$\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1}=o_p(1),$$
$$\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1}\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)=o_p\left(n^{-1/2}\right)$$

en probabilité. Ainsi, selon la quatrième égalité de (C.4), nous pouvons démontrer :

$$\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{1i},y_i\right)=o_p\left(n^{-1/2}\right),$$

et démontrons le troisième cas du théorème 2.

Bibliographie

Chatterjee, N., Chen, Y.H., Maas, P. et Carroll, R.J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111, 107-117.

Chen, J., et Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9(2), 385-406.

Chen, S., et Kim, J.K. (2014). Population empirical likelihood for nonparametric inference in survey sampling. *Statistica Sinica*, 24(1), 335-355.

Chen, Y., Li, P. et Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011-2021.

Chen, Y.H., et Chen, H. (2000). A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society: Series B*, 62, 449-460.

Csiszár, I., et Shields, P.C. (2004). Information theory and statistics: A tutorial. *Foundations and Trands in Communications and Information Theory*, 1, 417-528.

Elliott, M.R., et Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Engle, R.F., et McFadden, D.L. (1994). *Handbook of Econometrics: Volume IV*. Amsterdam.

Fuller, W.A. (2009). *Sampling Statistic*. Hoboken, New York: John Wiley & Sons, Inc.

Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029-1054.

Han, P., et Lawless, J.F. (2019). Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statistica Sinica*, 29(3), 1321-1342.

Hidiroglou, M. (2001). L'échantillonnage double. *Techniques d'enquête*, 27, 2, 157-169. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2001002/article/6091-fra.pdf.

Horn, R.A., et Johnson, C.R. (2012). *Matrix Analysis*. Cambridge university press, New York.

Imbens, G.W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics*, 20, 493-506.

Kim, J.K., et Rao, J.N.K. (2019). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.

Kim, J.K., et Wang, Z. (2019). Sampling techniques for big data analysis in finite population inference. *Revue Internationale de Statistique*, 87, S177-S191.

Kundu, P., Tang, R. et Chatterjee, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika*, 106(3), 567-585.

Lohr, S.L., et Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science*, 32, 293-312.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society: Series B*, 72, 27-48.

Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, 19, 1725-1747.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.

Qin, J., et Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22, 300-325.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83, 242-272.

Rivera-Rodriguez, C., Haneuse, S., Wang, M. et Spiegelman, D. (2020). Augmented pseudolikelihood estimation for two-phase studies. *Statistical Methods in Medical Research*, 29, 344-358.

Rivera-Rodriguez, C., Spiegelman, D. et Haneuse, S. (2019). On the analysis of two-phase designs in cluster-correlated data settings. *Statistics in Medicine*, 38, 4611-4624.

Robins, J.M., Rotnitzky, A. et Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866.

Sheng, Y., Sun, Y., Huang, C.-Y. et Kim, M.-O. (2021). Synthesizing external aggregated information in the presence of population heterogeneity: A penalized empirical likelihood approach. *Biometrics*.

Sheng, Y., Sun, Y., Huang, C.-Y. et Kim, M.-O. (2022). Synthesizing external aggregated information in the presence of population heterogeneity: A penalized empirical likelihood approach. *Biometrics*, 78(2), 679-690.

Shin, Y.E., Pfeiffer, R.M., Graubard, B.I. et Gail, M.H. (2020a). Weight calibration to improve efficiency for estimating pure risks from the additive hazards model with the nested case-control design. *Biometrics*, 1-13.

Shin, Y.E., Pfeiffer, R.M., Graubard, B.I. et Gail, M.H. (2020b). Weight calibration to improve the efficiency of pure risk estimates from case-control samples nested in a cohort. *Biometrics*, 76(4), 1087-1097.

Taylor, J.M.G., Choi, K. et Han, P. (2022). Data integration: Exploiting ratios of parameter estimates from a reduced external model. *Biometrika*, 1-16.

Wang, C.Y., Wang, S., Zhao, L.-P. et Ou, S.-T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association*, 92, 512-525.

Wang, H., et Kim, J.K. (2021). Propensity score estimation using density ratio model under item nonresponse. *arXiv preprint arXiv:2104.13469*.

Wang, L., Williams, M.L., Chen, Y. et Chen, J. (2020). Novel two-phase sampling designs for studying binary outcomes. *Biometrics*, 76, 210-223.

Wu, C., et Rao, J.N.K. (2006). Pseudo empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics/La revue canadienne de statistique*, 34(3), 359-375.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Xu, M., et Shao, J. (2020). Meta-analysis of independent datasets using constrained generalised method of moments. *Statistical Theory and Related Fields*, 4, 109-116.

Yang, S., et Kim, J.K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3, 625-650.

Yang, S., Zheng, D. et Wang, X. (2020). Elastic integrated analysis of randomized trial and real-world data for treatment heterogeneity estimation. *arXiv preprint arXiv:2005.10579v2*.

Yuan, K.-H., et Jennrich, R.I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, 65(2), 245-260.

Zhai, Y., et Han, P. (2022). Data integration with oracle use of external information from heterogeneous populations. *Journal of Computational and Graphical Statistics*, 0(0), 1-12.

Zhang, H., Deng, L., Wheeler, W., Qin, J. et Yu, K. (2021). Integrative analysis of multiple case-control studies. *Biometrics*, 1-12.

Zhao, P., Ghosh, M., Rao, J.N.K. et Wu, C. (2020). Bayesian empirical likelihood inference with complex survey data. *Journal of the Royal Statistical Society: Series B*, 82, 155-174.

Zubizarreta, J.R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110, 910-922.