Nº 11-522-x au catalogue
ISSN : 1709822-X

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

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

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

Date de diffusion : le 15 octobre 2021

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

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

Jae Kwang Kim, Hang J. Kim et Zhonglei Wang[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 de 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 se fonde sur une application nouvelle de la méthode de vraisemblance empirique. La méthode proposée est particulièrement attractive pour combiner des renseignements de plusieurs sources présentant différentes tendances d'information manquante. La méthode est appliquée à un exemple de données réelles combinant les données d'enquête de la Korean National Health and Nutrition Examination Survey (KNHANES, Enquête nationale coréenne sur la santé et la nutrition) et les mégadonnées du National Health Insurance Sharing Service (NHISS, Service national coréen de partage de l'assurance maladie).

Mots clés : mégadonnées; probabilité empirique; 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. 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) et Yang et Kim (2020) examinent les méthodes statistiques d'intégration des données aux fins d'inférence sur une population finie. De nombreuses méthodes (p. ex. Merkouris, 2010; Zubizarreta, 2015) cherchent principalement à estimer des moyennes ou des totaux de population. En effet, la combinaison de renseignements à des fins d'inférence analytique, comme l'analyse de régression, n'est pas pleinement étudiée dans la littérature publiée.

Dans cette étude, nous envisageons l'analyse de régression dans le contexte de l'intégration de données. Quand 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 avec covariables manquantes. Robins et coll. (1994) et Wang et coll. (1997) ont étudié l'estimation semi-paramétrique dans l'analyse de régression avec données de covariables manquantes selon l'hypothèse de covariables manquantes au hasard. Dans notre configuration d'intégration des données, la source de données externe avec covariable manquante peut être un recensement ou des mégadonnées, et il peut y avoir un biais de sélection dans les données externes.

Pour combiner des renseignements partiels de sources externes, nous utilisons l'idée de modèle de calage (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 au moyen d'une nouvelle application de la méthode de vraisemblance empirique (Owen, 1991; Qin et Lawless, 1994). Le modèle de travail réduit n'est pas nécessairement spécifié correctement, mais un bon modèle de travail peut améliorer l'efficacité de l'analyse qui en résulte. La méthode proposée est particulièrement attractive pour combiner des renseignements de plusieurs sources

[1]Jae Kwang Kim, Département de statistique, Iowa State University, États-Unis, 50011; Hang J. Kim, Division de la statistique et de la science des données, University of Cincinnati, États-Unis, 45221; Zhonglei Wang, Wang Yanan Institute for Studies in Economics, Xiamen University, Chine, 361005

de données présentant différentes tendances d'information manquante. Dans ce cas, il suffit de spécifier différents modèles de travail pour différentes tendances manquantes.

Dans une configuration similaire, Chatterjee et coll. (2016) ont également élaboré une méthode de calage fondée sur le maximum de vraisemblance contraint, qui utilise un modèle entièrement paramétrique pour la spécification de la vraisemblance et une contrainte élaborée à partir du modèle réduit pour l'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. 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 élargir l'applicabilité de la méthode proposée à de nombreux problèmes pratiques. En particulier, la méthode proposée s'applique directement aux données de l'échantillon d'enquête, qui est l'objet principal de notre article. Récemment, Xu et Shao (2020) ont élaboré une méthode d'intégration des données au moyen de la méthode généralisée de la technique des moments, mais leur méthode suppose implicitement que le modèle réduit est correctement spécifié. Sheng et coll. (2021) ont proposé une méthode de vraisemblance empirique pénalisée pour intégrer ce type d'information dans la configuration de régression logistique. Zhang et coll. (2021) ont également conçu un cadre de vraisemblance empirique rétrospective pour tenir compte du biais d'échantillonnage dans les études cas-témoins. Nous examinons une configuration de régression plus générale. La méthode de vraisemblance empirique que nous proposons est différente des leurs et ne nécessite pas que le modèle de travail réduit soit correctement spécifié.

Nous présenterons nos travaux dans l'article comme suit. Premièrement, nous proposons un cadre unifié d'intégration des sources de données externes dans l'analyse de régression. La méthode proposée utilise des hypothèses plus faibles que la méthode de Chatterjee et coll. (2016), ce qui donne des résultats d'estimation plus robustes. Deuxièmement, la méthode proposée a une large application, car elle peut facilement traiter plusieurs sources de données externes, comme le démontre la section 2.3. Elle s'applique également quand la source de données externe présente un biais de sélection. Dans l'application sur des données réelles de la section 3, nous démontrons que la méthode proposée peut utiliser des mégadonnées externes avec des probabilités de sélection inconnues en appliquant un ajustement de pondération du score de propension. Enfin, la méthode que nous proposons est facile à mettre en œuvre et pleinement justifiée en théorie. Le calcul est une application directe de la méthode type de vraisemblance empirique, facilement réalisable au moyen de logiciels existants.

2. Méthode proposée

2.1 Scénario de base

Nous considérons une population finie $\mathcal{U} = \{1, \dots, N\}$ de taille N. En lien avec le i^{e} enregistrement, soit y_i la variable d'intérêt de l'étude et $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$ le vecteur auxiliaire correspondant de longueur p. Nous souhaitons estimer un paramètre de population $\boldsymbol{\beta}_0$, qui résout $\boldsymbol{U}_1(\boldsymbol{\beta}) = \sum_{i\in\mathcal{U}} \boldsymbol{U}_1(\boldsymbol{\beta}; \boldsymbol{x}_i, y_i) = \boldsymbol{0}$ où $U_1(\boldsymbol{\beta}, \boldsymbol{x}, y)$ est une fonction d'estimation pré-spécifiée pour $\boldsymbol{\beta}$. Un exemple de fonction d'estimation est $\boldsymbol{U}_1(\boldsymbol{\beta}; \boldsymbol{x}_i, y_i) = \{y_i - m_1(\boldsymbol{x}_i; \boldsymbol{\beta})\}\, \boldsymbol{h}_1(\boldsymbol{x}_i; \boldsymbol{\beta})$, qui est implicitement basée sur un modèle de régression $E(Y_i|\boldsymbol{x}_i) = m_1(\boldsymbol{x}_i; \boldsymbol{\beta})$ au niveau de la super-population pour quelques $\boldsymbol{h}_1(\boldsymbol{x}_i; \boldsymbol{\beta})$ qui satisfont certaines conditions d'identification (p. ex. Kim et Rao, 2009). À partir de la population finie, on génère un échantillon probabiliste $S_1 \subset \mathcal{U}$ et on peut obtenir un estimateur $Z\ \widehat{\boldsymbol{\beta}}$ en résolvant

$$(1) \quad \widehat{U}_1(\boldsymbol{\beta}) = \sum_{i\in S_1} d_i \boldsymbol{U}_1(\boldsymbol{\beta}; \boldsymbol{x}_i, y_i) = \boldsymbol{0},$$

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

En plus de S_1, supposons que nous observons $\boldsymbol{x}_{i1}$ et y_i dans toute la population finie et que nous souhaitons intégrer ces renseignements supplémentaires pour améliorer l'efficacité de l'estimation de $\widehat{\boldsymbol{\beta}}$. Chen et Chen (2000) ont d'abord examiné ce problème dans le contexte des modèles d'erreur de mesure. Pour expliquer leur idée dans notre scénario, nous considérons d'abord un modèle « de travail » réduit,

$$(2) \quad E(Y_i|\boldsymbol{x}_{i1}) = m_2(\boldsymbol{x}_{i1};\boldsymbol{\alpha})$$

pour quelques valeurs $\boldsymbol{\alpha}$. Dans le modèle de travail (2), nous pouvons obtenir un estimateur $\widehat{\boldsymbol{\alpha}}$ à partir de l'échantillon actuel S_1 en résolvant $\widehat{\boldsymbol{U}}_2(\boldsymbol{\alpha}) = \sum_{i\in S_1} d_i \boldsymbol{U}_2(\boldsymbol{\alpha};\boldsymbol{x}_{i1},y_i) = \boldsymbol{0}$ où $\boldsymbol{U}_2(\boldsymbol{\alpha};\boldsymbol{x}_{i1},y_i) = \{y_i - m_2(\boldsymbol{x}_{i1};\boldsymbol{\alpha})\}\,\boldsymbol{h}_2(\boldsymbol{x}_{i1};\boldsymbol{\alpha})$ pour quelques $\boldsymbol{h}_2(\boldsymbol{x}_{i1};\boldsymbol{\alpha})$ satisfaisant des conditions semblables à celles imposées à $\boldsymbol{h}_1(\boldsymbol{x}_i;\boldsymbol{\beta})$. Notons que notre configuration envisage la situation où un sous-ensemble de données individuelles $(\boldsymbol{x}_{i1},y_i)$ est entièrement observé dans l'ensemble de la population finie $\mathcal{U}$. Par conséquent, on peut obtenir $\boldsymbol{\alpha}^*$ qui résoud $\sum_{i=1}^{N} \boldsymbol{U}_2(\boldsymbol{\alpha};\boldsymbol{x}_{i1},y_i) = \boldsymbol{0}$. Chen et Chen (2000) ont proposé d'utiliser $\widehat{\boldsymbol{\beta}}^* = \widehat{\boldsymbol{\beta}} + \widehat{Cov}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\alpha}})\{\hat{V}(\widehat{\boldsymbol{\alpha}})\}^{-1}(\boldsymbol{\alpha}^* - \widehat{\boldsymbol{\alpha}})$ comme estimateur efficace de $\boldsymbol{\beta}$, où $\hat{V}(\cdot)$ et $\widehat{Cov}(\cdot)$ désignent respectivement les estimateurs de variance et de covariance fondés sur le plan. Le modèle de travail (2) n'est pas nécessairement spécifié correctement, mais un bon modèle de travail peut améliorer l'efficacité de l'estimateur final.

On peut aussi adopter le maximum de vraisemblance contraint (MVC) comme dans Chatterjee et coll. (2016), qui avait d'abord été proposé dans un contexte d'échantillonnage non lié à une enquête. Dans une configuration d'échantillonnage d'enquête, nous pouvons interpréter Chatterjee et coll. (2016) comme une méthode d'estimation par le MVC quand $\boldsymbol{\beta}$ est un paramètre dans la distribution conditionnelle de Y_i étant donné $\boldsymbol{X}_i$ avec une densité $f(y_i|\boldsymbol{x}_i;\beta)$, et l'estimation par le MVC peut être exprimée comme le fait de trouver $\boldsymbol{\beta}$ qui maximise

$$(3) \quad l_p(\boldsymbol{\beta}) = \sum_{i\in S_1} d_i \log f(y_i|\boldsymbol{x}_i;\boldsymbol{\beta})$$

sujet à

$$(4) \quad \sum_{i\in S_1} d_i \int \boldsymbol{U}_2(\boldsymbol{\alpha}^*;\boldsymbol{x}_{i1},y) f(y|\boldsymbol{x}_i;\boldsymbol{\beta})\, dy = \boldsymbol{0}.$$

La contrainte (4) peut être comprise comme une contrainte pour que le paramètre $\boldsymbol{\beta}$ satisfasse $E\{\boldsymbol{U}_2(\boldsymbol{\alpha}^*;\boldsymbol{x}_{i1},Y_i)|\boldsymbol{x}_i;\boldsymbol{\beta}\} = \boldsymbol{0}$. 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 par le MVC n'est pas directement applicable à notre modèle de moyenne conditionnelle présenté dans (1), car la fonction de vraisemblance pour $\boldsymbol{\beta}$ n'est pas définie dans notre configuration. Néanmoins, on peut utiliser une fonction objective comme celle de la méthode généralisée des moments pour appliquer le problème d'optimisation contrainte, qui est asymptomatiquement équivalent à la méthode de vraisemblance empirique (Imbens, 2002). Chatterjee et coll. (2016) ont également constaté que la méthode par le MVC pouvait être formulée au moyen de la méthode de probabilité empirique de Qin et Lawless (1994) et Qin (2000). Toutefois, ils n'ont pas discuté explicitement de la façon de formuler le MVC comme application de la méthode de vraisemblance empirique.

2.2 Méthode proposée

Nous utilisons maintenant le cadre de vraisemblance empirique pour intégrer l'information auxiliaire. Le problème de calage classique peut être formulé comme la découverte des poids de calage $\boldsymbol{w} = \{w_i : i \in S_1\}$ basés sur une certaine fonction objective $Q(\boldsymbol{d},\boldsymbol{w})$ sujette à certaines contraintes de calage (Deville et Särndal, 1992) où $\boldsymbol{d} = \{d_i : i \in S_1\}$. Pour la fonction objective, nous pouvons soit utiliser la fonction de vraisemblance pseudo-empirique.

$$(5) \quad Q(\boldsymbol{d},\boldsymbol{w}) = \sum_{i\in S_1} d_i \log w_i$$

étudiée par Wu et Rao (2006), soit la fonction d'entropie maximale $Q(\boldsymbol{d},\boldsymbol{w}) = \sum_{i\in S_1} w_i \log(w_i/d_i)$ présentée dans Kim (2010). Notre contrainte de calage est

$$(6) \quad \sum_{i\in S_1} w_i \boldsymbol{U}_2(\boldsymbol{\alpha}^*;\boldsymbol{x}_{i1},y_i) = \boldsymbol{0},$$

où $\boldsymbol{\alpha}^*$ est l'information externe pour le modèle de travail réduit. Cette démarche ressemble à l'utilisation de (4), mais sans introduire la fonction de densité conditionnelle $f(y|\boldsymbol{x},\boldsymbol{\beta})$. Par conséquent, nous pouvons utiliser la méthode de calage du modèle suivante pour effectuer une estimation efficace de $\boldsymbol{\beta}$ comme suit : on utilise le modèle de travail réduit (2) pour obtenir $\boldsymbol{\alpha}^*$ à partir de la population finie; on trouve les poids de calage $\widehat{\boldsymbol{w}} = \{\widehat{w}_i : i \in S_1\}$ maximisant $Q(\boldsymbol{d},\boldsymbol{w})$ sujet à (6); et une fois qu'on obtient la solution $\widehat{\boldsymbol{w}}$ à partir du calage, on estime $\boldsymbol{\beta}$ en résolvant $\sum_{i\in S_1} \widehat{w}_i \boldsymbol{U}_1(\boldsymbol{\beta};\boldsymbol{x}_i,y_i) = \boldsymbol{0}$.

Si la donnée de référence $\boldsymbol{\alpha}^*$ 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 original et de l'échantillon externe pour obtenir l'estimation de la donnée de référence. Dans la pratique, nous n'avons pas nécessairement accès aux données brutes de l'échantillon externe, mais nous sommes souvent en mesure d'avoir ses statistiques sommaires. Supposons que l'échantillon externe fournit un estimateur ponctuel $\widehat{\boldsymbol{\alpha}}_2$ et son estimateur de variance $\boldsymbol{V}_2 = \hat{V}(\widehat{\boldsymbol{\alpha}}_2)$ pour le modèle de travail réduit de (2). Alors, on peut obtenir un estimateur de la donnée de référence $\boldsymbol{\alpha}^*$ par

$$(7) \quad \widehat{\boldsymbol{\alpha}}^* = (\boldsymbol{V}_1^{-1} + \boldsymbol{V}_2^{-1})^{-1}(\widehat{\boldsymbol{V}}_1^{-1}\widehat{\boldsymbol{\alpha}}_1 + \widehat{\boldsymbol{V}}_2^{-1}\widehat{\boldsymbol{\alpha}}_2)$$

où $\widehat{\boldsymbol{\alpha}}_1$ et $\boldsymbol{V}_1$ sont estimés avec l'échantillon interne S_1. Une fois $\widehat{\boldsymbol{\alpha}}^*$ obtenu par (7), il remplace $\bmalpha^*$ dans l'équation de calage de (6).

2.3 Intégration de données multiples

Intéressons-nous à une analyse de régression combinant de l'information partielle provenant d'échantillons externes. Pour expliquer l'idée, le tableau 2.3.1 présente un exemple de structure de données avec 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 2.3.1
Structure de données pour l'intégration des données d'enquête

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

Dans la configuration du tableau 2.3.1, supposons que nous sommes intéressés par l'estimation des paramètres dans le modèle de régression $E(Y|x_1,x_2) = m_1(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ où $m_1(\cdot)$ est connu mais $\boldsymbol{\beta} = (\beta_0,\beta_1,\beta_2)$ inconnu. On peut écrire l'équation d'estimation de $\boldsymbol{\beta}$ en utilisant l'échantillon A comme suit :

$$(8) \quad \widehat{\boldsymbol{U}}_a(\boldsymbol{\beta}) = \sum_{i\in A} d_{a,i}\{y_i - m(x_{i1},x_{i2};\boldsymbol{\beta})\}\boldsymbol{h}(x_{i1},x_{i2};\boldsymbol{\beta}) = \boldsymbol{0},$$

pour quelques $\boldsymbol{h}(x_{i1},x_{i2};\boldsymbol{\beta})$ de telle sorte que $\widehat{\boldsymbol{U}}_a(\boldsymbol{\beta})$ soit linéairement indépendant presque partout.

Nous souhaitons maintenant intégrer l'information partielle de l'échantillon B. Pour ce faire, supposons que nous ayons un modèle « de travail » $E(Y|x_1,z) = m_2(x_1,z;\boldsymbol{\alpha})$ pour quelques $\boldsymbol{\alpha}$. Notons qu'étant donné que (z_i,x_{i1},y_i) sont observés, nous pouvons utiliser l'échantillon B pour estimer $\boldsymbol{\alpha}$ en résolvant $\sum_{i\in B} d_{b,i}\boldsymbol{U}_b(\boldsymbol{\alpha};x_{i1},z_i,y_i) = \boldsymbol{0}$ pour quelques $\boldsymbol{U}_b$ satisfaisant $E\{\boldsymbol{U}_b(\boldsymbol{\alpha};x_1,z,Y|x_1,z)\} = \boldsymbol{0}$ dans le modèle de travail pour $E(Y|x_1,z)$.

De même, pour intégrer l'information partielle de l'échantillon C, nous supposons que nous avons un modèle « de travail » $E(Y|x_2,z) = m_3(x_2,z;\boldsymbol{\gamma})$ pour quelques $\boldsymbol{\gamma}$. Nous pouvons également construire une équation d'estimation sans biais $\sum_{i\in C} d_{c,i}\boldsymbol{U}_c(\boldsymbol{\gamma};x_{i2},z_i,y_i) = 0$ pour quelques $\boldsymbol{U}_c$ qui satisfont $E\{\boldsymbol{U}_c(\boldsymbol{\gamma};x_2,z,Y)|x_2,z\} = \boldsymbol{0}$ dans le modèle de travail pour $E(Y|x_2,z)$. Une fois que nous obtenons $\widehat{\boldsymbol{\alpha}}$ et $\widehat{\boldsymbol{\gamma}}$, nous pouvons utiliser cette information supplémentaire

pour améliorer l'efficacité de $\widehat{\boldsymbol{\beta}}$ dans (8). Pour intégrer l'information supplémentaire, nous pouvons la formuler comme maximisant $Q(\boldsymbol{d}_a, \boldsymbol{w}) = \sum_{i\in A} d_{a,i} \log w_i$ sujet à $\sum_{i\in A} w_i = N$ et

$$(9) \quad \sum_{i\in A} w_i \{\boldsymbol{U}_b(\widehat{\boldsymbol{\alpha}}; x_{i1}, z_i, y_i), \boldsymbol{U}_c(\widehat{\boldsymbol{\gamma}}; x_{i2}, z_i, y_i)\} = \mathbf{0}$$

où $\boldsymbol{d}_a$ et $\boldsymbol{w}$ sont des ensembles contenant les poids d'échantillonnage et les poids de calage pour ce qui est de l'échantillon A. La contrainte (9) intègre l'information supplémentaire. Une fois la solution $\widehat{w}_i$ obtenue, nous pouvons utiliser $\sum_{i\in A} \widehat{w}_i \{y_i - m(x_{i1}, x_{i2}; \boldsymbol{\beta})\} \boldsymbol{h}(x_{i1}, x_{i2}; \boldsymbol{\beta}) = \mathbf{0}$ pour estimer $\boldsymbol{\beta}$.

3. Étude par simulations

Pour évaluer les performances des estimateurs proposés sur un échantillon fini, nous avons réalisé une étude par simulations. Nous avons généré une population finie de taille $N = 100{,}000$, chaque enregistrement se composant des variables auxiliaires $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ et d'une variable réponse y_i. Nous supposons que $(\boldsymbol{x}_i, y_i)$ est disponible pour l'échantillon interne S_1 tandis que seul (x_{i1}, y_i) est disponible pour la population finie $\mathcal{U}$ ou un échantillon externe S_2.

Concernant les covariables, nous examinons deux scénarios pour la population finie : Scénario 1. $x_{i1} \sim N(3{,}1)$, $x_{i2} \sim N(11{,}6.5^2)$, et x_{i1} et x_{i2} sont indépendants; et Scénario 2. $x_{i1} \sim N(3{,}1)$ et $x_{i2} = x_{i1}^2 + N(0{,}1)$. Les paramètres de simulation sont choisis de façon à ce que la moyenne marginale et la variance de x_{i2} soient semblables dans les scénarios de dépendance et d'indépendance. La variable réponse y_i est générée au moyen d'une distribution de Bernoulli avec une probabilité de réussite $\Pr(Y_i = 1|x_{i1}, x_{i2}) = logit^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})$ avec les paramètres de simulation $(\beta_0, \beta_1, \beta_2) = (-0.5{,}0.1, -0.2)$. Nous considérons un échantillonnage de Poisson avec des probabilités d'inclusion satisfaisant $\pi_i \propto 0.9\, I(y_i = 1) + 0.1\, I(y_i = 0)$ pour produire un échantillon probabiliste S_1 de taille $n_1 = 5{,}000$.

Aux fins de la méthode proposée, nous considérons le modèle de travail réduit écrit par $U_2(\alpha; x_{i1}, y_i) = \{y_i - expit(\alpha_0 + \alpha_1 x_{i1})\}(1\ x_{i1})^{\mathrm{T}}$ où $expit(x) = \{1 + \exp(-x)\}^{-1}$. Nous comparons les performances de cinq méthodes : (i) l'estimateur avec l'échantillon probabiliste S_1 seulement, la solution de $\sum_{i\in S} d_i expit(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})(1\ x_{i1}\ x_{i2})^{\mathrm{T}} = \mathbf{0}$; (ii) l'estimateur proposé avec $\boldsymbol{\alpha}^*$ à partir de la population finie $\mathcal{U}$; (iii) l'estimateur proposé avec $\widehat{\boldsymbol{\alpha}}^*$ estimé à partir d'un échantillon externe S_2 de taille $n_2 = 50{,}000$; (iv) l'estimateur par le MVC s'appuyant sur (3) et (4) avec des renseignements partiels tirés de $\mathcal{U}$; et (v) l'estimateur par le MVC avec S_2.

Le tableau 3.1 montre les résultats de la simulation. Quand x_{i1} et x_{i2} sont générés indépendamment dans les données de simulation (Scénario 1), toutes les méthodes présentent des biais négligeables, et les quatre méthodes utilisant l'information supplémentaire sont plus efficaces pour l'estimation de β_0 et β_1 que l'utilisation de l'échantillon interne seulement (S_1 seulement). Quand x_{i1} et x_{i2} dans les données de simulation sont dépendants (Scénario 2), les estimateurs par le MVC (MVC-$\mathcal{U}$ et MVC-S_2) présentent des biais importants entraînant une EQM élevée et des couvertures d'intervalle de confiance incorrectes pour MVC-S_2. Dans les cas de covariables dépendantes, les estimateurs proposés présentent toujours des biais négligeables et des couvertures correctes des intervalles de confiance, et leurs erreurs quadratiques moyennes pour β_0 et β_1 sont également plus petites que celles de S_1-seulement. Notons que l'estimation de β_2 par les méthodes proposées n'apporte pas le gain d'efficacité attendu parce que les données externes se composent de x_{i1} seulement.

Tableau 3.1
Performances de la régression logistique selon un échantillonnage de Poisson mesurées par le biais de Monte Carlo (Biais), la racine de l'erreur quadratique moyenne (REQM) et une couverture de l'intervalle de confiance (IC) à 95 % mesurée aux fins de l'estimation avec l'échantillon interne seulement (S_1-seulement); méthodes proposées quand (y_i, x_{i1}) sont disponibles pour l'ensemble de la population (Prop $\mathcal{U}$) et pour un échantillon externe (Prop-S_2); et estimateurs par le MVC de type Chatterjee (MVC-$\mathcal{U}$ et MVC-S_2).

Scénario 1 : x_{i1} et x_{i2} sont indépendants

	β_0			β_1			β_2		
	Biais	REQM	IC	Biais	REQM	IC	Biais	REQM	IC
S_1 seulement	-0,002	0,129	0,950	0,001	0,037	0,933	0,000	0,007	0,953
Prop-$\mathcal{U}$	-0,003	0,097	0,944	0,001	0,025	0,934	0,000	0,007	0,952
MVC-$\mathcal{U}$	-0,024	0,082	--	0,006	0,020	--	0,000	0,006	--
Prop-S_2	-0,001	0,093	0,939	0,000	0,024	0,934	0,000	0,007	0,952
MVC-S_2	-0,022	0,076	0,959	0,006	0,018	0,971	0,000	0,006	0,996

Scénario 2 : x_{i1} et x_{i2} sont dépendants

	β_0			β_1			β_2		
	Biais	REQM	IC	Biais	REQM	IC	Biais	REQM	IC
S_1 seulement	0,002	0,174	0,947	0,000	0,126	0,949	0,000	0,024	0,946
Prop-$\mathcal{U}$	0,000	0,112	0,948	0,000	0,109	0,946	0,000	0,024	0,945
MVC-$\mathcal{U}$	0,504	0,511	--	-0,409	0,417	--	0,073	0,075	--
Prop-S_2	0,001	0,107	0,949	-0,001	0,108	0,948	0,000	0,024	0,945
MVC-S_2	0,505	0,511	0,002	-0,410	0,417	0,054	0,073	0,075	0,229

4. Étude d'application

4.1 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 la Korea National Health and Nutrition Examination Survey (KNHANES, Enquête nationale coréenne sur la santé et la nutrition). L'enquête annuelle comprend environ 5 000 personnes chaque année et recueille 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 la KNHANES est un échantillonnage stratifié utilisant l'âge, le sexe et la région 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 fournis aux utilisateurs de données avec les variables d'enquête.

Pour améliorer l'efficacité de l'analyse des données avec une KNHANES de taille $n_1 = 4\,929$, nous avons utilisé une base de données publique externe fournie par le National Health Insurance Sharing Service (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=un million de personnes, dont certaines variables sont un sous-ensemble de variables dans la KNHANES.

Ces structures de données, avec un petit n_1, un grand n_2, et des mégadonnées ayant un sous-ensemble de variables dans l'échantillon interne, conviennent bien au scénario présenté dans 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, leurs probabilités de sélection sont inconnues, de sorte que l'estimateur convergent par rapport au plan $\hat{\boldsymbol{\alpha}}_2$ de (7) n'est pas disponible. La section 4.2 aborde cette question en utilisant une méthode de pondération de la propension et la section 4.3 présente le résultat de l'analyse de l'étude d'application.

4.2 Pondération de la propension pour des données externes avec probabilité de sélection inconnue

Nous envisageons maintenant d'étendre la méthode proposée à une situation où l'échantillon externe S_2 se compose de mégadonnées dont les probabilités de sélection sont inconnues. Dans ce cas, le modèle de travail pour $E(Y_i|\boldsymbol{x}_{i1}) = m(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i1})$ peut ne pas se vérifier pour l'échantillon S_2. Néanmoins, nous pouvons tout de même résoudre

$$(10) \quad \sum_{i\in S_2}\{y_i - m(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_{i1})\}\boldsymbol{x}_{i1} = \mathbf{0}$$

pour obtenir $\hat{\alpha}_0$ et $\hat{\alpha}_1$. Si le mécanisme d'échantillonnage pour S_2 est ignorable ou non informatif, alors la solution de (10) est 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 (10) 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, $P(\delta_i = 1|\boldsymbol{x}_{i1}, y_i) = \pi(\boldsymbol{x}_{i1}, y_i; \boldsymbol{\phi})$, où $\delta_i = 1$ si $i \in S_2$ et zéro sinon.

Notons que nous pouvons exprimer $\pi(\boldsymbol{x}_{i1}, y_i)^{-1} = 1 + (N_0/N_1)r(\boldsymbol{x}_{i1}, y_i)$ où $\log r(\boldsymbol{x}_{i1}, y_i) = f(\boldsymbol{x}_{i1}, y_i|\delta_i = 0)/f(\boldsymbol{x}_{i1}, y_i|\delta_i = 1)$ 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 log-linéaire, $\log r(x_{i1}, y_i; \boldsymbol{\phi}) = \phi_0 + \phi_1 x_{i1} + \phi_2 y_i$. On obtient l'estimateur par l'entropie maximale de $\boldsymbol{\phi}$ en résolvant $(1/N_1)\sum_{i=1}^{N}\delta_i \exp(\phi_0 + \phi_1 x_{i1} + \phi_2 y_i)(1\ x_{i1}\ y_i)^{\mathrm{T}} = (1\ \hat{\bar{x}}_1\ \hat{\bar{y}})^{\mathrm{T}}$ où $(\hat{\bar{x}}_1, \hat{\bar{y}}) = \left(1/\widehat{N}_0\right)\{\sum_{i\in S_1} d_1(x_{i1} - y_i) - \sum_{i=1}^{N}\delta_i(x_{i1}, y_i)\}$, $\widehat{N}_0 = \sum_{i\in S_1} d_i - N_1$, et S_1 est l'échantillon interne. Après avoir obtenu $\widehat{\boldsymbol{\phi}}$, nous pouvons construire $\hat{\pi}(x_{i1}, y_i)$ et résoudre

$$(11) \quad \sum_{i\in S_2}\frac{1}{\pi(x_{i1}, y_i)}\{y_i - m(\alpha_0 + \alpha_1 x_{i1})\}(1\ x_{i1})^{\mathrm{T}} = \mathbf{0}$$

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

De plus, nous pouvons utiliser l'échantillon interne S_1 pour ajuster le même modèle de travail afin d'obtenir $\widehat{\boldsymbol{\alpha}}_1$. Ensuite, nous obtenons $\widehat{\boldsymbol{\alpha}}^*$ en utilisant (7) et nous appliquons la méthode de pondération par calage proposée pour combiner les renseignements des mégadonnées. Dans la pratique, il est difficile de calculer $\boldsymbol{V}_2$ dans (7), mais sa taille est négligeable si la taille de l'échantillon de S_2 est très grande. Dans ce cas, nous pouvons simplement utiliser $\widehat{\boldsymbol{\alpha}}^* = \widehat{\boldsymbol{\alpha}}_2$ dans le problème de calage.

4.3 Résultats de l'étude d'application : la KNHANES ou l'Enquête nationale coréenne sur la santé et la nutrition

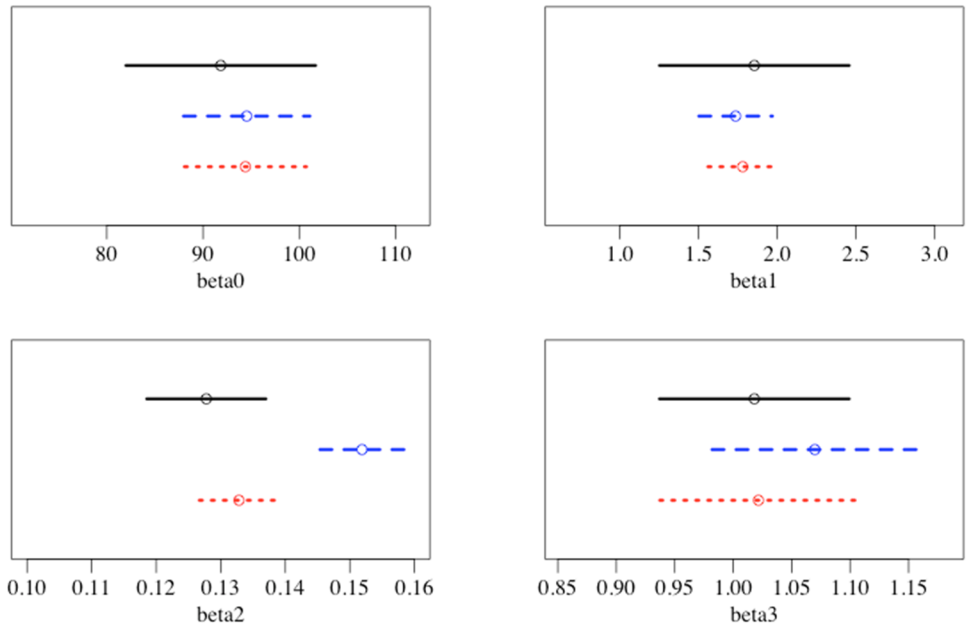
Dans cette étude d'application, nous utilisons $n_1 = 4\ 929$ enregistrements des données de la KNHANES qui n'ont pas de valeurs manquantes pour quatre variables : cholestérol total, hémoglobine, triglycéride et cholestérol à lipoprotéines de haute densité (LHD). À des fins de démonstration, nous supposons qu'un analyste souhaite réaliser l'analyse de régression suivante, $E(\text{Total Cholesterol}_i|\boldsymbol{x}_i) = \beta_0 + \beta_1\text{Hemoglobin}_i + \beta_2\text{Triglyceride}_i + \beta_3\text{LDH}_i$ pour $i \in S_1$. 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 LHD, ce qui ressemble au scénario de la section 3 où les covariables étaient fortement corrélées. Les données externes de grande taille consistent en $n_2 =$ un million d'enregistrements de données du NHISS, ayant des valeurs entièrement observées pour le cholestérol total, l'hémoglobine et le triglycéride. Pour relier l'échantillon externe à l'échantillon interne, nous supposons le modèle de travail $E(\text{Total Cholesterol}_i|\boldsymbol{x}_{i1}) = \alpha_0 + \alpha_1\text{Hemoglobin}_i + \alpha_2\text{Triglyceride}_i$ for $i \in S_1 \cup S_2$.

Dans cette étude d'application, nous mettons en œuvre les méthodes proposées sur l'échantillon externe où $\widehat{\boldsymbol{\alpha}}_2$ est utilisé au lieu de $\boldsymbol{\alpha}^*$ qui n'est pas disponible puisque nous n'avons pas des renseignements sur l'ensemble de la population. À partir de l'échantillon externe dont les probabilités de sélection sont inconnues, nous préparons deux

versions des méthodes proposées : (i) considérons S_2 comme un échantillon aléatoire simple, c'est-à-dire sans pondération de propension, et (ii) avec l'ajustement de pondération de propension présenté dans la section 4.2. Pour la pondération de propension, nous ajustons le modèle de ratio de densité log-linéaire aux données externes, $\log r(x_{i1}, y_i; \boldsymbol{\phi}) = \phi_0 + \phi_1 \text{Hemoglobin}_i + \phi_2 \text{Triglyceride}_i + \phi_3 \text{Total Cholesterol}_i$, calculons $\hat{\pi}(x_{i1}, y_i)$ étant donné $\widehat{\boldsymbol{\phi}}$, puis nous résolvons (11) pour obtenir $\widehat{\boldsymbol{\alpha}}_2$. Les performances des méthodes proposées sont comparées à la méthode de référence, qui utilise l'échantillon interne S_2 seulement pour obtenir des estimations des moindres carrés pondérées en tenant compte des poids d'échantillonnage.

La figure 4.3.1 montre les estimations ponctuelles et les intervalles de confiance de 95 % de $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ pour chaque méthode. Les méthodes proposées montrent des variances plus faibles 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 présente une différence systématique par rapport aux deux autres estimateurs. Quand l'ajustement de la pondération de propension est conjugué à la méthode proposée, son intervalle de confiance de β_2 est compris dans celui qu'on obtient en utilisant l'échantillon interne seulement. Ce résultat implique que le biais systématique dû au fait d'ignorer les probabilités d'échantillonnage est traité par l'ajustement de la pondération de propension. On n'attendait aucun gain d'efficacité dans l'estimation de β_3, car les données externes contiennent des renseignements sur x_{i1} (hémoglobine) et x_{i2} (triglycéride), et non sur x_{i3} (cholestérol LHD).

Figure 4.3.1
Comparaison de l'analyse de régression pour $E(\text{Total Cholesterol}_i | x_i) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\text{Hemoglobin}_i + \boldsymbol{\beta_2}\text{Triglyceride}_i + \boldsymbol{\beta_3} LHD_i$ utilisant les données internes de la KNHANES appuyées par les mégadonnées externes de la base de données du NHISS. Pour chaque panneau, les cercles sont des estimations ponctuelles et les lignes sont leurs intervalles de confiance à 95 % pour l'utilisation de l'échantillon interne 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).



5. Conclusion

L'intégration de sources de données externes à l'analyse de régression de l'échantillon interne est un problème pratique important. Nous nous sommes penchés sur ce problème en utilisant une nouvelle application de la pondération par calage du modèle (Wu et Sitter, 2001). La méthode proposée s'applique directement aux enquêtes avec échantillon 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 aux 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 dans le prolongement de notre recherche. Tout d'abord, une approche bayésienne peut être élaborée dans la même configuration. On peut en effet utiliser la méthode bayésienne de vraisemblance empirique de Zhao et coll. (2020) dans cette configuration. La méthode proposée peut aussi servir à combiner les données d'essais cliniques randomisés avec des mégadonnées réelles (Yang et coll., 2020). Ces extensions de la méthode seront présentées ailleurs. Il serait également intéressant de relier la méthode proposée à un plan de sondage à deux degrés (double), dont l'efficacité de plan et d'estimation a récemment fait l'objet d'études approfondies (Rivera-Rodriguez et coll., 2019; Wang et coll., 2020). La structure des données de l'échantillonnage à deux degrés, avec un échantillon au premier degré présentant un grand n et un petit p et un échantillon au second degré présentant un petit n et un grand p convient bien à la configuration supposée par l'approche par calage du modèle proposé.

Bibliographie

Chatterjee, N., Y.-H. Chen, P. Maas, et R. Carroll (2016), « Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-Level Information From External Big Data Sources », *Journal of the American Statistical Association*, 111, p. 107-117.

Chen, Y. H., et H. Chen (2000), « A Unified Approach to Regression Analysis Under Double-Sampling Designs », *Journal of the Royal Statistical Society: Series B*, 62, p. 449-460.

Deville, J.-C., et C.-E. Särndal (1992), « Calibration Estimators in Survey Sampling », *Journal of the American Statistical Association*, 87, p. 376-382.

Fuller, W. A. (2009), *Sampling Statistic*s, Hoboken, NJ: Wiley.

Hidiroglou, M. (2001), « L'échantillonnage double », *Techniques d'enquête*, 27, p. 143-154.

Imbens, G. W. (2002), « Generalized Method of Moments and Empirical Likelihood », *Journal of Business and Economic Statistics*, 20, p. 493-506.

Kim, J. K. (2010), « Estimation par calage en utilisant l'inclinaison exponentielle dans les enquêtes par sondage », *Techniques d'enquête*, 36, p. 145-155.

Kim, J. K., et J. N. K. Rao (2009), « A Unified Approach to Linearization Variance Estimation from Survey Data After Imputation for Item Nonresponse », *Biometrika*, 96, p. 917-932.

Kim, J. K., et Z. Wang (2019), « Sampling Techniques for Big Data Analysis in Finite Population Inference », *International Statistical Review*, 87, p. S177-S191.

Lohr, S. L. et T. E. Raghunathan (2017), « Combining Survey Data with Other Data Sources », *Statistical Science*, 32, p. 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, p. 27-48.

Owen, A. (1991), « Empirical Likelihood for Linear Models », *The Annals of Statistics*, 19, p. 1725-1747.

Qin, J. (2000), « Combining Parametric and Empirical Likelihoods », *Biometrika*, 87, p. 484-490.

Qin, J., et J. Lawless (1994), « Empirical Likelihood and General Estimating Equations », *The Annals of Statistics*, 22, p. 300-325.

Rivera-Rodriguez, C., D. Spiegelman, et S. Haneuse (2019), « On the Analysis of Two-phase Designs in Cluster-correlated Data Settings », *Statistics in Medicine*, 38, p. 4611-4624.

Robins, J. M., A. Rotnitzky, et L. P. Zhao (1994), « Estimation of Regression Coefficients When Some Regressors Are Not Always Observed », *Journal of the American Statistical Association*, 89, p. 846-866.

Sheng, Y., Y. Sun, C.-Y. Huang, et M.-O. Kim (2021), « Synthesizing External Aggregate Information in the Presence of Population Heterogeneity: A Penalized Empirical Likelihood Approach », *Biometrics*. DOI : 10.1111/biom.13429.

Wang, C. Y., S. Wang, L.-P. Zhao, et S.-T. Ou (1997), « Weighted Semiparametric Estimation in Regression Analysis with Missing Covariate Data », *Journal of the American Statistical Association*, 92, p. 512-525.

Wang, H., et J. K. Kim (2021), « Propensity Score Estimation Using Density Ratio Model Under Item Nonresponse », arXiv Preprint, arXiv: 2104.13469.

Wang, L., M. L. Williams, Y. Chen, et J. Chen (2020), « Novel Two-Phase Sampling Designs for Studying Binary Outcomes », *Biometrics*, 76, p. 210-223.

Wu, C., et J. Rao (2006), « Pseudo Empirical Likelihood Ratio Confidence Intervals for Complex Surveys », *Revue canadienne de statistique*, 34, p. 359-375.

Wu, C., et R. R. Sitter (2001), « A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data », *Journal of the American Statistical Association*, 96, p. 185-193.

Xu, M., et J. Shao (2020), « Meta-Analysis of Independent Datasets Using Constrained Generalised Method of Moments », *Statistical Theory and Related Fields*, 4, p. 109-116.

Yang, S., et J. K. Kim (2020), « Statistical Data Integration in Survey Sampling: A review », *Japanese Journal of Statistics and Data Science*, 3, p. 625-650.

Yang, S., D. Zheng, et X. Wang (2020), « Elastic Integrated Analysis of Randomized Trial and Real-World Data for Treatment Heterogeneity Estimation », arXiv Preprint, arXiv:2005.10579v2.

Yuan, K.-H. et R. I. Jennrich (1998), « Asymptotics of Estimating Equations Under Natural Conditions », *Journal of Multivariate Analysis*, 65, p. 245-260.

Zhang, H., L. Deng, W. Wheeler, J. Qin, et K. Yu (2021), « Integrative Analysis Of Multiple Case-Control Studies », *Biometrics*. https://doi.org/10.1111/biom.13461.

Zhao, P., M. Ghosh, J. Rao, et C. Wu (2020), « Bayesian Empirical Likelihood Inference with Complex Survey Data », *Journal of the Royal Statistical Society: Series B*, 82, p. 155-174.

Zubizarreta, J. R. (2015), « Stable Weights That Balance Covariates for Estimation with Incomplete Outcome Data », *Journal of the American Statistical Association*, 110, p. 910-922.