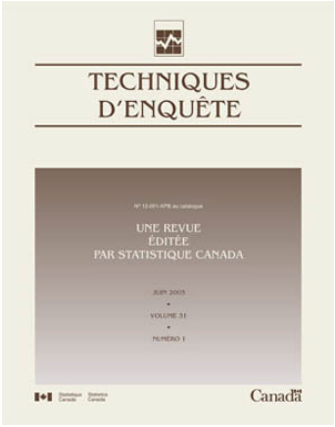
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Commentaires à propos de l'article « Contrôle de la divulgation statistique et avancées dans la protection officielle des renseignements : à la mémoire de Chris Skinner » : Note sur le lissage des poids dans l'échantillonnage



par Jae Kwang Kim et HaiYing Wang

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.

Techniques d'enquête, juin 2023
Vol. 49, nº 1, pp. 33-40
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article « Contrôle de la divulgation statistique et avancées dans la protection officielle des renseignements : à la mémoire de Chris Skinner » : Note sur le lissage des poids dans l'échantillonnage

Jae Kwang Kim et HaiYing Wang[1]

Résumé

Le lissage des poids est une technique utile pour améliorer l'efficacité des estimateurs fondés sur le plan exposés au risque de biais en raison d'une spécification erronée du modèle. Dans le prolongement du travail de Kim et Skinner (2013), nous proposons d'employer le lissage des poids pour construire la vraisemblance conditionnelle pour une inférence analytique efficace dans le cadre d'un échantillonnage informatif. La distribution bêta prime peut être utilisée pour construire un modèle de paramètres pour les poids dans l'échantillon. Un test du score est développé pour tester les erreurs de spécifications dans le modèle de pondération. Un estimateur de prétest s'appuyant sur le test du score peut être élaboré naturellement. L'estimateur de prétest est presque exempt de biais et peut être plus efficace que l'estimateur fondé sur le plan lorsque le modèle de pondération est correctement spécifié ou que les poids d'origine sont très variables. Une étude par simulation limitée est présentée pour étudier le rendement des méthodes proposées.

Mots-clés : Estimation de prétest; inférence analytique; méthode du maximum de vraisemblance conditionnelle; test du score.

1. Introduction

Supposons que la population finie de (x_i, y_i) est une réalisation indépendante et identiquement distribuée du modèle de superpopulation ayant une densité de $f(y \mid x; \theta)\, g(x)$, où θ est le paramètre d'intérêt et où la densité marginale de $g(.)$ n'est aucunement spécifié. À partir de la population finie, nous obtenons un échantillon probabiliste A ayant une probabilité d'inclusion de premier ordre connue de π_i. Nous observons (x_i, y_i) dans l'échantillon. Nous souhaitons estimer le paramètre du modèle θ à partir de l'échantillon complexe, qui est le problème principal dans le domaine de l'inférence analytique dans l'échantillonnage. Consulter Korn et Graubard (1999) et Fuller (2009, chapitre 6) pour obtenir des aperçus complets de l'inférence analytique dans l'échantillonnage.

Pour réaliser une estimation efficace, nous pouvons établir la fonction de vraisemblance conditionnelle à partir de l'échantillon comme suit :

$$L_c(\theta) = \prod_{i \in A} \frac{f(y_i \mid x_i; \theta)\, \tilde{\pi}(x_i, y_i)}{\int f(y \mid x_i; \theta)\, \tilde{\pi}(x_i, y)\, d\mu(y)} \tag{1.1}$$

où

$$\tilde{\pi}(x, y) = E(\pi \mid x, y) \tag{1.2}$$

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, Iowa, 50011, É.-U. Courriel : jkim@iastate.edu; HaiYing Wang, Department of Statistics, University of Connecticut, Storrs, Connecticut, 06269, É.-U.

est la probabilité d'inclusion conditionnelle et $\mu(\cdot)$ est la mesure dominante. Consulter la section 8.2 de Kim et Shao (2021) pour obtenir des précisions sur la méthode du maximum de vraisemblance conditionnelle.

Pour calculer la probabilité d'inclusion conditionnelle dans l'équation (1.2), nous pouvons utiliser la formule de Pfeffermann et Sverchkov (1999) :

$$E(\pi \mid x, y) = \frac{1}{E_s(w \mid x, y)}, \tag{1.3}$$

où $w = \pi^{-1}$ et $E_s(\cdot)$ est l'espérance sous la distribution échantillonnale, soit la distribution conditionnelle étant donné l'échantillon.

La probabilité d'inclusion conditionnelle obtenue au moyen de l'équation (1.3) peut être utilisée pour calculer les poids lissés $\tilde{w}_i = \{\tilde{\pi}(x_i, y_i)\}^{-1}$. Le lissage des poids peut réduire la variabilité du poids d'échantillonnage $w_i = \pi_i^{-1}$ lors de l'estimation des paramètres et peut donc mener à une estimation plus efficace, comme l'ont expliqué Beaumont (2008) et Kim et Skinner (2013). Pour calculer l'espérance conditionnelle $E_s(w \mid x, y)$, nous devons élaborer un modèle de régression pour w, qui peut être appelé un modèle de pondération.

Dans le présent article, nous examinons certaines classes paramétriques particulières de modèles de pondération. Dans la section 2, nous présentons un modèle de pondération s'appuyant sur la distribution bêta prime. Dans la section 3, nous proposons un test du score pour la spécification correcte du modèle dans le modèle de pondération. Dans la section 4, nous présentons les résultats d'une étude par simulation limitée. Enfin, dans la section 5, nous concluons par notre mot de la fin.

2. Modèle de pondération

Étant donné que les poids d'échantillonnage satisfont l'expression $w_i \geq 1 (i = 1, \ldots, n)$, on suppose que w_i^{-1} suit le modèle de distribution bêta : Bêta$(m(x_i, y_i)\phi, \{1 - m(x_i, y_i)\}\phi)$. Ainsi, la fonction de densité satisfait l'expression

$$f(w^{-1} \mid x, y) \propto (w^{-1})^{m\phi - 1} (1 - w^{-1})^{(1-m)\phi - 1},$$

et l'espérance conditionnelle et la variance sont

$$E(w^{-1} \mid x, y) = m(x, y), \quad \text{et} \quad V(w^{-1} \mid x, y) = \frac{m(x, y)\{1 - m(x, y)\}}{1 + \phi},$$

respectivement, où ϕ constitue le paramètre de précision. Le modèle logistique est un exemple de fonction moyenne :

$$m(x, y; \beta) = \frac{\exp(\beta_0 + \beta_1 x + \beta_2 y)}{1 + \exp(\beta_0 + \beta_1 x + \beta_2 y)}. \tag{2.1}$$

Il s'agit essentiellement d'un modèle de régression bêta. Pour obtenir plus de précisions sur la régression bêta, consulter Ferrari et Cribari-Neto (2004).

Malheureusement, l'approche de la régression bêta ne peut pas être appliquée directement, car le modèle de régression ne se vérifie pas nécessairement dans l'échantillon en raison de l'échantillonnage informatif. Pour éviter ce problème, nous pouvons dériver la distribution des données échantillonnées. N'oublions pas que si $X \sim \text{Bêta}(\alpha, \beta)$ alors $1-X$ suit $\text{Bêta}(\beta, \alpha)$ et $(1-X)/X$ suit une distribution bêta prime $\text{Bêta}'(\beta, \alpha)$. Par conséquent, $o = w - 1$ suit $\text{Bêta}'\left(\{1-m(x_i, y_i)\}\phi, m(x, y)\phi\right)$, et la fonction de densité est exprimée comme suit :

$$f(o \mid x, y) \propto o^{(1-m)\phi - 1}(1+o)^{-\phi}.$$

Selon le théorème de Bayes et $w^{-1} = (1+o)^{-1}$, la distribution échantillonnée de o satisfait

$$f_s(o \mid x, y) \propto f(o \mid x, y)\, P(\delta = 1 \mid x, y, w) = o^{(1-m)\phi - 1}(1+o)^{-\phi - 1}, \tag{2.2}$$

ce qui entraîne $o \mid (x, y, \delta = 1) \sim \text{Bêta}'\left(\{1-m(x, y)\}\phi, m(x, y)\phi + 1\right)$. Nous obtenons ainsi :

$$E_s(w \mid x, y) = 1 + E_s(o \mid x, y) = \frac{1}{m(x, y; \beta)} \tag{2.3}$$

et

$$\begin{aligned} \text{Var}_s(w \mid x, y) &= \frac{1-m(x, y)}{m(x, y)} \frac{1}{m(x, y)\cdot\phi - 1} \\ &\cong \frac{1-m(x, y)}{m(x, y)} \frac{1}{m(x, y)\cdot\phi} \end{aligned}$$

pour une taille suffisamment grande de ϕ. Ainsi, nous obtenons la méthode suivante d'estimation de moments de ϕ :

$$\hat{\phi} = \frac{1}{n}\sum_{i \in A} \frac{\{w_i \cdot m(x_i, y_i; \beta) - 1\}^2}{1 - m(x_i, y_i; \beta)} \tag{2.4}$$

qui dépend du paramètre inconnu β.

Nous pouvons utiliser les paramètres du modèle estimés de la procédure d'estimation itérative suivante.

1. Calculer

   $$\hat{\phi}^{(0)} = \frac{1}{n}\sum_{i \in A} \frac{(w_i/\overline{w} - 1)^2}{1 - 1/\overline{w}}$$

   en tant qu'estimateur initial de ϕ, où $\overline{w} = n^{-1}\sum_{i \in S} w_i$.
2. Au moyen de $\hat{\phi}^{(t)}$, calculer $\hat{\beta}^{(t)}$ en trouvant le maximiseur de

   $$\ell_c\left(\beta \mid \hat{\phi}^{(t)}\right) = \sum_{i \in S} \log f_s\left(o_i \mid x_i, y_i; \beta, \hat{\phi}^{(t)}\right)$$

   par rapport à β, où

$$f_s\left(o \mid x, y; \beta, \phi\right) = \frac{\Gamma\left(\phi+1\right)}{\Gamma\left(\phi-m\phi\right)\Gamma\left(m\phi+1\right)} o^{\phi-m\phi-1}\left(1+o\right)^{-\phi-1},$$

et $m = m\left(x, y; \beta\right)$.

3. Calculer $\hat{\phi}^{(t+1)}$ en appliquant l'équation (2.4) selon $\beta = \hat{\beta}^{(t)}$. Faire une mise à jour itérative de $\hat{\phi}$ et de $\hat{\beta}$ jusqu'à la convergence.

3. Test du score pour la spécification du modèle de pondération

La méthode de lissage des poids de la section 2 est justifiée selon l'hypothèse que le modèle de pondération est correctement spécifié. En pratique, il peut être recommandé de tester la validité du modèle de pondération avant d'utiliser l'estimateur fondé sur le modèle. Dans la présente section, nous considérons une version du test du score pour la spécification du modèle.

Supposons que $\hat{\theta}_c$ soit le maximiseur de la fonction de vraisemblance conditionnelle dans l'équation (1.1). Supposons que $\hat{\theta}_d$ soit l'estimateur fondé sur le plan de θ, que l'on obtient en maximisant la fonction du pseudo log-vraisemblance

$$\ell_p(\theta) = \sum_{i\in A}\frac{1}{\pi_i}\log f\left(y_i \mid \mathbf{x}_i; \theta\right). \tag{3.1}$$

L'estimateur par le pseudo maximum de vraisemblance a fait l'objet d'une discussion dans Chambers et Skinner (2003). Ainsi, nous pouvons mettre au point un test pour l'hypothèse nulle suivante :

$$E\left(\hat{\theta}_d\right) = E\left(\hat{\theta}_c\right). \tag{3.2}$$

Cependant, développer une statistique de test de type Wald pour l'hypothèse nulle de l'équation (3.2) peut être fastidieux, car la matrice de variance-covariance de $\hat{\theta}_d - \hat{\theta}_c$ doit être estimée.

Au lieu de tester l'équation (3.2), nous pouvons envisager de tester l'hypothèse nulle

$$H_0 : E\left\{\hat{S}_c\left(\theta_0\right)\right\} = 0, \tag{3.3}$$

où θ_0 est la vraie valeur du paramètre et $\hat{S}_c(\theta) = n^{-1}\partial \log L_c(\theta)/\partial\theta$ est la fonction de score obtenue à partir de la log-vraisemblance conditionnelle dans l'équation (1.1). Démonstration :

$$\hat{S}_c(\theta) = \frac{1}{n}\sum_{i\in A}\left[S\left(\theta; x_i, y_i\right) - E_s\left\{S\left(\theta; x_i, Y\right) \mid x_i\right\}\right],$$

où $S\left(\theta; x, y\right) = \partial \log f\left(y \mid x; \theta\right)/\partial\theta$ et

$$E_s\left\{S\left(\theta; x, Y\right) \mid x\right\} = \frac{\int S\left(\theta; x, y\right)\tilde{\pi}\left(x, y\right) f\left(y \mid x; \theta\right) dy}{\int \tilde{\pi}\left(x, y\right) f\left(y \mid x; \theta\right) dy}.$$

Dans certaines conditions de régularité (Binder, 1983), nous pouvons établir que

$$\sqrt{n}\left[\hat{S}_c(\theta) - E\left\{\hat{S}_c(\theta)\right\}\right] \xrightarrow{\mathcal{L}} N\left[0, \mathcal{I}_c(\theta)\right], \tag{3.4}$$

car $n \to \infty$, où $\overset{\mathcal{L}}{\to}$ désigne la convergence en distribution et

$$\begin{aligned} \mathcal{I}_c(\theta) &= -E\left\{\frac{\partial}{\partial \theta'} S_c(\theta)\right\} \\ &= n^{-1}\sum_{i=1}^{n}\left[E\left\{S_i S_i' \tilde{\pi}_i \mid \mathbf{x}_i;\theta\right\} - \frac{\left\{E\left(S_i \tilde{\pi}_i \mid \mathbf{x}_i;\theta\right)\right\}^{\otimes 2}}{E\left(\tilde{\pi}_i \mid \mathbf{x}_i;\theta\right)}\right]. \end{aligned} \tag{3.5}$$

La statistique de test proposée est

$$T\left(\hat{\theta}_d\right) = n\hat{S}_c\left(\hat{\theta}_d\right)' \left\{\mathcal{I}_c\left(\hat{\theta}_d\right)\right\}^{-1} \hat{S}_c\left(\hat{\theta}_d\right)$$

où $\hat{\theta}_d$ est l'estimateur par le pseudo maximum de vraisemblance de θ_0. Prenons note que

$$T\left(\hat{\theta}_d\right) = T\left(\theta_0\right) + o_p(1),$$

car $\hat{\theta}_d = \theta_0 + o_p(1)$, que le modèle de pondération soit vérifiable ou non. Selon l'hypothèse nulle dans l'équation (3.3), nous pouvons établir, au moyen de l'équation (3.4), que T converge vers la distribution $\chi^2(q)$, où $q = dim(\theta)$. Si l'hypothèse nulle est rejetée, cela signifie que $\tilde{\pi}(x, y)$ est incorrectement spécifiée dans la construction de la vraisemblance conditionnelle dans l'équation (1.1). Autrement, nous pouvons utiliser sans risque l'estimateur par le maximum de vraisemblance conditionnelle.

À proprement parler, la matrice d'information dans l'équation (3.5) ne tient pas compte de l'incertitude de $\hat{\beta}$ dans $\tilde{\pi}_i = \tilde{\pi}\left(x_i, y_i; \hat{\beta}\right)$. Pour intégrer l'incertitude dans $\hat{\beta}$, nous pouvons envisager une autre matrice d'information pour β. Le fait de ne pas tenir compte de l'incertitude dans $\hat{\beta}$ mènera à la surestimation de la variance et à un test prudent. Consulter l'étude par simulation dans la section suivante.

4. Étude par simulation

Pour mettre à l'épreuve notre théorie, nous réalisons une étude par simulation limitée. Dans la simulation, nous générons une population finie de taille $N = 10\,000$ et utilisons l'échantillonnage de Poisson pour sélectionner un échantillon de taille attendue de $n = 1\,000$. Nous répétons cette procédure indépendamment $B = 1\,000$ fois.

Dans chaque échantillon de Monte Carlo, nous générons $\left(x_i, y_i, \pi_i\right)$ pour $i = 1, \ldots, N$ où $x_i \sim (0, 2)$, $y_i = \theta_0 + \theta_1 x_i + e_i$, $(\theta_0, \theta_1) = (0{,}5; 0{,}5)$, $e_i \sim N\left(0; 0{,}5^2\right)$ et $\pi_i \mid x_i, y_i \sim \text{Bêta}\left(m\left(x_i, y_i\right)\phi, \left\{1 - m\left(x_i, y_i\right)\right\}\phi\right)$, où

$$m(x, y; \beta) = \frac{\exp\left(\beta_0 + \beta_1 x + \beta_2 y\right)}{1 + \exp\left(\beta_0 + \beta_1 x + \beta_2 y\right)} \tag{4.1}$$

selon $\beta_1 = 1$, $\beta_2 = 1$ et β_0 constituant différentes valeurs pour différents scénarios afin d'assurer que $n = 1\,000$. Dans l'étude par simulation, nous avons utilisé deux valeurs différentes de ϕ, $\phi = 100$ par rapport à $\phi = 1\,000$. La distribution des poids est moins asymétrique pour $\phi = 1\,000$.

Nous avons quatre plans d'échantillonnage différents :

Scénario 1. $\phi = 100;$ le modèle de pondération est correctement spécifié.

Scénario 2. $\phi = 100;$ la tranche inférieure de 30 % de π_i est multipliée par 0,25, c'est-à-dire que la tranche supérieure de 30 % de w_i dans l'ensemble de données complet est multipliée par 4. Par conséquent, le modèle de pondération de l'équation (4.1) n'est pas correctement spécifié.

Scénario 3. $\phi = 1\,000;$ le modèle de pondération est correctement spécifié.

Scénario 4. $\phi = 1\,000;$ la tranche inférieure de 30 % de π_i est multipliée par 4, c'est-à-dire que la tranche supérieure de 30 % de w_i dans l'ensemble de données complet est multipliée par 0,25. Par conséquent, le modèle de pondération de l'équation (4.1) n'est pas correctement spécifié.

Nous souhaitons estimer θ_0 et θ_1. Les trois estimateurs suivants sont pris en compte.

1. EPMV : l'estimateur par le pseudo maximum de vraisemblance $\hat{\theta}_d$ maximisant l'équation (3.1).
2. EMVC : l'estimateur par le maximum de vraisemblance conditionnelle $\hat{\theta}_c$ maximisant l'équation (1.1) au moyen de $\tilde{\pi}(x, y) = \{\tilde{w}(x, y)\}^{-1}$ et $\tilde{w}(x, y)$ constitue le poids lissé selon le modèle de pondération spécifié. Pour éviter les problèmes d'ordre numérique, nous estimons σ^2 selon une approche fondée sur le plan.
3. Prétest : l'estimateur de prétest obtenu grâce au test du score de la section 3. Plus précisément, l'estimateur de prétest $\hat{\theta}_{\text{pré}}$ pour lequel $\alpha = 0{,}05$ est défini comme

$$\hat{\theta}_{\text{pré}} = \begin{cases} \hat{\theta}_d & \text{si } T(\hat{\theta}) > q_{0,95}(\chi_2^2) \\ \hat{\theta}_c & \text{sinon,} \end{cases}$$

où $q_{0,95}(\chi_2^2)$ est le quantile 0,95 de la distribution $\chi^2(2)$.

Le tableau 4.1 présente les biais, les erreurs-types et la racine des erreurs quadratiques moyennes (REQM) des trois estimateurs à l'aide des échantillons de Monte Carlo. Les résultats de la simulation peuvent se résumer comme suit :

1. L'EPMV est presque sans biais pour tous les scénarios, mais il est moins efficace que les autres méthodes des scénarios 1 et 3, où le modèle de pondération est correctement spécifié.
2. L'EMVC est l'estimateur le plus efficace, mais il présente des biais importants lorsque le modèle de pondération n'est pas correctement spécifié. Les gains d'efficacité sont supérieurs pour une valeur ϕ plus petite, puisque la distribution des w_i est plus asymétrique et que l'avantage du lissage des poids est plus important.
3. L'estimateur de prétest est presque exempt de biais pour tous les scénarios et peut être plus efficace que l'EPMV lorsque le modèle de pondération est correctement spécifié (scénarios 1 et 3) ou que les poids d'origine sont très variables (scénario 2).

Tableau 4.1
Biais de Monte Carlo, erreurs-types (ET) de Monte Carlo et racine des erreurs quadratiques moyennes (REQM) de Monte Carlo des trois estimateurs en fonction des échantillons de Monte Carlo

		θ_0			θ_1		
Scénario	**Méthode**	**ET**	**Biais**	**REQM**	**ET**	**Biais**	**REQM**
1	EPMV	0,0768	-0,001	0,0768	0,0799	0,001	0,0800
	EMVC	0,0608	-0,001	0,0608	0,0425	0,001	0,0425
	Prétest	0,0701	0,006	0,0704	0,0672	-0,004	0,0673
2	EPMV	0,1198	-0,000	0,1198	0,1182	0,008	0,1185
	EMVC	0,0750	0,020	0,0777	0,0375	0,066	0,0764
	Prétest	0,1198	0,001	0,1198	0,1179	0,008	0,1182
3	EPMV	0,0651	0,000	0,0651	0,0645	0,000	0,0645
	EMVC	0,0525	0,002	0,0526	0,0413	-0,002	0,0413
	Prétest	0,0561	0,003	0,0563	0,0499	-0,003	0,0500
4	EPMV	0,0455	0,001	0,0456	0,0432	0,000	0,0432
	EMVC	0,0472	0,053	0,0713	0,0432	-0,127	0,1345
	Prétest	0,0456	0,001	0,0456	0,0433	0,000	0,0433

Note : EPMV signifie estimateur par le pseudo maximum de vraisemblance; EMVC signifie estimateur par le maximum de vraisemblance conditionnelle.

Les taux de rejet pour le test du score sont de 0,119, de 0,952, de 0,051 et de 0,997 pour les quatre scénarios, respectivement, où le niveau de signification est de $\alpha = 0{,}05$. Le taux de rejet élevé de 0,119 dans le scénario 1 est attribuable à l'effet de ne pas tenir compte de l'incertitude lors du lissage des poids. L'effet de ne pas tenir compte de l'incertitude lors du lissage des poids est négligeable dans le scénario 3, puisque l'effet du lissage des poids est moins important lorsque ϕ est grand. Le taux de rejet plus élevé indique que le test du score est plus conservateur lorsque l'on choisit l'EMVC en employant $\tilde{w}_i$ par rapport à l'EPMV.

5. Mot de la fin

Le présent article est dédié à la mémoire du professeur Chris Skinner. Le premier auteur a collaboré à divers projets avec Chris Skinner, et les premiers résultats de leurs recherches ont été publiés dans Kim et Skinner (2013). Lorsque Jae Kwang Kim a rendu visite à Chris Skinner à Southampton au cours de l'été 2011, ils ont d'abord consacré leurs efforts à l'inférence analytique selon l'échantillonnage informatif en se penchant sur les travaux de Pfeffermann et Sverchkov (1999), mais ils n'ont pas établi de lien avec le lissage des poids à ce moment-là. Ils se sont plutôt concentrés principalement sur la méthode de lissage des poids. Environ dix ans plus tard, nous présentons une méthode reliant le lissage des poids au cadre de vraisemblance.

Le lissage des poids est potentiellement utile, mais la correcte spécification du modèle est requise. L'estimateur de prétest s'appuyant sur le test du score de la section 3 peut être utilisé en pratique, car il compromet l'efficacité du lissage des poids et la robustesse de l'estimation fondée sur le plan. La façon d'estimer la variance de l'estimateur de prétest n'est pas examinée dans le présent article et sera étudiée à l'avenir.

Remerciements

Les auteurs remercient le rédacteur en chef et la rédactrice adjointe, Cynthia Bocci, pour leurs commentaires constructifs. Les recherches du premier auteur ont été appuyées par une subvention de l'Iowa Agriculture and Home Economics Experiment Station, à Ames, en Iowa. Les recherches du deuxième auteur ont été appuyées par une subvention de la Fondation nationale des sciences (CCF 2105571) et une subvention du College of Liberal Arts and Sciences de l'Université du Connecticut (Research Funding in Academic Themes).

Bibliographie

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95, 3, 539-553.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Chambers, R.L., et Skinner, C.J. (2003). *Analysis of Survey Data*. New York: John Wiley & Sons, Inc.

Ferrari, S.L.P., et Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31, 407-419.

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

Kim, J.K., et Shao, J. (2021). *Statistical Methods for Handling Incomplete Data*. CRC press, 2nd edition.

Kim, J.K., et Skinner, C.J. (2013). Weighting in survey analysis under informative sampling. *Biometrika*, 100, 358-398.

Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

Pfeffermann, D., et Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhyā, Series B*, 61, 166-186.