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

Techniques d’enquête

Utilisation d’un autre estimateur jackknife de la variance aux fins du calage des poids pour tenir compte de la non-réponse totale dans une enquête complexe

par Phillip S. Kott et Dan Liao

Date de diffusion : le 6 janvier 2022

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

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

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, décembre 2021
Vol. 47, nº 2, pp. 377-387
Statistique Canada, nº 12-001-X au catalogue

Utilisation d'un autre estimateur jackknife de la variance aux fins du calage des poids pour tenir compte de la non-réponse totale dans une enquête complexe

Phillip S. Kott et Dan Liao[1]

Résumé

La pondération par calage est un moyen statistiquement efficace de traiter la non-réponse totale. En supposant que le modèle (ou la sortie) de la réponse justifiant l'ajustement du poids de calage est exact, il est souvent possible de mesurer la variance des estimations de façon asymptotique et sans biais. Une des manières d'estimer la variance consiste à créer des poids de rééchantillonnage jackknife. Cependant, il arrive que la méthode classique de calcul des poids de rééchantillonnage jackknife pour les poids d'analyse calés échoue. Dans ce cas, il existe généralement une autre méthode de calcul des poids de rééchantillonnage jackknife. Cette méthode est décrite ici et appliquée à un exemple simple.

Mots-clés : Poids d'analyse; estimateur de la variance obtenu par linéarisation; estimateur jackknife de la variance avec suppression d'une UPE; poids de rééchantillonnage; asymptotiquement sans biais; modèle de réponse logistique borné.

1. Introduction

La pondération par calage est une méthode d'ajustement des poids dans la théorie de l'échantillonnage probabiliste consistant à forcer la somme pondérée de chaque variable d'un ensemble de variables d'enquête à égaler une cible déterminée. Quand cela se produit, on dit que les poids d'analyse satisfont à l'équation de calage. Plusieurs raisons conduisent à caler les poids d'analyse. La raison sur laquelle nous nous pencherons dans l'article est l'élimination du biais de sélection potentiel découlant de la non-réponse totale.

Dans la littérature sur l'échantillonnage d'enquête, il est courant de soutenir que l'ajustement du poids de calage d'un répondant estime implicitement l'inverse de sa probabilité de réponse (voir, par exemple, la section 5.1 de Fuller, 2009). Kott et Liao (2012) montrent que l'utilisation de la pondération par calage pour tenir compte de la non-réponse totale peut fournir une *double protection* contre le biais de non-réponse quand on estime un total de population. Cela signifie que si un modèle de résultats linéaire ou un modèle de sélection implicite se vérifie, l'estimateur obtenu est asymptotiquement sans biais dans un certain sens. Ils décrivent ensuite un estimateur de la variance obtenu par linéarisation pour un total estimé basé sur un échantillon stratifié à plusieurs degrés (ou à un degré) avec des poids d'analyse ajustés par calage.

Le bref traitement aux sections 2 et 3 de la pondération par calage pour la non-réponse et de l'estimation de la variance par linéarisation pour un estimateur calé d'un total de population est expliqué plus en détail dans Kott et Liao. On y trouve des démonstrations des diverses affirmations énoncées dans ces sections. Ils établissent la théorie étayant l'estimation de la variance par la méthode du jackknife.

1. Phillip S. Kott, statisticien principal de recherche, RTI International, Rockville, Maryland 20852, États-Unis. Courriel : pkott@rti.org; Dan Liao, statisticien de recherche, RTI International, Rockville, Maryland 20852, États-Unis.

En supposant un échantillon probabiliste stratifié à plusieurs degrés, un estimateur jackknife classique de la variance (avec suppression d'une UPE) crée des ensembles de poids de rééchantillonnage, un ensemble correspondant à chaque unité primaire d'échantillonnage (UPE) sélectionnée. Une UPE sélectionnée est supprimée à la fois et les poids de rééchantillonnage de ses éléments sous-échantillonnés sont fixés à zéro. Pour compenser, les poids de probabilité de rééchantillonnage des éléments restants dans la même strate que l'UPE abandonnée sont augmentés du facteur $n_h / (n_h - 1)$, où n_h est le nombre initial d'UPE sélectionnées à partir de la strate. Les poids de probabilité de rééchantillonnage sont calés d'une manière analogue aux poids d'analyse initiaux. La section 4 décrit ce jackknife et montre sa quasi-égalité avec l'estimateur de la variance obtenu par linéarisation presque sans biais pour un total de population.

L'avantage d'un jackknife avec suppression d'une UPE par rapport à un estimateur de la variance obtenu par linéarisation est qu'une fois les poids de rééchantillonnage calculés, l'estimation de la variance de la fonction lisse des totaux estimés (comme un coefficient de régression) est simple. Krewski et Rao (1981) proposent un traitement rigoureux du jackknife avec suppression d'une UPE et de ses propriétés.

Parfois, il n'y a pas de solution à l'équation de calage quand on commence par un ensemble de poids de probabilité de rééchantillonnage. La principale contribution du présent article se trouve dans le reste de la section 4, qui décrit et justifie une autre méthode de construction de poids de rééchantillonnage jackknife qui peut habituellement surmonter ce problème. Cette méthode a été introduite par Kott (2006) à d'autres fins.

La section 6 utilise une fonction d'ajustement des poids décrite à la section 5 pour illustrer la mise en œuvre de cette méthode. Elle compare ensuite favorablement les résultats de la méthode à ceux de deux méthodes concurrentes courantes. La section 7 traite d'une variante de l'autre méthode jackknife.

2. Pondération par calage

Supposons que nous avons un échantillon tiré aléatoirement S d'une population finie U. En l'absence de non-réponse (ainsi que d'erreur de couverture et d'erreur de mesure), la pondération par calage crée un ensemble de poids d'analyse, $\{w_k \mid k \in S\}$, qui ne dépend pas des valeurs d'enquête d'intérêt qui

1. sont proches des poids de probabilité inverse initiaux, $d_k = 1/\pi_k$ où π_k est la probabilité de sélection du k^{e} élément sélectionné;
2. satisfont à un ensemble d'équations de calage linéaires, une pour chaque composante de $\mathbf{z}_k$, un vecteur de variables auxiliaires avec des totaux de population connus :

$$\sum_{k \in S} w_k \mathbf{z}_k = \sum_{k \in U} \mathbf{z}_k$$

« Proches » signifie qu'à mesure que l'échantillon croît arbitrairement, la différence entre w_k et d_k disparaît dans la probabilité. Voir un traitement plus formel de la structure asymptotique supposée dans Isaki et Fuller (1982).

La plupart des enquêtes connaissent une non-réponse totale qui échappe au contrôle des statisticiens. On est obligé de supposer, explicitement ou implicitement, qu'il existe un type de modèle d'ajustement

pour tenir compte de la non-réponse. Un modèle de résultat (aussi appelé « modèle de prédiction ») sur une variable d'enquête d'intérêt suppose habituellement que le mécanisme de réponse/non-réponse, comme le plan de sondage, est ignorable. Un modèle de réponse suppose que le mécanisme de réponse se comporte comme une phase du sous-échantillonnage de Poisson (c'est-à-dire indépendamment). La double protection signifie que si le modèle de prédiction *ou* de réponse est spécifié correctement, l'estimateur sera presque (c'est-à-dire asymptotiquement) sans biais dans un certain sens. Ici, nous supposerons un modèle de réponse correctement spécifié.

Soit R le sous-ensemble de S contenant les répondants de l'enquête (par souci de simplicité, nous ne tenons pas compte de la possibilité de non-réponse partielle). L'échantillon des répondants peut être calé selon la population entière U :

$$\sum_{k\in R} w_k \mathbf{z}_k = \sum_{k\in U} \mathbf{z}_k \tag{2.1}$$

ou selon l'échantillon initial S :

$$\sum_{k\in R} w_k \mathbf{z}_k = \sum_{k\in S} d_k \mathbf{z}_k. \tag{2.2}$$

Nous supposons un modèle de réponse dans lequel la probabilité de réponse pour chaque $k\in U$, p_k, est une fonction indépendante ayant la forme $p_k = p(\boldsymbol{\gamma}^T\mathbf{x}_k)$, où $p(.)$ est une fonction monotonique lisse, et à la fois le vecteur connu $\mathbf{x}_k$ et le vecteur de paramètre inconnu $\boldsymbol{\gamma}$ ont le même nombre de composantes que $\mathbf{z}_k$. Dans la majorité de la littérature $\mathbf{x}_k$ est égal à $\mathbf{z}_k$, mais la plus grande partie de la théorie reste vraie si ce n'est pas le cas.

Supposons un vecteur $\mathbf{g}$ tel que l'insertion de $w_k = d_k / p(\mathbf{g}^T\mathbf{x}_k)$ résout soit l'équation de calage dans (2.1) soit dans (2.2), alors $\mathbf{g}$ est un estimateur convergent pour $\boldsymbol{\gamma}$. Kott et Liao (2017) décrivent ce qu'il faut faire si le nombre de composantes dans $\mathbf{x}_k$ est plus petit que leur nombre dans $\mathbf{z}_k$.

La fonction $f(\mathbf{g}^T\mathbf{x}_k) = 1/p(\mathbf{g}^T\mathbf{x}_k)$ est appelée *fonction d'ajustement des poids*. Le théorème de la valeur moyenne nous dit que sous des conditions faibles $f(\mathbf{g}^T\mathbf{x}_k) - f(\boldsymbol{\gamma}^T\mathbf{x}_k) \approx f'(\mathbf{g}^T\mathbf{x}_k)(\mathbf{g}-\boldsymbol{\gamma})^T\mathbf{x}_k$. Par conséquent, à mesure que l'échantillon des répondants croît de façon arbitraire $f(\mathbf{g}^T\mathbf{x}_k)$ converge vers $f(\boldsymbol{\gamma}^T\mathbf{x}_k) = 1/p_k$ et $\mathbf{g}$ converge vers $\boldsymbol{\gamma}$.

3. Estimation de la variance par linéarisation

Lors du calage de l'échantillon des répondants selon l'échantillon complet avec (2.2), l'estimateur par calage pour un total de population, $t = \sum_R w_k y_k$ peut être exprimé comme suit :

$$\begin{aligned}
t &= \sum_{k\in S} d_k \mathbf{z}_k^T\mathbf{b} + \sum_{k\in R} d_k f(\mathbf{g}^T\mathbf{x}_k)(y_k - \mathbf{z}_k^T\mathbf{b}) \\
&\approx \sum_{k\in S} d_k \mathbf{z}_k^T\mathbf{b} + \sum_{k\in R} d_k f(\boldsymbol{\gamma}^T\mathbf{x}_k)(y_k - \mathbf{z}_k^T\mathbf{b}) + \sum_{k\in R} d_k f'(\mathbf{g}^T\mathbf{x}_k)\left[(\mathbf{g}-\boldsymbol{\gamma})^T\mathbf{x}_k\right](y_k - \mathbf{z}_k^T\mathbf{b}) \\
&= \sum_{k\in S} d_k \mathbf{z}_k^T\mathbf{b} + \sum_{k\in R} d_k f(\boldsymbol{\gamma}^T\mathbf{x}_k)(y_k - \mathbf{z}_k^T\mathbf{b}) + (\mathbf{g}-\boldsymbol{\gamma})^T \sum_{k\in R} d_k f'(\mathbf{g}^T\mathbf{x}_k)\mathbf{x}_k(y_k - \mathbf{z}_k^T\mathbf{b}) \\
&= \sum_{k\in S} d_k \mathbf{z}_k^T\mathbf{b} + \sum_{k\in R} d_k p_k^{-1}(y_k - \mathbf{z}_k^T\mathbf{b}),
\end{aligned} \tag{3.1}$$

où

$$\mathbf{b} = \left[\sum_{k\in R} d_k f'(\mathbf{g}^T\mathbf{x}_k)\,\mathbf{x}_k\mathbf{z}_k^T\right]^{-1}\sum_{k\in R} d_k f'(\mathbf{g}^T\mathbf{x}_k)\,\mathbf{x}_k y_k.$$

L'étape clé ici est que $\mathbf{b}$ a été défini de telle sorte que $\sum_R d_k f_k'(\mathbf{g}^T\mathbf{x}_k)\,\mathbf{x}_k(y_k - \mathbf{z}_k^T\mathbf{b}) = 0.$ Notons que $f'(.)$ dans $\mathbf{b}$ est la *dérivée* de la fonction d'ajustement de la pondération.

Soit $\mathbf{b}^*$ la limite de probabilité de $\mathbf{b}$ quand l'échantillon de répondants (d'UPE) croît arbitrairement. La variance de t selon le plan initial et le modèle de sélection équivaut presque à la variance de $\sum_S d_k q_k^*$, où

$$q_k^* = \mathbf{z}_k^T\mathbf{b}^* + p_k^{-1}(y_k - \mathbf{z}_k^T\mathbf{b}^*)\,I_k,$$

et $I_k = 1$ quand k est un répondant d'unité et 0 autrement.

Pour de nombreux plans, on peut calculer approximativement q_k^* en remplaçant $\mathbf{b}^*$ par $\mathbf{b}$ et p_k^{-1} par $f(\mathbf{g}^T\mathbf{x}_k)$, et la variance de $\sum_S d_k q_k$ est estimée selon le plan initial comme si les $q_k = \mathbf{z}_k^T\mathbf{b} + f(\mathbf{g}^T\mathbf{x}_k)(y_k - \mathbf{z}_k^T\mathbf{b})\,I_k$ étaient des constantes. Lors du calage de l'échantillon des répondants selon la population avec l'équation (2.1), $\sum_S d_k\mathbf{z}_k^T\mathbf{b}$ dans l'équation (3.1) est remplacé par $\sum_U \mathbf{z}_k^T\mathbf{b}$, ce qui ne contribue pas à la variance, donc $q_k = f(\mathbf{g}^T\mathbf{x}_k)(y_k - \mathbf{z}_k^T\mathbf{b})\,I_k$. D'une façon ou d'une autre, le remplacement de q_k^* par q_k a tendance à sous-estimer les variances avec des échantillons *finis* (le remplacement est asymptotiquement ignorable), car $e_k = (y_k - \mathbf{z}_k^T\mathbf{b})^2$ tend à être plus petit que $e_k^* = (y_k - \mathbf{z}_k^T\mathbf{b}^*)^2$.

Étant donné un échantillon probabiliste stratifié à plusieurs degrés avec n_h UPE échantillonnées parmi les H strates, soit S_{hj} le sous-échantillon d'éléments dans chaque UPE j dans la strate h. Un estimateur obtenu par linéarisation presque sans biais pour la variance de t est

$$v(t) = \sum_{h=1}^{H}\frac{n_h}{n_h - 1}\left[\sum_{j=1}^{n_h}\left(\sum_{k\in S_{hj}} d_k q_k\right)^2 - \frac{1}{n_h}\left(\sum_{a=1}^{n_h}\sum_{\kappa\in S_{ha}} d_\kappa q_\kappa\right)^2\right], \qquad (3.2)$$

où $q_k = \alpha\mathbf{z}_k^T\mathbf{b} + f(\mathbf{g}^T\mathbf{x}_k)\,e_k I_k$, et $\alpha = 1$ quand l'échantillon des répondants est calé par rapport à l'échantillon initial et 0 quand l'échantillon des répondants est calé par rapport à la population. Comme c'est souvent le cas en pratique et comme on le voit ici, l'équation (3.2) suppose qu'on perd peu en traitant la sélection d'UPE dans les strates comme si elle avait été tirée avec remise, ce qui évite la nécessité d'une correction de population finie.

4. Estimation de la variance par la méthode du jackknife

Soit S_{h+} tous les éléments échantillonnés dans la strate h. La réplique jackknife (avec suppression d'une UPE) hj^{e} classique pour un total estimé $t = \sum_R w_k y_k$ est

$$t^{(hj)} = \sum_{k \in R} w_k^{(hj)} y_k = \sum_{k \in S} d_k^{(hj)} f(g^{(hj)T} \mathbf{x}_k) I_k y_k, \tag{4.1}$$

où $d_k^{(hj)} = 0$ quand $k \in S_{hj}$

$d_k^{(hj)} = [n_h / (n_h - 1)] d_k$ quand $k \in S_{h+}$ mais $k \notin S_{hj}$

$d_k^{(hj)} = d_k$ autrement,

et $\mathbf{g}^{(hj)}$ résout l'équation de calage de rééchantillonnage :

$$\sum_{k \in R} d_k^{(hj)} f(\mathbf{g}^{(hj)T} \mathbf{x}_k) \mathbf{z}_k = \sum_{k \in U} \mathbf{z}_k \text{ quand l'échantillon des répondants est calé selon } U, \text{ ou}$$

$$\sum_{k \in R} d_k^{(hj)} f(\mathbf{g}^{(hj)T} \mathbf{x}_k) \mathbf{z}_k = \sum_{k \in S} d_k^{(hj)} \mathbf{z}_k \text{ quand l'échantillon des répondants est calé selon } S$$

pour chaque hj. Observons que $t^{(hj)}$ est une estimation du total de la population $\sum_U y_k$ sans suppression de l'UPE hj.

L'estimateur jackknifede la variance avec suppression d'une UPE pour t est

$$\text{var}_J(t) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (t^{(hj)} - t)^2. \tag{4.2}$$

Soit $e_k^* = (y_k - \mathbf{z}_k^T \mathbf{b}^*)$. Par conséquent,

$$\begin{aligned} t^{(hj)} - t &= \sum_{k \in S} \left\{ (d_k^{(hj)} - d_k) \alpha \mathbf{z}_k^T \mathbf{b}^* + \left[d_k^{(hj)} f(\mathbf{g}^{(hj)T} \mathbf{x}_k) - d_k f(\mathbf{g}^T \mathbf{x}_k) \right] I_k e_k^* \right\} \\ &\approx \sum_{k \in S} (d_k^{(hj)} - d_k) \left[\alpha \mathbf{z}_k^T \mathbf{b}^* + f(\mathbf{g}^T \mathbf{x}_k) I_k e_k^* \right]. \end{aligned}$$

Nous pouvons alors conclure que $\text{var}_J(t)$ est à peu près égal au jackknife avec suppression d'une UPE de $\sum_S d_k q_k^*$, où $q_k^* = \alpha \mathbf{z}_k^T \mathbf{b}^* + f(\mathbf{g}^T \mathbf{x}_k) I_k e_k^*$, d'après un échantillon stratifié à plusieurs degrés avec H strates et n_h UPE dans la strate h. Avec un peu d'algèbre, on montre que le jackknife avec suppression d'une UPE pour $\sum_S d_k q_k^*$ est égal à $\text{var}(t)$ dans l'équation (3.2) avec q_k^* remplaçant q_k parce que

$$\sum_{k \in S} (d_k^{(hj)} - d_k) q_k^* = \sum_{h=1}^{H} \sum_{k \in S_{h+}} (d_k^{(hj)} - d_k) q_k^*,$$

où

$$\begin{aligned} \sum_{k \in S_{h+}} (d_k^{(hj)} - d_k) q_k^* &= \sum_{k \in S_{h+}} \frac{1}{n_h - 1} d_k q^* - \frac{n_h}{n_h - 1} \sum_{k \in S_{hj}} d_k q^* \\ &= -\frac{n_h}{n_h - 1} \left(\sum_{k \in S_{hj}} d_k q^* - \frac{1}{n_h} \sum_{k \in S_{h+}} d_k q^* \right). \end{aligned}$$

Notons que la contribution à l'estimateur jackknife de la variance pour la hj^{e} réplique provient principalement de la hj^{e} UPE.

Observons que le petit biais à la baisse dans les échantillons finis causé par le remplacement de q_k par q_k^* dans $\text{var}(t)$ ne s'applique pas à $\text{var}_J(t)$ dans l'équation (4.2). Ce dernier peut avoir une légère tendance à être biaisé à la hausse dans les échantillons finis parce que $\mathbf{g}^{(hj)}$ et $\mathbf{g}$, bien qu'ils soient tous deux des estimateurs convergents de $\boldsymbol{\gamma}$, n'ont pas besoin d'être exactement égaux.

Un problème se pose parfois avec le calcul de l'estimateur jackknife de la variance $\text{var}_J(t)$ dans la pratique. Ce problème se produit quand $f(.)$ est tel que, bien qu'il y ait une valeur $\mathbf{g}$ satisfaisant l'équation de calage dans (2.1) ou (2.2), aucune valeur $\mathbf{g}^{(hj)}$ ne satisfait son analogue pour au moins une réplique jackknife hj. Quand cela se produit, on peut suivre une suggestion de Kott (2006) et calculer $w_k^{(hj)}$ dans l'équation (4.1) avec la solution suivante :

$$\tilde{w}_k^{(hj)} = d_k^{(hj)} f(g^T \mathbf{x}_k) + \left[\mathbf{c}^{(hj)} - \sum_{\kappa \in R} d_\kappa^{(hj)} f(g^T \mathbf{x}_\kappa) \mathbf{z}_\kappa^T \right] \left[\sum_{\kappa \in R} d_\kappa^{(hj)} f'(g^T \mathbf{x}_\kappa) \mathbf{x}_\kappa \mathbf{z}_\kappa^T \right]^{-1} d_k^{(hj)} f'(g^T \mathbf{x}_k) \mathbf{x}_k, \quad (4.3)$$

où $\mathbf{c}^{(hj)}$ est la cible de calage pour la hj^{e} réplique :

$$\mathbf{c}^{(hj)} = \sum_{\kappa \in U} \mathbf{z}_\kappa^T \text{ quand l'échantillon des répondants est calé selon } U,$$

et

$$\mathbf{c}^{(hj)} = \sum_{\kappa \in S} d_\kappa^{(hj)} \mathbf{z}_\kappa^T \text{ quand l'échantillon des répondants est calé selon } S.$$

Selon le plan de sondage $\sum_R \tilde{w}_k^{(hj)} \mathbf{z}_k = \mathbf{c}^{(hj)}$.

Si l'on suppose $e_k = y_k - \mathbf{z}_k^T \mathbf{b}$, on peut voir qu'avec $\tilde{t}^{(hj)} = \sum_R \tilde{w}_k^{(hj)} y_k$,

$$\begin{aligned} \tilde{t}^{(hj)} - t &= \sum_{k \in S} (d_k^{(hj)} - d_k) \left[\alpha \mathbf{z}_k^T \mathbf{b} + f(\mathbf{g}^T \mathbf{x}_k) I_k e_k \right] \\ &\approx \sum_{k \in S} (d_k^{(hj)} - d_k) \left[\alpha \mathbf{z}_k^T \mathbf{b}^* + f(\mathbf{g}^T \mathbf{x}_k) I_k e_k^* \right] \end{aligned}$$

de sorte que l'autre estimateur jackknife de la variance $\text{var}_{AJ}(t)$ calculé avec $\tilde{t}^{(hj)}$ au lieu de $t^{(hj)}$ est presque sans biais. Notons que la seule restriction possible sur le calcul de $\tilde{w}_k^{(hj)}$ est que $\sum_R d_k^{(hj)} f'(g^T \mathbf{x}_k) \mathbf{x}_k \mathbf{z}_k^T$ soit non singulier.

Observons que l'équation (4.3) peut être réécrite ainsi :

$$\tilde{w}_k^{(hj)} = \tilde{d}_k^{(hj)} \tilde{f}(\tilde{\mathbf{g}}^{(hj)T} \tilde{\mathbf{x}}_k) \quad (4.4)$$

où $\tilde{d}_k^{(hj)} = d_k^{(hj)} f(g^T \mathbf{x}_k)$,

$$\tilde{f}(\tilde{\mathbf{g}}^{(hj)T} \tilde{\mathbf{x}}_k) = 1 + \tilde{\mathbf{g}}^{(hj)T} \tilde{\mathbf{x}}_k,$$

$$\tilde{\mathbf{g}}^{(hj)T} = \left[\mathbf{c} - \sum_{\kappa \in R} \tilde{d}_{\kappa}^{(hj)} \mathbf{z}_{\kappa}^{T} \right] \left[\sum_{\kappa \in R} \tilde{d}_{\kappa}^{(hj)} \tilde{\mathbf{x}}_{\kappa} \mathbf{z}_{\kappa}^{T} \right]^{-1},$$

et

$$\tilde{\mathbf{x}}_k = \frac{f'(\mathbf{g}^T \mathbf{x}_k)}{f(\mathbf{g}^T \mathbf{x}_k)} \mathbf{x}_k.$$

Cette équation traite la hj^{e} réplique comme l'échantillon complet. La fonction d'ajustement des poids $\tilde{f}(.)$ est linéaire et $\tilde{f}_k$ n'est pas limité à des valeurs positives même si les f_k le sont. De plus, notons que même quand $\mathbf{x}_k = \mathbf{z}_k$, $\tilde{\mathbf{x}}_k$ n'est pas égal à $\mathbf{z}_k$, sauf si $f(.) = \exp(.)$.

5. Le modèle de réponse logistique (borné)

Jusqu'à maintenant, nous n'avons pas spéficié de fonction de réponse, $p(.) = 1/f(.)$. Envisageons maintenant un modèle de réponse logistique (ou logit) borné de forme :

$$p(\boldsymbol{\gamma}^T \mathbf{x}_k) = \frac{1 + \exp(\boldsymbol{\gamma}^T \mathbf{x}_k)/U}{L + \exp(\boldsymbol{\gamma}^T \mathbf{x}_k)}. \tag{5.1}$$

où $1 \le L < U$. Quand $L = 1$ et U est infini, il s'agit d'un modèle de réponse logistique standard, où la probabilité de réponse peut varier de 0 à 1, excluant les extrémités. Pour les valeurs finies de L et U, la probabilité de réponse bornée se situe entre $1/U$ et $1/L$. La valeur de la fonction d'ajustement varie donc de L à U. En pratique, L est généralement fixé à 1, tandis que U est souvent fixé aussi bas que l'échantillon le permet pour que l'équation de calage se vérifie.

Les procédures de calage en langage SUDAAN® (Research Triangle Institute, 2012), WTADJUST lorsque $\mathbf{x}_k = \mathbf{z}_k$ et WTADJX sinon, correspondent à une fonction équivalente d'ajustement des poids :

$$f(\mathbf{g}^T \mathbf{x}_k) = \frac{L + B \exp(A \mathbf{g}^T \mathbf{x}_k)}{1 + B \exp(A \mathbf{g}^T \mathbf{x}_k)/U}, \tag{5.2}$$

où $A = \frac{U-L}{(C-L)(U-C)}$, $B = U\frac{C-L}{U-C}$, et $L < C < U$.

Le choix de C permet de déterminer que $\mathbf{g}$ satisfait l'équation de calage, mais n'affecte pas la valeur de l'ajustement des poids en soi, $f_k = f(\mathbf{g}^T \mathbf{x}_k)$. Par conséquent, C peut être toute valeur entre L et U. Quand $L = 1$, $C = 2$ et U est infini, $\mathrm{A} = \mathrm{B} = 1$.

Un petit calcul révèle avec la fonction d'ajustement des poids dans l'équation (5.2) :

$$f_k' = f'(g^T \mathbf{x}_k) = \frac{(U - f_k)(f_k - L)}{(U - C)(C - L)},$$

qui est nécessaire pour calculer l'équation (4.3) ou (4.4). Le modèle exponentiel général dans les procédures de calage SUDAAN permet à L, C et U de varier d'un élément à l'autre, une souplesse difficile à interpréter dans la modélisation des réponses et qui n'est pas prise en compte ici.

Ce qui sera utile ici, bien que ce ne soit pas à des fins de modélisation, est la possibilité que L dans l'équation (5.2) soit 0 et que U soit infini. Lors de la résolution itérative d'une équation de calage pour $\mathbf{g}$ avec $f(\mathbf{g}^T\mathbf{x}_k) = \exp(\mathbf{g}^T\mathbf{x})$ au moyen de la méthode de Newton, les procédures de calage SUDAAN commencent par résoudre $\mathbf{g}_1$ dans l'équation de calage avec $f(\mathbf{g}_1^T\mathbf{x}_k) = 1 + \mathbf{g}_1^T\mathbf{x}_k$, ce qui est un résultat utile lors du calcul de poids de l'autre jackknife. (Les programmes fixent la première itération de l'ajustement des poids à $f_{k1} = \exp(\mathbf{g}_1^T\mathbf{x}_k)$, à partir de quoi $1 + \mathbf{g}_1^T\mathbf{x}_k$ se calcule facilement.)

6. Exemple de simulation

La population MU281 de municipalités dans Särndal, Swensson et Wretman (1992; les données de la version légèrement révisée se trouvent à l'adresse http://lib.stat.cmu.edu/datasets/mu284; une des municipalités a été accidentellement abandonnée dans cette analyse) a été augmenté d'un indicateur (RESP) pour déterminer si un élément (municipalité) répondrait s'il était échantillonné. Les probabilités de réponse par élément ont été générées à l'aide d'une fonction logistique de l'une des covariables de l'ensemble de données (le log de la population de 1975 de l'élément en milliers). La probabilité moyenne de réponse était d'environ 70 %.

Un échantillon aléatoire simple stratifié de 10 éléments toutes les 8 strates a été simulé 1 716 fois. Dans chaque échantillon simulé, les éléments avec RESP = 1 ont été traités comme des répondants, et l'échantillon des répondants a été calé selon l'échantillon complet au moyen de la fonction d'ajustement des poids de l'équation (5.2) avec une borne inférieure de 1 et une borne supérieure de 5. Dans le modèle de calage, les deux composantes de $\mathbf{x}_k$ étaient 1 et le log de la population de 1975 de l'élément en milliers; $\mathbf{z}_k$ a été réglé comme étant égal à $\mathbf{x}_k$. Dans 1 225 des 1 716 simulations, les échantillons des répondants ont été calés avec succès sur les deux composantes (c'est-à-dire qu'ils satisfaisaient l'équation de calage de 2.2) et ont produit des erreurs-types par linéarisation.

Les moyennes estimées (ratios de deux totaux estimés) et les erreurs-types (racines carrées des variances estimées) ont été calculées pour quatre variables :

P85	Population de 1985 (en milliers).
RM4T85	Recettes fiscales municipales de 1985 (en millions de couronnes).
ME84	Nombre d'employés municipaux en 1984.
REV84	Valeurs immobilières selon l'évaluation de 1984 (en millions de couronnes).

Bien que la procédure SUDAAN WTADJUST puisse calculer les erreurs-types lors de l'utilisation d'un jackknife avec suppression d'une UPE, elle ne peut pas le faire quand le calage d'une ou plusieurs répliques échoue. C'est pourquoi deux versions des erreurs-types du jackknife avec suppression d'une UPE classique ont été calculées au moyen d'une macro créée par les auteurs. Dans l'une d'elles, on utilise les poids « calés » imparfaits de la dernière itération pour les répliques ayant échoué. Dans l'autre, les

répliques dont le calage a échoué ont été abandonnées, et l'estimateur jackknife de la variance modifié suivant a été calculé :

$$\text{var}_J^*(t) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h^*} \sum_{j=1}^{n_h^*} (t^{(hj)} - t)^2, \tag{6.1}$$

où n_h^* est le nombre de répliques dans la strate h qui ont été calées avec succès. Cet estimateur jackknife de la variance révisé a été proposé par Rust (1985) quand les répliques sont abandonnées au hasard, ce qui n'est pas le cas ici. Le code SUDAAN pouvant être appelé par SAS (SAS Institute Inc., 2015) utilisé dans l'analyse pour une seule simulation est disponible sur demande auprès des auteurs.

Parmi les 1 225 échantillons analysables, dans 867 simulations toutes les répliques utilisant la méthode du jackknife avec suppression d'une UPE classique avaient un calage réussi, tandis que les 358 autres simulations avaient au moins une réplique dont le calage ne réussissait pas après 50 itérations (la valeur par défaut est de 10). Le tableau 6.1 présente la moyenne des résultats dans les deux situations. Quand aucune réplique classique n'échouait, les erreurs-types de la méthode jackknife classique et de substitution sont proches (en moyenne) et légèrement supérieures à celles produites par la linéarisation comme le prédit la théorie (notons que les deux versions de jackknife classique avec suppression d'une UPE sont identiques).

Tableau 6.1
Erreurs-types fondées sur les méthodes de jackknife lors du calage de la non-réponse avec un modèle logistique borné

	Variable	Moyenne estimée	Erreur-type de la méthode de linéarisation	Erreur-type de l'autre méthode de jackknife	Erreur-type de la méthode de jackknife classique comprenant les répliques ayant échoué	Erreurs-types de la méthode de jackknife classique avec abandon des répliques ayant échoué
867 simulations sans échec de calage pour les répliques de la méthode classique	P85	22,41	2,06	2,09	2,11	2,11
	RMT85	167,70	17,31	17,72	17,86	17,86
	ME84	1 215,87	124,67	127,27	128,15	128,15
	REV84	2 425,52	212,83	217,00	219,67	219,67
358 simulations avec au moins un échec de calage pour une réplique de la méthode classique	P85	22,78	2,24	2,31	2,95	2,04
	RMT85	170,48	18,93	19,47	24,39	16,60
	ME84	1 236,75	135,94	139,65	175,99	121,31
	REV84	2 451,95	239,27	239,18	296,77	208,45

Quand le calage d'au moins une réplique échoue pour le jackknife classique avec suppression d'une UPE, les erreurs-types de l'autre méthode jackknife sont de nouveau proches des erreurs-types de la méthode de linéarisation, même si le calage échoue dans 114 de ces 358 simulations en raison d'une

(quasi) singularité dans au moins une des répliques. Il est toutefois clair que l'inclusion des répliques ayant échoué surestime l'erreur-type et leur abandon la sous-estime nettement par rapport à la linéarisation. Il semblerait que l'autre estimateur jackknife de la variance produise l'ensemble de poids de rééchantillonnage le plus utile dans cette situation.

Le tableau 6.1 compare les erreurs-types des jackknifes concurrents aux erreurs-types fondées sur la linéarisation plutôt qu'aux erreurs-types empiriques parce que la correction de la population finie a été ignorée. De plus, l'ajustement du modèle de réponse logistique borné dans les simulations n'était pas le modèle de réponse non borné servant à générer les réponses.

7. Discussion

Il y a une petite chance (environ 1,5 % dans nos simulations) que l'équation (4.4) donne des poids de rééchantillonnage négatifs. Or les procédures prédéfinies de nombreux progiciels statistiques (comme SAS) ne peuvent pas traiter les poids négatifs. Par conséquent, il se peut qu'il faille calculer les totaux estimés calculés à partir des poids de rééchantillonnage sans l'aide d'une procédure prédéfinie.

Il n'est pas nécessaire d'avoir accès à SUDAAN pour calculer des poids jackknife autres pour les estimateurs par calage. Les routines *gencalib* dans le progiciel « Sampling » de R (Tillé et Matei, 2016) peuvent réaliser le calage non seulement sous un modèle de réponse logistique borné, mais aussi selon un calage linéaire. Bien qu'il y ait des macros SAS équivalentes de WTADJUST, à notre connaissance, il n'existe actuellement aucune macro de pondération par calage SAS accessible au public utilisable quand $\tilde{\mathbf{x}}_k$ dans l'ajustement des poids (équation 4.4) n'est pas égal à $\mathbf{z}_k$. Espérons que ce sera bientôt le cas.

Remerciements

Les auteurs tiennent à remercier les rédacteurs dont les suggestions ont aidé à améliorer la qualité de cette note.

Bibliographie

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

Isaki, C., et Fuller, W. (1982). Survey design under the regression super-population model. *Journal of the American Statistical Association*, 77, 89-96.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Technique d'enquête*, 32, 2, 149-160. Article accessible a l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006002/article/9547-fra.pdf.

Kott, P., et Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration weighting routine. *Survey Research Methods*, 6, 105-111.

Kott, P., et Liao, D. (2017). Calibration weighting for nonresponse that is not missing at random: Allowing more calibration than response-model variables. *Journal of Survey Statistics and Methodology*, 5, 159-174.

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

Research Triangle Institute (2012). *SUDAAN Language Manual, Release 11.0.* RTI International, Research Triangle Park, NC.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 1, 381-397.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

SAS Institute Inc. (2015). SAS/STAT® 14.1 User's Guide. SAS Institute Inc., Cary, NC.

Tillé, Y., et Matei, A. (2016). Package 'sampling' disponible en ligne sur http://cran.r-project.org/web/packages/sampling/sampling.pdf.