Traitement des données influentes dans le cas d'un sondage à deux phases avec une application au traitement de la non-réponse

Jean-François Beaumont, *Statistics Canada*
Cyril Favre Martinoz, Crest-Ensai
David Haziza, Université de Montréal, Crest-Ensai

6 novembre 2012

Sommaire

Sommaire

Unités influentes

- Généralement, les unités avec un poids élevé et/ou une valeur élevée
- Les unités influentes sont très présentes dans le cas des enquêtes entreprises, lorsqu'on observe des variables d'intérêt comme le chiffre d'affaire, qui sont très asymétriques et dont les queues de distribution sont lourdes
- Les unités influentes sont des unités légitimes, on considère que les erreurs de mesures ont été supprimées
- La plupart des estimateurs classiques (HT,Greg) sont sensibles à la présence d'unités influentes
- Enfin, l'impact de ces unités peut être limité en utilisant un plan adapté : par exemple, avec un sondage aléatoire simple stratifié avec une strate exhaustive

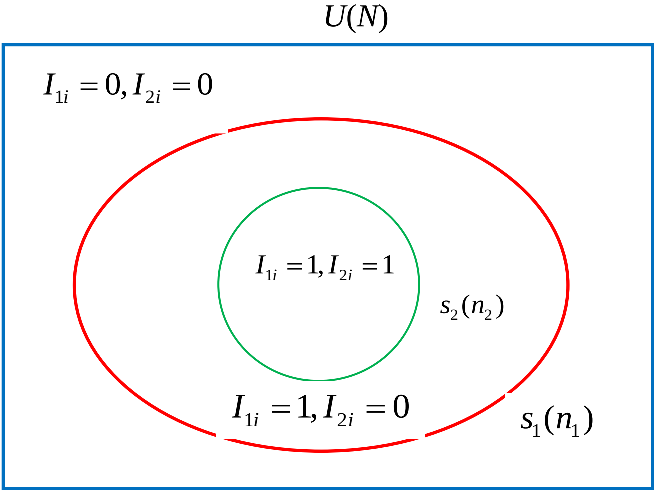
Unités influentes

- Cependant même avec un bon plan de sondage, il reste des unités influentes, par exemple des stratum jumpers
- En présence d'unités influentes, les estimateurs classiques demeurent sans biais, mais leur variance peut être élevée
- Réduire l'impact de ces unités influentes, conduit à des estimateurs plus stables, mais biaisés
- Il faut trouver le bon compromis entre biais et variance

Plan de sondage à deux phases

- U : population de taille N
- s_1 : échantillon issu de la première phase de tirage de taille n_1
- s_2 : échantillon issu de la deuxième phase de tirage à partir de s_1, de taille n_2
- I_{1i} : indicatrice d'appartenance à l'échantillon s_1 de l'unité i
- I_{2i} : indicatrice d'appartenance à l'échantillon s_2 de l'unité i
- Vecteurs des indicatrices $\mathbf{I}_1 = (I_{11}, \cdots, I_{1N})'$ et $\mathbf{I}_2 = (I_{21}, \cdots, I_{2N})'$
- Probabilité d'inclusion de première phase de l'unité i : $\pi_{1i} = P(I_{1i} = 1)$
- Probabilité d'inclusion de deuxième phase de l'unité i : $\pi_{2i}(\mathbf{I}_1) = P(I_{2i} = 1 | \mathbf{I}_1; I_{1i} = 1)$

Plan de sondage à deux phases



Plan de sondage à deux phases

- Un plan de sondage à deux phases est dit invariant si $P(\mathbf{I}_2|\mathbf{I}_1) = P(\mathbf{I}_2)$
- Dans ce cas, on a : $\pi_{2i}(\mathbf{I}_1) = \pi_{2i}$
- Exemple d'invariance : On empile deux sondages aléatoires simples sans remise de taille n_1 et n_2 fixé à priori
- Exemple de non invariance :
  - On effectue un sondage aléatoire simple sans remise en première phase
  - En deuxième phase, on effectue un tirage proportionnel à la variable x récoltée au cours de la première phase :

$$\pi_{2i}(\mathbf{I}_1) = n_2 \frac{x_i}{\sum_{i \in s_1} x_i},$$

- Dans la suite, on supposera que le plan de sondage à deux phases considéré est invariant

Estimation

- On s'intéresse à l'estimation du total de la variable d'intérêt y,

$$Y = \sum_{i \in U} y_i$$

- Les valeurs de la variable y ne sont connues que pour $i \in s_2$
- En l'absence de non-réponse, on utilise l'estimateur par double dilation :

$$\hat{Y}_{DE} = \sum_{i \in s_2} \frac{y_i}{\pi_{1i}\pi_{2i}} = \sum_{i \in s_2} \frac{y_i}{\pi_i^*}$$

Estimation

- On s'intéresse à l'estimation du total de la variable d'intérêt y,

$$Y = \sum_{i \in U} y_i$$

- Les valeurs de la variable y ne sont connues que pour $i \in s_2$
- En l'absence de non-réponse, on utilise l'estimateur par double dilation :

$$\hat{Y}_{DE} = \sum_{i \in s_2} \frac{y_i}{\pi_{1i}\pi_{2i}} = \sum_{i \in s_2} \frac{y_i}{\pi_i^*}$$

- L'estimateur $\hat{Y}_{DE}$ est sans biais sous le plan :

$$E_1 E_2(\hat{Y}_{DE} | \mathbf{I}_1) = Y$$

Erreur d'échantillonnage

- On s'intéresse à l'erreur d'échantillonnage totale de $\hat{Y}_{DE}$:

$$\hat{Y}_{DE} - Y$$

- Comment quantifier l'influence d'une unité sur l'erreur d'échantillonnage ?
  Dans le cas d'un tirage avec une seule phase, on peut utiliser le biais conditonnel ; Moreno-Rebollo, Munoz-Reyez and Munoz-Pichardo (1999), Beaumont, Haziza and Ruiz-Gazen (2011).

Erreur d'échantillonnage

- On s'intéresse à l'erreur d'échantillonnage totale de $\hat{Y}_{DE}$:

$$\hat{Y}_{DE} - Y$$

- Comment quantifier l'influence d'une unité sur l'erreur d'échantillonnage ?
  Dans le cas d'un tirage avec une seule phase, on peut utiliser le biais conditonnel ; Moreno-Rebollo, Munoz-Reyez and Munoz-Pichardo (1999), Beaumont, Haziza and Ruiz-Gazen (2011).
- Comment construire un estimateur robuste en présence d'unités influentes ?
  Dans le cas d'un tirage avec une seule phase, on peut se réferer à l'article de Beaumont, Haziza and Ruiz-Gazen (2011).

Sommaire

Le biais conditionnel

- Pour commencer on distingue trois types d'unités :
  - $i \in s_2$: unité échantillonnée
  - $i \in s_1 - s_2$: unité échantillonnée en première phase, mais pas en deuxième
  - $i \in U - s_1$: unité non échantillonnée

Le biais conditionnel

- Pour commencer on distingue trois types d'unités :
  - $i \in s_2$: unité échantillonnée
  - $i \in s_1 - s_2$: unité échantillonnée en première phase, mais pas en deuxième
  - $i \in U - s_1$: unité non échantillonnée
- On ne peut réduire l'impact que des unités échantillonnées (i.e., des unités appartenant à s_2)

Le biais conditionnel

- Pour commencer on distingue trois types d'unités :
  - $i \in s_2$: unité échantillonnée
  - $i \in s_1 - s_2$: unité échantillonnée en première phase, mais pas en deuxième
  - $i \in U - s_1$: unité non échantillonnée
- On ne peut réduire l'impact que des unités échantillonnées (i.e., des unités appartenant à s_2)
- On peut calculer le biais conditionnel, pour une unité échantillonnée $i \in s_2$:

$$B_i^{DE}(I_{1i} = 1, I_{2i} = 1) \quad = \quad E_1 E_2(\hat{Y}_{DE} - Y | \mathbf{I}_1, I_{1i} = 1, I_{2i} = 1)$$

Le biais conditionnel

- Pour commencer on distingue trois types d'unités :
  - $i \in s_2$: unité échantillonnée
  - $i \in s_1 - s_2$: unité échantillonnée en première phase, mais pas en deuxième
  - $i \in U - s_1$: unité non échantillonnée
- On ne peut réduire l'impact que des unités échantillonnées (i.e., des unités appartenant à s_2)
- On peut calculer le biais conditionnel, pour une unité échantillonnée $i \in s_2$:

$$B_i^{DE}(I_{1i} = 1, I_{2i} = 1) = E_1 E_2(\hat{Y}_{DE} - Y | \mathbf{I}_1, I_{1i} = 1, I_{2i} = 1)$$

Le biais conditionnel

- Pour un plan de sondage à deux phases quelconque :

$$B_i^{DE}(I_{1i}=1, I_{2i}=1) = \underbrace{\sum_{j \in U} \left(\frac{\pi_{ij}^*}{\pi_i^* \pi_j^*} - 1 \right) y_j}_{\text{Influence totale de l'unité } i}$$

- Pour un empilement de deux sondages aléatoires simples sans remise de taille n_1 et n_2 : $\pi_i^* = \frac{n_1}{N} \times \frac{n_2}{n_1} = \frac{n_2}{N}$

$$B_i^{DE}(I_{1i}=1, I_{2i}=1) = \frac{N}{(N-1)}(\frac{N}{n_2} - 1)(y_i - \bar{Y})$$

Le biais conditionnel

- Pour un plan de sondage à deux phases quelconque :

$$B_i^{DE}(I_{1i}=1, I_{2i}=1) = \underbrace{\sum_{j\in U}\left(\frac{\pi^*_{ij}}{\pi^*_i\pi^*_j}-1\right)y_j}_{\text{Influence totale de l'unité } i}$$

- Pour un empilement de deux sondages aléatoires simples sans remise de taille n_1 et n_2 : $\pi_i^* = \frac{n_1}{N}\times\frac{n_2}{n_1} = \frac{n_2}{N}$

$$B_i^{DE}(I_{1i}=1, I_{2i}=1) = \frac{N}{(N-1)}(\frac{N}{n_2}-1)(y_i-\bar{Y})$$

- Plan quelconque/ Tirage Poissonien :

$$B_i^{DE}(I_{1i}=1, I_{2i}=1) = \sum_{j\in U}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}-1\right)y_j + \pi_{1i}^{-1}\left(\pi_{2i}^{-1}-1\right)y_i$$

Le biais conditionnel

- Pour un plan de sondage à deux phases quelconque :

$$B_i^{DE}(I_{1i}=1, I_{2i}=1) \quad = \quad \underbrace{\sum_{j\in U}\left(\frac{\pi^*_{ij}}{\pi^*_i\pi^*_j}-1\right)y_j}_{\text{Influence totale de l'unité } i}$$

- Pour un empilement de deux sondages aléatoires simples sans remise de taille n_1 et n_2 : $\pi^*_i = \frac{n_1}{N}\times\frac{n_2}{n_1}=\frac{n_2}{N}$

$$B_i^{DE}(I_{1i}=1, I_{2i}=1) \quad = \quad \frac{N}{(N-1)}(\frac{N}{n_2}-1)(y_i-\bar{Y})$$

- Plan quelconque/ Tirage Poissonien :

$$B_i^{DE}(I_{1i}=1, I_{2i}=1) = \sum_{j\in U}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}-1\right)y_j + \pi_{1i}^{-1}\left(\pi_{2i}^{-1}-1\right)y_i$$

Le biais conditionnel : Propriétés

- Il s'agit d'une mesure d'influence qui tient compte du plan de sondage
- Il peut s'interpréter, pour certains plans, commme la contribution de chaque unité à l'erreur d'échantillonnage totale
- Si $\pi_i^* = 1$, alors $B_i^{DE}(I_{1i} = 1, I_{2i} = 1) = 0$
- Il se peut que le biais conditionel soit inconnu, il suffit de l'estimer

Construction d'un estimateur robuste

- En suivant la démarche de Beaumont, Haziza and Ruiz-Gazen (2011), on construit :

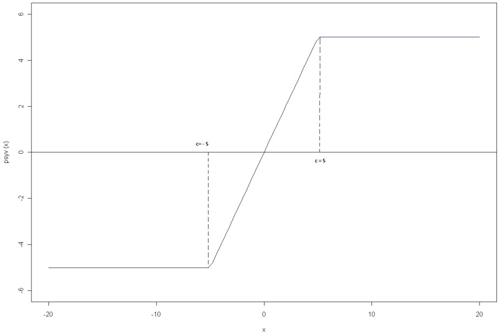
$$\hat{Y}_{DE}^{R} = \hat{Y}_{DE} - \sum_{i \in s_2} \hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1) + \sum_{i \in s_2} \psi \left\{ \hat{B}_i^{DE}(I_{1i} = 1, I_{2i} = 1) \right\}$$

- Avec la fonction de Huber ψ :

$$\psi(t) = \begin{cases} c & \text{si } t > c \\ t & \text{si } |t| \leq c \\ -c & \text{si } t < -c \end{cases}$$

- c : tuning constant
- Cas particulier : Une seule phase ; i.e., $I_{2i} = 1$ pout tout $i \Rightarrow \hat{Y}_{DE}^{R}$ est égale à l'estimateur robuste proposé par Beaumont, Haziza et Ruiz-Gazen (2011)

Construction d'un estimateur robuste



Sommaire

Notation

- s_2 : échantillon des répondants

Notation

- s_2 : échantillon des répondants
- n_2 : nombre d'unités répondantes (aléatoire)

Notation

- s_2 : échantillon des répondants
- n_2 : nombre d'unités répondantes (aléatoire)
- I_{2i} : unité de réponse de l'unité i

Notation

- s_2 : échantillon des répondants
- n_2 : nombre d'unités répondantes (aléatoire)
- I_{2i} : unité de réponse de l'unité i
- π_{2i} : probabilité de réponse de l'unité i inconnue

Notation

- s_2 : échantillon des répondants
- n_2 : nombre d'unités répondantes (aléatoire)
- I_{2i} : unité de réponse de l'unité i
- π_{2i} : probabilité de réponse de l'unité i inconnue
- On suppose que les individus répondent indépendamment les uns les autres (équivalent à un tirage Poissonien en deuxième phase)

Notation

- s_2 : échantillon des répondants
- n_2 : nombre d'unités répondantes (aléatoire)
- I_{2i} : unité de réponse de l'unité i
- π_{2i} : probabilité de réponse de l'unité i inconnue
- On suppose que les individus répondent indépendamment les uns les autres (équivalent à un tirage Poissonien en deuxième phase)
- Si les π_{2i} étaient connues, on construirait l'estimateur repondéré pour la non-réponse :

$$\tilde{Y}_{PSA} = \sum_{i \in s_2} \frac{y_i}{\pi_{1i}\pi_{2i}}$$

Notation

- s_2 : échantillon des répondants
- n_2 : nombre d'unités répondantes (aléatoire)
- I_{2i} : unité de réponse de l'unité i
- π_{2i} : probabilité de réponse de l'unité i inconnue
- On suppose que les individus répondent indépendamment les uns les autres (équivalent à un tirage Poissonien en deuxième phase)
- Si les π_{2i} étaient connues, on construirait l'estimateur repondéré pour la non-réponse :

$$\tilde{Y}_{PSA} = \sum_{i \in s_2} \frac{y_i}{\pi_{1i}\pi_{2i}}$$

- Influence de l'unité i sur cet estimateur :

$$B_i^{PSA}(I_{1i}=1, I_{2i}=1) = \underbrace{\sum_{j \in U} \left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1 \right) y_j}_{\text{Influence de l'unité } i \text{ sur l'erreur d'échantillonnage}} + \underbrace{\pi_{1i}^{-1} \left(\pi_{2i}^{-1} - 1 \right) y_i}_{\text{l'influence de l'unité } i \text{ sur l'erreur de non-réponse}}$$

Modèle de non-réponse

- En pratique, les probabilités de réponse π_{2i} sont inconnues
- On pose le modèle paramétrique suivant : $\pi_{2i} = m\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$,

Modèle de non-réponse

- En pratique, les probabilités de réponse π_{2i} sont inconnues
- On pose le modèle paramétrique suivant : $\pi_{2i} = m\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$, où
  - m(.) est une fonction de lien connu
  - $\mathbf{x}_i$ est un vecteur de variable auxiliaire connu pour toutes les unités de l'échantillon (répondants et non répondants)
  - $\boldsymbol{\alpha}$ est un vecteur de paramètres inconnus

Modèle de non-réponse

- En pratique, les probabilités de réponse π_{2i} sont inconnues
- On pose le modèle paramétrique suivant : $\pi_{2i} = m\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$, où
  - m(.) est une fonction de lien connu
  - $\mathbf{x}_i$ est un vecteur de variable auxiliaire connu pour toutes les unités de l'échantillon (répondants et non répondants)
  - $\boldsymbol{\alpha}$ est un vecteur de paramètres inconnus
- Exemple : le modèle logistique

$$\pi_{2i} = \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)}{1 + \exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)}$$

Modèle de non-réponse

- En pratique, les probabilités de réponse π_{2i} sont inconnues
- On pose le modèle paramétrique suivant : $\pi_{2i} = m\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$, où
  - m(.) est une fonction de lien connu
  - $\mathbf{x}_i$ est un vecteur de variable auxiliaire connu pour toutes les unités de l'échantillon (répondants et non répondants)
  - $\boldsymbol{\alpha}$ est un vecteur de paramètres inconnus
- Exemple : le modèle logistique

$$\pi_{2i} = \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)}{1 + \exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)}$$

- Probabilité de réponse estimée de l'unité i : $\hat{\pi}_{2i} = m\left(\mathbf{x}_i, \hat{\boldsymbol{\alpha}}\right)$

Modèle de non-réponse

- En pratique, les probabilités de réponse π_{2i} sont inconnues
- On pose le modèle paramétrique suivant : $\pi_{2i} = m\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$, où
  - m(.) est une fonction de lien connu
  - $\mathbf{x}_i$ est un vecteur de variable auxiliaire connu pour toutes les unités de l'échantillon (répondants et non répondants)
  - $\boldsymbol{\alpha}$ est un vecteur de paramètres inconnus
- Exemple : le modèle logistique

$$\pi_{2i} = \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)}$$

- Probabilité de réponse estimée de l'unité i : $\hat{\pi}_{2i} = m\left(\mathbf{x}_i, \hat{\boldsymbol{\alpha}}\right)$

Méthodes des groupes de réponses homogènes

- Etape 1 :
  On range les individus selon la méthode des scores, par probabilité estimée croissante
- Etape 2 :
  On classe les individus dans un nombre fixé de groupes, de même taille
- Etape 3 :
  On calcule les probabilités de réponse dans chaque groupe g :
- Si $i \in g$

$$\hat{\pi}_{2i} = \frac{\sum_{i \in sr_g} \pi_{1i}^{-1}}{\sum_{i \in s_g} \pi_{1i}^{-1}}$$

- Estimateur repondéré par groupe :

$$\hat{Y}_{PSA} = \sum_{i \in s_2} \frac{y_i}{\pi_{1i}\hat{\pi}_{2i}}$$

Méthodes des groupes de réponses homogènes

- Etape 1 :
  On range les individus selon la méthode des scores, par probabilité estimée croissante
- Etape 2 :
  On classe les individus dans un nombre fixé de groupes, de même taille
- Etape 3 :
  On calcule les probabilités de réponse dans chaque groupe g :
- Si $i \in g$

$$\hat{\pi}_{2i} = \frac{\sum_{i \in sr_g} \pi_{1i}^{-1}}{\sum_{i \in s_g} \pi_{1i}^{-1}}$$

- Estimateur repondéré par groupe :

$$\hat{Y}_{PSA} = \sum_{i \in s_2} \frac{y_i}{\pi_{1i}\hat{\pi}_{2i}}$$

Modèle de non-réponse

- On peut montrer que :

$$\hat{Y}_{PSA} - \hat{Y}_L = O_p(n^{-1}),$$

où $\hat{Y}_L$ est la version linéarisée de $\hat{Y}_{PSA}$.

Modèle de non-réponse

- On peut montrer que :

$$\hat{Y}_{PSA} - \hat{Y}_L = O_p(n^{-1}),$$

où $\hat{Y}_L$ est la version linéarisée de $\hat{Y}_{PSA}$.

- Biais conditionnel asymptotique d'une unité répondante sur l'erreur d'échantillonnage et l'erreur de non réponse est :

$$B_i^{PSA}(I_{1i} = 1, I_{2i} = 1) \approx E_1 E_2(\hat{Y}_L - Y | \mathbf{I}_1, I_{1i} = 1, I_{2i} = 1)$$

$$\begin{aligned} B_i^{PSA}(I_{1i} = 1, I_{2i} = 1) &\approx \sum_{j \in U} \left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1 \right) (y_j - \bar{y}_U) \\ &+ \pi_{1i}^{-1} \left(\pi_{2i}^{-1} - 1 \right) (y_i - \bar{y}_g) \end{aligned}$$

Estimateur robuste dans le cas de non-réponse

- Une version robuste de $\hat{Y}_{PSA}$

$$\begin{aligned}\hat{Y}^{R}_{PSA} &= \hat{Y}_{PSA} - \sum_{i \in s_2} \hat{B}^{PSA}_i (I_{1i}=1, I_{2i}=1) \\ &+ \sum_{i \in s_2} \psi \left\{ \hat{B}^{PSA}_i (I_{1i}=1, I_{2i}=1) \right\}\end{aligned}$$

- Problème : détermination de la tuning constante
- On utilise un critère de type min-max pour la déterminer et on obtient :

$$\hat{Y}^{R}_{PSA} = \hat{Y}_{PSA} - \frac{\min_{i \in S}(\hat{B}^{PSA}_i) + \max_{i \in S}(\hat{B}^{PSA}_i)}{2}$$

Sommaire

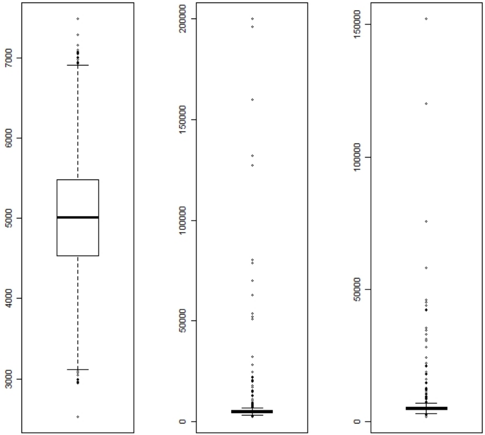
Echantillonnage et Simulation de la Non-réponse

- On a généré trois populations de taille $N = 5000$ avec une variable aléatoire vectorielle d'intérêt notée Y et une variable auxiliaire X.
- On a effectué $P = 5000$ tirages aléatoires simples sans remise de taille n_1 en première phase.
- Les probabilités de réponse ont été générées de la façon suivante :

$$\pi_{2i} = \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)}{1 + \exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)}$$

- La moyenne des probabilités de réponse est de 0.7.
- On a mis en oeuvre l'estimation robuste avec la méthode « min-max » pour deux tailles d'échantillons.

Résultats pour l'estimation d'un total



Résultats pour l'estimation d'un total

$$RV_{MC}\left(\hat{\theta}^{R}_{PSA}, \hat{\theta}_{PSA}\right) = \frac{\frac{1}{P}\sum_{p=1}^{P}\left(\hat{\theta}^{R}_{PSA} - E_{MC}\left(\hat{\theta}^{R}_{PSA}\right)\right)^2}{\frac{1}{P}\sum_{p=1}^{P}\left(\hat{\theta}_{PSA} - E_{MC}\left(\hat{\theta}_{PSA}\right)\right)^2},$$

et

$$RE_{MC}\left(\hat{\theta}^{R}_{PSA}, \hat{\theta}_{PSA}\right) = \frac{\frac{1}{P}\sum_{p=1}^{P}\left(\hat{\theta}^{R}_{PSA} - \theta\right)^2}{\frac{1}{P}\sum_{p=1}^{P}\left(\hat{\theta}_{PSA} - \theta\right)^2}.$$

Résultats pour l'estimation d'un total

Pop	Taille n_1	$RB_{MC}\left(\hat{\theta}_p^{RPSA}\right)$ (%)	$RV_{MC}\left(\hat{\theta}_p^{RPSA}, \hat{\theta}_p^{PSA}\right)$	RE_{MC}
1	300	0.02	1.0	0.98
	500	0.04	1.0	0.98
2	300	-0.98	0.52	0.58
	500	-0.60	0.61	0.64
3	300	-1.33	0.50	0.52
	500	-1.16	0.59	0.63

Sommaire

Quelques remarques

- On peut généraliser les résultats aux cas de plans à deux phases non invariants
- On peut étudier d'autres estimateurs, commme l'estimateur par calage
- Estimation de l'erreur quadratique de cet estimateur

Bibliographie I

J.F. Beaumont, D. Haziza, and A. Ruiz-Gazen.
A unified approach to robust estimation in finite population sampling.
En révision, 2011.

R.L. Chambers.
Outlier robust finite population estimation.
Journal of the American Statistical Association, pages 1063–1069, 1986.

P.N. Kokic and P.A. Bell.
Optimal winsorizing cutoffs for a stratified finite population estimator.
Journal of official statistics-Stockholm, 10 :419–419, 1994.

Bibliographie II

J.L. Moreno-Rebollo, A. Munoz-Reyes, M.D. Jimenez-Gamero, and J. Munoz-Pichardo.
Influence diagnostic in survey sampling : Estimating the conditional bias.
Metrika, 55(3) :209–214, 2002.

J.L. Moreno-Rebollo, A. Muñoz-Reyes, and J. Muñoz-Pichardo.
Miscellanea. influence diagnostic in survey sampling : conditional bias.
Biometrika, 86(4) :923–928, 1999.

J. Munoz-Pichardo, J. Munoz-Garcia, J.L. Moreno-Rebollo, and R. Pino-Mejias.
A new approach to influence analysis in linear models.
Sankhyā : The Indian Journal of Statistics, Series A, pages 393–409, 1995.