Traitement des unités influentes dans les enquêtes en présence de non-réponse totale

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

2 juillet 2013

Sommaire

Sommaire

Le contexte

- On suppose qu'on observe des unités potentiellement influentes dans notre échantillon
- On se trouve en présence de non-réponse totale : certaines unités de notre échantillon n'ont pas répondu
- On suppose que les individus répondent indépendamment les uns des autres
- On va modéliser cette non-réponse comme une phase supplémentaire de l'échantillonnage de Poisson
- L'objectif est de quantifier l'influence des unités présentes dans l'échantillon des répondants
- Construire un estimateur robuste en présence de non-réponse totale

Le contexte

- En présence de non-réponse, on peut adopter une approche en deux étapes

Le contexte

- En présence de non-réponse, on peut adopter une approche en deux étapes
- Approche en deux étapes :

$$\underbrace{d_i}_{\text{poids d'échantillonnage}} \longrightarrow \underbrace{w_i^* = d_i/\hat{p}_i}_{\text{poids ajustés pour la non-réponse}} \longrightarrow \underbrace{w_i}_{\text{poids de calage}}$$

Le contexte

- En présence de non-réponse, on peut adopter une approche en deux étapes
- Approche en deux étapes :

$$\underbrace{d_i}_{\text{poids d'échantillonnage}} \longrightarrow \underbrace{w_i^* = d_i/\hat{p}_i}_{\text{poids ajustés pour la non-réponse}} \longrightarrow \underbrace{w_i}_{\text{poids de calage}}$$

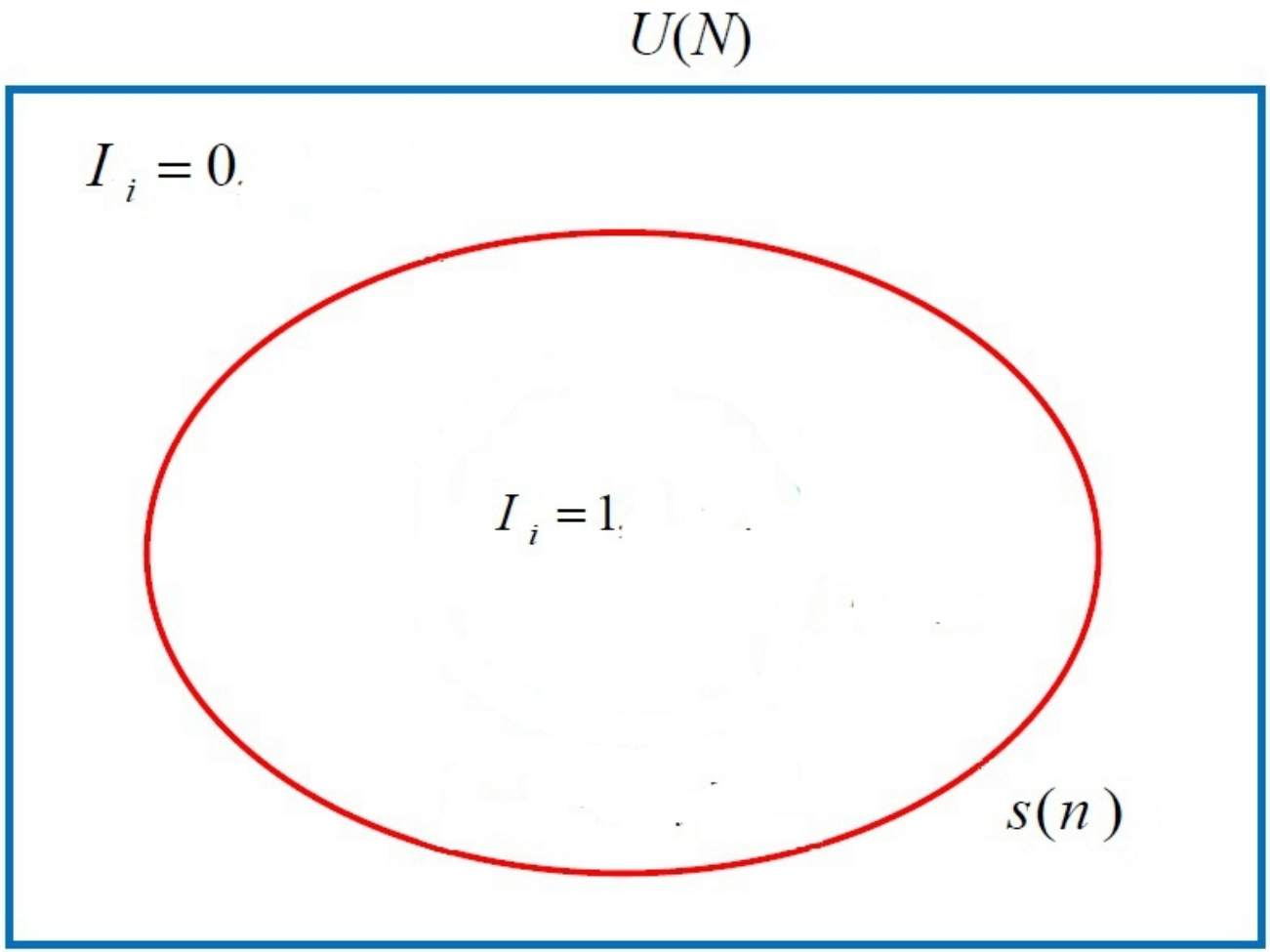
  - première étape : réduction du biais de non-réponse
  - deuxième étape : assurer la cohérence entre les estimations produites à partir de l'échantillon et les vrais totaux connus sur la population, et améliorer si possible la précision des estimateurs

Sommaire

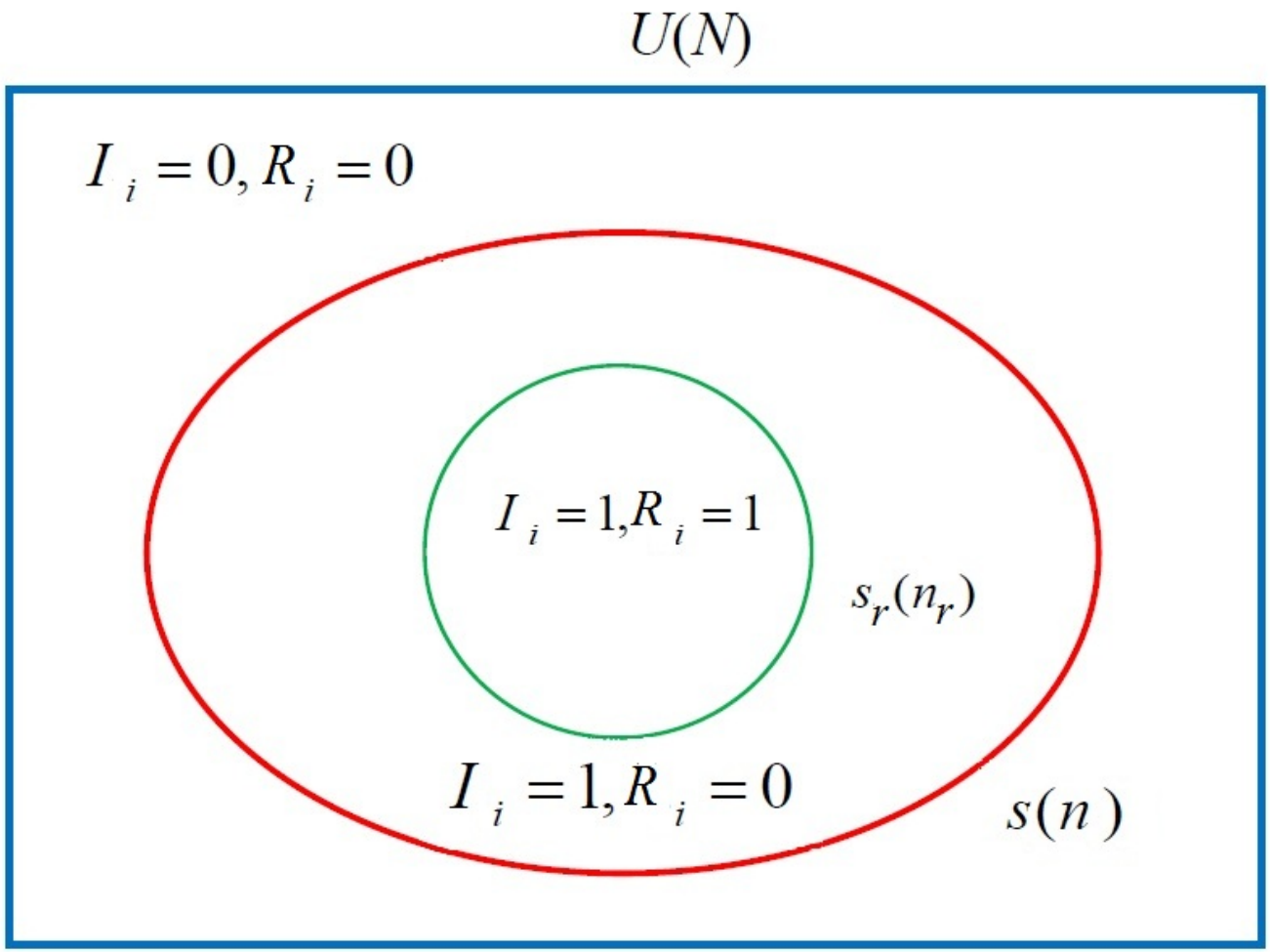
Notation

- U : population de taille N
- s : échantillon issu du mécanisme d'échantillonnage de taille n
- s_r : échantillon issu du mécanisme de non-réponse, de taille n_r
- I_i : indicatrice d'appartenance à l'échantillon s de l'unité i
- R_i : indicatrice d'appartenance à l'échantillon des répondants s_r de l'unité i
- Vecteurs des indicatrices $\mathbf{I} = (I_1, \cdots, I_N)'$
- Probabilité d'inclusion de l'unité i : $\pi_i = P(I_i = 1)$
- Probabilité de réponse de l'unité i : $p_i = P(R_i = 1)$

Représentation graphique du mécanisme



Représentation graphique du mécanisme



La configuration

Définition

Une configuration est un quadruplet qui consiste en :
(1) une variable d'intérêt
(2) un paramètre de la population
(3) un plan de sondage
(4) un estimateur

Dans cet exposé, on va se placer dans les trois configurations suivantes :

- (Chiffre d'affaires, total , Plan de sondage quelconque + Tirage Poissonien, Estimateur repondéré pour la non-réponse (p_i connues))
- (Chiffre d'affaires, total , Plan de sondage quelconque + Tirage Poissonien, Estimateur repondéré pour la non-réponse (p_i inconnues))
- (Chiffre d'affaires, total , Plan de sondage quelconque + Tirage Poissonien, Estimateur repondéré pour la non-réponse suivi d'un calage)

Estimation

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

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

- En l'absence de non-réponse, on utiliserait l'estimateur par dilatation (Horvitz-Thompson) :

$$\hat{Y}_\pi = \sum_{i \in s} d_i y_i \text{ avec } d_i = \frac{1}{\pi_i}$$

- Les valeurs de la variable y ne sont connues que pour $i \in s_r$
- Si les probabilités de réponses p_i étaient connues, on construirait l'estimateur repondéré pour la non-réponse :

$$\tilde{Y}_{PSA} = \sum_{i \in s_r} \frac{d_i}{p_i} y_i$$

Sommaire

Erreur totale

- On s'intéresse à l'erreur totale de $\tilde{Y}_{PSA}$: $\tilde{Y}_{PSA} - Y$

$$\underbrace{\tilde{Y}_{PSA} - Y}_{\text{erreur totale}} = \underbrace{\left(\hat{Y}_{\pi} - Y\right)}_{\text{erreur d'échantillonnage}} + \underbrace{\left(\tilde{Y}_{PSA} - \hat{Y}_{\pi}\right)}_{\text{erreur de non-réponse}} \quad (1)$$

- Une unité influente aura potentiellement un impact sur l'erreur d'échantillonnage et l'erreur de non-réponse

Erreur totale

- On s'intéresse à l'erreur totale de $\tilde{Y}_{PSA}$: $\tilde{Y}_{PSA} - Y$

$$\underbrace{\tilde{Y}_{PSA} - Y}_{\text{erreur totale}} = \underbrace{\left(\hat{Y}_{\pi} - Y\right)}_{\text{erreur d'échantillonnage}} + \underbrace{\left(\tilde{Y}_{PSA} - \hat{Y}_{\pi}\right)}_{\text{erreur de non-réponse}} \quad (1)$$

- Une unité influente aura potentiellement un impact sur l'erreur d'échantillonnage et l'erreur de non-réponse
- Comment mesurer l'impact de ces unités sur les deux types d'erreur ?

Erreur totale

- On s'intéresse à l'erreur totale de $\tilde{Y}_{PSA}$: $\tilde{Y}_{PSA} - Y$

$$\underbrace{\tilde{Y}_{PSA} - Y}_{\text{erreur totale}} = \underbrace{\left(\hat{Y}_{\pi} - Y\right)}_{\text{erreur d'échantillonnage}} + \underbrace{\left(\tilde{Y}_{PSA} - \hat{Y}_{\pi}\right)}_{\text{erreur de non-réponse}} \quad (1)$$

- Une unité influente aura potentiellement un impact sur l'erreur d'échantillonnage et l'erreur de non-réponse
- Comment mesurer l'impact de ces unités sur les deux types d'erreur ? Dans le cas d'une seule phase : biais conditionnel ; Moreno-Rebollo, Munoz-Reyez and Munoz-Pichardo (1999), Beaumont, Haziza and Ruiz-Gazen (2013).

Erreur totale

- On s'intéresse à l'erreur totale de $\tilde{Y}_{PSA}$: $\tilde{Y}_{PSA} - Y$

$$\underbrace{\tilde{Y}_{PSA} - Y}_{\text{erreur totale}} = \underbrace{\left(\hat{Y}_{\pi} - Y\right)}_{\text{erreur d'échantillonnage}} + \underbrace{\left(\tilde{Y}_{PSA} - \hat{Y}_{\pi}\right)}_{\text{erreur de non-réponse}} \quad (1)$$

- Une unité influente aura potentiellement un impact sur l'erreur d'échantillonnage et l'erreur de non-réponse
- Comment mesurer l'impact de ces unités sur les deux types d'erreur ? Dans le cas d'une seule phase : biais conditionnel ; Moreno-Rebollo, Munoz-Reyez and Munoz-Pichardo (1999), Beaumont, Haziza and Ruiz-Gazen (2013).
- Comment construire un estimateur robuste à partir de cette mesure d'influence ? Dans le cas d'une seule phase : Beaumont, Haziza and Ruiz-Gazen (2013).

Le biais conditionnel

- Pour commencer on distingue trois types d'unités :
- $i \in s_r$: unité échantillonnée et répondante
- $i \in s - s_r$: unité échantillonnée n'ayant pas répondu
- $i \in U - s$: unité non échantillonnée
- On ne peut réduire l'impact que des unités répondantes (i.e., des unités appartenant à s_r)

Le biais conditionnel

- Pour commencer on distingue trois types d'unités :
- $i \in s_r$: unité échantillonnée et répondante
- $i \in s - s_r$: unité échantillonnée n'ayant pas répondu
- $i \in U - s$: unité non échantillonnée
- On ne peut réduire l'impact que des unités répondantes (i.e., des unités appartenant à s_r)
- On peut calculer le biais conditionnel, pour une unité échantillonnée et répondante $i \in s_r$:

$$
\begin{aligned}
B_i^{PSA}(I_i = 1, R_i = 1) &= E_p E_r(\tilde{Y}_{PSA} - Y | \mathbf{I}, I_i = 1, R_i = 1) \\
&= E_p(\hat{Y}_\pi - Y | \mathbf{I}, I_i = 1) + E_p E_r(\tilde{Y}_{PSA} - \hat{Y}_\pi | \mathbf{I}, I_i = R_i = 1) \\
&= \underbrace{\sum_{j \in U} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_j}_{\text{Influence de l'unité } i \text{ sur l'erreur d'échantillonnage}} + \underbrace{\pi_i^{-1} \left(p_i^{-1} - 1 \right) y_i}_{\text{Influence de l'unité } i \text{ sur l'erreur de non-réponse}}
\end{aligned}
$$

Le biais conditionnel : propriétés

- Il s'agit d'une mesure d'influence qui tient compte du plan de sondage et du mécanisme de non-réponse
- Il peut s'interpréter, pour certains plans, commme la contribution de chaque unité à l'erreur totale
- Si $\pi_i = 1$, alors la partie du biais conditionnel associée à l'erreur d'échantillonnage est nulle
- Si $p_i = 1$, alors la partie du biais conditionnel associée à l'erreur de non-réponse est nulle
- Il se peut que le biais conditionel soit inconnu, il suffit de l'estimer

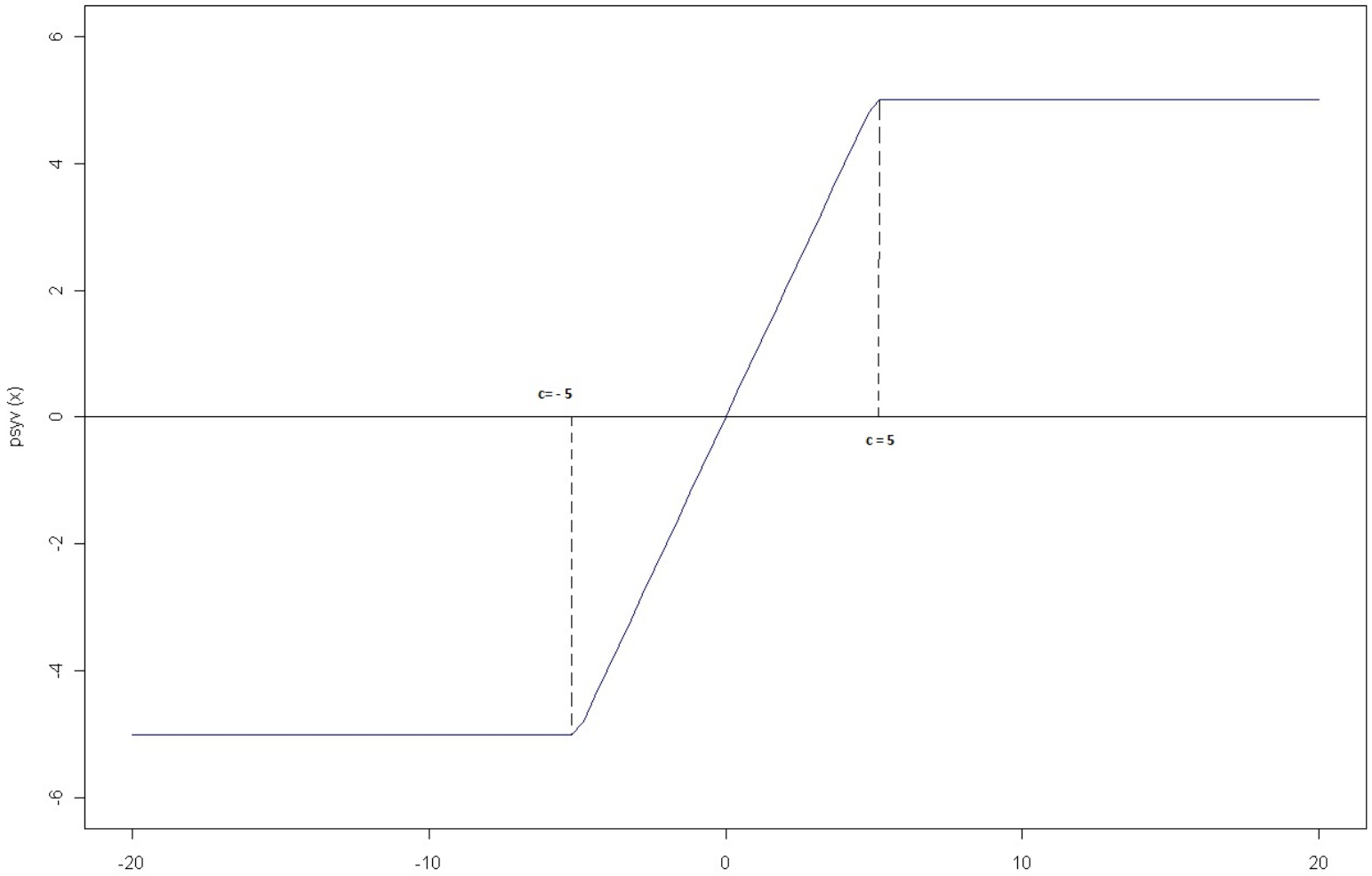
Construction d'un estimateur robuste

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

$$\hat{\tilde{Y}}_{PSA}^{R} = \tilde{Y}_{PSA} - \sum_{i \in s_r} \hat{B}_i^{PSA}(I_i = 1, R_i = 1) + \sum_{i \in s_r} \psi_c \left\{ \hat{B}_i^{PSA}(I_i = 1, R_i = 1) \right\}$$

- 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

Représentation graphique de la fonction de Huber



Construction d'un estimateur robuste

- Comme dans Beaumont et al. (2013), on s'intéresse à déterminer la valeur de c qui minimise le plus grand biais conditionnel estimé dans l'échantillon de l'estimateur $\hat{\tilde{Y}}^R_{PSA}$
- Formellement, on cherche la valeur de c qui minimise

$$\max_{i \in s_r} \left\{ (|\hat{B}_i^{RPSA}(I_i = 1, R_i = 1)|) \right\},$$

où $\hat{B}_i^{RPSA}(I_i = 1, R_i = 1)$ désigne le biais conditionnel estimé de l'estimateur $\hat{\tilde{Y}}^R_{PSA}$ associé à l'unité échantillonnée i.

- On obtient alors un estimateur robuste :

$$\hat{\tilde{Y}}^R_{PSA} = \tilde{Y}_{PSA} - \frac{1}{2}(\hat{B}_{min} + \hat{B}_{max})$$

où $\hat{B}_{min} = min_{i \in s_r}(\hat{B}_i^{PSA}(I_i = 1, R_i = 1))$

Sommaire

Modèle de non-réponse

- En pratique, les probabilités de réponse p_i sont inconnues
- On découpe notre population en G sous-populations (ou sous-groupes) supposées homogènes au sens de la non-réponse par une procédure adaptée
- On peut définir les G groupes en croisant toutes les modalités des variables explicatives de la non-réponse
- On peut utiliser une méthode des scores.

Modélisation des probabilités de réponse

- On pose le modèle paramétrique suivant : $p_i = \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)}$

Modélisation des probabilités de réponse

- On pose le modèle paramétrique suivant : $p_i = \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)}$ où
- $\mathbf{x}_i$ est un vecteur de variables auxiliaires connu pour toutes les unités de l'échantillon (répondantes et non répondantes)
- $\boldsymbol{\alpha}$ est un vecteur de paramètres inconnus
- La probabilité de réponse estimée de l'unité i est donnée par :

$$\hat{p}_i = \frac{\exp\left(\mathbf{x}_i'\hat{\boldsymbol{\alpha}}\right)}{1+\exp\left(\mathbf{x}_i'\hat{\boldsymbol{\alpha}}\right)}$$

Modélisation des probabilités de réponse

- On pose le modèle paramétrique suivant : $p_i = \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)}$ où
- $\mathbf{x}_i$ est un vecteur de variables auxiliaires connu pour toutes les unités de l'échantillon (répondantes et non répondantes)
- $\boldsymbol{\alpha}$ est un vecteur de paramètres inconnus
- La probabilité de réponse estimée de l'unité i est donnée par :

$$\hat{p}_i = \frac{\exp\left(\mathbf{x}_i'\hat{\boldsymbol{\alpha}}\right)}{1+\exp\left(\mathbf{x}_i'\hat{\boldsymbol{\alpha}}\right)}$$

- Le total Y est alors estimé par :

$$\hat{Y}_{PSA} = \sum_{i\in s_r} \frac{d_i}{\hat{p}_i} y_i = \sum_{i\in s_r} d_i^* y_i \text{ avec } d_i^* = \frac{d_i}{\hat{p}_i}$$

Méthode des scores

- 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 G de groupes, de même taille
- Etape 3 :
  On calcule les probabilités de réponse dans chaque groupe g :
- Si $i \in g$

$$\hat{p}_i = \frac{\sum_{i \in s_{rg}} \pi_i^{-1}}{\sum_{i \in s_g} \pi_i^{-1}}$$

- Estimateur repondéré par groupe :

$$\hat{Y}_{PSA} = \sum_{i \in s_r} \frac{d_i}{\hat{p}_i} y_i = \sum_{g=1}^{G} \sum_{i \in s_{rg}} \frac{d_i}{\hat{p}_g} y_i$$

Biais conditionnel de l'estimateur ajusté pour la non-réponse

- Rappel avec les p_i connues :

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

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

$$\begin{aligned} B_i^{PSA}(I_i = 1, R_i = 1) \quad &\approx \quad \sum_{j \in U} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_j \\ &+ \quad \pi_i^{-1} \left(p_i^{-1} - 1 \right) (y_i - \bar{y}_g) \end{aligned}$$

Estimateur robuste dans le cas de non-réponse

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

$$\begin{aligned}\hat{Y}^R_{PSA} &= \hat{Y}_{PSA} - \sum_{i \in s_r} \hat{B}_i^{PSA}(I_i = 1, R_i = 1) \\ &+ \sum_{i \in s_r} \psi_c \left\{ \hat{B}_i^{PSA}(I_i = 1, R_i = 1) \right\}\end{aligned}$$

- On cherche la valeur de c qui minimise le plus grand biais conditionnel estimé dans l'échantillon de l'estimateur $\hat{Y}^R_{PSA}$
- $\hat{Y}^R_{PSA} = \hat{Y}_{PSA} - \frac{1}{2}(\hat{B}_{min} + \hat{B}_{max})$

Implémentation

- L'implémentation de l'estimateur robuste peut s'effectuer en modifiant les valeurs de y ou en modifiant les poids des unités échantillonnées, d_i^*.
- En terme de valeurs modifiées :

$$\hat{Y}_{PSA}^{R}(c) = \sum_{i \in s_r} d_i^* \tilde{y}_i,$$

avec

$$\tilde{y}_i = y_i - \phi_i \frac{\hat{B}_i^{PSA}(I_i = 1, R_i = 1)}{d_i^*}$$

et

$$\phi_i = 1 - \frac{\psi_c \left\{ \hat{B}_i^{PSA}(I_i = 1, R_i = 1) \right\}}{\hat{B}_i^{PSA}(I_i = 1, R_i = 1)}.$$

- En remarquant que $0 \leq \psi_c(z)/z \leq 1$ alors, $0 \leq \phi_i \leq 1$.

Implémentation

- En terme de poids modifiés :

$$\hat{Y}^R_{PSA}(c) = \sum_{i \in s_r} \tilde{d}^*_i y_i,$$

avec

$$\tilde{d}^*_i = d^*_i - \phi_i \frac{\hat{B}^{PSA}_i(I_i = 1, R_i = 1)}{y_i}.$$

et

$$\phi_i = 1 - \frac{\psi_c \left\{ \hat{B}^{PSA}_i(I_i = 1, R_i = 1) \right\}}{\hat{B}^{PSA}_i(I_i = 1, R_i = 1)}.$$

Extension pour le calage

Approche en deux étapes :

$$\underbrace{d_i}_{\text{poids d'échantillonnage}} \longrightarrow \underbrace{w_i^* = d_i/\hat{p}_i}_{\text{poids ajustés pour la non-réponse}} \longrightarrow \underbrace{w_i}_{\text{poids de calage}}$$

Extension pour le calage

Approche en deux étapes :

$$\underbrace{d_i}_{\text{poids d'échantillonnage}} \longrightarrow \underbrace{w_i^* = d_i/\hat{p}_i}_{\text{poids ajustés pour la non-réponse}} \longrightarrow \underbrace{w_i}_{\text{poids de calage}}$$

- Après la première phase de répondération, en pratique, on effectue un calage afin de garantir une certaine cohérence et améliorer la précision de l'estimateur.
- Notons $\hat{Y}_{cal} = \sum_{i \in s_r} w_i y_i$,
- On peut calculer de façon approchée le biais conditionnel associé à cet estimateur $\hat{Y}_{cal}$ par une linéarisation
- Utiliser ce biais conditionnel pour construire un estimateur robuste $\hat{Y}_{cal}^R$

Sommaire

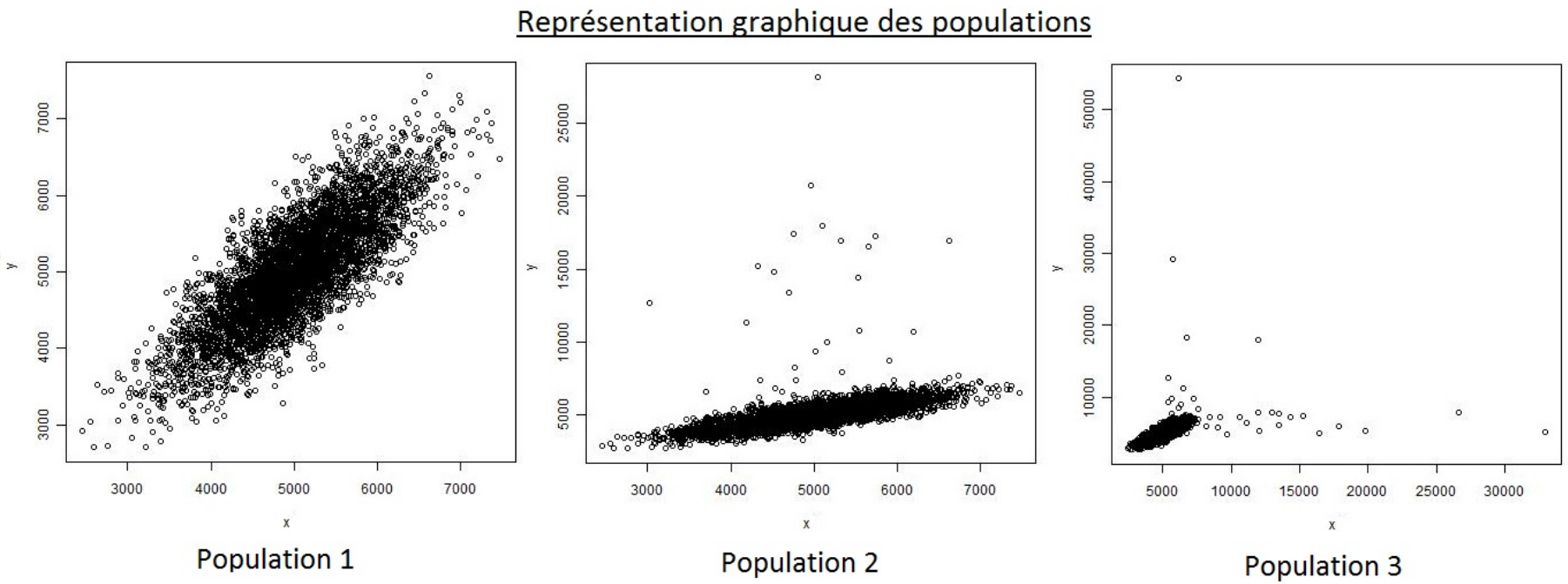
Echantillonnage et simulation de la Non-réponse

- 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 en première phase.
- Les probabilités de réponse ont été générées de la façon suivante :

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

- La moyenne des probabilités de réponse est de 0.7
- On a effectué la méthode des scores avec G=10 groupes de réponse homogènes
- On a mis en oeuvre l'estimation robuste avec la méthode « min-max » pour deux tailles d'échantillons.

Réprésentation graphique des populations générées



Calcul des critères d'efficacité

$$RB_{MC}\left(\hat{\theta}^R_{PSA}\right) = \frac{\frac{1}{P}\sum_{p=1}^{P}\left(\hat{\theta}^R_{PSA} - \theta\right)}{\theta} \times 100.$$

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

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

Pop	Taille n	$RB_{MC}\left(\hat{\theta}^{R}_{PSA}\right)(\%)$	$RV_{MC}\left(\hat{\theta}^{R}_{PSA},\hat{\theta}_{PSA}\right)$	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

- L'estimateur est très facile à implémenter
- En absence d'unités influentes, il est à peine moins efficace que l'estimateur non robuste
- 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.