Risque d'amplification du biais de l'estimateur par calage généralisé en présence de non-réponse

Éric Lesage

Insee

Mars-Avril 2015

Journées de méthodologie statistique
Insee, Paris

Travail réalisé en collaboration avec David Haziza (Département de Mathématique de l'Université de Montréal) et Xavier D'Haultfoeuille (Crest-Ensae).

Sommaire

Introduction I

Le traitement de la non-réponse totale par calage est un domaine de recherche déjà ancien mais qui reste très actif :

- Dupont (1996), Deville (1998), Le Guennec et Sautory (2004), Deville (2004) et Sautory (2003).
- Lundström et Särndal (1999), Särndal et Lundström (2005, 2010, 2011).
- Kott (2006), Chang et Kott (2008), Kott et Chang (2008), Kott et Liao (2012).
- Osier (2012), Lesage (2012), Lesage, Haziza et D'haultfoeuille (2015), et Haziza et Lesage (2015)

Introduction II

- Nécessité de modéliser le mécanisme de non-réponse pour choisir la méthode de calage appropriée
  - choix des variables instrumentales, des variables de calage et de la fonction de calage.
- Risques d'amplification du biais et de la variance de l'estimateur par calage généralisé.
  - Importance du choix des variables de calage $\boldsymbol{x}$

Contexte I

- U : population finie de taille N
- $y_1, \ldots, y_J$: J variables d'intérêt collectées par l'enquête
  - $\boldsymbol{y}_i^\top = (y_{i,1}, \ldots, y_{i,J})$ vecteur des caractéristiques associées à l'unité $i \in U$
- Objectif : estimer les totaux dans la population finie

$$t_{yj} = \sum_{i \in U} y_{i,j} \qquad j = 1, \ldots, J$$

- $x_1, \ldots, x_P$: P variables auxiliaires
  - $\boldsymbol{x}_i^\top = (x_{i,1}, \ldots, x_{i,P})$ vecteur associé à l'unité $i \in U$

Contexte II

- Échantillon s sélectionné selon un plan de sondage probabiliste $p(s)$
  - $\mathcal{S}$: ensemble des échantillons possibles
  - I_i : variable indicatrice d'échantillonnage
  - π_i : probabilité d'inclusion de l'unité i :

$$\pi_i = \mathbb{E}(I_i \mid \boldsymbol{Y}_i = \boldsymbol{y}_i,\ \boldsymbol{X}_i = \boldsymbol{x}_i) = \mathbb{E}(I_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i)$$

  - $d_i = \pi_i^{-1}$: poids d'échantillonnage
- Mécanisme de non-réponse
  - $s_r \subset s$: sous ensemble des répondants
  - R_i : variable indicatrice de réponse

Définition : Estimateur par calage (Deville et Särndal, 1992)

$$\hat{t}_C = \sum_{i \in s} w_i y_i$$

- Poids de calage

$$w_i = d_i \times F(\hat{\lambda}^\top \boldsymbol{x}_i)$$

- Fonction de calage
  - $F(\cdot)$ une fonction dérivable et monotone,
  - $F(0) = 1$ et $F^{'}(0) = 1$.
- Équations de calage

  $\hat{\lambda}$ est un vecteur de paramètres solution de :

$$\sum_{i \in s} d_i F(\hat{\lambda}^\top \boldsymbol{x}_i)\boldsymbol{x}_i = \boldsymbol{t}_x.$$

- Exemple de fonctions de calage :
  - méthode linéaire : $F(\hat{\lambda}^\top \boldsymbol{x}_i) = 1 + \hat{\lambda}^\top \boldsymbol{x}_i$,
  - méthode exponentielle : $F(\hat{\lambda}^\top \boldsymbol{x}_i) = \exp(\hat{\lambda}^\top \boldsymbol{x}_i)$.

Définition : Estimateur par calage généralisé (Deville, 1998)

$$\hat{t}_C = \sum_{i \in s} w_i y_i,$$

- Poids de calage : $w_i = d_i \times F\left(\hat{\lambda}^\top z_i\right)$,
- z_i vecteur des instruments de calage, connu pour $i \in s$,
- $F(.)$ fonction de calage,
- Équations de calage :

$$\sum_{i \in s} d_i F\left(\hat{\lambda}^\top z_i\right) x_i = \mathbf{t_x},$$

x_i est le vecteur des variables de calage
t_x est le vecteur des marges de calage.

Remarque importante : il n'est pas nécessaire de connaître t_z.

Les variables explicatives de la non-réponse et de la variable d'intérêt y

- On s'intéresse à **une** variable d'intérêt y_{j0} qu'on note y pour simplifier
- $\boldsymbol{z}$: vecteur de Q variables explicatives de la non-réponse qui sont également liées à la variable d'intérêt y
  - $\boldsymbol{z}_i^\top = (z_{i,1}, \ldots, z_{i,Q})$ vecteur associé à l'unité $i \in U$
  - Nature des variables dans $\boldsymbol{z}$: des x_p^o, des x_p^* et des y_j !
  - Il suffit que $\boldsymbol{z}_i$ soit connu pour les unités $i \in s_r$ et on n'est pas obligé de connaitre les totaux sur la population $t_{z_{i,q}}$!

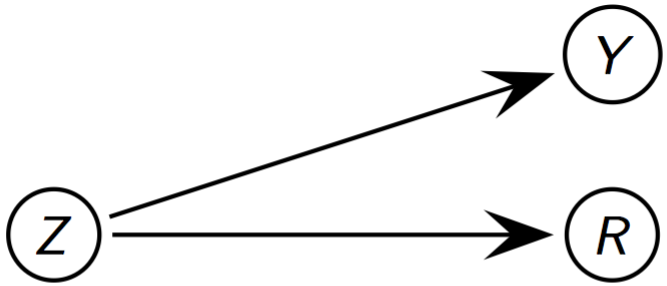


Figure: Relation entre les variables Y, Z and R

Les variables explicatives de la non-réponse et de la variable d'intérêt y

- La variable y n'a pas d'effet propre dans le modèle de non-réponse après avoir tenu compte des variables $\boldsymbol{z}$:

$$Y_i \perp R_i \mid \boldsymbol{Z}_i \tag{1}$$

- On peut dire que y est exclue des variables explicatives de la non-réponse et la relation (1) est appelée relation d'exclusion de y pour le modèle de non-réponse
- Modèle de non réponse considéré :

$$\mathbb{E}(R_i \mid \boldsymbol{Z}_i,\ Y_i,\ I_i = 1) = \mathbb{E}(R_i \mid \boldsymbol{Z}_i,\ I_i = 1) = h(\gamma\ ;\ \boldsymbol{Z}_i) > 0,$$

où γ est un vecteur de même dimension que $\boldsymbol{z}$ et $h(\cdot)$ une fonction suffisamment régulière.

- $p_i = h(\gamma\ ;\ \boldsymbol{Z}_i)$: probabilité de répondre de l'unité $i \in s_r$

Estimation des probabilités de réponse p_i

- Rappel : dans l'approche de repondération en deux étapes pour traiter la non-réponse
  - Le mécanisme de non-réponse est traité comme une ultime phase de sondage
  - $\hat{p}_i = h(\hat{\gamma}\ ;\ \boldsymbol{Z}_i)$: estimateur de p_i, où $\hat{\gamma}$ est un estimateur convergent de γ
  - "*Propensity score adjusted estimator*"

$$\hat{t}_{PSA} = \sum_{i \in s_r} \pi_i^{-1}\ \hat{p}_i^{-1}\ y_i$$

- On va montrer
  - on peut estimer les probabilités de réponse p_i à partir d'un système d'équations estimantes qui sont similaires à des équations de calage généralisées
  - l'estimateur par calage généralisé est un estimateur de type PSA.

Estimation de γ

- A partir de notre modèle de non-réponse, on a :

$$\mathbb{E}(I_i\ R_i \mid \boldsymbol{Z}_i,\ Y_i) = \mathbb{E}(I_i \mid \boldsymbol{Z}_i, ; Y_i)\ h(\gamma\ ;\ \boldsymbol{Z}_i)$$

- Sans perte de généralité on peut écrire

$$\mathbb{E}(I_i\ R_i \mid \boldsymbol{Z}_i,\ Y_i) = \pi_i\ h(\gamma\ ;\ \boldsymbol{Z}_i)$$

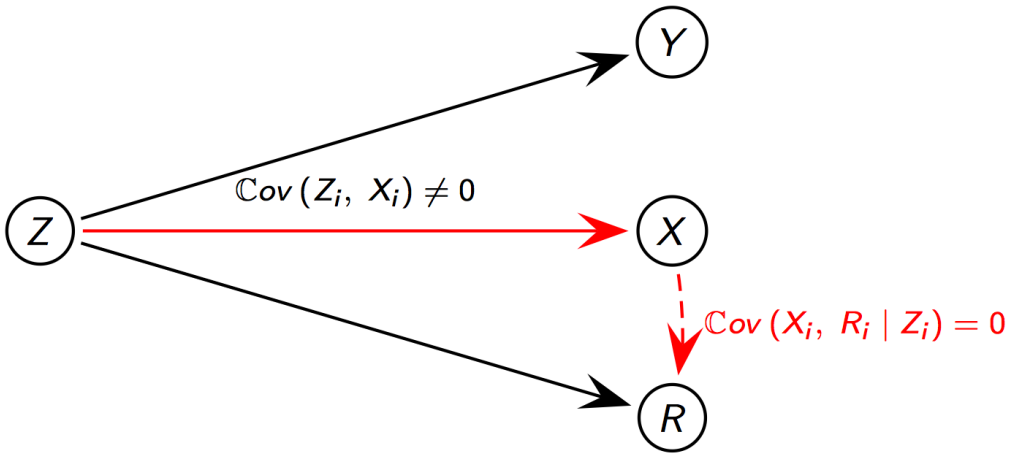
- Deux types d'équations de moments du modèle de non-réponse
  1. $\mathbb{E}(\pi_i^{-1}\ h^{-1}(\gamma\ ;\ \boldsymbol{z}_i)\ \boldsymbol{z}_i\ I_i\ R_i) = \mathbb{E}(\boldsymbol{z}_i)$
  2. $\mathbb{E}(\{R_i - \pi_i\ h(\gamma\ ;\ \boldsymbol{z}_i)\}\ \boldsymbol{z}_i\ I_i\ R_i) = \mathbf{0}$
- Deux types d'équations estimantes (contre partie empirique)
  1. $\sum_{i \in s_r} \pi_i^{-1}\ h^{-1}(\hat{\gamma}\ ;\ \boldsymbol{z}_i)\ \boldsymbol{z}_i = \boldsymbol{t_z}$
  2. $\sum_{i \in s} \{R_i - \pi_i\ h(\hat{\gamma}\ ;\ \boldsymbol{z}_i)\}\ \boldsymbol{z}_i = \mathbf{0}$

Estimation de γ : problème de covariables inobservées

- Si $z_{i,q}$ n'est pas une variable auxiliaire (x^o ou x^*) on ne peut résoudre aucune des 2 équations estimantes précédentes.
- La solution est de trouver une variable auxiliaire, $x_{i,q}$, qui serve de proxy pour $z_{i,q}$.
- C'est une solution semblable à celle utilisée pour contourner le problème d'endogénéité en économétrie...
- ... toutefois, notre problème de départ n'est pas la présence d'une variable endogène mais d'une variable explicative $z_{i,q}$ partiellement observée.
- On verra que, comme en économétrie, l'utilisation d'une variable instrumentale (notre proxy) fait peser le risque d'avoir un estimateur dont le biais est amplifié par la faible corrélation entre la variable explicative $z_{i,q}$ et sa proxy $x_{i,q}$.

Variable proxy de z

- Considérons simplement que
  - $\boldsymbol{z}_i^\top = (1,\ z_i)$
- Soit x la variable proxy de z
  - Variable auxiliaire x^* (ou x^0, on utilise alors $\hat{t}_{x,\pi}$ à la place de t_x)
  - Corrélée à z
  - x n'intervient pas dans le modèle de non-réponse : $X \perp R \mid \boldsymbol{Z}$



Estimation de γ : équations de moment avec variable instrumentale

- Modèle de non-réponse

$$\mathbb{E}(I_i \, R_i \mid \boldsymbol{Z}_i, \, X_i) = \pi_i \, h(\gamma \, ; \, \boldsymbol{Z}_i),$$

  - $\boldsymbol{x}_i^\top = (1, \, x_i)$ et $\boldsymbol{t_x}^\top = (N, t_x)$

- Nouvelles équations de moments

$$\mathbb{E}(\pi_i^{-1} \, h^{-1}(\gamma \, ; \, \boldsymbol{z}_i) \, I_i \, R_i \, \boldsymbol{x}_i) = \mathbb{E}(\boldsymbol{x}_i)$$

- Équations estimantes (contre partie empirique)

$$\sum_{i \in s_r} \pi_i^{-1} \, h^{-1}(\hat{\gamma} \, ; \, \boldsymbol{z}_i) \, \boldsymbol{x}_i = \boldsymbol{t_x}$$

Lien avec le calage généralisé

- Le système précédent correspond à un estimateur par calage généralisé où
  - $i \in s_r$, échantillon des répondants
  - $\boldsymbol{x}$: vecteur des variables de calage
  - $\boldsymbol{t_x}$: vecteur des marges de calage
  - $\boldsymbol{z}$: vecteur des instruments de calage
  - Le vecteur des paramètres γ correspond au vecteur $\boldsymbol{\lambda}$
  - $F(\boldsymbol{\lambda}^\top \boldsymbol{Z}_i) = h^{-1}(\boldsymbol{\lambda}\ ;\ \boldsymbol{Z}_i)$ où $F(\cdot)$ est une fonction de calage (monotone et continûment dérivable)
- La recherche de l'estimateur $\hat{t}_{PSA} = \sum_{i \in s_r} \pi_i^{-1}\, \hat{p}_i^{-1}\, y_i$ nous amène dans ce cas à l'estimateur par calage généralisé $\hat{t}_C = \sum_{i \in s_r} \pi_i^{-1}\, F(\hat{\boldsymbol{\lambda}}^\top \boldsymbol{z}_i)\, y_i$
- Voyons maintenant les propriétés de l'estimateur $\hat{t}_C$...

Convergence de l'estimateur par calage généralisé

$$\hat{t}_C = \sum_{i \in s_r} \pi_i^{-1} \, F(\hat{\lambda}^\top \boldsymbol{z}_i) \, y_i$$

- Sous les conditions
  - √ Modèle de non réponse : $\mathbb{E}(R_i \mid \boldsymbol{Z}_i, \; I_i = 1) = F^{-1}(\gamma^\top \boldsymbol{Z}_i) > 0$,
  - √ Relation d'exclusion sur y : $Y \perp R \mid \boldsymbol{Z}$
  - √ Relation d'exclusion sur x : $X \perp R \mid \boldsymbol{Z}$
    - et quelques conditions techniques
- On a :
  1. $\hat{\lambda} \xrightarrow{P} \gamma$
  2. $F(\hat{\lambda}^\top \boldsymbol{z}_i) p_i - 1 \xrightarrow{P} 0$
  3. $(\hat{t}_C - t_y)/N \xrightarrow{P} 0.$
- Remarque : ces résultats sont vrais quelque soit le modèle qui lie y et $\boldsymbol{Z}$!
- Variance amplifiée par la faible corrélation entre z et x (Osier, 2012)

Non convergence de l'estimateur par calage généralisé

- Sous les conditions
  - ✓ Modèle de non réponse : $\mathbb{E}(R_i \mid \boldsymbol{Z}_i,\ I_i = 1) = F^{-1}(\boldsymbol{\gamma}^\top \boldsymbol{Z}_i) > 0,$
  - ✓ Relation d'exclusion sur y : $Y \perp R \mid \boldsymbol{Z}$
  - X Relation d'exclusion sur x non vérifiée : $X \not\perp R \mid \boldsymbol{Z}$
    en l'occurrence :

$$\mathbb{E}(R_i \mid \boldsymbol{Z}_i,\ \boldsymbol{X}_i,\ I_i = 1) \neq \mathbb{E}(R_i \mid \boldsymbol{Z}_i,\ I_i = 1)$$

- On obtient :
  - $F(\hat{\boldsymbol{\lambda}}^\top \boldsymbol{z}_i)p_i - 1$ ne converge pas vers 0
  - L'estimateur $\hat{t}_C$ n'est pas convergent
  - Le biais est amplifié par la faible corrélation entre z et x

Seconde expression du biais

Proposition :

Le biais approché de $\hat{t}_C$ peut s'écrire :

$$\begin{aligned} Biais(\hat{t}_C) &\approx -\sum_{k\in\mathcal{U}}(1-p_iF_i)\left(y_i - \boldsymbol{z}_i^\top \boldsymbol{\beta}_{pf}\right) \\ &+ \frac{\beta_{pf,1}}{\alpha_{pf,1}} \times \sum_{k\in\mathcal{U}}(1-p_iF_i)\left(x_i - \boldsymbol{z}_i^\top \boldsymbol{\alpha}_{pf}\right), \end{aligned}$$

- $\left(y_i - \boldsymbol{z}_i^\top \boldsymbol{\beta}_{pf}\right)$ résidu de la régression de y sur $\boldsymbol{z}$
- $\left(x_i - \boldsymbol{z}_i^\top \boldsymbol{\alpha}_{pf}\right)$ résidu de la régression de x sur $\boldsymbol{z}$
- $\alpha_{pf,1}$ corrélation entre x et z

Amplification de la variance de l'estimateur $\hat{t}_C$

Proposition :

La variance liée à la non-réponse approchée de $\hat{t}_C$ est :

$$\begin{aligned} A\mathbb{V}ar_q\left(\hat{t}_C\right) &= \sum_{k\in\mathcal{U}} p_i(1-p_i)F_i^2 \\ &\times \left\{\left(y_i - \boldsymbol{z}_i^\top \boldsymbol{\beta}_{pf}\right) - \frac{\beta_{pf,1}}{\alpha_{pf,1}}\left(x_i - \boldsymbol{z}_i^\top \boldsymbol{\alpha}_{pf}\right)\right\}^2. \end{aligned}$$

- Variance due à la non-réponse est faible si :
  - les résidus des modèles de régression $\left(y_i - \boldsymbol{z}_i^\top \boldsymbol{\beta}_{pf}\right)$ et $\left(x_i - \boldsymbol{z}_i^\top \boldsymbol{\alpha}_{pf}\right)$ sont petits,
  - $\alpha_{pf,1}$ est fort.
- Variance due à la non-réponse instable si : $\alpha_{pf,1}$ faible (i.e., une faible corrélation entre x et z).

Cas particulier des modèles de régression linéaire entre y et le vecteur $\boldsymbol{z}$ et entre x et $\boldsymbol{z}$

- Convergence de l'estimateur par calage généralisé si :
  - X ~~Modèle de non réponse : $\mathbb{E}(R_i \mid \boldsymbol{Z}_i,\ I_i = 1) = F^{-1}(\gamma^\top \boldsymbol{Z}_i) > 0$,~~
  - √ Relation d'exclusion sur y : $Y \perp R \mid \boldsymbol{Z}$
  - √ Relation d'exclusion sur x : $X \perp R \mid \boldsymbol{Z}$
  - √ Modèle de régression linéaire entre y et $\boldsymbol{z}$

$$\mathbb{E}(Y_i \mid Z_i) = \boldsymbol{\beta}^\top Z_i$$

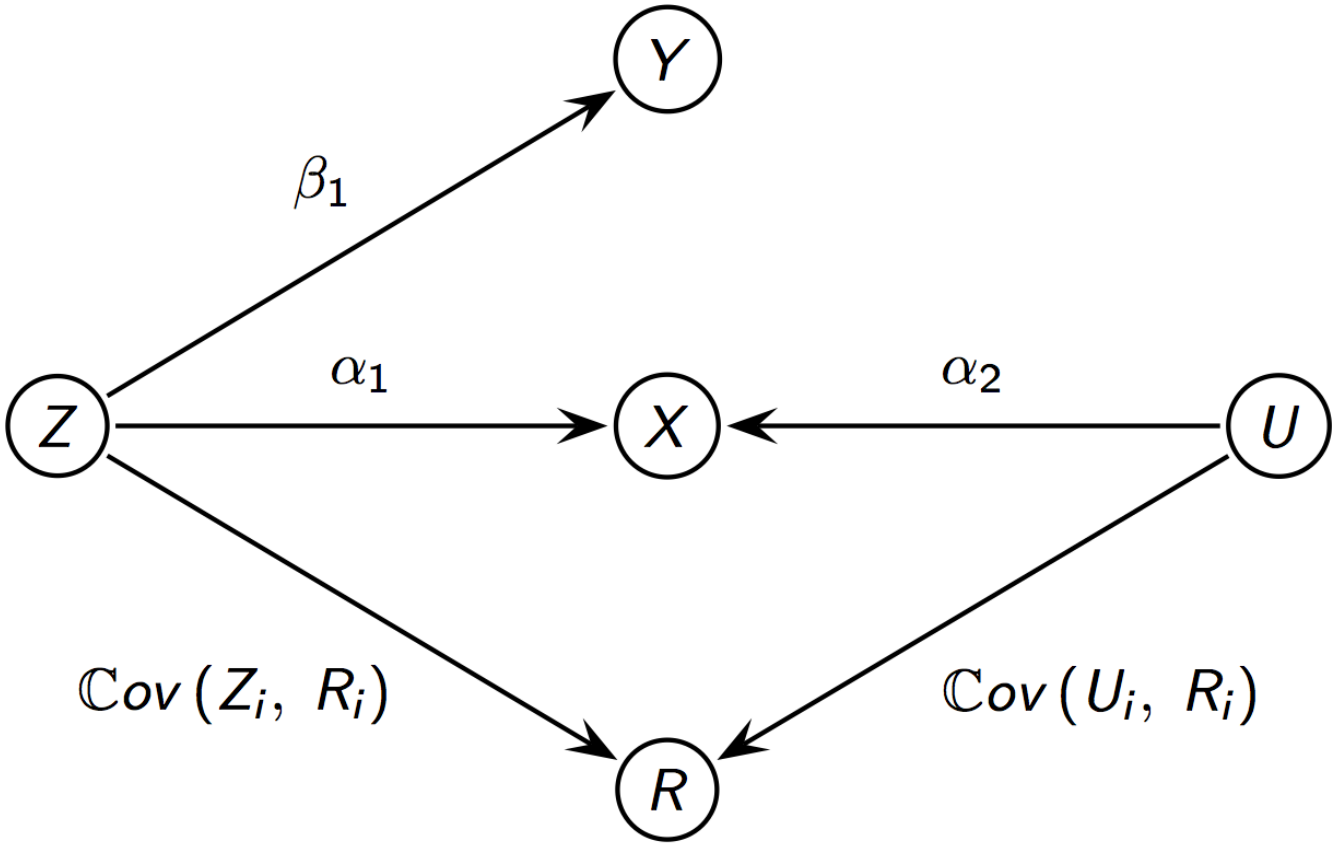
  - √ Modèle de régression linéaire entre x et $\boldsymbol{z}$

$$\mathbb{E}(Y_i \mid Z_i) = \boldsymbol{\alpha}^\top Z_i$$

Étude par simulation

On génère une population finie de taille $N = 1\ 000$

- instrument de calage z
- variable inobservée u.
- z et u : générées selon une loi uniforme sur $[-\sqrt{3}, \sqrt{3}]$.
- variable d'intérêt y,
  - $y_i = 10 + 2z_i + \varepsilon_i^y$,
  - où $\varepsilon_i^y \sim \mathcal{N}(0, 1)$, $R^2 \approx 80\%$.
- Jeu de variables proxy $x^{(\alpha_1, \alpha_2)}$
  - $X_i^{(\alpha_1, \alpha_2)} = \alpha_1 z_i + \alpha_2 u_i + \sigma_{(\alpha_1, \alpha_2)} \varepsilon_i^{(\alpha_1, \alpha_2)}$,
  - où $\varepsilon^{(\alpha_1, \alpha_2)} \sim \mathcal{N}(0, 1)$,
  - $\alpha_1 \in \{0.2, 0.3, 0.5, 0.7\}$,
  - $\alpha_2 \in \{0, 0.1, 0.3, 0.5\}$.


Y
β_1
α_1
α_2
Z
X
U
$\mathbb{C}ov(Z_i, R_i)$
$\mathbb{C}ov(U_i, R_i)$
R

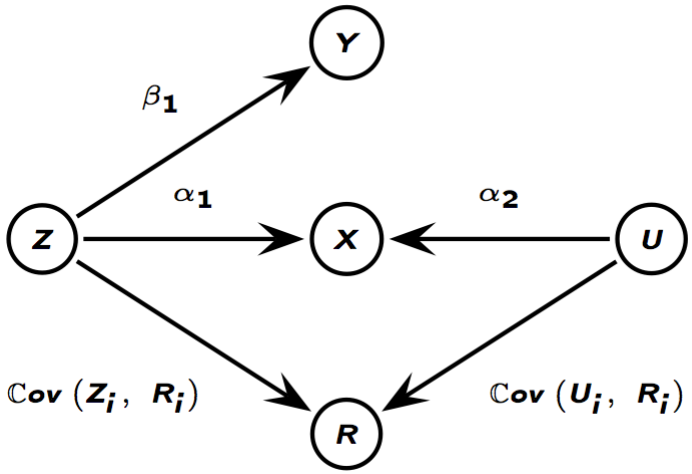
Étude par simulation

- On considère un recensement $n = N = 1\ 000$.
- Probabilité de répondre de l'unité i :
  - $\mathrm{logit}(\mathrm{p_i}) = 1.5\mathrm{z_i} + \mathrm{u_i}$
  - Taux de réponse moyen : 50%
- R_i pour $i \in U$ sont générées indépendamment selon des lois de Bernoulli de paramètre p_i.
- Ce processus est répété $K = 10\ 000$ fois, menant à $K = 10\ 000$ ensembles de répondants.
- On calcule l'estimateur par calage $\hat{t}_C(\alpha_1, \alpha_2)$ avec **la méthode linéaire**, pour les différentes valeurs de α_1 et α_2.

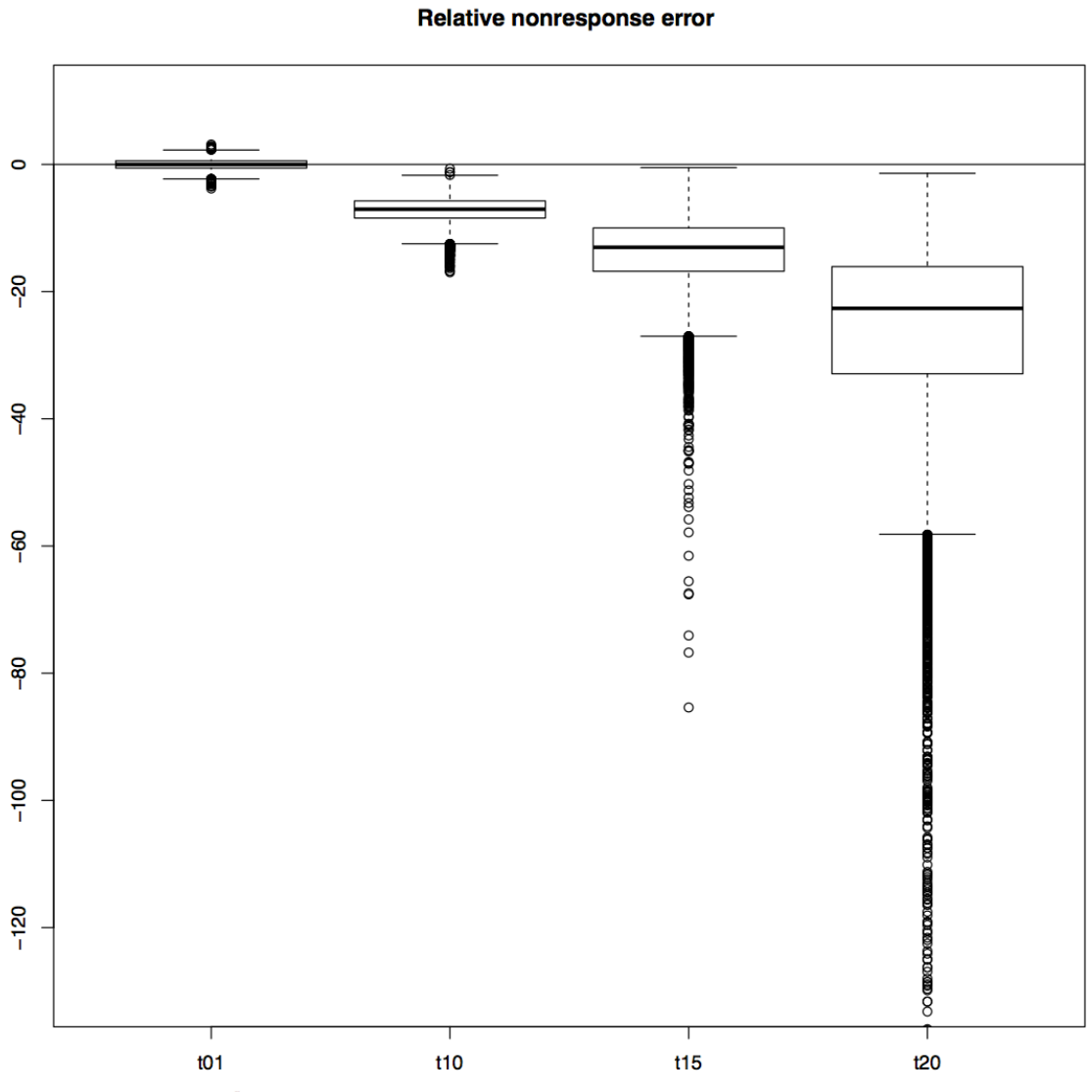
Résultats des simulations

α_1	$\alpha_2 = 0$	$\alpha_2 = 0.1$	$\alpha_2 = 0.3$	$\alpha_2 = 0.5$
0.7	0.02	−0.9	−2.8	−4.9
	(0.9)	(0.9)	(1.0)	(1.1)
0.5	−0.1	−1.3	−4.1	−7.2
	(1.4)	(1.5)	(1.7)	(2.1)
0.3	−0.2	−2.4	−7.5	−14.0
	(2.6)	(3.0)	(4.1)	(5.9)
0.2	−0.6	−4.5	−13.8	−27.4
	(4.6)	(15.6)	(61.9)	(65.6)

Table: Biais relatif Monte Carlo et CV Monte Carlo (en %)



Boxplot des erreurs relatives (en %) pour différentes paires (α_1, α_2)



Seconde étude par simulation

- On considère que la variable liée à la non-réponse est la variable d'intérêt y.
- La variable proxy $x^{(\alpha_1,\alpha_2)}$ est généré selon le modèle

$$X_i^{(\alpha_1,\alpha_2)} = \alpha_1 \mathbb{V}ar(y)^{-1}(y_i - \beta_0) + \alpha_2 u_i + \sigma_{(\alpha_1,\alpha_2)}\varepsilon_i^{(\alpha_1,\alpha_2)}$$

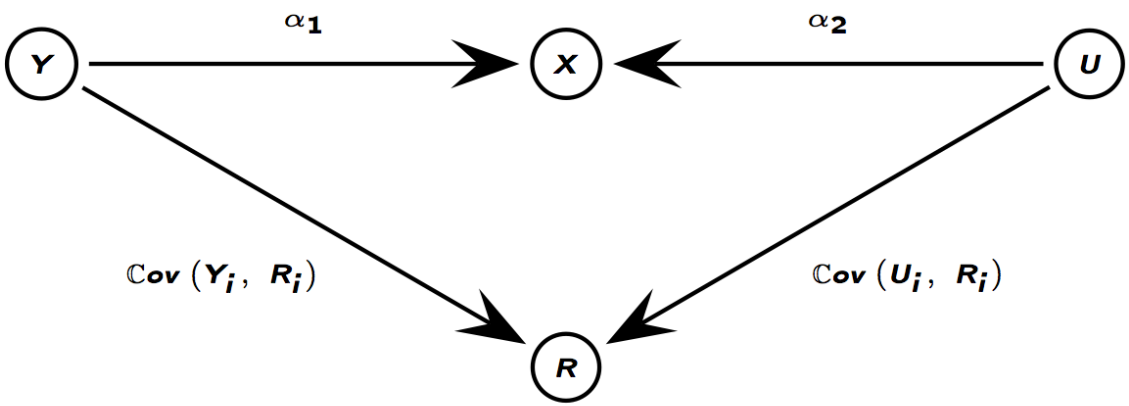
- Probabilité de répondre de l'unité i :

$$\text{logit}(p_i) = -5 + 0.5y_i + u_i.$$

- Taux de réponse moyen : 50%
- On calcule $\hat{t}_C(\alpha_1, \alpha_2)$ avec $z = y$.

Résultats des simulations

$\alpha_1\alpha_2$	0	0.1	0.3	0.5
0.7	-0.01 (0.8)	-1.0 (0.8)	-3.3 (0.9)	-5.7 (1.0)
0.5	-0.1 (1.4)	-1.5 (1.5)	-4.8 (1.7)	-8.3 (2.0)
0.3	-0.2 (2.6)	-2.7 (2.9)	-8.7 (4.0)	-15.5 (5.5)
0.2	-0.5 (5.0)	-4.7 (7.1)	-15.0 (25.2)	-30.2 (39.1)



Conclusion

- L'utilisation d'un estimateur par calage généralisé pour traiter la non-réponse requiert une bonne connaissance du modèle de non-réponse i.e. :
  1. des variables "explicatives" de la non-réponse, $\boldsymbol{z}$;
  2. de la forme de la fonction du modèle linéaire généralisée (qui permet de choisir la fonction de calage).

$$\mathbb{E}(R_i \mid \boldsymbol{Z}_i) = F^{-1}(\boldsymbol{\lambda}^\top \boldsymbol{Z}_i)$$

- et des variables instrumentales $\boldsymbol{x}$ du modèle de non-réponse qui peuvent servir de proxy pour $\boldsymbol{z}$:
  1. qui n'interviennent pas dans le modèle de non-réponse (absence de biais)

$$\mathbb{E}(R_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i) = \mathbb{E}(R_i \mid \boldsymbol{Z}_i)$$

  2. et qui soient fortement corrélées aux variables instrumentales (non-amplification du biais et de la variance).

Merci de votre attention.

bibliographie partielle I

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976).
Some results on generalized difference estimation and generalized regression estimation for finite populations.
Biometrika, 63(3), 615-620.

Chang, T. and Kott, P. (2008).
Using calibration weighting to adjust for nonresponse under a plausible model.
Biometrika, 95(3) :555–571.

Deville, J.-C. and Särndal, C.-E. (1992).
Calibration estimators in survey sampling.
Journal of the American Statistical Association, 87(418) :376–382.

Deville, J.-C. (1998).
La correction de la non-réponse par calage ou par échantillonnage équilibré.
In Actes du colloque de la Société Statistique du Canada, Sherbrooke, Canada.

Deville, J.-C. (2004).
La correction de la non-réponse par calage généralisé.
Actes des journées de méthodologie statistique, pages 4–20.

Dupont, F. (1996).
Calage et redressement de la non-réponse totale.
In Actes des journées de méthodologie statistique, 15 et 16 décembre 1993, number 56.
INSEE-Méthodes.

bibliographie partielle II

Kott, P. S. (2006).
Using calibration weighting to adjust for nonresponse and coverage errors.
Survey Methodology, 32(2) :133.

Kott, P. S. and Chang, T. (2010).
Using calibration weighting to adjust for nonignorable unit nonresponse.
Journal of the American Statistical Association, 105(491) :1265–1275.

Kott, P. S. and Liao, D. (2012).
Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine.
In Survey Research Methods, volume 6, pages 105–111.

Lesage, E. (2012).
Correction de la non-réponse non ignorable par une approche modèle.
In Actes des Journées de Méthodologie Statistique de l'INSEE.

Le Guennec, J. et Sautory, O. (2004).
Correction de la non-réponse par calage généralisé : une expérimentation.
Actes des journées de méthodologie statistique, 16 et 17 décembre 2002

Lundström, S. and Särndal, C.-E. (1999).
Calibration as a standard method for treatment of nonresponse.
Journal of Offcial Statistics, 15(2) :305–327.

bibliographie partielle III

Osier, G.

Dealing with non-ignorable non-response using generalised calibration : A simulation study based on the luxemburgish household budget survey.

Economie et Statistiques, Working papers du STATEC, (65).

Sautory, O. (2003).

Calmar 2 : une nouvelle version du programme calmar de redressement d'échantillon par calage.

In Recueil : Symposium de Statistique Canada.

Särndal, C.-E. (2011).

Three factors to signal non-response bias with applications to categorical auxiliary variables.

International Statistical Review, 79(2) :233–254.

Särndal, C.-E. and Lundström, S. (2010).

Design for estimation : Identifying auxiliary vectors to reduce nonresponse bias.

Survey Methodology, 36 :131–144.

Särndal, C.-E. and Lundström, S.(2005).

Estimation in surveys with nonresponse.

Wiley Hoboken, NJ.