Discussion sur l'utilisation du calage en présence de non-réponse

Éric Lesage

Insee, Division Etudes Territoriales

15 Mars 2016

Séminaire de **M**éthodologie **S**tatistique
Miscellanées sur le calage
Insee, PARIS

Travail réalisé en collaboration avec David Haziza
Département de mathématiques et de statistique, Université de Montréal

Plan de la présentation

Cette présentation correspond à un travail réalisé en collaboration avec David Haziza et publié dans la revue Journal of Official Statistics (Haziza et Lesage, 2016).

Sommaire :

- La non-réponse dans l'enquête
- L'approche de pondération en deux étapes pour traiter la non-réponse
- L'approche de pondération en une étape pour traiter la non-réponse
- Les résultats d'une étude par simulation
- Le lien entre le modèle de réponse et les équations de calage

Pondération de base en présence de données complètes

- 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

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

- π_i : probabilité d'inclusion de l'unité i,
- $d_i = \pi_i^{-1}$: poids d'échantillonnage
- Lorsqu'on applique le système de pondération de base à une variable y, on obtient l'estimateur de Horvitz-Thompson :

$$\hat{t}_{y\pi} = \sum_{i \in s} d_i y_i = \sum_{i \in s} \frac{y_i}{\pi_i}$$

Non-réponse

- Une partie des personnes échantillonnées ne répond pas (non réponse totale)
  - $s_r \subset s$: sous ensemble des répondants
  - R_i : variable indicatrice de réponse
- Mécanisme de réponse :
  - On fait l'hypothèse que cette non-réponse n'est pas liée directement aux variables d'intérêt
  - Cas MAR (Missing At Random conditionally to $\boldsymbol{Z}$)
  - Il existe des variables explicatives de la non-réponse, notées $\boldsymbol{z}$, qui sont également liées à des variables d'intérêt

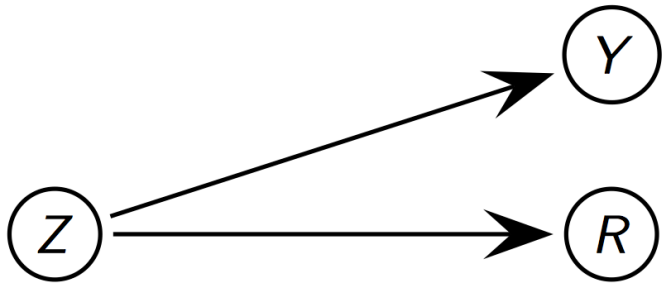


Figure: Relation entre les variables Y, Z and R

Pondération en deux étapes

$$\underbrace{d_i}_{\text{Poids de base}} \longrightarrow \underbrace{\widetilde{w}_i}_{\text{(1) Poids ajustés pour la non-réponse}} \longrightarrow \underbrace{w_i}_{\text{(2) Poids calés (ou finaux)}}$$

- Approche traditionnelle dans les agences statistiques
- Première étape :
  - Traiter la non-réponse.
  - Assimiler le mécanisme de non réponse à une seconde phase d'échantillonnage avec un tirage de Poisson de paramètres $(p_1, \ldots, p_N)$ liés à la propension à répondre des individus.
  - Estimer les paramètres p_i à l'aide d'un modèle de réponse.
- Deuxième étape :
  - Calage
  - Garantir la cohérence
  - Réduire, si possible, la variance des estimateurs

Pondération en deux étapes : la première étape

- On postule un modèle de non-réponse :

$$\mathbb{E}(R_i \mid \boldsymbol{Y}_i = \boldsymbol{y}_i,\ \boldsymbol{Z}_i = \boldsymbol{z}_i) \quad = \quad f(\boldsymbol{z}_i, \gamma),$$

où

  - $\boldsymbol{z}_i$ est un vecteur de variables auxiliaires disponibles pour les répondants et les non-répondants ;
  - Exemples de variables z : paradonnées, variables sur la base de sondage
  - γ est un vecteur de paramètres inconnus.

- On définit les paramètres du tirage de Poisson de la façon suivante : $p_i = f(\boldsymbol{z}_i, \gamma)$ (probabilité de réponse conditionnelle de l'unité i).
- Système de pondération ajusté pour la non-réponse $\{\widetilde{w}_i ; i \in s_r\}$:

$$\widetilde{w}_i = d_i \times \frac{1}{\widehat{p}_i},$$

  - $\widehat{p}_i{}^{-1}$: facteur d'ajustement pour la non-réponse pour l'unité i.

Pondération en deux étapes : la première étape

- Lorsqu'on applique le système de pondération ajusté pour la non-réponse à une variable y, on obtient l'estimateur ajusté (Propensity Score Adjusted estimator)

$$\widehat{t}_{PSA} = \sum_{i \in s_r} \widetilde{w}_i y_i = \sum_{i \in s_r} \left(d_i \times \frac{1}{\widehat{p}_i} \right) y_i.$$

- Estimation paramétrique des p_i :
  - Repose sur des hypothèses fortes à propos de la forme de $f(\mathbf{z}_i; .)$.
  - Exemple : modèle de régression logistique

$$p_i = \left\{ 1 + \exp(\mathbf{z}_i^\top \boldsymbol{\gamma}) \right\}^{-1}$$

  - Presque jamais utilisée en pratique
    - Certaines probabilités estimées pourraient être très petites, ce qui conduirait à des facteurs $\widehat{p}_i^{-1}$ extrêmes $\rightarrow$ poids dispersés $\rightarrow$ estimateurs instables
    - Vulnérable à la mauvaise spécification du modèle paramétrique (la fonction $f(\mathbf{z}_i; .)$).

Estimation non-paramétrique des p_i

- Repose sur des hypothèses faibles à propos de la forme de $f(\mathbf{z}_i; .)$
  - Robuste à la non-inclusion de termes quadratiques, cubiques, ... et de variables croisées,...
- Les méthodes non-paramétriques
  - Méthodes par noyaux et polynômes locaux (Da Silva et Opsomer, 2006, 2009)
  - Arbres de régression (Phipps et Toth, 2012)
  - Groupes de réponse homogène résultant du croisement de variables qualitatives (post-strates)
  - Classes de repondération basées sur des probabilités de réponse estimées (méthode des scores) ; par exemple, Little (1986), Eltinge et Yansaneh (1997) et Haziza et Beaumont (2007).
    1. Obtenir des estimations préliminaires, $\tilde{p}_i$, des p_i en utilisant une régression logistique
    2. Diviser l'échantillon en classes de manière à ce que les unités dans une classe aient des $\tilde{p}_i$ approximativement égales
    3. Estimer p_i pour une unité dans une classe par le taux de réponse observé dans la même classe $\rightarrow \widehat{p}_i$
    4. Calculer les poids ajustés pour la non-réponse : $\widetilde{w}_i = d_i/\widehat{p}_i$.

Pondération en deux étapes : la première étape

- Choix du vecteur **z** :
  - Inclure dans le modèle de non-réponse les variables qui sont à la fois liées à la probabilité de réponse et aux variables d'intérêt
  - Si une variable z est liée à la probabilité de réponse mais pas liée aux variables d'intérêt, il n'est pas nécessaire de l'inclure dans le modèle de non-réponse car elle ne sera pas utile pour réduire le biais mais elle risque de contribuer à faire augmenter la variance de $\widehat{t}_{PSA}$.
- Si le modèle de non-réponse est bien spécifié, alors

$$\widehat{t}_{PSA} = \sum_{i \in s_r} \widetilde{w}_i y_i = \sum_{i \in s_r} \left(d_i \times \frac{1}{\widehat{p}_i} \right) y_i$$

est asymptotiquement sans biais, quelle que soit la variable y.

Pondération en deux étapes : la deuxième étape

- Étape de calage :
  - Soit $\mathbf{x}_i^* = (x_{1i}^*, \ldots, x_{Li}^*)^\top$ un vecteur de variables auxiliaires de dimension L associé à l'unité i.
  - On supposera que $\mathbf{x}_i^*$ est observée pour tout $i \in s_r$.
  - On supposera connu le vecteur des totaux au niveau de la population,

$$t_{\mathbf{x}^*} = \sum_{i \in U} \mathbf{x}_i^*.$$

- Objectif : déterminer un système de pondération calé (final), $\{w_i; i \in s_r\}$, qui satisfasse les équations de calage

$$\sum_{i \in s_r} w_i \mathbf{x}_i^* = \sum_{i \in U} \mathbf{x}_i^*.$$

- Poids initiaux : $\widetilde{w}_i$
- Poids finaux :

$$w_i = \widetilde{w}_i \times F(\widehat{\boldsymbol{\lambda}}^\top \mathbf{x}_i^*) = d_i \times \frac{1}{\widehat{p}_i} \times F(\widehat{\boldsymbol{\lambda}}^\top \mathbf{x}_i^*).$$

Pondération en deux étapes : la deuxième étape

Estimateur par calage :

$$\widehat{t}_{yC,2} = \sum_{i \in s_r} w_i y_i$$

- Asymptotiquement sans biais quelle que soit la variable y si $\widehat{t}_{PSA}$ asymptotiquement est sans biais.
- Toutes les fonctions de distance sont asymptotiquement équivalentes
- Comment choisir la fonction de calage $F(.)$ et les variables de calage $\mathbf{x}_i^*$?

Pondération en une étape

$$\underbrace{d_i}_{\text{Poids de base}} \longrightarrow \underbrace{w_i}_{\text{Poids calés (ou finaux)}}$$

- Trois objectifs simultanés :
  1. Assurer la cohérence entre les estimations issues d'une enquête et les totaux connus au niveau de la population
  2. Améliorer la précision des estimateurs
  3. Réduire les erreurs non dues à l'échantillonnage (erreurs de non-réponse et erreurs de couverture)

- Bibliographie : Dupont (1993), Deville et Dupont (1993), Lundström et Särndal (1999) et Särndal et Lundström (2005)

Pondération en une étape

- Approche en une étape :

$$\underbrace{d_i}_{\text{Poids de base}} \longrightarrow \underbrace{w_i}_{\text{Poids calés (ou finaux)}}$$

- Lundström et Särndal (1999) :

  "No explicit response model is needed, in contrast of the conventional methods"

- Peut-on réellement faire l'économie d'une modélisation des probabilités de réponse ?

La nature des variables de calage en présence de non-réponse

- Info-U : un vecteur d'information auxiliaire $\mathbf{x}_i^*$ de taille J^* est disponible pour $i \in s_r$ (au moins) et le vecteur des totaux, $t_{\mathbf{x}^*} = \sum_{i \in U} \mathbf{x}_i^*$, est connu.
- Info-s : un vecteur d'information auxiliaire $\mathbf{x}_i^o$ de taille J^o pour $i \in s$ (répondantes et non-répondantes). Seul le vecteur des totaux estimés au niveau de l'échantillon s, $\widehat{t}_{\mathbf{x}^o \pi} = \sum_{i \in s} d_i \mathbf{x}_i^o$, est disponible.
- On distingue au final 3 niveaux d'information auxiliaire :

	x_i^o	$x_i^*(s)$	$x_i^*(s_r)$	y_i
$i \in s_r$	$\surd$	$\surd$	$\surd$	$\surd$
$i \in s$	$\surd$	$\surd$	X	X
$\hat{t}_{.,\pi}$	$\surd$	$\surd$	X	X
$t_.$	X	$\surd$	$\surd$	X

Pondération en une étape

- Vecteur $\mathbf{x}_i^*$: variables garantissant la cohérence (par ex., groupe d'âge, sexe, région) et pouvant être liées à la non-réponse
- Vecteur $\mathbf{x}_i^o$: variables liées à la non-réponse (par ex., paradonnées)
- Pour $i \in s_r$, on construit le vecteur

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_i^* \\ \mathbf{x}_i^o \end{pmatrix}$$

de taille $L = L^* + L^o$.

- On définit cette fois le vecteur des marges de calage par :

$$\boldsymbol{t}_{\mathbf{x}} = \begin{pmatrix} t_{\mathbf{x}^*} \\ \widehat{t}_{\mathbf{x}^o\pi} \end{pmatrix}.$$

Définition : Estimateur par calage en présence de non-réponse

$$\hat{t}_{yC,1} = \sum_{i \in s_r} w_i \, y_i$$

- Poids de calage

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

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

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

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

Pondération en une étape

- Avant de donner l'expression du biais, supposons qu'il n'y ait pas de variables $\mathbf{x}^*$.
- On compare deux estimateurs :

$$\widehat{t}_{PSA} = \sum_{i \in s_r} d_i \times \frac{1}{\widehat{p}_i} \times y_i \quad \text{et} \quad \widehat{t}_{yC,1} = \sum_{i \in s_r} d_i \times F(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_i) \times y_i$$

- Si on veut que $\widehat{t}_{yC,1}$ soit asymptotiquement sans biais, quelle que soit la variable y, il suffirait que

$$F(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_i) = \widehat{p}_i^{-1}.$$

Biais de non-réponse de $\widehat{t}_{yC,1}$:

$$Biais(\widehat{t}_{yC,1}) \approx -\sum_{i \in U} (1 - p_i F_i)\left(y_i - \mathbf{x}_i^\top \boldsymbol{B}_{p,f}\right),$$

où :

- $p_i = \mathbb{E}(R_i \mid \boldsymbol{Y}_i = \boldsymbol{y}_i,\ \boldsymbol{X}_i = \boldsymbol{x}_i)$
- $F_i \equiv F\left(\lambda_N^\top \mathbf{x}_i\right)$
  - $\boldsymbol{\lambda}_N$ est le vecteur de paramètres solution de : $\sum_{i \in U} F(\boldsymbol{\lambda}_N^\top \boldsymbol{x}_i)\, \boldsymbol{x}_i = \boldsymbol{t}_{\boldsymbol{x}}$.
  - $\hat{\boldsymbol{\lambda}}_r - \boldsymbol{\lambda}_N$ converge vers $\mathbf{0}$,
- $\boldsymbol{B}_{p,f} = \left(\sum_{i \in U} p_i f_i \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \left(\sum_{i \in U} p_i f_i \mathbf{x}_i y_i\right)$,
  - avec $f_i \equiv F'\left(\lambda_N^\top \mathbf{x}_i\right)$.

Biais de non-réponse de $\widehat{t}_{yC,1}$:

$$Biais(\widehat{t}_{yC,1}) \approx -\sum_{i \in U}(1 - p_i F_i)\left(y_i - \mathbf{x}_i^\top \boldsymbol{B}_{p,f}\right),$$

- Le biais est approximativement égal à 0 si :
  - les résidus $e_i = \left(y_i - \mathbf{x}_i^\top \boldsymbol{B}_{p,f}\right)$ ne sont pas liés à $\delta_i = p_i F_i - 1$.
    Par exemple, cette condition est satisfaite si on a un modèle de régression linéaire entre y et $\boldsymbol{x}$:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \qquad \mathbb{E}(\epsilon_i | \mathbf{x}_i) = 0,$$

  - ou si $F(\hat{\boldsymbol{\lambda}}_r^\top \boldsymbol{x}_i)^{-1}$ est un estimateur convergent de p_i :

$$F_i^{-1} = p_i.$$

Correspondance entre la fonction de calage et le modèle de réponse

Quand peut-on avoir $F_i = p_i^{-1}$?

- Méthode linéaire :

$$F(\boldsymbol{\lambda}^\top \mathbf{x}_i) = 1 + \boldsymbol{\lambda}^\top \mathbf{x}_i \qquad \rightarrow \qquad p_i^{-1} = 1 + \boldsymbol{\lambda}^\top \mathbf{x}_i$$

- Méthode exponentielle :

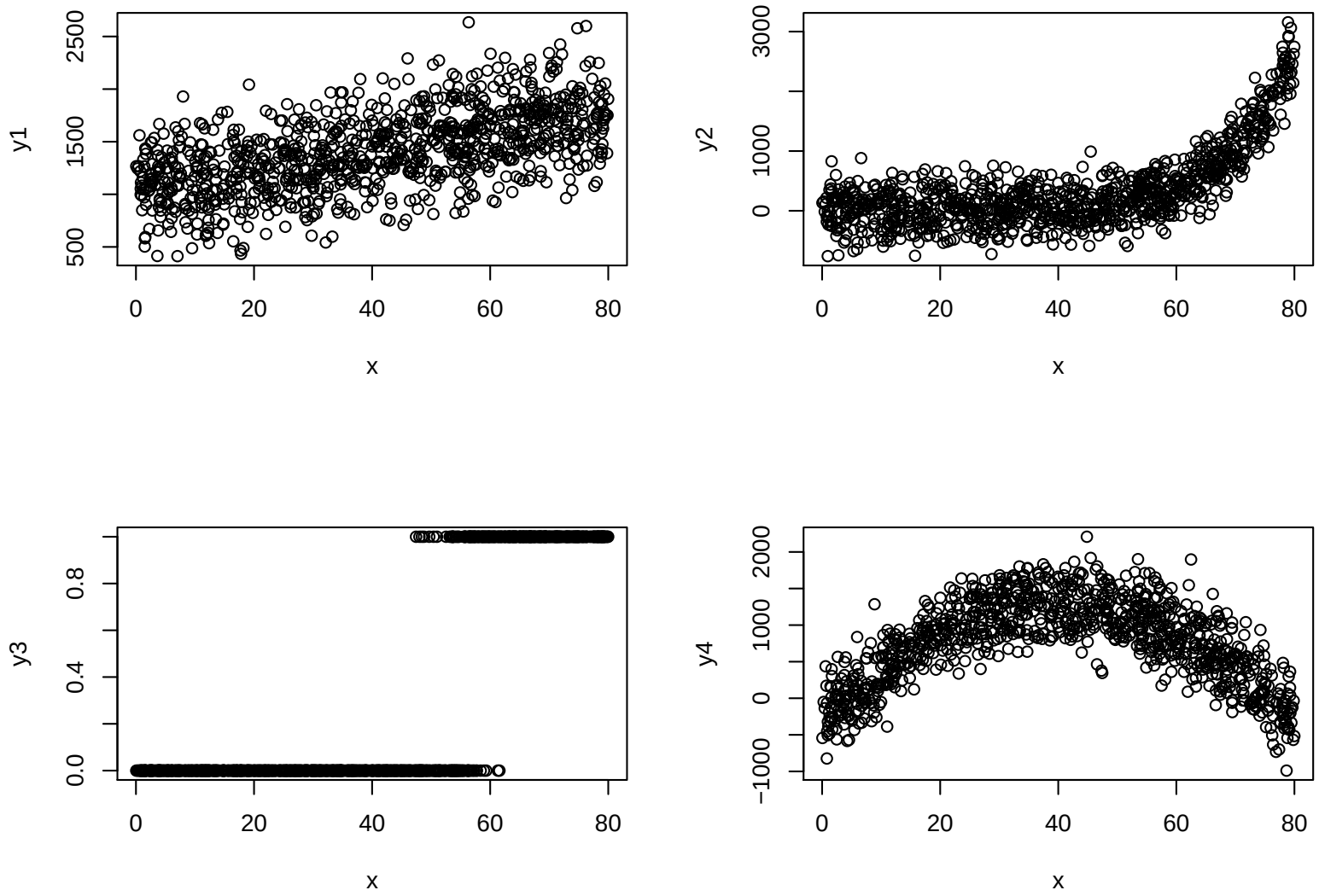
$$F(\boldsymbol{\lambda}^\top \mathbf{x}_i) = \exp(\boldsymbol{\lambda}^\top \mathbf{x}_i) \qquad \rightarrow \qquad p_i^{-1} = \exp(\boldsymbol{\lambda}^\top \mathbf{x}_i)$$

- En choisissant une fonction de calage, on postule (implicitement) une relation entre la probabilité de réponse et le vecteur de variables auxiliaires $\mathbf{x}$.
- Qu'en est-il si la fonction de calage ne correspond pas au modèle de réponse ? Le biais peut-il être important ?

Étude par simulation

- On a généré une population de taille $N = 1\ 000$, avec une variable auxiliaire x et 4 variables d'intérêt y_1, y_2, y_3 et y_4.
- Les valeurs de x ont été générées à partir d'une loi uniforme $[0, 80]$.
- Relation entre les variable d'intérêt et x :
  - y_1 et x : linéaire
  - y_2 et x : exponentielle
  - y_3 et x : logistique
  - y_4 et x : quadratique
- Cas de recensement ; i.e., $n = N = 1000$.

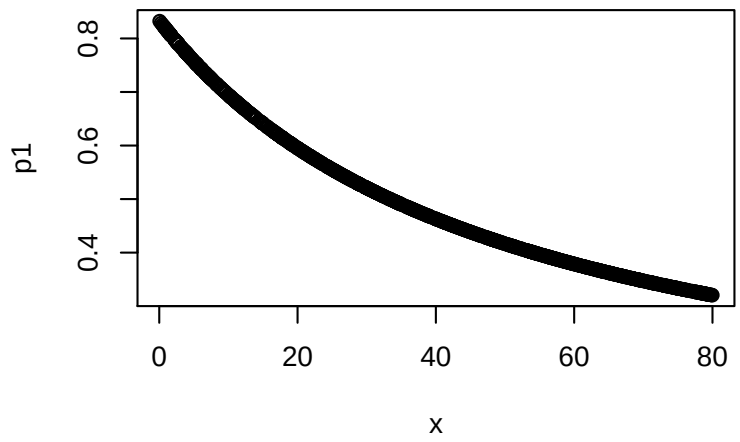
Relation entre les variables d'intérêt et x

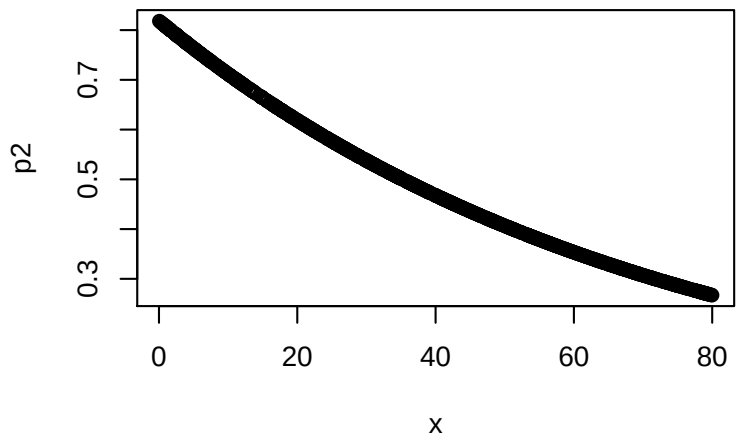


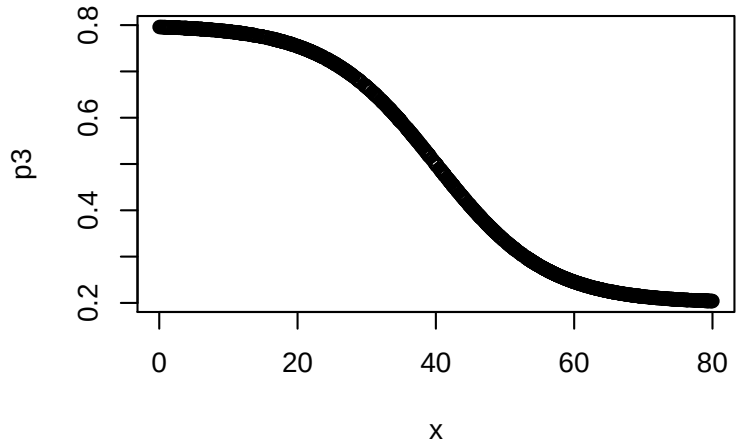
Étude par simulation

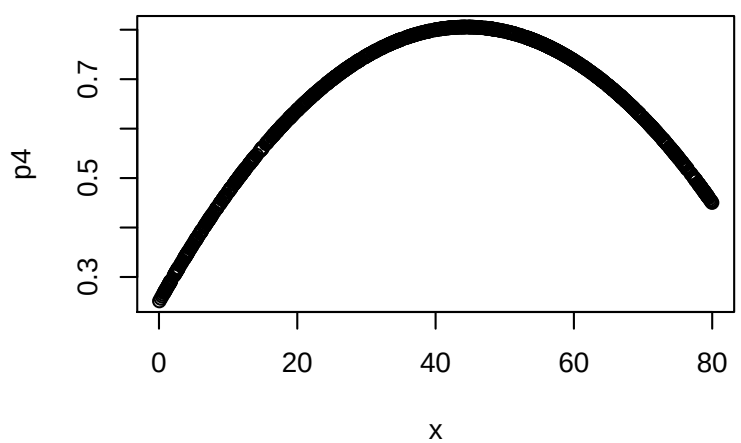
- Dans chaque population, on a généré de la non-réponse à la variable y selon 4 mécanismes de non-réponse
- Taux global de réponse : 50%.
  - p_1 : inverse linéaire $\rightarrow p_i = (1.2 + 0.024\ x_i)^{-1}$.
  - p_2 : exponentielle $\rightarrow p_i = \exp(-0.2 - 0.014 x_i)$.
  - p_3 : logistique $\rightarrow p_i = 0.2 + 0.6\ \{1 + \exp(-5 + x_i/8)\}^{-1}$.
  - p_4 : quadratique $\rightarrow 0.7 + 0.45\ (x_i/40 - 1)^2 + 0.0025\ x_i$.
- $K = 5\ 000$ itérations

Mécanismes de non-réponse









Étude par simulation

- Objectif : estimer t_{y1}, t_{y2}, t_{y3} et t_{y4}.
- On a calculé trois estimateurs de t_{yj}, $j = 1, ..., 4$.

(i) L'estimateur non-ajusté :

$$\widehat{t}_{jun} = N\bar{y}_{jr}.$$

(ii) L'estimateur calé en une étape :

$$\widehat{t}_{jC,1} = \sum_{i \in s_r} d_i \times F(\hat{\boldsymbol{\lambda}}_r^{\top} \mathbf{x}_i) y_{ji}$$

au moyen de plusieurs fonctions de calage $F(.)$ avec $\mathbf{x}_i = (1, x_i)^{\top}$.

(iii) L'estimateur PSA :

$$\widehat{t}_{jPSA} = \sum_{i \in s_r} d_i \frac{1}{\widehat{p}_i} y_{ji},$$

où les $\widehat{p}_i$ sont obtenues au moyen de la méthode des scores avec 20 classes.

Étude par simulation

Mesures Monte Carlo :

- Biais relatif Monte Carlo (en %) :

$$RB_{MC}(\widehat{t}) = \frac{100}{K} \sum_{k=1}^{K} \frac{(\widehat{t}_{(k)} - t_y)}{t_y}.$$

- Racine carrée de l'erreur quadratique moyenne relative Monte Carlo :

$$RRMSE_{MC}(\widehat{t}) = 100 \times \frac{\left\{ K^{-1} \sum_{k=1}^{K} (\widehat{t}_{(k)} - t)^2 \right\}^{1/2}}{t_y}.$$

$$Biais(\widehat{t}_{yC,1}) \approx -\sum_{i \in U}(1 - p_i F_i)\left(y_i - \mathbf{x}_i^\top \boldsymbol{B}_{p,f}\right)$$

	$\widehat{t}_{un}$	$\widehat{t}_{yC,1}$ $F(u) = 1 + u$	$\widehat{t}_{yC,1}$ $F(u) = \exp(u)$	$\widehat{t}_{yC,1}$ $F(u)$=logit	$\widehat{t}_{PSA}$
y_1 (lin.)	-4.1 (4.2)	0.0 (0.7)	0.0 (0.7)	0.0 (0.7)	-0.0 (0.8)
y_2 (exp.)	-28.1 (28.7)	-0.1 (5.5)	2.8 (6.1)	3.3 (6.4)	-0.1 (3.0)
y_3 (logit)	-27.5 (27.9)	-0.1 (3.4)	1.7 (3.6)	2.1 (3.8)	-0.1 (2.3)
y_4 (quad.)	-4.8 (5.3)	0.1 (2.8)	-2.0 (3.3)	-2.4 (3.5)	-0.1 (1.4)

Table – p_i-inverse linéaire : $p_i = (1.2 + 0.024\ x_i)^{-1}$

$$Biais(\widehat{t}_{yC,1}) \approx -\sum_{i \in U}(1 - p_i F_i)\left(y_i - \mathbf{x}_i^\top \boldsymbol{B}_{p,f}\right)$$

	$\widehat{t}_{un}$	$\widehat{t}_{yC,1}$ $F(u) = 1 + u$	$\widehat{t}_{yC,1}$ $F(u) = \exp(u)$	$\widehat{t}_{yC,1}$ $F(u)$=logit	$\widehat{t}_{PSA}$
y_1 (lin.)	-4.9 (4.9)	-0.0 (0.8)	0.0 (0.8)	0.0 (0.8)	-0.0 (0.8)
y_2 (exp.)	-35.1 (35.5)	-4.0 (7.1)	-0.0 (5.8)	0.7 (5.9)	-0.1 (3.2)
y_3 (logit)	-33.8 (34.1)	-2.5 (4.3)	0.0 (3.3)	0.6 (3.3)	-0.1 (2.3)
y_4 (quad.)	-3.6 (4.3)	2.9 (4.2)	0.0 (2.7)	-0.6 (2.8)	-0.0 (1.6)

Table – p_i-exponentielle : $p_i = \exp(-0.2 - 0.014 x_i)$

$$Biais(\widehat{t}_{yC,1}) \approx -\sum_{i \in U}(1 - p_i F_i)\left(y_i - \mathbf{x}_i^\top \boldsymbol{B}_{p,f}\right)$$

	$\widehat{t}_{un}$	$\widehat{t}_{yC,1}$ $F(u) = 1 + u$	$\widehat{t}_{yC,1}$ $F(u) = \exp(u)$	$\widehat{t}_{yC,1}$ $F(u)$=logit	$\widehat{t}_{PSA}$
y_1 (lin.)	-7.3 (7.3)	-0.3 (0.9)	-0.2 (0.9)	-0.2 (0.9)	-0.1 (0.9)
y_2 (exp.)	-51.5 (51.7)	-10.0 (12.3)	-0.4 (7.0)	0.9 (7.1)	-0.2 (3.7)
y_3 (logit)	-53.4 (53.5)	-12.1 (12.9)	-5.6 (6.7)	-4.5 (5.8)	-0.3 (3.0)
y_4 (quad.)	-1.0 (2.3)	11.7 (12.3)	4.3 (5.3)	3.1 (4.3)	-0.0 (1.8)

Table – p_i-logistique : $p_i = 0.2 + 0.6\ \{1 + \exp(-5 + x_i/8)\}^{-1}$

$$Biais(\widehat{t}_{yC,1}) \approx -\sum_{i \in U}(1 - p_i F_i)\left(y_i - \mathbf{x}_i^\top \boldsymbol{B}_{p,f}\right)$$

	$\widehat{t}_{un}$	$\widehat{t}_{yC,1}$ $F(u) = 1 + u$	$\widehat{t}_{yC,1}$ $F(u) = \exp(u)$	$\widehat{t}_{yC,1}$ $F(u)$=logit	$\widehat{t}_{PSA}$
y_1 (lin.)	1.3 (1.4)	-0.2 (0.5)	-0.2 (0.5)	-0.2 (0.5)	0.0 (0.5)
y_2 (exp.)	-8.3 (9.4)	-19.7 (19.9)	-19.2 (19.5)	-19.0 (19.3)	-0.4 (2.3)
y_3 (logit)	-0.5 (3.2)	-11.8 (12.0)	-11.5 (11.7)	-11.4 (11.6)	-0.1 (1.0)
y_4 (quad.)	13.1 (13.2)	13.7 (13.8)	13.4 (13.5)	13.2 (13.4)	0.2 (1.1)

Table – p_i-quadratique : $p_i = 0.7 + 0.45\,(x_i/40 - 1)^2 + 0.0025\,x_i$

Pondération en une étape

- En présence de non-réponse, les fonctions de calage ne sont pas équivalentes.
- On avait posé la question : peut-on réellement faire l'économie d'une modélisation des probabilités de réponse dans le contexte d'une pondération en une étape ?
- À la lumière des résultats, la réponse est non, en général
  - Exception : variables catégorielles que l'on croise (cas d'une post-stratification)$\rightarrow$ la fonction de calage n'est pas importante !
  - Cohérent avec la *Remarque* 10.1 dans Särndal et Lundström (2005)

Comparaison

Pondération en deux étapes

- Avantages
  - Permet de bien séparer les deux étapes de pondération : impact de chacune des étapes
  - Utilisation de méthodes non-paramétriques (par ex., classes de repondération)
- Désavantages
  - Plus complexe à implanter

Pondération en une étape

- Avantages
  - Simplicité
- Désavantages
  - Choix de la fonction de calage "caché"
  - Méthodes essentiellement paramétriques
  - Ne permet pas bien d'apprécier l'impact de la non-réponse

Pondération en une étape : équations estimantes

- Faisons le chemin à l'envers !
- On peut montrer qu'en partant du modèle paramétrique de non-réponse, on retrouve les équations de calage qui jouent le rôle d'équations estimantes.

Estimation de γ

1. A partir de notre modèle de non-réponse, on a, sans perte de généralité :

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

qu'on peut réécrire : $\mathbb{E}\left\{\frac{I_i\ R_i}{\pi_i\ f(\gamma\ ;\ \boldsymbol{Z}_i)} \mid \boldsymbol{Z}_i,\ Y_i\right\} = 1.$

2. Équations de moments du modèle de non-réponse

$$\mathbb{E}\left\{\frac{I_i\ R_i}{\pi_i\ f(\gamma\ ;\ \boldsymbol{Z}_i)} \times\ \boldsymbol{Z}_i\right\} = \mathbb{E}(\boldsymbol{Z}_i)$$

3. Équations estimantes (contre partie empirique)

$$\sum_{i \in s_r} \pi_i^{-1}\ f(\hat{\gamma}\ ;\ \boldsymbol{z}_i)^{-1}\ \boldsymbol{z}_i = \boldsymbol{t_z}.$$

Lien avec le calage généralisé

- Le système précédant correspond à un estimateur par calage où
  - $\boldsymbol{z}_i$: vecteur des variables de calage
  - $\boldsymbol{t}_z$: vecteur des marges de calage
  - Le vecteur des paramètres γ correspond au vecteur $\boldsymbol{\lambda}$
  - $F(\boldsymbol{\lambda}^\top \boldsymbol{Z}_i) = f^{-1}(\boldsymbol{\lambda}\ ;\ \boldsymbol{Z}_i)$ 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 $\hat{t}_{yC,1} = \sum_{i \in s_r} \pi_i^{-1}\ F(\hat{\boldsymbol{\lambda}}_r^{\ \top} \boldsymbol{z}_i)\ y_i$.

Merci de votre attention.

Des questions ?