Approche modèle pour l'estimation en présence de non-réponse non-ignorable en sondage

Journées de Méthodologie Statistique

Eric Lesage

Crest-Ensai

25 janvier 2012

Comment traiter la non-réponse ?

- Le mécanisme de non-réponse constitue une sélection aléatoire mais non contrôlée par un plan de sondage
- L'échantillon des répondants est-il représentatif de la population U ?
- Peut-on ignorer ce mécanisme ? Et si oui, comment ?

La non-réponse corrélée à la variable d'intérêt

- Un mécanisme de non-réponse lié à la variable d'intérêt peut biaiser l'estimateur de son total
- Des exemples de l'enquête emploi ou de l'enquête patrimoine

Le contexte

- On se place dans le cas d'une enquête exhaustive
- U est une population finie de taille N
- Les éléments de U sont repérés par l'indice k
- Pour chaque élément k on observe les réalisations (x_k, y_k, z_k, r_k) du vecteur aléatoire (X_k, Y_k, Z_k, R_k)
- Les vecteurs (X_k, Y_k, R_k) sont i.i.d

La non-réponse ignorable

- Elle peut être MCAR ou MAR
- MCAR : *Missing completely at random* ; la non-réponse n'est pas corrélée à la variable d'intérêt
- MAR : *Missing at random* ; la non-réponse n'est corrélée à la variable d'intérêt qu'à travers des covariables de Y_k observées. Conditionnellement à ces variables, la non-réponse n'est pas corrélée à la variable d'intérêt
- Exemple des groupes homogènes de repondération Sexe*groupes d'âge

La non-réponse ignorable, cas MAR

- Modèle de régression linéaire de Y_k sur U :

$$Y_k = \alpha_0 + \alpha_1 X_k + \varepsilon_k$$

où $\mathbb{E}(\varepsilon_k / X_k) = 0$

- Modèle de réponse,

$$\mathbb{E}(R_k / X_k) = \rho(X_k)$$

- Non-réponse MAR :

$$\mathbb{E}(\varepsilon_k R_k / X_k) = 0 \Longleftrightarrow \mathbb{E}(Y_k R_k / X_k) = \mathbb{E}(Y_k / X_k)\mathbb{E}(R_k / X_k)$$

- Modèle de régression linéaire de Y_k sur l'échantillon des répondants s_r :

$$Y_k = \alpha_0 + \alpha_1 X_k + \varepsilon_k = \alpha' \mathbf{X_k} + \varepsilon_k$$

où $\mathbb{E}(\varepsilon_k R_k / X_k) = 0$, $\alpha' = (\alpha_0, \alpha_1)$ et $\mathbf{X_k}' = (1, X_k)$

- L'estimateur par les MCO

$$\hat{\alpha} = \left[\frac{1}{N} \sum_{k \in U} R_k \mathbf{X_k} \mathbf{X_k}' \right]^{-1} \left[\frac{1}{N} \sum_{k \in U} R_k \mathbf{X_k} Y_k \right]$$

converge asymptotiquement vers α

La non-réponse ignorable, cas MAR

Pour estimer le total $t_y = \sum_{k \in U} y_k$,
on peut utiliser une estimateur par la régression :

$$\hat{t}_y = \mathbf{t_x}' \hat{\alpha}$$

où $\mathbf{t_x}' = (N, t_x = \sum_{k \in U} x_k)$

Campus de Ker Lann, rue Blaise Pascal - BP 37203 - 35172 Bruz Cedex - Tél : 02.99.05.32.32 - Fax : 02.99.05.32.05
www.ensai.com

La non-réponse non-ignorable

- On étudie un cas de non-réponse non-ignorable proche du cas précédent
- Cette fois-ci, une des variables explicatives dans le modèle modèle sur Y_k n'est pas une variable auxiliaire
- C'est une variable observée uniquement sur l'échantillon des répondants
- C'est donc une autre variable d'intérêt du modèle

Notations

- Variables auxiliaires dont les totaux sont connus sur U
  $\mathbf{X_k}' = (1, X_{k,1}, X_{k,2})$
- Variables explicatives de la non-réponse
  $\mathbf{Z_k}' = (1, X_{k,1}, Z_{k,2})$
- Variables d'intérêt
  $\mathbf{Y_k}' = (Y_k, Z_{k,2})$
- La variable $X_{k,2}$ est corrélée positivement à la variable $Z_{k,2}$

Modèle de non-réponse

$$E(R_k/\mathbf{Z_k}, \mathbf{X_k}, \mathbf{Y_k}) = E(R_k/\mathbf{Z_k}) = \rho(\mathbf{Z_k})$$

Ou :

$$E(R_k\mathbf{X_k}/\mathbf{Z_k}) = E(R_k/\mathbf{Z_k})E(\mathbf{X_k}/\mathbf{Z_k})$$
$$E(R_k\mathbf{Y_k}/\mathbf{Z_k}) = E(R_k/\mathbf{Z_k})E(\mathbf{Y_k}/\mathbf{Z_k})$$

Remarque : la variable $X_{k,2}$ est corrélée à $Z_{k,2}$ mais n'a pas d'effet explicatif direct sur la non-réponse.

Modèle de la variable d'intérêt

- On suppose que Y_k suit un modèle de régression linéaire

$$Y_k = \alpha_0 + \alpha_1 X_{k,1} + \alpha_2 Z_{k,2} + \varepsilon_k$$
$$\mathbb{E}(\varepsilon_k / \mathbf{Z_k}, \mathbf{X_k}) = 0$$

- Pour autant, on ne peut pas proposer un estimateur Greg de t_y car on ne dispose pas de la valeur du total t_{z_2}

Modèle de régression à variable instrumentale

- On écrit un nouveau modèle "dégradé" pour Y_k en utilisant la corrélation entre $X_{k,2}$ et $Z_{k,2}$:

$$Y_k = \beta_0 + \beta_1 X_{k,1} + \beta_2 X_{k,2} + \tau_k$$

$$\mathbb{E}(\tau_k / \mathbf{Z_k}) = 0$$

- D'un point de vue statistique, il s'agit d'un modèle de régression à variable instrumentale
- Ce modèle est moins bien ajusté que le modèle initial, par contre, il reste identifiable sur s_r car $\mathbb{E}(\tau_k R_k / \mathbf{Z_k}) = 0$

Modèle de régression à variable instrumentale

En effet, sur s_r on prendra l'estimateur :

$$\begin{aligned}\hat{\beta}^{\mathbf{VI}} &= \left[\frac{1}{N}\sum_{k\in U} c_k R_k \mathbf{Z_k}\mathbf{X_k}'\right]^{-1}\left[\frac{1}{N}\sum_{k\in U} c_k R_k \mathbf{Z_k} Y_k\right] \\ &= \beta + \left[\frac{1}{N}\sum_{k\in U} c_k R_k \mathbf{Z_k}\mathbf{X_k}'\right]^{-1}\left[\frac{1}{N}\sum_{k\in U} c_k R_k \mathbf{Z_k} \tau_k\right],\end{aligned}$$

où c_k s'interprète comme un poids de l'élément k et est une fonction de $\mathbf{Z_k}$: $c_k = f(\mathbf{Z_k})$.

Estimateur IVGreg

- $\hat{\beta}^{\mathbf{VI}}$ converge asymptotiquement vers β
- On peut utiliser cette fois-ci un estimateur par la régression (instrumentale) :

$$\begin{aligned} \hat{t}_y^{VI} &= \mathbf{t_x}' \hat{\beta}^{VI} \\ &= \sum_{k \in U} c_k R_k \mathbf{t_x}' \left[\sum_{l \in U} c_l R_l \mathbf{Z_l} \mathbf{X_l}' \right]^{-1} \mathbf{Z_k} Y_k \end{aligned}$$

Estimateur IVGreg : estimateur linéaire

$$\hat{t}_y^{VI} = \sum_{k \in U} R_k w_k^{VI} Y_k,$$

où

$$\begin{aligned} w_k^{VI} &= c_k \mathbf{t_x}' \left[\sum_{l \in U} c_l R_l \mathbf{Z_l} \mathbf{X_l}' \right]^{-1} \mathbf{Z_k} \\ &= c_k \left(1 + \left(\mathbf{t_x}' - \sum_{k \in U} c_k R_k \mathbf{X_k}' \right) \left[\sum_{l \in U} c_l R_l \mathbf{Z_l} \mathbf{X_l}' \right]^{-1} \mathbf{Z_k} \right) \end{aligned}$$

est le poids d'enquête de l'élément k.

Simulations Monte Carlo

On génère une population de taille $N = 1000$

- Variables explicatives de la non-réponse
  $Z_{k,1} \sim$ gamma$(20, 20)$ (observée sur s_r)
  $Z_{k,2} \sim U[0, 600]$ (observée sur s_r)
  $Z_{k,3} \sim U[0, 600]$, (non observée)
- Variables auxiliaires
  $X_{k,1} = Z_{k,1}$
  $X_{k,2} = 0,5(Z_{k,2} + U_{k,2})$, où $U_{k,2} \sim \mathfrak{N}(0, 150^2)$
  $X_{k,3} = 0,5(Z_{k,2} + Z_{k,3})$

Simulations Monte Carlo

La variable Y_k est générée par le modèle linéaire :

$$Y_k = 100 + 20X_{k,1} + 20Z_{k,2} + E_k.$$

où $E \sim \mathfrak{N}(0, 2000^2)$

Simulations Monte Carlo

- $K = 1000$ simulations du mécanisme de réponse
- R_k suit une loi de Bernoulli de paramètre :

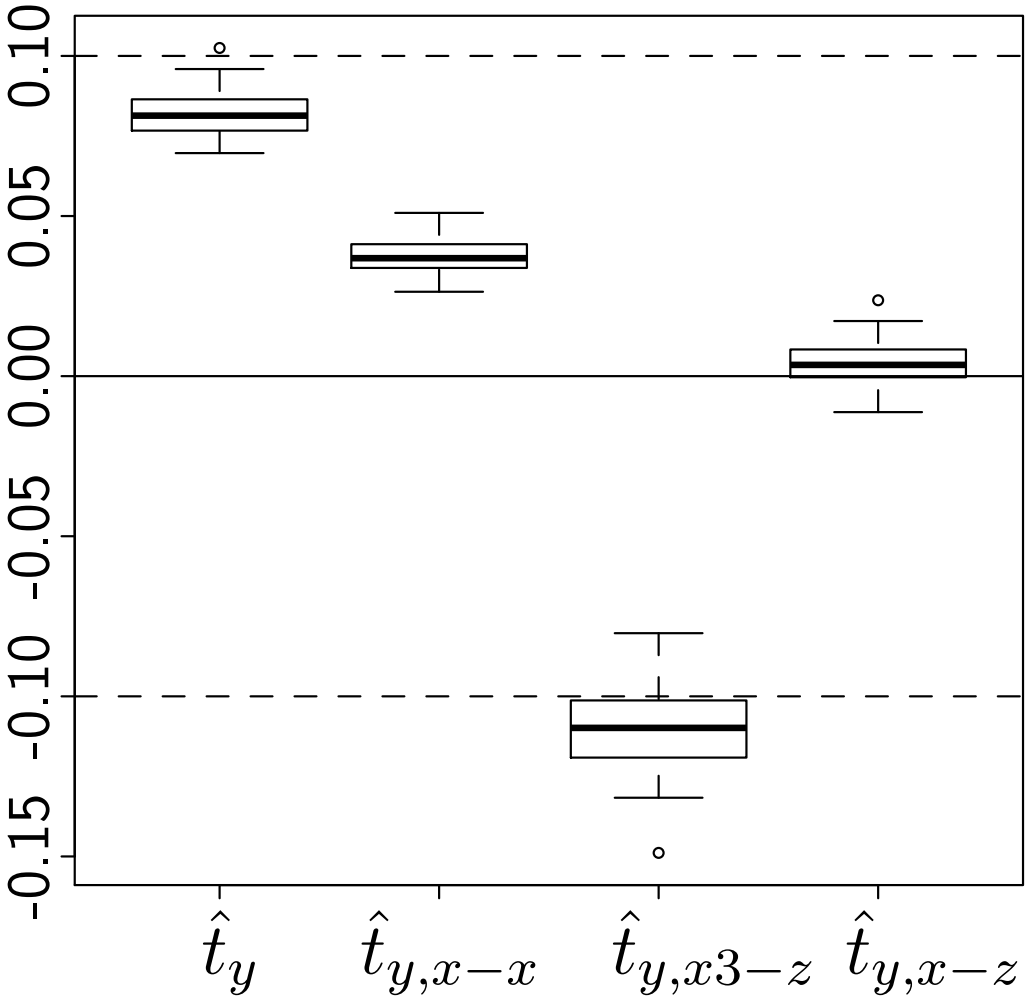
$$p_k = 0.01 + 0.9 \left(\frac{\exp(-7 + 0.005 z_{k,1} + 0.010 z_{k,2} + 0.010 z_{k,3})}{1 + \exp(-7 + 0.005 z_{k,1} + 0.010 z_{k,2} + 0.010 z_{k,3})} \right)$$

Simulations Monte Carlo

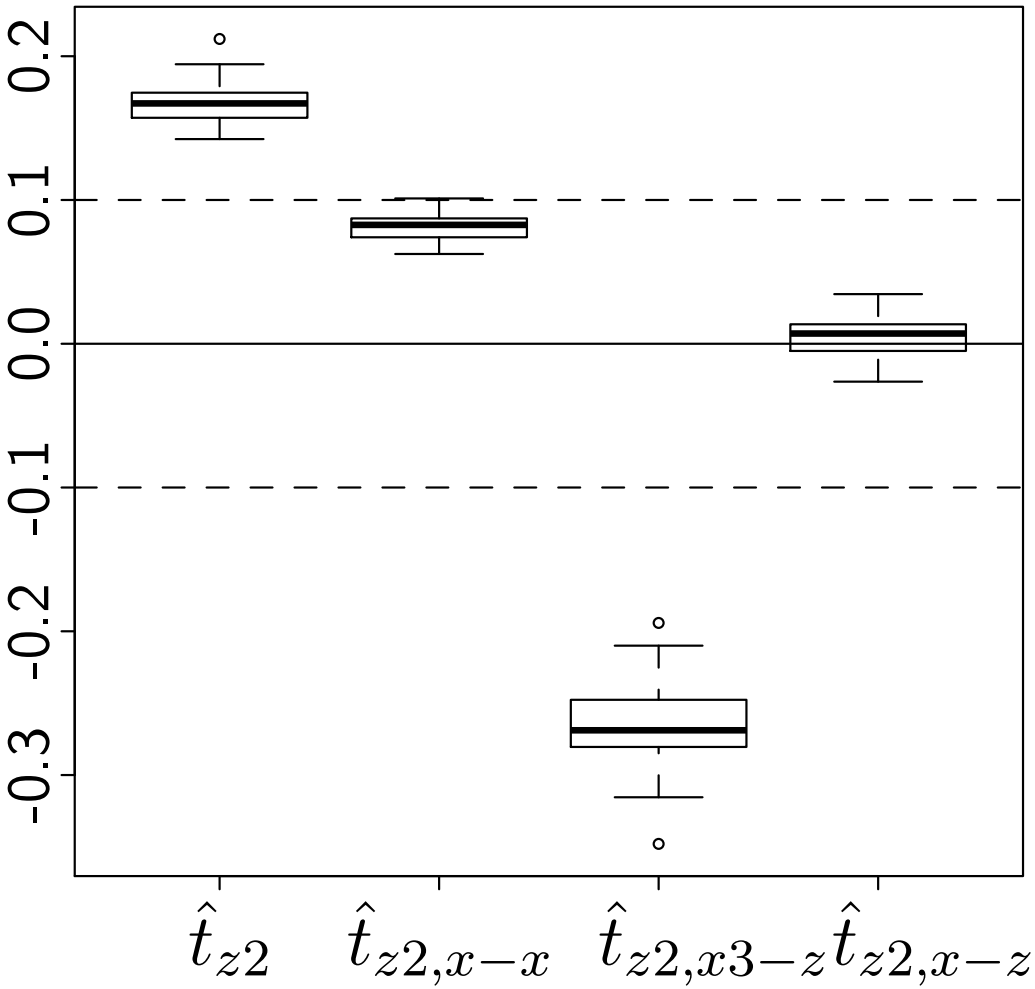
On compare quatre estimateurs du total t_y

- $\hat{t}_y$: moyenne des valeurs de y sur s_r multipliée par N
- $\hat{t}_{y,x-x}$: estimateur par la régression habituel
- $\hat{t}_{y,x3-z}$: estimateur par la régression instrumentale avec X_3 comme variable proxy de Z_2
- $\hat{t}_{y,x-z}$: estimateur par la régression instrumentale avec Z_1 et Z_2 comme instruments et X_2 comme variable proxy de Z_2

Biais relatifs



Biais relatifs - estimateurs de t_{z_2}



- Beaumont, J.-F. (2000). Une méthode d'estimation en présence de non-réponse non-ignorable. Techniques d'enquêtes, vol 26, pp 145-151.
- Deville, J.-C. (2004). La correction de la non-réponse par calage généralisé. Actes des journées de méthodologie satistique, 16 et 17 décembre 2002, INSEE Méthodes.
- Fuller, A.F. (2009). Sampling Statistics. Wiley, 371.
- Gautier, E. (2005). Eléments sur les mécanismes de la sélection dans les enquêtes et sur la non-réponse non-igorable. Actes des journées de méthodologie satistique, INSEE.
- Särndal, C.E. and Sixten L. (2005). Estimation in Surveys with Nonresponse. Wiley.