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

Éric Lesage
CREST(ENSAI) and IRMAR, Campus de Ker Lann, F-35172 BRUZ
Université européenne de Bretagne, France

13 janvier 2012

1 Introduction

Cet article traite de l'estimation d'un total d'une variable d'intérêt Y en présence de valeurs manquantes dues à la non-réponse d'une partie de l'échantillon.

Lorsque la variable d'intérêt n'est pas liée au comportement de non-réponse, on peut estimer le total de y, noté t_y, au moyen des estimateurs habituels. On est dans un cas de valeurs manquantes générées par un processus que l'on considère comme complètement aléatoire (MCAR : missing completely at random).
Lorsqu'on observe une partie des variables explicatives de la non-réponse (ces variables sont des variables auxiliaires dont on connaît au moins le total) et que la variable d'intérêt dépend de ces variables mais non des autres variables explicatives inobservées de la non-réponse alors on peut utiliser un estimateur de type Greg (estimateur par la régression) pour estimer t_y. Cette situation est qualifiée de situation MAR (Missing at random), i.e. que conditionnellement aux variables observées explicatives de la non-réponse, la variable y n'est pas liée à la non-réponse. En d'autres termes, le modèle de y en fonction des variables observées explicatives de la non-réponse est le même sur la population totale et sur la population des répondants.

Les cas MCAR et MAR sont bien connus. Ils correspondent à ce qu'on appelle de la non-réponse ignorable. Toutes les autres situations font partie de ce qu'on appelle la non-réponse non-ignorable (voir par exemple Särndal, 2005). C'est une de ces situations que nous allons étudier dans cette note. On étudie un cas proche du cas MAR, celui où une des variables qui expliquent le comportement de réponse n'est observée que partiellement, à savoir sur l'échantillon des répondants. Cette variable peut d'ailleurs être la variable d'intérêt de l'étude elle-même. On peut citer l'exemple de l'enquête emploi où la position en activité des individus interrogés est une cause de non-réponse mais aussi une des variables d'intérêt de l'enquête.

L'approche utilisée dans cet article peut être qualifiée "d'approche modèle" par opposition à l'approche par quasi-échantillonnage (quazi-randomisation) qui s'appuie

sur l'estimation des probabilités de réponse des individus de la population. Dans cette dernière approche, le mécanisme de sélection des répondants est assimilé à une ultime phase d'échantillonnage. Beaumont(2000) et Gautier(2005) ont proposé des méthodes d'estimation qui reposent sur une estimation par maximum de vraisemblance. Deville(2004) a utilisé le calage généralisé pour corriger la non-réponse non-ignorable.

L'originalité de cet article est qu'on pose un modèle non-paramétrique pour la non-réponse et qu'on ne cherche pas à estimer les probabilités de réponse des individus de l'enquête. On utilise par contre une variable instrumentale du modèle de non-réponse pour écrire un nouveau modèle sur la variable y.

On se placera dans le cas d'une enquête exhaustive. Ce cadre permet de simplifier les écritures sans réduire la portée des résultats. Il est facile de transposer les formules en cas d'une enquête par sondage.

2 Définition de la régression linéaire à variables instrumentales

On va utiliser par la suite un modèle de régression linéaire à variables instrumentales (voir par exemple Fuller(2009)). Ce modèle diffère du modèle classique de régression linéaire par la condition qui porte sur le terme d'erreur. Ce type de modèle est utilisé par exemple en économétrie pour résoudre des problèmes d'endogénéité des variables explicatives par rapport au terme d'erreur dans un modèle linéaire.

$$\begin{aligned} Y_k &= \alpha_0 + \alpha_1 X_{k,1} + \alpha_2 X_{k,2} + \varepsilon_k \\ E(\varepsilon_k / Z_{k,1} Z_{k,2}) &= 0, \end{aligned}$$

où $Z_{k,1}$ est corrélé à $X_{k,1}$ et $Z_{k,2}$ est corrélé à $X_{k,2}$.

3 Le modèle de non-réponse

Soit U, une population finie de taille N. Les éléments de U sont indexés par l'indice k.

On note R_k l'indicatrice qui prend la valeur 1 quand l'individu k répond, $\mathbf{X_k}' = (1, X_{k,1}, X_{k,2})$ 3 variables auxiliaires dont les totaux sont connus sur U, $\mathbf{Y_k}' = (Y_{k,1}, Y_{k,2})$ deux variables d'intérêt et $\mathbf{Z_k}' = (1, X_{k,1}, Y_{k,2})$ le vecteur de variables explicatives de la non-réponse.
On suppose que la variable $X_{k,2}$ est corrélée positivement à la variable $Y_{k,2}$.
On considère le modèle de non-réponse suivant :

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

On peut également considérer une condition un peu moins forte d'orthogonalité conditionnelle à $\mathbf{Z_k}$ plutôt que d'indépendance conditionnelle de R_k et $\mathbf{X_k}$, et de R_k et $\mathbf{Y_k}$.

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

La distinction entre des variables observées (même partiellement) explicatives de la non-réponse (le vecteur Z_k) et les autres variables est un travail important qui s'avère, dans la pratique, plus compliqué qu'il n'y paraît. Les variables sont divisées en deux groupes :

1. les variables explicatives de la non réponse : $X_{k,1}, Y_{k,2}$ (le vecteur $\mathbf{Z}_k$) ;
2. les variables non corrélées à la non réponse (conditionnellement à $\mathbf{Z_k}^t = (1, X_{k,1}, Y_{k,2})$) : $X_{k,2}, Y_{k,1}$.

Remarques :
– $X_{k,2}$ est une variable instrumentale de $Y_{k,2}$ dans le modèle de réponse, elle pourrait être utilisée dans des équations de moment pour estimer les paramètres du modèle s'il est paramétrique.
– si on connaissait la forme de $\rho(\mathbf{Z_k})$ on pourrait prédire les valeurs des probabilités de réponse pour chaque individu.

4 Le modèle de la variable d'intérêt Y_1

On suppose que Y_1 suit un modèle de régression linéaire.

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

Le mécanisme de sélection est exogène pour ce modèle. En effet, le terme d'erreur est une combinaison linéaire de variables qui sont orthogonales a R_k conditionnellement à Z_k, donc $\mathbb{E}(R_k\varepsilon_k/\mathbf{Z_k}) = \mathbb{E}(R_k/\mathbf{Z_k})\,\mathbb{E}(\varepsilon_k/\mathbf{Z_k}) = 0$.
En d'autres termes, le mécanisme de réponse est ignorable pour l'estimation des paramètres α du modèle sur la variable d'intérêt Y_1.

5 Lien entre la variable Y_2 et sa variable instrumentale X_2

Pour autant, on ne peut pas proposer un estimateur Greg de t_{y_1} car on ne dispose pas de la valeur de t_{y2}. On contourne le problème en cherchant un autre modèle pour y_1. On va retirer la variable Y_2 du modèle de régression linéaire de Y_1 et la remplacer par une variable **proxy** qui doit être :

1. une variable auxiliaire
2. une variable bien corrélée à y_2
3. une variable qui n'est pas corrélée à la non-réponse conditionnellement à $\mathbf{Z_k}$. Autrement dit, cette variable ne contient pas d'information sur la non-réponse additionnelle par rapport à $\mathbf{Z_k}$.

Dans notre cas, ce sera la variable X_2 qui sera la variable proxy associée à Y_2. On remarque que c'était une variable qui est une variable instrumentale pour le modèle de non-réponse.

6 Modèle de regression linéaire à variables instrumentales

Si on suppose qu'il existe un modèle de régression linéaire :

$$\begin{aligned} X_{k,2} = \gamma_0 + \gamma_1 X_{k,1} + \gamma_2 Y_{k,2} + \nu_k \\ \mathbb{E}(\nu_k/\mathbf{Z_k}) = 0 \\ \mathbb{V}(\nu_k/\mathbf{Z_k}) = \sigma_\nu^2 \end{aligned}$$

On peut alors écrire un nouveau modèle pour Y_1 :

$$Y_{k,1} = \beta_0 + \beta_1 X_{k,1} + \beta_2 X_{k,2} + \tau_k \tag{3}$$

$$\begin{aligned} \tau_k = \epsilon_k - \frac{\alpha_2}{\gamma_2}\nu_k \\ \mathbb{E}(\tau_k/\mathbf{Z_k}) = 0 \\ \mathbb{V}(\tau_k/\mathbf{Z_k}) = \sigma_\tau^2 \end{aligned}$$

D'un point de vue statistique, il s'agit d'un modèle de régression à variable instrumental. Ce modèle est bien entendu moins bien ajusté que le modèle initial. Par contre, il reste identifiable sur s_r. En effet, sur s_r on prendra l'estimateur :

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

où c_k s'interprète comme un poids de l'élément k et est une fonction de $\mathbf{Z_k}$: $c_k = f(\mathbf{Z_k})$. On peut supposer pour l'instant que $c_k = 1$. On verra par la suite qu'on peut utiliser un poids de calage quelconque.

On suppose que les vecteurs de variables $(X_{k,1}, X_{k,2}, Y_{k,1}, Y_{k,2}, R_k)'$ sont iid et que $\mathbb{E}\left(\frac{1}{N} c_k R_k \mathbf{Z_k}\mathbf{X_k}'\right)$ est une matrice inversible.

L'estimateur $\hat{\beta}^{\mathbf{VI}}$ est asymptotiquement sans biais, car $\left[\frac{1}{N}\sum_{k\in U} c_k R_k \mathbf{Z_k}\tau_k\right]$ converge, en probabilités, vers :

$$\mathbb{E}\left(c_k R_k \mathbf{Z_k}\tau_k\right) = \mathbb{E}\left(\mathbb{E}\left(c_k R_k \mathbf{Z_k}\tau_k/\mathbf{Z_k}\right)\right) = \mathbb{E}\left(c_k \mathbf{Z_k}\mathbb{E}\left(R_k/\mathbf{Z_k}\right)\mathbb{E}\left(\tau_k/\mathbf{Z_k}\right)\right) = 0.$$

7 Estimateur linéaire de totaux IVREG

On définit l'estimateur par la regression instrumentale pour le total t_{y1} par :

$$\begin{aligned}\hat{t}_{y1}^{VI} &= \mathbf{t_x}'\hat{\beta}^{VI} \\ &= \sum_{k\in U} c_k R_k \mathbf{t_x}' \left[\sum_{k\in U} c_k R_k \mathbf{Z_k}\mathbf{X_k}'\right]^{-1} \mathbf{Z_k} Y_{k1}\end{aligned}$$

Une des équations normales du modèle de régression à variables instrumentales est :

$$\sum_{k\in s_r} c_k (Y_{k,1} - \mathbf{X_k}'\hat{\beta}^{VI}) \times 1 = 0,$$

où s_r est l'échantillon des répondants, car le vecteur des instruments contient la variable constante égale à 1 : $\mathbf{Z_k}' = (1, X_{k,1}, Y_{k,2})$.
En utilisant cette identité, on peut donner une autre formule pour l'estimateur par régression instrumentale :

$$\hat{t}_{y1}^{VI} = \sum_{k\in U} c_k R_k Y_{k,1} + \left(\mathbf{t_x} - \sum_{k\in U} c_k R_k \mathbf{X_k}\right)' \hat{\beta}^{VI}$$

$$\hat{t}_{y1}^{VI} = \sum_{k\in U} R_k c_k \left(1 + \left(\mathbf{t_x}' - \sum_{k\in U} c_k R_k \mathbf{X_k}'\right)\left[\sum_{k\in U} c_k R_k \mathbf{Z_k}\mathbf{X_k}'\right]^{-1} \mathbf{Z_k}\right) Y_{k1}$$

L'estimateur $\hat{t}_{y1}^{VI}$ peut donc s'écrire :

$$\hat{t}_{y1}^{VI} = \sum_{k\in U} R_k w_k^{VI} Y_{k1},$$

où

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

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

Ce poids peut être utilisé pour estimer n'importe quelle variable d'intérêt qui n'a pas d'effet explicatif propre de la non réponse (i.e. qui est non corrélée à R_k condtionnellement à $\mathbf{Z_k}$).

C'est le cas de la variable Z_2. On peut donc estimer t_{Z_2} par $\hat{t}_{z2}^{VI} = \sum_{k\in U} w_k^{VI} R_k Z_{k,2}$.

Remarques :

– On retrouve des résultats similaires à ceux obtenus dans une approche GREG. Notamment, l'estimateur a la propriété de calage pour les variables auxiliaires. En outre si on peut choisir $c_k = \frac{1}{\rho(\mathbf{Z_k})}$ alors l'estimateur $\hat{\beta}^{VI}$ converge vers le vecteur des paramètres estimé sur U même si le modèle est faux et les poids w_k convergent vers les inverses des probabilités de réponse $w_k^{VI} = c_k = \frac{1}{\rho(\mathbf{Z_k})}$.
– Si on prend un système de poids de calage (généralisé) pour les c_k (qui est fonction de $\mathbf{Z_k}$) tel que $\mathbf{t_x} - \sum_{k\in U} c_k R_k \mathbf{X_k} = 0$ alors on a directement $w_k^{VI} = c_k$. Ce qui signifie qu'il n'est pas besoin de faire l'estimation des paramètres du modèle pour avoir l'estimation ponctuelle $\hat{t}_{y1}^{VI}$.

8 Etude par simulation

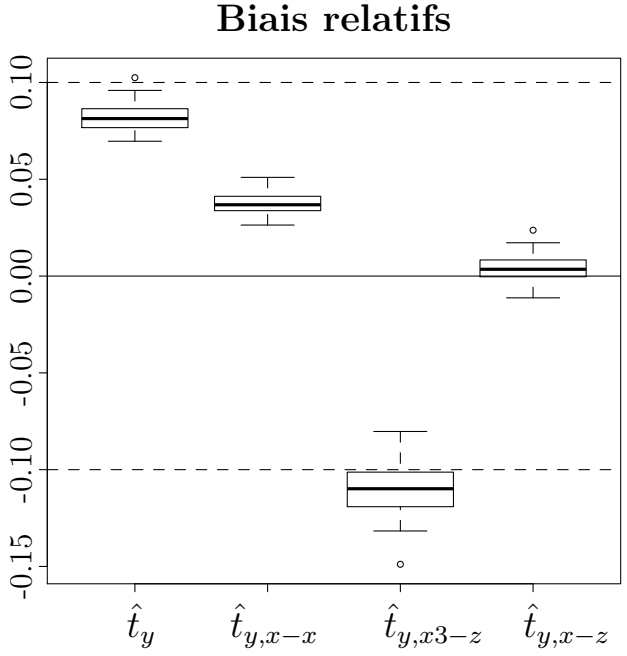

FIGURE 1 – Comparaison de l'efficacité des estimateurs de t_y

On compare les performances de plusieurs estimateurs en présence de non-réponse au moyen de simulations par méthodes de Monte Carlo.
On génère une population de taille $N = 1000$, une variable Z_1 par une loi gamma(20,20), une variable Z_2 et une variable Z_3 par des lois uniformes sur l'intervalle $[0, 600]$, une variable U_2 par une loi normale $\mathfrak{N}(0, 150^2)$ et une variable E_k par une loi normale $\mathfrak{N}(0, 2000^2)$. On pose ensuite que $X_1 = z_1$, $X_2 = 0,5(Z_2 + U_2)$ et que $X_3 = 0,5(Z_2 + Z_3)$. La variable Z_3 n'est pas observée, ni sur la population totale, ni sur les répondants. Les variables Z_1 et Z_2 sont observées sur l'échantillon des répondants et les variables X_1, X_2 et X_3 sont des variables auxiliaires dont les totaux sont connus sur U.

La variable Y suit le modèle :

$$Y_k = 100 + 20X_1 + 20Z_2 + E_k.$$

On peut calculer $\mu_{X_1} = 401,1$, $\mu_{X_2} = 150,9$ et $\mu_Y = 14\ 105$.

On répète ensuite $K = 1000$ fois le mécanisme de réponse en utilisant le modèle : R_k suit une loi de Bernoulli de paramètre :

$$p_k = 0.01 + 0.9 \left(\frac{exp(-7 + 0.005Z1 + 0.010Z2 + 0.010Z3)}{1 + exp(-5 + 0.005Z1 + 0.010Z2)} \right).$$

On compare quatre estimateurs du total μ_y
- $\hat{t}_y$ qui est simplement la moyenne des valeurs de y sur l'échantillon des répondants,
- $\hat{t}_{y,x-x}$ qui est l'estimateur par la regression habituel (estimateur Greg),
- $\hat{t}_{y,x3-z}$ qui est l'estimateur par la regression instrumentale avec X_3 comme variable proxy de Z_2
- et enfin $\hat{t}_{y,x-z}$qui est l'estimateur par la regression instrumentale avec $Z1$ et $Z2$ comme instruments et X_2 comme variable proxy de Z_2.

La figure(1) donne une représentation graphique du biais relatif de chacun des quatre estimateurs. On voit que la difficulté réside dans le choix de la variable proxy de Y_2. Si on se trompe, le biais peut être plus fort que celui de l'estimateur Greg.

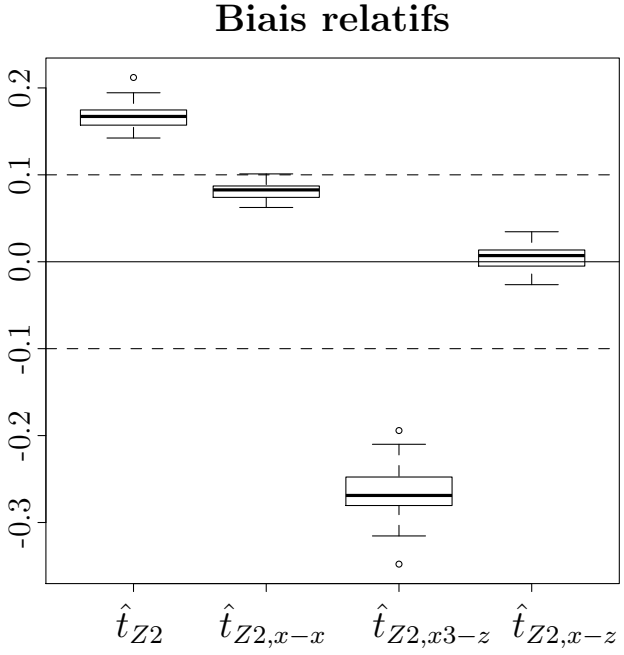

FIGURE 2 – Comparaison de l'efficacité des estimateurs de t_{Z_2}

9 Conclusion

On a montré que, sous certaines hypothèses, on pouvait corriger le biais de sélection causé par le fait que la variable d'intérêt est liée à la non-réponse. On a appliqué notre modèle pour corriger de la non-réponse totale, ce qui a conduit au final à une repondération. Bien entendu, le modèle (3) peut être utilisé pour faire de l'imputation. Des travaux de recherche sont actuellement en cours sur ce sujet et seront publiés prochainement.

Références

[1] Beaumont, J.-F. (2000). Une méthode d'estimation en présence de non-réponse non-ignorable. Techniques d'enquêtes, vol 26, pp 145-151.

[2] 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.

[3] Fuller, A.F. (2009). Sampling Statistics. Wiley, 371.

[4] 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.

[5] Särndal, C.E. and Sixten L. (2005). Estimation in Surveys with Nonresponse. Wiley.