Correction de la non-réponse non-ignorable à l'aide du calage généralisé avec des variables latentes

Alina Matei et Giovanna Ranalli

Institut de statistique, Université de Neuchâtel, Suisse
IRDP, Neuchâtel et Université de Perugia

Rennes, 7 Novembre, 2012
Colloque francophone sur les sondages

European Social Survey (Suisse, 2010)

- Êtes-vous intéressés par la politique ?
- Durant les 12 derniers mois, avez-vous contacté un homme politique ou un responsable politique fédéral, cantonal, etc. ?
- Y a-t-il un parti politique dont vous vous sentiez plus proche que des autres partis ?
- Êtes-vous membre d'un parti politique ?

Motivation

- On a plusieurs variables d'intérêt $y_1, y_2, \ldots, y_k$.
- On est en présence de non-réponse (totale) non-ignorable.
- On aimerait calculer un estimateur calé qui utilise un système unique de poids (donc qui ne dépend pas nécessairement d'une unique variable d'intérêt).

Plan de la présentation

- notations;
- calage généralisé et problème de la non-réponse;
- variables latentes (latent trait models/item response models);
- méthode proposée et exemples.

Notations

- Soit $U = \{1, \ldots, k, \ldots, N\}$ une population finie.
- L'unité k est l'unité de référence.
- Un échantillon $s \subseteq U$ (de taille n) est tiré sans remise en utilisant un plan $p(s)$.
- On a $\pi_k = Pr(k \in s) = \sum_{s \ni k, s \in \mathcal{S}} p(s)$ et également le poids $d_k = 1/\pi_k$.
- Les variables auxiliaires sont notées $\mathbf{x}_k$.
- Le but est d'estimer le total de la population $T = \sum_{k \in U} y_k$, où y_k est la variable d'intérêt.
- L'estimateur calé du total $T = \sum_{k \in U} y_k$ est

$$\widehat{T} = \sum_{k \in s} w_k y_k.$$

Types de non-réponse

On a deux types de non-réponse :

- *non-réponse totale* - l'absence complète d'information sur une unité;
- *non-réponse partielle* - une absence d'information sur une unité limitée à certaines variables seulement.

Correction de la non-réponse totale

- On considère que la non-réponse totale est présente et on note par $r \subseteq s$ l'ensemble des répondants.
- On note le plan de sondage $q(r|s)$ et on a

$$q(r|s) \geq 0, \text{ pour tout } r \in \mathcal{R}_s \text{ et } \sum_{r \in \mathcal{R}_s} q(r|s) = 1,$$

où $\mathcal{R}_s = \{r|r \subseteq s\}$.

- La variable d'intérêt est connue uniquement sur l'ensemble r.
- On définit l'indicateur de réponse $R_k = 1$ si $k \in r$ and 0 sinon et la probabilité de réponse $p_k = Pr(R_k = 1|k \in s)$.
- Les unités répondent indépendamment les unes des autres et indépendamment de s

$$q(r|s) = \prod_{k \in r} p_k \prod_{k \in s \setminus r} (1 - p_k).$$

Calage, calage généralisé et non-réponse totale

- Le modèle de non-réponse peut être écrit

$$q(r|s,\gamma) = \prod_{k\in r} F_k^{-1}(\gamma) \prod_{k\in \bar{r}} (1 - F_k^{-1}(\gamma)),$$

où

$$\sum_{k\in r} \mathbf{x}_k d_k F_k(\gamma) = \sum_{k\in r} \mathbf{x}_k d_k F(\gamma^T \mathbf{x}_k) = \sum_{k\in U} \mathbf{x}_k,$$

et $F_k(\gamma) = F(\gamma^T \mathbf{x}_k)$ et $p_k = F_k(\gamma)^{-1}$.

- Dans le calage généralisé

$$\sum_{k\in r} \mathbf{x}_k d_k F(\gamma^T \mathbf{z}_k) = \sum_{k\in U} \mathbf{x}_k,$$

où le vecteur $\mathbf{z}_k$ est connu uniquement sur r et doit être très corrélé avec $\mathbf{x}_k$.

- Les variables $\mathbf{z}$ sont des *variables instrumentales*.

Formes différentes de probabilités de réponse

$$\sum_{k\in r} w_k \mathbf{x}_k = \sum_{k\in r} d_k F(\mathbf{h}^T \mathbf{z}_k)\mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k,$$

avec $p_k = 1/F(\boldsymbol{\gamma}^T \mathbf{z}_k)$ et $\widehat{p}_k = 1/F(\mathbf{h}^T \mathbf{z}_k)$, où $\mathbf{h}$ est un estimateur consistent de $\boldsymbol{\gamma}$.

- Ajustement linéaire : $w_k = d_k(1 + \mathbf{h}^T \mathbf{z}_k), p_k = 1/(1 + \boldsymbol{\gamma}^T \mathbf{z}_k)$.
- Ajustement de type raking ratio : $w_k = d_k \exp(\mathbf{h}^T \mathbf{z}_k), p_k = 1/\exp(\boldsymbol{\gamma}^T \mathbf{z}_k)$,
- Ajustement logistique : $w_k = d_k(1 + \exp(\mathbf{h}^T \mathbf{z}_k)), p_k = 1/(1 + \exp(\boldsymbol{\gamma}^T \mathbf{z}_k))$, où on cale sur $\sum_{k\in U} \mathbf{x}_k - \sum_{k\in r} \mathbf{x}_k d_k$ au lieu de $\sum_{k\in U} \mathbf{x}_k$.
- Ajustement exponentiel généralisé (Folsom and Singh, 2000):

$$F(\mathbf{h}^T \mathbf{z}_k) = \frac{L(U-C) + U(C-L)\exp(A\mathbf{h}^T \mathbf{z}_k)}{(U-C) + (C-L)exp(A\mathbf{h}^T \mathbf{z}_k)} \in (L, U),$$

où $A = (U-L)/((C-L)(U-C))$ et $L \geq 1, 1 < U \leq \infty, U > C > L$, ($C = 1$ et $L < 1 < U$ dans le papier de Deville et Särndal, 1992). Pour $L = 1, C = 2$ et $U = \infty, F(\mathbf{h}^T \mathbf{z}_k)$ approche $1 + \exp(\mathbf{h}^T \mathbf{z}_k)$ et pour $C = 1, L = 0, U = \infty$, $\exp(\mathbf{h}^T \mathbf{z}_k)$.

Variables latentes

- L'hypothèse de base est la suivante : *un petit ensemble de variables latentes explique la dépendance des variables observées ou manifestes*.
- Des variables inobservées comme l'intelligence, la compétence en mathématique, l'attitude envers la politique, la préférence des consommateurs, qui ne sont pas mesurées directement, peuvent être quantifiées en utilisant des variables latentes.
- Les modèles à trait latent (latent trait models ou Item Response models) forment une classe de modèles qui font le lien (en général) entre des variables observées (binaires ou discrètes) et une seule variable latente.

Données binaires

- Chaque unité $k \in r$ doit répondre à un certain nombre de variables (M).
- Considérons qu'on dispose des variables binaires $yy_{k\ell}, \ell \in \{1, \ldots, m\}, m \leq M$.
- Supposons que les $yy_{k\ell}$ sont liées à une variable latente (continue ou discrète) qui indique un degré de soutien pour une *attitude*.
- On utilise ici un modèle à trait latent avec une seule variable latente, notée θ_k, qui est calculé pour chaque $k \in r$.

yy_1	yy_2	yy_3	yy_4
0	0	0	1
1	1	1	1
1	1	1	0
1	1	1	1
1	1	1	1
1	0	1	1
0	0	0	0
0	0	0	0

Calcul de θ_k (cas continu)

- Soit
$$q_{kj} = Pr(yy_{kj} = 1|\theta_k), j \in \{1, \ldots, m\}.$$
- Le modèle à trait latent est
$$q_{kj} = \frac{\exp(\beta_{j0} + \beta_{j1}\theta_k)}{1 + \exp(\beta_{j0} + \beta_{j1}\theta_k)}. \quad (1)$$
- Généralement $\theta_k \sim N(0, 1)$.
- Le modèle (1) est essentiellement une régression logistique excepté le fait que la variable θ_k ne peut pas être observée.
- Le modèle (1) permet l'estimation de β_{j0}, β_{j1} et θ_k par la méthode du maximum de vraisemblance.
- En principe, on peut ajouter des variables auxiliaires dans le modèle (1)
$$q_{kj} = \frac{\exp(\beta_{j0} + \beta_{j1}\theta_k + \beta_{j2}x_k)}{1 + \exp(\beta_{j0} + \beta_{j1}\theta_k + \beta_{j2}x_k)}. \quad (2)$$

Calcul de θ_k (cas discret)

- On parle de l'analyse des classes latentes.
- Les variables observées sont souvent dichotomiques et on postule l'existence d'une variable latente également qualitative à k-modalités (les classes latentes).
- Le nombre de classes esr établi sur la base des critères comme AIC, BIC etc.
- L'analyse des classes latentes peut se faire également avec des variables auxiliaires.

La méthode proposée

- Nous travaillons dans le cas de la non-réponse non-ignorable.
- Dans ce cas, Deville (2000) a proposé d'utiliser la variable d'intérêt y_{kj} comme variable instrumentale dans le calage généralisé pour réduire le biais de la non-réponse.
- Pourtant, un système différent de pondération doit être calculé pour chaque variable d'intérêt.
- Supposons que $\mathbf{y}_k = (y_{k1}, \ldots, y_{km})'$ est le vecteur des variables d'intérêt pour l'unité k.
- Nous proposons d'utiliser des variables latentes comme variables instrumentales au lieu de y_{kj} dans la méthode du calage généralisé.
- Ces variables latentes peuvent être calculées dans deux cas.

Cas I

- en utilisant $\mathbf{y}_k$ comme variables observées pour calculer la variable latente

y_1	y_2	y_3	y_4
0	0	0	1
1	1	1	1
1	1	1	0
1	1	1	1
1	1	1	1
1	0	1	1
0	0	0	0
0	0	0	0

Cas II

- en utilisant les indicateurs de réponse à chaque item comme variables observées (dans le cas où la non-réponse partielle est également présente et où on peut mettre en évidence une variable latente de type *attitude de réponse à l'enquête*).

t_1	t_2	t_3	t_4
0	0	0	1
1	1	1	1
1	1	1	0
1	1	1	1
1	1	1	1
1	0	1	1
0	0	0	0
0	0	0	0

Exemple 1

- On utilise une base de données extraite de l'enquête *Attitudes Sociales Britanniques* faite en 1986.
- On a une population de $N = 379$ individus qui ont dû répondre aux quatre questions suivantes sur l'avortement :
  Est-ce que la loi devrait autoriser l'avortement dans les cas suivants ?
  1. La femme décide de son propre chef.
  2. Le couple ne souhaite pas avoir d'enfant.
  3. La femme n'est pas mariée et ne veut pas épouser l'homme.
  4. Le couple ne peut pas se permettre d'autres enfants.
- Les données ont été analysées dans Bartholomew et all. (2002).

Deux scénarios possibles

On est en présence de non-réponse totale non-ignorable :

1. la probabilité de réponse est modélisée à l'aide d'une variable d'intérêt (par exemple y_2);
2. la probabilité de réponse dépend du sujet même de l'enquête et dans ce cas il est plus probable que cette probabilité dépende de l'attitude envers l'avortement, θ.

- La variable d'intérêt est y_2 (la deuxième variable).
- Dans la population, $\mathbf{y_k}$ détermine la variable latente *attitude envers l'avortement* θ_k (continue):

$$y_{kj} = \mu_j + \Lambda_{kj}\theta_k + e_{kj}, j = 1, \ldots, 4 \tag{3}$$

  où $e_{kj} \sim N(0, \sigma_j^2)$ et $\theta_k \sim N(0, 1)$ sont indépendantes.
- Une variable auxiliaire x_k a été générée en utilisant l'expression

$$x_k = 1 + \sum_{j=1}^{4} y_{kj} + \varepsilon_k, \varepsilon_k \sim N(0, 1). \tag{4}$$

- $cor(\mathbf{x}, \mathbf{y_2}) = 0.74, cor(\mathbf{x}, \boldsymbol{\theta}) = 0.85, cor(\mathbf{y_2}, \boldsymbol{\theta}) = 0.85$
- Nous avons comparé les estimateurs calés

$$\sum_{k \in r} y_{k2} d_k F(y_{k2} h_1) \text{ et } \sum_{k \in r} y_{k2} d_k F(\theta_k h_2)$$

  à l'aide de simulation de Monte Carlo.
- 10'000 échantillons de taille $n = 100$ ont été tirés de la population.
- On a utilisé les R packages : ltm (Rizopoulos), poLCA (Linzer et Louis), sampling(Tillé et Matei) pour calage généralisé.

Scénario 1 : Résultats I

- $N = 379, n = 100$.
- La non-réponse a été créée en utilisant un plan de Poisson avec les probabilités

$$p_k = Pr(R_k = 1|y_{k2}) = \frac{1}{1 + 2y_{k2}},$$

qui donnent une taille moyenne de r égale à 60.

- On a utilisé un ajustement linéaire tronqué avec les bornes $L = 0.5$ et $U = 5$.

total pop.	poids	estimateur	biais	var	$\sqrt{EQM}$
$Y_2 = 225$	calage_g_y2	224.67	-0.33	503.83	22.45
	calage_g_latente	196.17	-28.83	301.19	33.65
	calage_x	178.07	-46.93	268.57	49.71
$Y_1 = 166$	calage_g_y2	166.60	0.6	405.69	20.15
	calage_g_latente	166.48	0.48	307.77	17.55
	calage_x	150.45	-15.55	284.71	22.95

Scénario 1 : Résultats II

- on utilise le même contexte qu'avant, sauf que la non-réponse a été créée en utilisant un plan de Poisson avec les probabilités

$$p_k = Pr(R_k = 1|y_{k2}) = \frac{1}{\exp(0.5y_{k2} + 0.2)},$$

qui donnent une taille moyenne de r égale à 62.

- On a utilisé un ajustement de type raking ratio.

total pop.	poids	estimateur	biais	var	$\sqrt{EQM}$
$Y_2 = 225$	calage_g_y2	224.78	- 0.17	416.66	20.41
	calage_g_latente	220.21	-11.87	253.35	19.86
	calage_x	216.72	-20.98	219.54	25.69
$Y_1 = 166$	calage_g_y2	165.78	- 0.35	344.35	18.56
	calage_g_latente	167.42	3.27	326.43	18.36
	calage_x	162.97	- 7.89	285.61	18.65

Scénario 1 : Résultats III

Des résultats similaires ont été obtenus en utilisant

$$p_k = \frac{(U-C)+(C-L)*exp(A*y_{k2}/3)}{L*(U-C)+U*(C-L)*exp(A*y_{k2}/3)},$$

(Folsom and Singh, 2000) avec une moyenne de 0.66, où $L = 1, U = 3, C = 1/0.85 = 1.17, A = (U-L)/((C-L)*(U-C)), n = 100, N = 379$

total pop.	poids	estimateur	biais	var	$\sqrt{EQM}$
$Y_2 = 225$	calage_g_y2	220.23	-4.77	270.85	17.14
	calage_g_latente	212.08	-12.92	201.35	19.19
	calage_x	205.54	-19.46	192.31	23.89
$Y_1 = 166$	calage_g_y2	164.59	-1.41	297.97	17.32
	calage_g_latente	168.09	2.09	289.34	17.14
	calage_x	159.44	-6.56	259.03	17.38

Scénario 2 : Résultats

- on utilise le même contexte qu'avant, sauf que la non-réponse a été créée en utilisant un plan de Poisson avec les probabilités

$$p_k = Pr(R_k = 1|\theta_1) = \frac{1}{\exp(0.5\theta_1 + 0.2)},$$

qui donnent la taille moyenne de r égale à 63.

- On a utilisé un ajustement de type raking ratio.

total pop.	poids	estimateur	biais	var	$\sqrt{EQM}$
$Y_2 = 225$	calage_g_y2	235.25	10.25	363.20	21.64
	calage_g_latente	225.12	0.12	219.00	14.80
	calage_x	217.89	-7.11	183.57	15.30
$Y_1 = 166$	calage_g_y2	160.60	-5.40	319.34	18.67
	calage_g_latente	165.96	-0.04	312.11	17.67
	calage_x	156.05	-9.95	267.75	19.15

Exemple II

- Nous utilisons des données d'une enquête similaire : *Attitudes Sociales Britanniques* faite en 1983.
- Nous disposons également d'une variable auxiliaire réelle : âge des répondants (sept tranches d'âge; $\mathbf{x}_{n_r \times 7}$).
- La variable latente est calculée à partir d'une analyse en classes en prenant en compte également les tranches d'âge (c'est une variable discrète à deux modalités).
- Il y a des corrélations faibles entre la variable d'intérêt (y_2) et les variables auxiliaires et entre la variable latente et les variables auxiliaires (entre 0.01 et 0.12).

Remarques

- Supposons qu'on ait un seul instrument z. Le problème majeur dans le calage généralisé est la différence de dimensions entre $\mathbf{x}$ et $\mathbf{z}$ (ou $(\mathbf{1z})$).
- Pour ce type de problème trois solutions sont envisageables :
  1. appliquer la méthode de Chang et Kott (2008); Kott et Chang (2010);

$$\sum_{k \in r} w_k \mathbf{x}_k^T \mathbf{A} = \mathbf{T}_x^T \mathbf{A},$$

où $\mathbf{x}_{n_r \times p}, \mathbf{z}_{n_r \times q}, p > q, \mathbf{A}_{p \times q}$,

$$\mathbf{A} = N^{-1} \mathbf{V}^{-2} \sum_{k \in r} d_k F'(\mathbf{z}_k^T \mathbf{h}) \mathbf{x}_k \mathbf{z}_k^T,$$

et

$$\mathbf{V} = N^{-2} \sum_{k \in r} \left(d_k^2 F(\mathbf{z}_k^T \mathbf{h})^2 - d_k F(\mathbf{z}_k^T \mathbf{h}) \right) \mathbf{x}_k \mathbf{x}_k^T.$$

  2. ajouter des colonnes de $\mathbf{x}$ comme variables instrumentales ou générer des nouvelles variables à ajouter comme variables instrumentales, de telle manière que le nombre de variables auxiliaires soit égal au nombre d'instruments.
  3. réaliser un double calage quand le nombre de colonnes de $\mathbf{x}$ est grand.

Double calage

1. sélectionner parmi toutes les variables auxiliaires la variable la plus corrélée avec $\mathbf{z}$: $\widetilde{x}$;
2. réaliser un calage généralisé en utilisant $\mathbf{z}$ comme instrument et $\widetilde{x}$ comme variable auxiliaire.
3. réaliser un calage sur toutes les variables auxiliaires.

Remarques :

- L'avantage est d'obtenir des poids finaux de la forme

$$d_k F(h_1^T z_k) F(\mathbf{h}_2^T \mathbf{x}_k).$$

Scénario 1 : Résultats I - y_2 est une variable 0/1

- $N = 1437, n = 100$.
- La non-réponse a été créée en utilisant un plan de Poisson avec les probabilités

$$p_k = Pr(R_k = 1|y_{k2}) = \frac{1}{\exp(y_{k2}/2)},$$

  qui donnent une taille moyenne de r égale à 80.
- $cor(y_2, \theta) = 0.92, cor(y_4, \theta) = 0.65, cor(y_2, y_4) = 0.59$
- On a utilisé un ajustement de type raking ratio.

total pop.	poids	estimateur	biais	var	$\sqrt{EQM}$
$Y_2 = 703$	calage_g_y2	702.75	-0.25	25312.59	159.10
	calage_g_latente	590.40	-112.60	8669.91	146.12
	calage_x	532.76	-170.24	6214.11	187.60
$Y_4 = 686$	calage_g_y2	686.23	0.23	13734.07	117.19
	calage_g_latente	627.42	-58.58	7684.33	105.43
	calage_x	585.21	-100.79	6301.86	128.30

Scénario 1 : Résultats I - y_2 est une variable 1/2

- $N = 1437, n = 100$.
- La non-réponse a été créée en utilisant un plan de Poisson avec les probabilités

$$p_k = Pr(R_k = 1|y_{k2}) = \frac{1}{\exp(0.2y_{k2})},$$

qui donnent une taille moyenne de r égale à 74.

- On a utilisé un ajustement de type raking ratio.

total pop.	estimateur	valeur	biais	var	$\sqrt{EQM}$
$Y_2 = 2140$	calage_g_y2	2140.1	-0.1	9825.9	99.1
	calage_g_latente	2133.8	6.2	9362.8	96.9
	calage_x	2070.9	69.1	7080.1	108.8
$Y_4 = 2123$	calage_g_y2	2122.9	0.1	8254.5	90.8
	calage_g_latente	2127.2	-4.2	8474.4	92.1
	calage_x	2081.0	-42.0	7078.8	94.0

Conclusions

- Nous avons proposé une méthode pour réduire le biais de la non-réponse dans le cas où une variable latente peut être calculée à partir de variables d'intérêt.
- La méthode a l'avantage de créer un système unique de poids.
- L'estimateur calé en utilisant une variable latente donne de bon résultats dans les études réalisées.
- Plus de recherches doivent être faites pour estimer la variance en présence d'une variable latente θ_k.