

Correction de la non réponse par calage généralisé : une simulation

Le principe

- enquête exhaustive, l'unité sélectionnée est le logement
- l'enquête se déroule 6 ans après la mise à jour de la base de sondage
- un quart de la population ne répond pas
- le comportement de réponse apparaît corrélé aux variables X contenues dans la base de sondage
- ces mêmes variables sont collectées dans l'enquête auprès des répondants : variables Z
- on suppose que les variables Z sont de meilleurs facteurs explicatifs de la non-réponse

La population de référence

Enquête PCV d'octobre 1996

Base de sondage	1990		1996	
			Répondants (résidences principales)	Absents (non-répondants ou logements vides)
RP 1990	Résidences principales	6735	5105	1630
	Logements vides	692	279	413
BSLN		592	417	175
Total		8019	5801	2218

La simulation

- Tirage d'échantillons de répondants parmi la population des 5105 ménages ayant répondu à l'enquête PCV de 1996
- Recherche des facteurs explicatifs de la non-réponse :
 - taille du ménage (personnes seules/autres)
 - lieu de résidence (Paris/autre)
 - activité du chef de ménage (inactifs/actifs)
 - nationalité du chef de ménage (étrangers/français)

- Modèle de réponse

16 groupes homogènes de réponse, définis par le croisement des modalités des variables ci-dessus :

groupe de réponse : G_h

effectif dans U : N_h

effectif dans s : M_h

- Estimation du modèle de réponse

$$\text{Prob}\{k \in r / k \in s_h^{90}\} = \frac{M_h^{90}}{N_h^{90}} = f_h^{90}$$

La simulation

- Génération des échantillons
 - Stratification par groupes de réponse définis par le croisement des modalités des variables observées dans l'enquête :
 - Z_1 = taille du ménage en 1996
 - Z_2 = lieu de résidence
 - Z_3 = activité du chef de ménage en 1996
 - Z_4 = nationalité du chef de ménage en 1996
 - Sondage aléatoire simple dans les groupes de réponse G_h^{96} avec les taux f_h^{90} observés parmi les 6735 ménages interrogés dans l'enquête de 1996
 - 1014 échantillons de taille 3910

Le modèle de calage

- Les résultats de l'échantillon sont calés sur les totaux, dans la population de référence, des variables explicatives du comportement de réponse

$$[X] = \begin{bmatrix} SEUL90 \\ NONSEUL90 \\ INACTIF90 \\ PARIS90 \\ ETRAN90 \end{bmatrix}$$

- 1^{er} modèle : calage classique

$$[x_k] = \begin{bmatrix} seul_k^{90} \\ nonseul_k^{90} \\ inactif_k^{90} \\ paris_k^{90} \\ etran_k^{90} \end{bmatrix}$$

- 2^{ème} modèle : calage généralisé avec dissociation des variables X et Z où les variables Z sont celles corrélées au comportement de réponse mais relevées dans l'enquête

$$[x_k] = \begin{bmatrix} seul_k^{90} \\ nonseul_k^{90} \\ inactif_k^{90} \\ paris_k^{90} \\ etran_k^{90} \end{bmatrix} \quad [z_k] = \begin{bmatrix} seul_k^{96} \\ nonseul_k^{96} \\ inactif_k^{96} \\ paris_k^{96} \\ etran_k^{96} \end{bmatrix}$$

Les indicateurs

Pour une variable d'intérêt \mathcal{Y} :

○ total dans la population : $Y = \sum_{k \in U} Y_k$

○ estimateur Horvitz-Thomson dans l'échantillon s :

$$\hat{Y}_s = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_h \sum_{k \in s_h} \frac{y_k}{f_h}$$

• Moyenne empirique

$$\hat{Y}_{emp} = \frac{1}{1014} \sum_{s=1}^{1014} \hat{Y}_s$$

• Erreur quadratique moyenne empirique

$$EQM_{emp}(\hat{Y}) = \frac{1}{1013} \sum_{s=1}^{1014} (\hat{Y}_s - Y)^2$$

• Variance empirique

$$\hat{V}_{emp} = \frac{1}{1013} \sum_{s=1}^{1014} (\hat{Y}_s - \hat{Y}_{emp})^2$$

• Biais empirique

$$Biais_{emp}^2 = (\hat{Y}_{emp} - Y)^2$$

(Rappel : $EQM = Variance + Biais^2$)

MOYENNE EMPIRIQUE DES ESTIMATEURS DES TOTAUX

$$\hat{Y}_{emp} = \frac{1}{1014} \sum_{s=1}^{1014} \hat{Y}_s$$

Variables d'intérêt ayant une corrélation forte avec les variables de calage

	Population	Calage simple	Calage généralisé
nombre d'allocations perçues	2667	2604	2672
chef de ménage actif	2775	2857	2768
chef de ménage retraité	1430	1358	1437
perçoit un salaire	2997	3078	2991

Variables d'intérêt ayant une corrélation moyenne avec les variables de calage

	Population	Calage simple	Calage généralisé
a des notions d'anglais	2557	2598	2554
aucun chômage depuis 5 ans	2003	2049	2001
inscrit sur une liste électorale	4278	4278	4282
travail de nuit régulier	251	256	251
chef de ménage retiré des affaires	311	297	313
fréquence des relations non familiales	1442	1442	1442

BIAIS EMPIRIQUE (au carré) DE L'ESTIMATEUR DU TOTAL

$$Biais^2 = \left(\hat{Y}_{emp} - Y \right)^2$$

Variables d'intérêt ayant une corrélation forte avec les variables de calage

	Calage simple	Calage généralisé
nombre d'allocations perçues	3972	22
chef de ménage actif	6738	56
chef de ménage retraité	5112	50
perçoit un salaire	6509	31

Variables d'intérêt ayant une corrélation moyenne avec les variables de calage

	Calage simple	Calage généralisé
a des notions d'anglais	1698	10
aucun chômage depuis 5 ans	2158	6
inscrit sur une liste électorale	0.2	18
travail de nuit régulier	26	0.1
chef de ménage retiré des affaires	206	4
fréquence des relations non familiales	0.8	0.5

ERREUR QUADRATIQUE MOYENNE EMPIRIQUE DE L'ESTIMATEUR DU TOTAL

$$EQM_{emp}(\hat{Y}) = \frac{1}{1013} \sum_{s=1}^{1014} (\hat{Y}_s - Y)^2$$

Variables d'intérêt ayant une corrélation forte avec les variables de calage

	Calage simple	Calage généralisé
nombre d'allocations perçues	4188	690
chef de ménage actif	6908	950
chef de ménage retraité	5271	966
perçoit un salaire	6720	676

Variables d'intérêt ayant une corrélation moyenne avec les variables de calage

	Calage simple	Calage généralisé
a des notions d'anglais	2036	470
aucun chômage depuis 5 ans	2397	492
inscrit sur une liste électorale	67	262
travail de nuit régulier	89	70
chef de ménage retiré des affaires	298	145
fréquence des relations non familiales	809	827

CALAGE GÉNÉRALISÉ : VARIANCE ESTIMÉE PAR LA VARIANCE DES RESIDUS DE REGRESSION

Variabes d'intérêt ayant une corrélation forte avec les variables de calage

	variance vraie	moyenne des estimateurs de variance
nombre d'allocations perçues	681	667
chef de ménage actif	871	851
chef de ménage retraité	912	882
perçoit un salaire	667	648

Variabes d'intérêt ayant une corrélation moyenne avec les variables de calage

	variance vraie	moyenne des estimateurs de variance
a des notions d'anglais	459	454
aucun chômage depuis 5 ans	511	494
inscrit sur une liste électorale	278	268
travail de nuit régulier	72	71
chef de ménage retiré des affaires	134	133
fréquence des relations non familiales	863	844