APPLICATION DU CALAGE GÉNÉRALISÉ A LA CORRECTION DE LA NON-RÉPONSE : UNE EXPÉRIMENTATION

Josiane LE GUENNEC [(*)], Olivier SAUTORY[(**)]

[(*)] Cepe - Ensai
[(**)] Cepe - Insee

Le calage généralisé permet de redresser la non-réponse dans une enquête même lorsque les caractéristiques individuelles les plus corrélées à la non-réponse ne sont connues que dans l'échantillon des personnes ayant répondu au questionnaire.

Parmi les situations de ce type figure le cas où la non-observation des variables auxiliaires de calage chez les non-répondants résulte du décalage temporel entre les informations contenues dans la base de sondage et la réalisation du sondage. Les enquêtes de l'INSEE auprès des ménages, dont les échantillons sont tirés dans le dernier recensement de population, entrent dans cette catégorie dès que le recensement s'éloigne dans le temps.

La comparaison entre l'échantillon des répondants et celui des non-répondants fait généralement ressortir chez ces derniers une prédominance de personnes vivant seules, dans une grande ville et plus encore à Paris, de personnes âgées inactives et des personnes étrangères plus récemment arrivées en France. Ces caractéristiques : âge, activité, nationalité, nombre de personnes habitant le logement, taille de la commune de résidence, sont relevées lors du recensement, donc présentes dans la base de sondage. Néanmoins, lorsque l'enquêteur passe en 2002, c'est évidemment la situation en 2002 de la personne interrogée qui est susceptible d'expliquer son comportement vis-à-vis de l'enquête, non celle du ménage habitant le même logement en 1999, qui a pu changer. Tout se passe comme si les variables Z explicatives de la non-réponse (valeurs au moment de l'enquête) étaient observées sur les seuls répondants, et leurs totaux Z dans la population inconnus.

On a testé le calage généralisé pour redresser la non-réponse dans cette configuration. Ces tests ont pris la forme d'une série de simulations, dont les résultats sont présentés ici. Toutes ont pris pour cadre l'enquête permanente sur les conditions de vie des ménages (PCV) réalisée en octobre 1996.

1. La non-réponse dans l'enquête PCV de 1996

Le questionnaire de cette enquête s'intéresse aux conditions d'emploi, à la formation, aux relations sociales et aux pratiques de loisir des personnes de 15 ans ou plus.

L'échantillon initial comprend 7999 ménages, tirés principalement dans le fichier du recensement de la population de 1990, et de façon complémentaire dans celui des logements neufs (BSLN) construits depuis 1990. Répondants et non-répondants à l'enquête de 1996 se répartissent ainsi :

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[1]	5801	2218

Afin de construire des simulations basées sur un modèle réaliste de réponse, on a recherché les facteurs corrélés à la non-réponse effectivement rencontrée dans l'enquête de 1996. Celle-ci ne peut être analysée que pour les ménages dont on connaît les caractéristiques dans la base de sondage avant tirage. C'est pourquoi on a exclu les logements tirés dans la BSLN ainsi que ceux tirés dans le recensement de 1990, mais vides à l'époque. L'étude a donc porté sur les 6735 logements de l'échantillon initial occupés en 1990.

Sur ces 6735 ménages interrogés, 1630 sont absents du fichier des répondants, soit un taux de non-réponse apparent de 24 %. Ce chiffre comprend nécessairement, outre les non-répondants véritables, des logements devenus inoccupés en 1996. Aucune information dans le fichier de l'enquête ne permettant de les distinguer, tous sont assimilés à des non-répondants.

L'analyse discriminante a conduit à retenir 4 variables binaires corrélées à la non-réponse :

- la taille du ménage (personnes seules/autres ménages)
- l'activité du chef de ménage (actifs/inactifs)
- le lieu de résidence (agglomération parisienne/autres communes)
- la nationalité du chef de ménage (français/étranger)

La fonction discriminante estimée conduit cependant à mal classer 45 % environ des non-répondants.

Le croisement des 8 modalités décrites ci-dessus définit des groupes homogènes de réponse. Au total, on a 16 groupes non vides, dans lesquels on a calculé les taux de non-réponse effectivement obtenus dans l'enquête. Pour ce calcul, un ménage est ici classé dans le groupe G_h^{90} correspondant aux caractéristiques des occupants du même logement en 1990, d'après la base de sondage :

$$\text{taux de réponse} = r_h = \frac{\text{Rep}(G_h^{90})}{\text{Ech}(G_h^{90})}$$

[1] Parmi les 5801 répondants, certains logements résultent de l'éclatement de logements initialement tirés, d'où un total supérieur à l'échantillon de base de 7999 logements.

où Rep est le nombre de répondants et Ech l'effectif de l'échantillon initial dans le groupe homogène de réponse *h*.

On a observé les taux de non-réponse suivants dans les marges de cette distribution :

- personnes seules : 34,7 %
- inactifs : 29,6 %
- à Paris : 33,5 %
- étrangers : 34,8 %.

2. Le modèle

Afin d'isoler, dans un calage, l'effet dû au strict redressement de la non-réponse, on étudie celle-ci dans une enquête exhaustive, et non dans une enquête par sondage.

L'échantillon PCV de 1996 étant pris comme population de référence, la non-réponse est simulée par tirage d'échantillons aléatoires de répondants dans l'ensemble des ménages pour lesquels on dispose à la fois des informations de la base de sondage et des réponses à l'enquête.

2.1 Le modèle de réponse

Les 5105 ménages ayant répondu à l'enquête de 1996 constituent la population de référence pour simuler la non-réponse dans une enquête exhaustive.

Les variables discriminantes identifiées ci-dessus et présentes dans la base de sondage sont également relevées dans l'enquête de 1996. Pour chaque ménage de cette population, on dispose donc de la valeur des variables *X* : activité, taille du ménage, nationalité et commune de résidence en 1990 et de la valeur des variables Z constituées des mêmes variables observées en 1996. Chaque ménage peut être classé dans un groupe G_h^{90} correspondant aux caractéristiques relevées en 1990 dans la base de sondage et dans un groupe G_h^{96} correspondant à ses caractéristiques véritables d'après l'enquête en 1996.

On suppose que le vrai modèle de réponse est basé sur la valeur des variables Z au moment de l'enquête. C'est pourquoi les échantillons de répondants sont simulés par le tirage d'échantillons stratifiés selon les groupes G_h^{96}, par sondage aléatoire simple sans remise dans les groupes. Les taux de sondage par groupe utilisés sont les r_h calculés précédemment (voir ci-dessus). On a donc dans chaque groupe :

$$M_h = r_h \times N(G_h^{96}) \text{ répondants}$$

où $N(G_h^{96})$ est l'effectif total du groupe *h* dans l'enquête et r_h le taux de réponse calculé précédemment.

Il s'ensuit que les taux de réponse f_{h90} dans les groupes G_h^{90} définis par les valeurs des variables X dans la base de sondage (c'est-à-dire en 1990) sont distincts des taux de réponse r_h. Alors que ceux-ci sont fixes, f_{h90} varie d'un échantillon à l'autre.

La permanence dans le temps des ménages dans leurs logements est néanmoins suffisante pour que la stratification construite sur les variables Z assure une corrélation, même affaiblie, entre l'appartenance à l'échantillon et la valeur des variables X.

Cette méthode nous a donné des échantillons de 3910 répondants. On a simulé 1014 échantillons.

2.2 Le modèle de calage

Deux calages ont été testés, qui utilisent les variables explicatives du comportement de réponse : le premier avec la méthode classique, soit le vecteur $\mathbf{X}$ des valeurs de la base de sondage, le second avec le calage généralisé dissociant les vecteurs $\mathbf{X}$ (valeurs de la base de sondage) et $\mathbf{Z}$ (valeurs dans l'enquête de 1996).

Test n°1 : $$[X] = \begin{bmatrix} SEUL90 \\ NONSEUL90 \\ INACTIF90 \\ PARIS90 \\ ETRAN90 \end{bmatrix}$$

Test n°2 : $$[X] = \begin{bmatrix} SEUL90 \\ NONSEUL90 \\ INACTIF90 \\ PARIS90 \\ ETRAN90 \end{bmatrix} \qquad [Z] = \begin{bmatrix} SEUL96 \\ NONSEUL96 \\ INACTIF96 \\ PARIS96 \\ ETRAN96 \end{bmatrix}$$

Dans les deux cas, les estimateurs ont été calés sur les totaux X dans la population des variables explicatives de la non-réponse, et non sur les effectifs des groupes homogènes de réponse définis par le croisement des modalités concernées, le nombre de groupes étant jugé trop important.

2.3 Les variables d'intérêt

L'effet des différents modes de redressement a été mesuré par l'estimation des totaux $\hat{Y}$ des variables suivantes (les variables quantitatives sont repérées par une *, les autres variables sont des variables indicatrices 1/0) :

Variables bien corrélées à l'ensemble des facteurs de non-réponse :

- chef de ménage actif
- chef de ménage retraité
- perception d'un salaire
- nombre d'allocations perçues (*) : RMI, allocation de chômage, retraite

Variables moyennement corrélées à l'ensemble des facteurs de non-réponse :

- aucune période de chômage déclarée
- ne travaille jamais la nuit
- a des notions d'anglais
- inscription sur une liste électorale
- chef de ménage retiré des affaires

Variables peu corrélées aux facteurs de non-réponse, ou corrélées à un seul de ces facteurs :

- travail de nuit, régulier ou occasionnel
- période de chômage au cours des 5 dernières années
- fréquence des rencontres en-dehors de la famille (*)

- nombre d’adhésions à une association autre que de retraités ou du 3ème âge (*)
- nombre d’adhésions à une association du 3ème âge ou de retraités (*).

3. Les résultats des simulations

3.1 Les indicateurs

L'efficacité des modèles de calage testés est jugée à travers la moyenne empirique des estimateurs des variables d'intérêt et leur erreur quadratique moyenne empirique :

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

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

où $\hat{Y}_s$ est l'estimateur du total de la variable Y dans l'échantillon *s* et Y la vraie valeur de ce total dans la population de référence.

Rappelons que l'erreur quadratique moyenne, dans le cas d'un estimateur biaisé (et c'est le cas ici en raison de la non-réponse), est égale à la somme de la variance et du carré du biais :

$$EQM\left(\hat{Y}\right) = E\left[\hat{Y} - E\left(\hat{Y}\right)\right]^2 + \left[E\left(\hat{Y}\right) - Y\right]^2$$

On présente également la variance empirique et le carré du biais empirique calculé comme la différence entre la moyenne empirique et la vraie valeur dans la population.

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

3.2 Les résultats

Le calage généralisé, qui utilise en variables instrumentales les facteurs de non-réponse observés à la date de l'enquête, réduit sensiblement le biais d'échantillonnage, par rapport au calage n'utilisant que les variables de la base de sondage et par rapport à un redressement uniforme de la non-réponse.

La moyenne empirique de la distribution approche mieux la vraie valeur dans la population (tableau 1) et le biais moyen est très fortement réduit (tableau 2). Il en résulte une meilleure précision globale, l'erreur quadratique moyenne diminuant sensiblement (tableau 3).

L'amélioration est d'autant plus marquée que les variables d'intérêt sont plus corrélées aux facteurs de non-réponse. Pour les autres, le calage généralisé améliore l'estimateur, ou en donne un équivalent au calage ordinaire.

Ce résultat s'obtient au prix d'un accroissement de la variabilité de l'estimateur autour de sa moyenne. L'EQM diminue, mais la variance est en général plus importante (tableau 4). Le calage généralisé améliore la précision de l'estimation en réduisant le biais. En conséquence, les intervalles de confiance estimés avec la variance empirique recouvrent mieux la vraie valeur des paramètres (tableau 5).

Les tableaux qui suivent donnent les estimations obtenues avec une méthode linéaire de calage. Le calage avec les fonctions exponentielle ou logit a donné des résultats équivalents, tant en moyenne qu'en intervalle de confiance.

4. L'estimateur de variance

Plaçons-nous dans un premier temps dans le cas d'une enquête par sondage. Le gestionnaire d'enquête doit apprécier la précision de ses résultats en estimant la variance dans son échantillon unique.

4.1 Calage simple sans non-réponse

De façon générale, avec une fonction de calage linéaire, l'estimateur obtenu après calage est égal à l'estimateur redressé par régression de la variable d'intérêt Y sur les *n* observations dans l'échantillon des variables de calage X :

$$\hat{Y}_{cal} = \hat{Y}_p + \hat{B}'\left(X - \hat{X}_p\right) = \hat{B}'X + \hat{U}_p$$

où : $\hat{B} = (X'DX)^{-1}X'DY$

$$\hat{Y}_p = \sum_{k\in s}\frac{y_k}{p_k}$$

$$\hat{X}_p = \sum_{k\in s}\frac{x_k}{p_k}$$

$$\hat{U}_p = \sum_{k\in s}\frac{\hat{u}_k}{p_k} = \sum_{k\in s}\frac{y_k - \hat{B}'x_k}{p_k}$$

La matrice $[X]$ (n,p) est celle des *n* observations dans l'échantillon des variables X et le vecteur Y contient les *n* observations dans l'échantillon de la variable d'intérêt Y. $[D]$ est la matrice diagonale (n,n) contenant les poids de sondage initiaux avant calage : $d_k = \frac{1}{p_k}$. $\hat{u}_k$ est l'écart entre la vraie valeur y_k de la variable d'intérêt observée sur l'individu k et la valeur prédite par le modèle pour cet individu.

Une approximation de la vraie variance de l'estimateur calé est donnée par celle des résidus individuels de la régression ci-dessus, mesurés avec les coefficients $\tilde{B}$ estimés dans la population.

En l'absence de non-réponse, elle a pour expression :

$$V\left(\hat{Y}_{cal}\right) \approx V\left(\sum_{k\in s}\frac{u_k}{p_k}\right) = V\left(\sum_{k\in s}\frac{y_k - \tilde{B}'x_k}{p_k}\right) = \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{u_k}{p_k}\frac{u_l}{p_l}$$

$$\tilde{B} = (X'X)^{-1}X'Y$$

où $[X]$ et Y sont respectivement la matrice des variables de calage et le vecteur des variables d'intérêt dans la population.

Elle est estimée par :

$$\hat{V}\left(\hat{Y}_{cal}\right) = \hat{V}\left(\sum_{k\in s}\frac{\hat{u}_k}{p_k}\right) = \hat{V}\left(\sum_{k\in s}\frac{y_k - \hat{B}'x_k}{p_k}\right) = \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{p_{kl}}g_k\frac{\hat{u}_k}{p_k}g_l\frac{\hat{u}_l}{p_l} \qquad (1)$$

avec : $g_k = w_k \pi_k$, les w_k étant les poids de calage.

Ou de façon alternative par :

$$\hat{V}\left(\hat{Y}_{cal}\right)=\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{p_{kl}}\left(\frac{\hat{u}_k}{p_k}\right)\left(\frac{\hat{u}_l}{p_l}\right)$$

où : $\hat{u}_k = y_k - \hat{B}' x_k$, $\hat{B}$ étant estimé dans l'échantillon.

4.2 Calage simple avec non-réponse

En présence de non-réponse, lorsque les variables de calage expliquent le comportement de réponse, le calage corrige simultanément le biais dû à l'échantillonnage initial et la non-réponse par re-pondération. La non-réponse peut être assimilée à une deuxième phase du sondage effectuée par tirage stratifié avec sondage aléatoire simple dans les groupes homogènes de réponse, au taux f_h égal au taux de réponse observé dans le groupe *h*.

L'estimateur du total de la variable Y devient alors :

$$\hat{Y}_{cal} = \hat{Y}_{rep} + \left(X - \hat{X}_{p^*}\right)' \hat{B}_1 = X\hat{B}_1 + \hat{U}_{p^*}$$

où : $\hat{Y}_{rep} = \sum_h \sum_{k\in r_h} \frac{y_k}{p_k f_h}$

$$\hat{X}_{p^*} = \sum_h \sum_{k\in s_h} \frac{x_k}{p_k f_h}$$

$$p^*_k = p_k f_h$$

$$\hat{B}_1 = \left(\sum_h \frac{1}{f_h}\sum_{k\in r_h}\frac{x_k x_k^{'}}{p_k}\right)^{-1}\sum_h \frac{1}{f_h}\sum_{k\in r_h}\frac{x_k y_k}{p_k} = (\mathsf{X}'\mathsf{D}^*\mathsf{X})^{-1}\,\mathsf{X}'\mathsf{D}^*\,\mathsf{Y}$$

$$\mathsf{D}^* = Diag\left(\frac{1}{p_k f_h}\right)$$

$[\mathsf{X}]$ étant la matrice des variables de calage dans l'échantillon des répondants.

La variance de l'estimateur après calage s'estime encore comme celle des résidus de la régression de Y sur $[\mathsf{X}]$, mais selon l'expression propre au sondage en deux phases :

$$V\left(\hat{Y}_{cal}\right) \approx V_1\left(E_2\left(\hat{U}_{p^*}\right)\right) + E_1\left(V_2\left(\hat{U}_{p^*}\right)\right) = \sum_U\sum_U \Delta_{kl}\frac{\tilde{u}_k}{p_k}\frac{\tilde{u}_l}{p_l} + E_1\left[\left(\sum_h n_h^2 \frac{1-f_h}{m_h} S_h^2\left[\frac{\tilde{u}_k}{p_k}\right]\right)/\mathsf{s}\right]$$

où : $\tilde{u}_k = y_k - \tilde{B}_1^{'} x_k$ avec $\tilde{B}_1$ estimé dans la population, et m_h est le nombre de répondants dans l'échantillon de taille n_h dans le groupe de réponse *h*.

Estimateur de la variance :

$$\hat{V}\left(\hat{Y}_{cal}\right)=\sum_{r}\sum_{r}\frac{\Delta_{kl}}{p^{*}_{kl}}\frac{\hat{u}_k}{p_k}\frac{\hat{u}_l}{p_l}+\sum_{h}n_h^2\frac{1-f_h}{m_h}s_h^2\left[\frac{\hat{u}_k}{p_k}\right] \qquad (2)$$

où : $\hat{u}_k = y_k - \hat{B}_1' x_k$ avec $\hat{B}_1$ estimé sur l'échantillon de répondants

$$s_h^2\left[\frac{\hat{u}_k}{p_k}\right]=\frac{1}{m_h-1}\sum_{k\in r_h}\left(\frac{\hat{u}_k}{p_k}-\bar{\hat{u}}_h\right)^2, \quad \bar{\hat{u}}_h=\frac{1}{m_h}\sum_{k\in h_h}\frac{\hat{u}_k}{p_k}$$

4.3 Calage généralisé

Avec le calage généralisé et la fonction de calage linéaire, l'estimateur $\hat{Y}$ est égal à l'estimateur redressé au moyen d'une régression instrumentale de la variable Y sur les variables X, utilisant les variables Z en instrument, ce qui nous donne :

$$\hat{Y}_{calgen}=\hat{Y}_p+\hat{B}'\left(X-\hat{X}_p\right)$$

où : $\hat{\mathrm{B}}=\left(\mathrm{Z}'\mathrm{D}\,\mathrm{X}\right)^{-1}\left(\mathrm{Z}'\mathrm{D}\,\mathrm{Y}\right)$ et $\mathrm{D}=Diag\left(\frac{1}{p_k}\right)$ en l'absence de non-réponse

$\hat{\mathrm{B}}=\left(\mathrm{Z}'\mathrm{D}^{*}\mathrm{X}\right)^{-1}\left(\mathrm{Z}'\mathrm{D}^{*}\mathrm{Y}\right)$ et $\mathrm{D}^{*}=Diag\left(\frac{1}{p_k f_h}\right)$ en présence de non-réponse

Comme précédemment, la variance est estimée par celle des résidus individuels de la régression.

4.4 Résultats

Dans cette simulation, l'enquête initiale étant exhaustive et la non-réponse assimilée à un sondage aléatoire simple stratifié par groupe homogène de réponse, on estime la variance des résidus par la formule habituelle d'un plan équiprobable sans remise :

$$\hat{V}\left(\hat{Y}_{cal}\right)=\sum_{h}N_h^2\frac{1-f_h}{M_h}s_h^2 \qquad (3)$$

avec : $f_h=\frac{M_h}{N_h}$ = taux de réponse dans le groupe homogène de réponse G_h^{96}

$$s_h^2=\frac{1}{M_h-1}\sum_{k\in G_h}\left(\hat{u}_k-\bar{\hat{u}}_h\right)^2 \qquad (4)$$

$$\bar{\hat{u}}_h = \frac{1}{M_h} \sum_{k \in G_h} \hat{u}_k$$

Les formules (2) et (3) appellent les remarques suivantes :

- dans le cas d'une enquête exhaustive réelle et non plus d'une simulation, N_h, effectif du groupe de réponse G_h^{96} dans la population, est inconnu. Il faudrait donc l'estimer par la somme des poids de calage w_k dans le groupe h : $\hat{N}_h = \sum_{k \in G_h} w_k$.

  D'où : $$\hat{V}\left(\hat{Y}_{cal}\right) = \sum_h \hat{N}_h^2 \frac{1 - \hat{f}_h}{M_h} s_h^2$$

(5)

- dans le cas d'une enquête par sondage, l'effectif m_h de répondants dans le groupe homogène de réponse G_h^{96} est observé mais aléatoire. L'effectif échantillonné n_h et par conséquent le taux de réponse f_h dans le groupe G_h^{96} sont inconnus et peuvent être estimés par :

$$\hat{f}_h = \frac{\hat{M}_h}{\hat{N}_h} = \frac{\sum_{k \in r_h} \frac{1}{p_k}}{\sum_{k \in r_h} w_k} \qquad \hat{n}_h = \frac{m_h}{\hat{f}_h}$$

Commentaires

Dans cette simulation de 1014 échantillons, on a estimé la variance dans chaque échantillon en utilisant pour N_h la vraie valeur dans la population. Par ailleurs, on a mesuré la variance vraie des estimateurs, à partir des résidus de la régression dans la population de référence (et donc du coefficient $\tilde{B}$) (tableau 6).

On voit qu'il y a une bonne convergence entre la variance vraie des estimateurs et la variance empirique de la distribution présentée dans le tableau 4. Dans le cas présent, l'estimateur obtenu par calage simple est biaisé, puisqu'il ne correspond pas exactement au vrai modèle de réponse. La variance des résidus approche bien la variance de l'estimateur, mais non l'erreur quadratique incluant le biais. Cet écart est très fortement réduit avec l'estimateur par calage généralisé, celui-ci étant quasiment sans biais.

On a calculé les intervalles de confiance avec les variances estimées par celle des résidus, dans chaque échantillon simulé. Le tableau 7 montre, pour chaque variable d'intérêt, le nombre d'échantillons ne recouvrant pas la vraie valeur dans la population selon le redressement effectué. On voit qu'il est très fortement réduit dans la distribution obtenue avec le calage généralisé.

Tableau 1 - MOYENNE EMPIRIQUE[2] DES ESTIMATEURS DES TOTAUX

Corrélation forte avec les variables auxiliaires

Calage variable	Population	Echantillon non redressé	Redressement uniforme[3]	Calage simple	Calage généralisé
nombre d'allocations perçues	2667	1972.00	2574.70	2603.98	2671.54
chef de ménage actif	2775	2220.08	2898.60	2857.08	2767.53
chef de ménage retraité	1430	1016.70	1327.43	1358.51	1437.09
perçoit un salaire	2997	2385.19	3114.17	3077.68	2991.42
ne perçoit pas de salaire	2108	1524.81	1990.83	2027.32	2113.58

Corrélation moyenne avec les variables auxiliaires

Calage variable	Population	Echantillon non redressé	Redressement uniforme	Calage simple	Calage généralisé
a des notions d'anglais	2557	1995.02	2604.75	2598.21	2553.81
sans notions d'anglais	2548	1914.98	2500.25	2506.79	2551.19
aucun chômage depuis 5 ans	2003	1588.23	2073.63	2049.45	2000.55
inscrit sur une liste électorale	4278	3282.82	4286.14	4277.54	4282.25
non inscrit sur une liste électorale	827	627.18	818.86	827.46	822.75
travail de nuit régulier	251	199.14	260.00	256.07	250.64
chef de ménage retiré des affaires	311	225.96	295.02	296.65	312.93

Corrélation faible avec les variables auxiliaires

Calage variable	Population	Echant.non redressé	Redress. uniforme	Calage simple	Calage généralisé

[2] $\hat{\bar{Y}}_{emp} = \frac{1}{1014}\sum_{s=1}^{1014}\hat{Y}_s$, où $\hat{Y}_s$ est l'estimateur du total de la variable Y dans l'échantillon s

[3] Echantillon repondéré de façon uniforme par l'inverse du taux de réponse : n/m

nbre d'adhésions à une association	7273.00	5643.36	7368.12	7232.97	7301.49
nbre d'adhés. à une assoc. du 3ème age	1213.00	913.62	1192.84	1180.75	1219.29
une période de chômage dans l'année	82.00	65.56	85.60	86.04	81.70
chômage une fois il y a plus d'un an	198.00	155.66	203.23	203.80	198.69
plusieurs périodes de chômage il y a plus d'un an	81.00	63.40	82.77	82.75	80.67
travail de nuit occasionnel	552.00	437.61	571.36	564.10	550.43
travail de nuit=jamais	1559.00	1234.51	1611.82	1599.87	1558.53
chef de ménage chômeur	293.00	227.08	296.48	299.13	291.74
chef de ménage étudiant	60.00	41.37	54.01	56.46	59.72
chef de ménage femme au foyer	71.00	53.45	69.79	70.94	71.33
chef de ménage autre inactif	163.00	123.77	161.59	163.96	162.65
fréquence des relations non familiales	1442.67	1105.16	1442.92	1441.77	1441.99

Tableau 2 - BIAIS EMPIRIQUE[4] (AU CARRÉ) DE L'ESTIMATEUR DU TOTAL

Corrélation forte avec les variables auxiliaires

Variable	Echantillon non redressé	Redressement uniforme	Calage simple	Calage généralisé
nombre d'allocations perçues	483020.89	8519.38	3971.71	20.6233
chef de ménage actif	307935.24	15276.13	6737.74	55.8751
chef de ménage retraité	170818.36	10520.96	5111.17	50.3234
perçoit un salaire	374313.48	13727.72	6509.41	31.1909
ne perçoit pas de salaire	40108.67	13727.72	6509.41	31.1909

Corrélation moyenne avec les variables auxiliaires

Variable	Echantillon non redressé	Redressement uniforme	Calage simple	Calage généralisé
a des notions d'anglais	315819.61	2280.39	1698.09	10.1814
sans notions d'anglais	400716.47	2280.39	1698.09	10.1814
aucun chômage depuis 5 ans	172036.79	4988.76	2157.91	6.0165
inscrit sur une liste électorale	990374.36	66.31	0.21	18.0364
non inscrit sur une liste électorale	39929.81	66.31	0.21	18.0364
travail de nuit régulier	2689.56	81.02	25.67	0.1263
chef de ménage retiré des affaires	7231.88	255.40	205.96	3.7125

Corrélation faible avec les variables auxiliaires

Variable	Echantillon non redressé	Redressement uniforme	Calage simple	Calage généralisé

[4] $Biais^2 = \left(\hat{\bar{Y}}_{emp} - Y \right)^2$, où $\hat{\bar{Y}}_{emp}$ est la moyenne empirique et Y le total dans la population.

nbre d'adhésions à une association	2655729.87	9047.73	1602.56	811.753
nbre d'adhés. à une assoc. du 3ème âge	89629.97	406.30	1040.25	39.615
une période de chômage dans l'année	270.17	12.97	16.33	0.089
chômage une fois il y a plus d'un an	1792.86	27.36	33.62	0.475
chômage plusieurs fois il y a plus d'un an	309.82	3.15	3.06	0.106
travail de nuit occasionnel	13084.07	374.85	146.53	2.468
travail de nuit=jamais	105290.65	2789.46	1670.08	0.224
chef de ménage chômeur	4345.98	12.09	37.54	1.579
chef de ménage étudiant	347.19	35.88	12.53	0.078
chef de ménage femme au foyer	307.84	1.46	0.00	0.107
chef de ménage autre inactif	1539.21	1.98	0.92	0.124
fréquence des relations non familiales	113912.57	0.07	0.80	0.459

Tableau 3 - ERREUR QUADRATIQUE MOYENNE EMPIRIQUE[5] DE L'ESTIMATEUR DU TOTAL

Corrélation forte avec les variables auxiliaires

variable	Echantillon non redressé	Redressement uniforme	Calage simple	Calage généralisé
nombre d'allocations perçues	483606.47	8713.19	4187.94	690.018
chef de ménage actif	308309.92	15411.71	6907.65	950.294
chef de ménage retraité	171052.96	10643.82	5270.66	965.726
perçoit un salaire	374793.30	13929.31	6720.16	675.990

Corrélation moyenne avec les variables auxiliaires

variable	Echantillon non redressé	Redressement uniforme	Calage simple	Calage généralisé
a des notions d'anglais	316323.63	2610.37	2036.45	469.676
aucun chômage depuis 5 ans	172341.64	5223.87	2397.42	492.408
inscrit sur une liste électorale	991443.64	222.54	167.17	262.132
travail de nuit régulier	2728.72	143.33	89.19	69.812
chef de ménage retiré des affaires	7289.58	341.84	297.72	144.962

Corrélation faible avec les variables auxiliaires

variable	Echantillon non redressé	Redressement uniforme	Calage simple	Calage généralisé
nbre d'adhésions à une association	2723885.25	120769.45	109560.33	118665.34
nbre d'adhés. à une assoc. du 3ème age	92069.87	4415.09	4954.25	4419.04
une période de chômage dans l'année	282.29	33.19	37.75	21.00
chômage une fois il y a plus d'un an	1824.04	77.52	86.48	56.56
chômage plusieurs fois il y a plus d'un an	323.56	26.05	26.71	24.05
travail de nuit occasionnel	13172.09	503.25	279.40	162.11
travail de nuit=jamais	105533.84	3029.59	1917.18	386.09
chef de ménage chômeur	4394.17	86.92	117.30	95.13
chef de ménage étudiant	359.44	56.21	34.94	27.60
chef de ménage femme au foyer	320.57	22.64	22.01	25.61
chef de ménage autre inactif	1567.24	47.17	47.44	57.22
fréquence des relations non familiales	14474.68	766.58	809.42	826.9

[5] $EQM_{emp}\left(\hat{Y}\right)=\frac{1}{1013}\sum_{s=1}^{1014}\left(\hat{Y}_s-Y\right)^2$, où $\hat{Y}_s$ est l'estimateur du total de la variable Y dans l'échantillon s et Y la vraie valeur de ce total dans la population.

Tableau 4 - VARIANCE EMPIRIQUE[6] DES ESTIMATEURS

Corrélation forte avec les variables auxiliaires

variable	Redressement uniforme	Calage simple	Calage généralisé
nombre d'allocations perçues	185.401	212.310	669.374
chef de ménage actif	120.507	163.264	894.364
chef de ménage retraité	112.467	154.446	915.353
perçoit un salaire	188.039	204.324	644.768

Corrélation moyenne avec les variables auxiliaires

variable	Redressement uniforme	Calage simple	Calage généralisé
a des notions d'anglais	327.728	336.685	459.484
aucun chômage depuis 5 ans	230.178	237.386	486.386
inscrit sur une liste électorale	156.167	166.957	244.077
travail de nuit régulier	62.225	63.491	69.686
chef de ménage retiré des affaires	86.191	91.562	141.246

Corrélation faible avec les variables auxiliaires

variable	Redressement uniforme	Calage simple	Calage généralisé
nbre d'adhésions à une association	111712.79	107956.19	117852.78
nbre d'adhésions à une assoc. du 3ème age	4008.39	3912.97	4379.38
une période de chômage dans l'année	20.21	21.40	20.91
une période de chômage il y a plus d'un an	50.13	52.83	56.08
chômage plusieurs fois il y a plus d'un an	22.90	23.64	23.94
travail de nuit occasionnel	128.03	132.72	159.64
travail de nuit=jamais	237.37	245.45	385.86
chef de ménage chômeur	74.83	79.73	93.55
chef de ménage étudiant	20.29	22.40	27.52
chef de ménage femme au foyer	21.18	22.00	25.51
chef de ménage autre inactif	45.19	46.52	57.10
fréquence des relations non familiales	766.52	808.62	826.48

[6] $\hat{V}_{emp} = \frac{1}{1013}\sum_{s=1}^{1014}\left(\hat{Y}_s - \hat{\bar{Y}}_{emp}\right)^2$, où $\hat{Y}_s$ est l'estimateur du total de la variable Y dans l'échantillon s et $\hat{\bar{Y}}_{emp}$ la moyenne empirique de la distribution.

Tableau 5 - INTERVALLES DE CONFIANCE À 95 % (avec la variance empirique)

Corrélation forte avec les variables auxiliaires

	Calage simple			Calage généralisé		
Variable	borne inf.	borne sup.	vraie valeur hors IC	borne inf.	borne sup.	vraie valeur hors IC
nombre d'allocations perçues	2575.42	2632.54	*	2620.83	2722.25	
chef de ménage actif	2832.04	2882.13	*	2708.91	2826.14	
chef de ménage retraité	1334.15	1382.87	*	1377.79	1496.39	
perçoit un salaire	3049.66	3105.70	*	2941.65	3041.18	
ne perçoit pas de salaire	1999.30	2055.34	*	2063.82	2163.35	

Corrélation moyenne avec les variables auxiliaires

	Calage simple			Calage généralisé		
Variable	borne inf.	borne sup.	vraie valeur hors IC	borne inf.	borne sup.	vraie valeur hors IC
a des notions d'anglais	2562.24	2634.17	*	2511.80	2595.82	
sans notions d'anglais	2470.83	2542.76	*	2509.18	2593.20	
aucun chômage depuis 5 ans	2019.25	2079.65	*	1957.32	2043.77	
inscrit sur une liste électorale	4252.21	4302.86		4251.63	4312.87	
non inscrit sur liste électorale	802.14	852.79		792.13	853.37	
travail de nuit régulier	240.45	271.68		234.28	267.01	
chef de ménage retiré des affaires	277.89	315.40		289.63	336.22	

Corrélation moyenne avec les variables auxiliaires

	Calage simple			Calage généralisé		
Variable	borne inf.	borne sup.	vraie valeur hors IC	borne inf.	borne sup.	vraie valeur hors IC
nbre d'adhésions à une assoc.	6588.98	7876.96		6628.63	7974.35	
adhés. à une assoc.du 3ème âge	1058.14	1303.35		1089.59	1349.00	
chômage une fois dans l'année	76.97	95.11		72.74	90.67	
chômage une fois il y a plus d'un an	189.55	218.04		184.01	213.37	
chômage plusieurs fois il y a plus d'un an	73.22	92.28		71.08	90.26	

travail de nuit occasionnel	541.52	586.69		525.67	575.19
travail de nuit=jamais	1569.16	1630.57	*	1520.03	1597.03
chef de ménage chômeur	281.63	316.63		272.79	310.70
chef de ménage étudiant	47.18	65.74		49.44	70.00
chef de ménage femme au foyer	61.75	80.13		61.43	81.23
chef de ménage autre inactif	150.59	177.33		147.84	177.46
fréquence des relations non familiales	1386.04	1497.51		1385.64	1498.34

Tableau 6 - VARIANCE ESTIMÉE PAR LA VARIANCE DES RÉSIDUS DE RÉGRESSION

Corrélation forte avec les variables auxiliaires

	Calage simple		Calage généralisé	
Variable	variance vraie	moyenne des estimateurs	variance vraie	moyenne des estimateurs
nombre d'allocations perçues	232.797	229.514	680.717	666.621
chef de ménage actif	167.673	165.187	871.236	851.215
chef de ménage retraité	172.233	168.417	911.668	880.498
perçoit un salaire	215.616	210.334	666.593	647.853

Corrélation moyenne avec les variables auxiliaires

	Calage simple		Calage généralisé	
Variable	variance vraie	moyenne des estimateurs	variance vraie	moyenne des estimateurs
a des notions d'anglais	343.926	338.910	458.542	453.994
aucun chômage depuis 5 ans	252.048	248.606	510.892	494.126
inscrit sur une liste électorale	185.969	181.790	278.181	268.334
travail de nuit régulier	64.925	64.079	71.801	70.626
chef de ménage retiré des affaires	101.531	100.868	133.634	133.176

Corrélation faible avec les variables auxiliaires

	Calage simple		Calage généralisé	
Variable	variance vraie	moyenne des estimateurs	variance vraie	moyenne des estimateurs
nbre d'adhésions à une association	112734.98	112461.93	113172.29	113016.55
nbre d'adhés. à une assoc. du 3ème âge	4112.89	4096.09	4210.78	4196.14
une période de chômage dans l'année	20.86	20.74	21.77	21.64
chômage une fois il y a plus d'un an	53.50	52.60	57.12	56.11

chômage plusieurs fois il y a plus d'un an	24.05	23.51	24.80	24.20
travail de nuit occasionnel	129.53	127.92	153.52	150.65
travail de nuit=jamais	251.42	247.98	401.30	390.50
chef de ménage chômeur	84.34	82.66	98.61	96.17
chef de ménage étudiant	28.99	27.09	29.95	27.97
chef de ménage femme au foyer	24.54	23.91	26.40	25.72
chef de ménage autre inactif	51.14	50.05	59.91	58.54
fréquence des relations non familiales	863.04	841.60	863.30	843.63

Tableau 7 - INTERVALLES DE CONFIANCE À 95 % CALCULÉS AVEC LA VARIANCE ESTIMÉE

NOMBRE D'ÉCHANTILLONS AYANT LA VRAIE VALEUR HORS IC

Corrélation forte avec les variables auxiliaires

variable	Calage simple	Calage généralisé
nombre d'allocations perçues	1002	51
chef de ménage actif	1014	64
chef de ménage retraité	1013	57
perçoit un salaire	1014	51

Corrélation moyenne avec les variables auxiliaires

variable	Calage simple	Calage généralisé
a des notions d'anglais	618	64
aucun chômage depuis 5 ans	852	53
inscrit sur une liste électorale	39	57
travail de nuit régulier	94	51
chef de ménage retiré des affaires	301	60

Corrélation faible avec les variables auxiliaires

variable	Calage simple	Calage généralisé
nbre d'adhésions à une association	55	61
nbre d'adhésions à une association du 3ème age	104	63
une période de chômage dans l'année	135	51
une période de chômage il y a plus d'un an	112	51
plusieurs périodes de chômage il y a plus d'un an	50	49
chômage =non concerné	988	54
travail de nuit occasionnel	199	61
travail de nuit=jamais	753	54
travail de nuit=non concerné	986	52
chef de ménage chômeur	102	53
chef de ménage étudiant	98	58
chef de ménage militaire	39	39
chef de ménage femme au foyer	42	49
chef de ménage autre inactif	43	55
fréquence des relations non familiales	66	66