Faut-il pondérer ?

...Ou l'éternelle question de l'économètre confronté à un problème de sondage

Laurent Davezies et Xavier D'Haultfœuille

Juin 2009

Résumé

Ce papier précise dans quels cas les estimations effectuées sur données d'enquête doivent être pondérées. Nous montrons que les estimateurs pondérés sont plus robustes que les estimateurs non pondérés, même s'ils sont moins précis lorsque ces derniers sont valides. Dans certains cas, la comparaison des deux estimateurs permet de tester la compatibilité entre les hypothèses concernant le mécanisme de sélection des observations (i.e., le plan de sondage et la non-réponse) et celles portant sur le modèle économétrique considéré. Enfin, quelques méthodes de calcul de précision des estimateurs pondérés sont présentées.

1 Introduction

La situation du chargé d'études utilisant des données issues d'une enquête est quelque peu schizophrène. D'un côté, il est prêt à estimer les modèles économétriques qu'on lui a enseigné, où toutes les observations se valent puisque l'échantillonnage est supposé i.i.d [1].De l'autre, il ne peut ignorer que ces données sont issues d'un sondage, et qu'à ce titre, il devrait utiliser un estimateur basé sur des poids de sondage. Comment réconcilier ces deux approches, alors que leur formalisme semble incompatible ?

L'objet de cette note est de montrer que même si l'on reste dans le cadre pratique de données i.i.d., les variables de poids sont en général utiles voire indispensables. Ainsi, nous montrons

1. La question de l'utilisation de poids est peu abordée dans la littérature économétrique. Le papier se rapprochant le plus de la note ci-présente est celui de Wooldridge (2007). Voir également Lerman et Manski (1977) ou Imbens (1992) pour un traitement des "choice based samples", qui sont abordés dans l'exemple 6 ci-dessous.

qu'en général, les estimateurs pondérés sont plus robustes que les estimateurs non pondérés, dans le sens qu'ils convergent sous des hypothèses moins restrictives. La contrepartie, comme toujours en statistique, est une perte de précision. La note met également en exergue l'importance des variables utilisées dans le calage et éventuellement dans le plan de sondage lorsque le tirage de l'échantillon est à probabilités inégales. Ceci signifie deux choses : d'une part l'économètre se doit de connaître ces variables [2], et de l'autre, le choix des variables utilisées dans le redressement a une influence sur les outils économétriques utilisables par la suite.

Cette note s'articule comme suit. La deuxième partie rappelle la définition des pondérations en théorie des sondages, et définit le cadre d'analyse retenu ici. La troisième présente les résultats de convergence des estimateurs pondérés et non pondérés, suivant les hypothèses retenues sur le mécanisme de sélection et la nature du paramètre estimé. La quatrième partie indique comment calculer la précision d'estimateurs pondérés et les pièges à éviter dans ce cadre. Pour conclure, une feuille de route résume succintement les différentes étapes de la démarche présentée. Les détails techniques des résultats sont renvoyés en annexe.

2 Définition des pondérations et cadre retenu

2.1 Les pondérations en sondage

Dans le cadre standard de sondages, on considère une population notée $U = \{1, ..., N\}$ dans laquelle on tire aléatoirement un échantillon $\mathcal{S} \subset U$ [3]. En l'absence de non-réponse, la variable usuelle de pondération de l'individu k est définie par

$$w_k = \frac{1}{P(k \in \mathcal{S})}.$$

Ces poids sont connus puisqu'ils sont définis au moment du tirage de l'échantillon. Ils permettent de calculer l'estimateur de Horvitz-Thompson $\widehat{t}$ d'un total $t_y = \sum_{k \in U} y_k$ d'une variable y :

$$\widehat{t} = \sum_{k \in \mathcal{S}} w_k y_k.$$

Rappelons que l'intérêt d'un tel estimateur est d'être sans biais, et ce quelles que soient les valeurs de $(y_k)_{k \in U}$. En pratique, une partie de l'échantillon est non-répondante. On note

2. En général, ces informations sont disponibles dans la documentation de l'enquête. Dans le cas contraire, on pourra retrouver ces informations à partir de notes du service producteur.

3. La présentation donnée ici est très succinte, pour davantage de détails sur le formalisme de la théorie des sondages, on se référera par exemple à Tillé (2001).

alors $\mathcal{R} \subset \mathcal{S}$ l'échantillon des répondants. Dans ce cas la variable de poids à utiliser pour calculer l'estimateur de Horvitz-Tompson serait $1/P(k \in \mathcal{R})$. Cependant, cette probabilité est inconnue puisqu'on ne connaît pas la probabilité de réponse des individus. La procédure habituelle consiste alors à estimer les poids par un modèle de non réponse et/ou par calage [4]. On a alors :

$$w_k = \frac{1}{\widehat{P}(k \in \mathcal{R})}.$$

2.2 Le modèle adopté ici

Pour relier ces variables de poids aux modèles économétriques, on adopte dans la suite un formalisme un peu différent. Plus précisément, on considère un modèle de "superpopulation" où les valeurs individuelles sur la population U sont des réalisations de variables i.i.d. Par exemple, $(y_1, ..., y_N)$ sont les réalisations des variables aléatoires i.i.d. $(Y_1, ..., Y_N)$, de même loi que Y. On suppose que l'échantillon des répondants $\mathcal{R}$ est défini par $\mathcal{R} = \{i \in U / D_i = 1\}$, où $(D_1, ..., D_N)$ sont également i.i.d., de même loi que D. On a $D = S \times R$ où S est l'indicatrice d'appartenance à l'échantillon initial et R est l'indicatrice de réponse à l'enquête. S est supposée dépendre de variables $\widetilde{X}_1$ [5]. On note $\widetilde{X}_2$ les variables expliquant R, c'est-à-dire les variables ayant été incluses dans le modèle de non-réponse et/ou le calage. Soit alors $\widetilde{X} = (\widetilde{X}_1, \widetilde{X}_2)$ l'ensemble des variables qui influent sur D. On considérera que la variable de poids W est alors définie par

$$W = \frac{1}{\widehat{P}(D = 1 | \widetilde{X})}.$$

Deux remarques s'imposent à ce stade. Premièrement, le modèle d'échantillonnage que l'on considère ici correspond, en théorie des sondages, à un plan poissonnien, c'est-à-dire à un plan où chaque individu est tiré indépendamment des autres. Cette hypothèse permet de s'assurer que l'on reste dans un cadre i.i.d. standard en économétrie. Elle est cependant peu réaliste pour un certain nombre de sondages, en particulier les sondages aréolaires. Notons que si cette hypothèse n'est pas vérifiée, les estimateurs de variance proposés en section 4 ne seront pas convergents ; en revanche la convergence des estimateurs présentés en section 3 ne sera pas affectée. Deuxièmement, on supposera par la suite que W tend vers

4. Soit $\widetilde{x}_{2k}$ des variables auxilaires dont les totaux $t_{\widetilde{x}_2}$ sur la population entière sont connus. Le calage a pour but de déterminer les poids w_k les plus proches possibles de $1/P(k \in \mathcal{S})$ tels que les estimateurs des totaux $\sum_{k \in \mathcal{S}} w_k \widetilde{x}_{2k}$ soient exacts, i.e. égaux aux totaux $t_{\widetilde{x}_2}$. Pour plus de détails, on se référera à Deville et Sarndal (1992) ou Sautory (1993) sur le calage, et à Deville (2002) sur le calage généralisé.

5. Ainsi, $\widetilde{X}_1$ inclut par exemple l'effectif en tranches dans les enquêtes annuelles d'entreprises de l'INSEE.

$1/P(D=1|\widetilde{X})$. Cette hypothèse n'est pas anodine car la plupart du temps, l'estimation de cette probabilité repose sur une forme paramétrique (comme un logit ou un probit dans un modèle de non-réponse). Une mauvaise spécification conduit en général à des estimateurs non-convergents.

3 Pondérer ou non, telle est la question.

On s'intéresse maintenant à l'estimation d'un paramètre θ dépendant potentiellement d'une variable expliquée Y et de variables explicatives X, avec a priori $X \neq \widetilde{X}$. Pour toutes variables aléatoires (U,V), on note F_U (resp. $F_{U|V}$) la loi de U (resp. la loi conditionnelle de U sachant V) et $U \perp\!\!\!\perp V$ pour "U est indépendante de V". Le paramètre θ est donc fonction de $F_{X,Y}$.

3.1 Les variables $\widetilde{X}$ sont les facteurs pertinents de la sélection

On suppose tout d'abord que les concepteurs d'enquêtes ont réussi à capter, à travers $\widetilde{X}$, les facteurs pertinents de la sélection de l'échantillon. Cette hypothèse se traduit formellement par :

H0. $(Y,X) \perp\!\!\!\perp D \mid \widetilde{X}$.

Considérons par exemple une enquête ménages à probabilités de tirage égales et où la non-réponse a été corrigée par calage sur le type de ménage (variable TYPMEN) et l'âge de la personne de référence (variable AGEPR). Dans ce cas on a $\widetilde{X}_1 = \emptyset$ et $\widetilde{X}_2 =$(TYPMEN, AGEPR). Si l'on s'intéresse à un modèle où le salaire Y de la personne de référence est expliqué par le nombre d'années d'études X, l'hypothèse H0 stipule qu'à type de ménage et âge de la personne de référence fixée, les répondants et non-répondants de l'enquête ont la même distribution de salaires et du nombre d'années d'études. Ceci exclut par exemple que le salaire ait un effet "propre" (i.e. à âge et type de ménage fixé) sur la probabilité de réponse à l'enquête.

3.1.1 Un cas particulier : on peut pondérer, mais on ne doit pas.

Outre H0, on suppose vérifiée l'hypothèse suivante.

H1. θ dépend uniquement de $F_{Y|X}$ et $\widetilde{X} \subset X$.

Résultat 1 : sous les hypothèses H0 et H1, les estimateurs seront convergents que l'on pondère ou non. L'estimateur non pondéré sera cependant plus précis.

La deuxième partie de l'hypothèse H1 suppose simplement que les variables $\widetilde{X}$ utilisées pour calculer les pondérations sont incluses dans l'estimation. Cela nécessite naturellement que l'économètre connaisse ces variables. Notons qu'il n'est pas toujours souhaitable d'inclure l'ensemble des $\widetilde{X}$ dans les explicatives du modèle. En effet, certaines composantes de $\widetilde{X}$ peuvent être endogènes pour le problème étudié. C'est le cas par exemple si $\widetilde{X}$ inclut la CSP et que l'on cherche à estimer une équation de salaire. Le cas où $\widetilde{X} \not\subset X$ est traité dans la section suivante.

Illustrons maintenant, à travers des exemples, la première partie de l'hypothèse H1.

Exemple 1 : régression linéaire. On suppose ici

$$Y = X'\theta + \varepsilon, \qquad E(\varepsilon|X) = 0 \tag{3.1}$$

Dans ce cas, θ dépend uniquement de $F_{Y|X}$ puisque

$$\theta = \frac{\partial E(Y|X=x)}{\partial x}.$$

Considérons maintenant une régression linéaire instrumentale, i.e. telle que $E(\varepsilon|Z) = 0$ [6]. Alors sous la condition de rang que les différentes fonctions $(E(X_k|Z))_{k=1,\dots,K}$ sont linéairement indépendantes (où X_k désigne la k-ième composante de X), θ est défini par l'équation

$$E(Y|Z) = E(X|Z)'\theta.$$

Ainsi, θ ne dépend que de $F_{Y,X|Z}$. Lorsque $\widetilde{X} \subset Z$, les estimateurs pondérés et non pondérés seront alors convergents, et l'estimateur non pondéré sera plus précis.

Exemple 2 : modèle logit / probit. Ici

$$Y = \mathbb{1}\{X'\theta + \varepsilon \geq 0\}$$

où $-\varepsilon \perp\!\!\!\perp X$, de fonction de répartition F. Dans ce cas,

$$\theta = \frac{\partial}{\partial x}\left[F^{-1}\left(E(Y|X=x)\right)\right].$$

Plus généralement que le modèle logit / probit, considérons un modèle conditionnel paramétrique où $F_{Y|X}$ dépend d'un paramètre η, i.e. s'écrit $F_{Y|X,\eta}$. Si $\theta = f(\eta)$, alors on peut montrer [7] que θ ne dépend que de $F_{Y|X}$.

6. Notons que l'hypothèse H0 s'écrit ici $(Y, X, Z) \perp\!\!\!\perp D|\widetilde{X}$.

7. Pour cela, il faut également supposer que le modèle est identifiable, i.e., $F_{Y|X,\eta} = F_{Y|X,\eta'}$ implique $\eta = \eta'$.

Exemple 3 : régression non paramétrique. On s'intéresse ici à la fonction $\theta(x) = E(Y|X = x)$. $\theta(.)$ dépend évidemment uniquement de $F_{Y|X}$.

3.1.2 Le cas général : on doit pondérer.

On considère maintenant les situations où H1 n'est plus satisfaite, H0 restant vérifiée.

Résultat 2 : Supposons que H0 soit satisfaite, mais pas H1. Dans ce cas, les estimateurs pondérés seront convergents, mais pas les estimateurs non pondérés en général.

L'intuition, dans ce cas, est que θ dépendra de la distribution des X. Or la distribution de X dans l'échantillon ne "correspond" pas à celle de la population (par exemple, les diplômés du supérieur sont sous représentés dans l'échantillon des répondants car leur probabilité de réponse est plus faible). L'estimateur non pondéré sera, de ce fait, non convergent en général.

Exemple 4 : statistiques simples. On s'intéresse par exemple à $\theta = E(Y)$. Ce paramètre ne dépend pas uniquement de $F_{Y|X}$, mais aussi de la distribution des X (puisque $E(Y) = E(E(Y|X))$). Ici, l'estimateur pondéré n'est autre que l'estimateur de Horvitz-Thompson. Ceci vaut en fait pour toute fonction de la distribution de Y telle que la variance, les quantiles...

Exemple 1 (suite). Revenons sur le modèle linéaire, mais avec une hypothèse plus faible que (3.1) :

$$Y = X'\theta + \varepsilon, \qquad E(X\varepsilon) = 0 \tag{3.2}$$

L'hypothèse $E(\varepsilon|X) = 0$ revient à supposer que la meilleure approximation en X de Y est linéaire et vaut $X'\theta$, alors que $E(X'\varepsilon) = 0$ revient juste à supposer que $X'\theta$ est la meilleure approximation linéaire en X de Y. Cette dernière hypothèse est donc beaucoup plus faible [8]. Dans ce cas θ dépend en général de la loi jointe de (X, Y) et non uniquement de la loi de $Y|X$. Ainsi, l'estimateur non pondéré sera non convergent en général. Ceci peut être illustré par l'exemple suivant. Considérons le modèle simple :

$$\ln(\text{salaire}) = a + b \times \text{ années d'éducation} + \varepsilon$$

Supposons qu'en fait, les rendements de l'éducation soient décroissants [9]. Le paramètre

8. En particulier cette hypothèse n'est pas testable alors que $E(\varepsilon|X) = 0$ l'est.

9. Ce n'est pas tant la décroissance qui importe ici, mais le fait que les rendements sont non constants avec le nombre d'année d'éducation.

b correspond alors à une sorte de rendement moyen sur l'ensemble de la population (cf. graphique 1) [10]. Supposons maintenant que la probabilité de répondre à l'enquête diminue avec le nombre d'années d'étude. Dans ce cas l'échantillon comprendra davantage de personnes peu éduquées, pour qui le rendement est élevé. Un estimateur non pondéré surestimera alors en général b. L'estimateur pondéré, en revanche, accentuera l'importance des quelques individus très éduqués présents dans l'enquête, et conduira à un estimateur convergent.

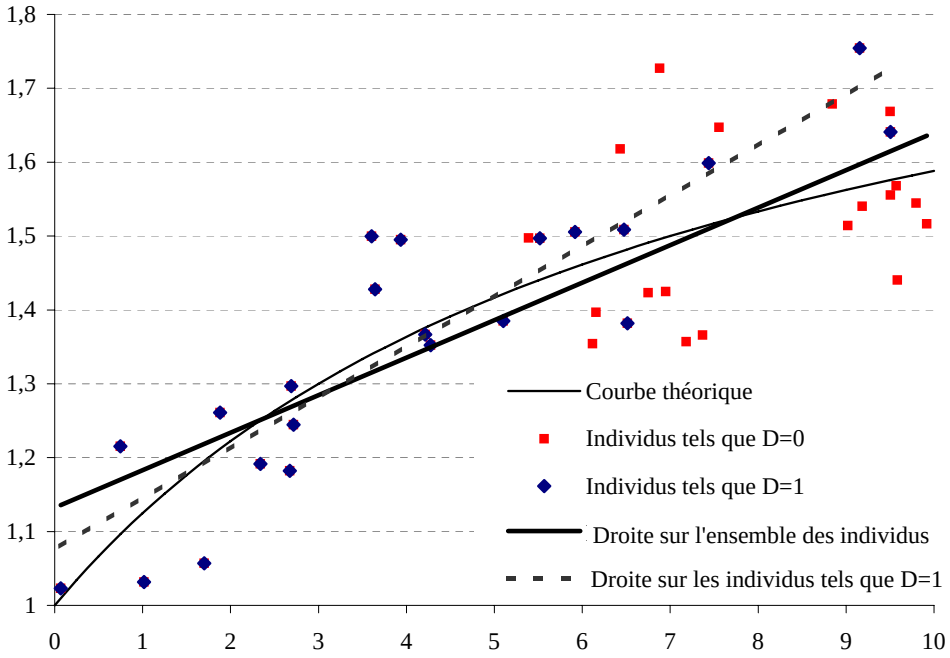

FIGURE 1: Exemple 1, estimation des rendements de l'éducation.

Notons qu'il est également nécessaire de pondérer lorsque $E(\varepsilon|X) = 0$ mais $\widetilde{X} \not\subset X$. Ceci vaut également pour la régression instrumentale décrite précédemment. Dès lors que $\widetilde{X} \not\subset Z$, les deux étapes de cette régression doivent être pondérées.

Exemple 2 (suite). On s'intéresse, dans le cadre des modèles logit ou probit, à l'effet marginal moyen $\widetilde{\theta}_k$ de la variable X_k (on note les autres variables X_{-k}). $\widetilde{\theta}_k$ est alors défini par

$$\widetilde{\theta}_k = E\left[\frac{\partial E(Y|X)}{\partial x_k}\right].$$

que l'on peut réécrire

$$\widetilde{\theta}_k = E\left[F'(X'\theta)\right]\theta_k$$

10. Ceci n'est pas tout à fait exact car en fait, $b \neq E(\partial E(Y|X)/\partial x)$ en général.

où θ_k est la k-ième composante de θ. Ainsi, $\widetilde{\theta}_k$ dépend également de la loi de X, à travers le premier terme. Dans ce cas, on peut adopter l'approche hybride d'estimer θ sans poids, puis utiliser ceux-ci dans l'estimation de $E\left[F'(X'\theta)\right]$.

Exemple 5 : matching. Le même argument s'applique à l'estimation des "effets de traitement" par matching. Supposons que l'on veuille mesurer l'efficacité d'une politique publique. Soit $T = 1$ si l'individu est concerné par la politique (i.e.,"traité"), $T = 0$ sinon. Supposons que l'on veuille mesurer (par exemple) l'effet moyen sur les traités

$$\Delta^{TT} = E(Y_1 - Y_0|T = 1)$$

où Y_1 (resp. Y_0) est la variable d'intérêt qu'aurait l'individu s'il était traité (resp. non traité). On observe seulement $Y = TY_1 + (1-T)Y_0$ mais on fait l'hypothèse d'indépendance conditionnelle :

$$Y_0 \perp\!\!\!\perp T|X.$$

Cette hypothèse est nécessaire pour que les estimateurs de matching soient convergents. On a, sous cette condition,

$$\Delta^{TT} = E(Y - E(Y|T = 0, X)|T = 1).$$

Le terme $E(Y|T = 0, X)$ correspond à l'exemple 3 de la régression non-paramétrique [11]. Il peut donc être estimé sans les poids, par matching ou par un estimateur à noyau par exemple. En revanche, Δ^{TT} est une moyenne simple (de la variable $Y - E(Y|T = 0, X)$) qui correspond à l'exemple 4. Pour l'estimer de manière convergente, il faut donc pondérer.

Exemple 6 : "choice-based sample". Lorsque la variable d'intérêt binaire Y intervient dans l'échantillonnage, on parle de "choice-based sample" (cf. Lerman et Manski, 1977). C'est le cas par exemple si l'on s'intéresse à la probabilité d'être cadre à partir d'un échantillon où les cadres sont surreprésentés [12]. On a alors $Y \subset \widetilde{X}$, et on ne peut dans ce cas avoir $\widetilde{X} \subset X$ [13]. L'estimateur non pondéré sera donc en général non convergent, contrairement à l'estimateur pondéré.

3.2 Certains facteurs de la sélection ont été omis de $\widetilde{X}$

On peut identifier deux situations principales où l'hypothèse H0 n'est pas vérifiée :

11. Il s'agit en effet de la régression non-paramétrique sur la sous-population des individus tels que $T = 0$.

12. cf. l'enquête ECMOSS (Enquête sur le Coût de la Main d'Oeuvre et la Structure des Salaires) de l'INSEE.

13. En revanche, l'hypothèse H0 peut tout à fait rester valide.

1. Le concepteur d'enquête a été "négligent", par exemple en redressant la non-réponse sur un nombre insuffisant de variables. Dès lors, $\widetilde{X}$ n'inclut pas tous les facteurs pertinents de la sélection. En revanche, les X considérés par l'économètre incluent bien l'ensemble de ces facteurs. Dans ce cas, on aura

   H0'. $Y \perp\!\!\!\perp D|X$.

   Lorsque $\widetilde{X} \subset X$, cette hypothèse est moins forte que H0 puisque cette dernière implique H0' (cf. annexe). Lorsque $\widetilde{X} \not\subset X$, en revanche, aucune des deux hypothèses n'est plus restrictive a priori, et seul le contexte peut permettre de trancher entre les deux.

2. Même conditionnellement à $(X, \widetilde{X})$, le processus de sélection est lié à Y. Il est possible, dans une enquête sur le patrimoine des ménages, que les ménages fortunés soient plus récalcitrants à répondre toutes choses égales par ailleurs. Dans ce cas, ni H0 ni H0' ne seront satisfaites.

Considérons ces deux situations successivement.

3.2.1 Cas 1

Dans ce cas, une première solution est de reconstruire une nouvelle variable de poids $W' = 1/P(D = 1|X)$. L'hypothèse H0 sera satisfaite si l'on remplace $\widetilde{X}$ par X et les résultats de la section précédente s'appliquent.

La construction de cette variable peut cependant s'avérer fastidieuse. Elle est même impossible en l'absence d'information auxiliaire (i.e., sur les non-répondants ou via une autre source comme des fichiers administratifs) sur X. Dans ce cas, on a le résultat suivant.

Résultat 3 : Sous l'hypothèse H0' et si θ dépend uniquement de $F_{Y|X}$, un estimateur non pondéré de θ sera convergent et "efficace". En revanche, les paramètres dépendant de la loi jointe de (X, Y) ne pourront pas être estimés de manière convergente en général.

Par ailleurs, l'estimateur pondéré (en utilisant les poids W) sera convergent mais inefficace lorsque $\widetilde{X} \subset X$. Il sera non convergent en général lorsque $\widetilde{X} \not\subset X$.

3.2.2 Cas 2

Dans cette dernière situation, les estimateurs pondérés et non pondérés seront en général non convergents. Pour résoudre ce problème, il existe deux grandes classes de solutions :

1. Les modèles de sélection ;
2. Les méthodes de calage généralisé.

Ces deux solutions nécessitent l'existence d'instruments Z vérifiant des relations d'exclusion et disponibles sur toute la population (c'est-à-dire qu'on les observe sur les répondants et soit sur les non-répondants, soit dans une autre source comme un fichier administratif). Elles diffèrent quant à la nature de la relation d'exclusion. Dans le premier cas, on suppose que Z joue sur l'appartenance à l'échantillon des répondants mais pas directement sur Y :

$$Y \perp\!\!\!\perp Z|X \tag{3.3}$$

Par exemple, Z pourrait correspondre à des caractéristiques d'enquêteur. Ces caractéristiques sont susceptibles d'expliquer la non-réponse, mais ne jouent pas directement sur Y. On peut alors obtenir des estimateurs convergents en corrigeant de la sélection par des modèles à la Heckman. Pour plus de détails sur les modèles de sélection, on se référera par exemple à Wooldridge (2002).

La méthode de calage généralisé (Deville, 2002), quant à elle, repose sur l'hypothèse que Z joue sur Y, mais est indépendant de la sélection conditionnellement à $(\widetilde{X}_2, Y)$:

$$Z \perp\!\!\!\perp D|\widetilde{X}_2, Y \tag{3.4}$$

Par exemple, dans une enquête sur la santé, la sélection peut être corrélée à un indicateur de santé Y car les personnes hospitalisées ne sont pas interrogées. On pourra utiliser comme instrument Z, si cette variable est connue, le nombre de médecins dans la zone. Cette variable est en effet corrélée a priori avec la santé des individus, mais pas directement à la réponse à l'enquête. De façon générale, si les concepteurs d'enquête pensent qu'un phénomène de ce type est en jeu, ils ont intérêt à utiliser le calage généralisé en incluant dans $\widetilde{X}_2$ l'ensemble des facteurs déterminant la sélection. On est alors ramené au cadre de la partie 3.1, et tous les résultats correspondants sont valides.

On peut mettre en œuvre le calage généralisé sous SAS grâce à la macro calmar2 [14], avec la syntaxe suivante [15] :

```
%calmar2(datamen=table_entree, marmen=table_marge, poids=inv_prob,
        nonrep=oui, datapoi=table_sortie, poidsfin=W)
```

14. Au moment où cette note a été écrite, la version la plus récente de la macro calmar2 ne permettait pas d'estimer des poids en cas de sur-identification, c'est à dire lorsque la dimension de Z est supérieure à celle de Y.

15. Cette syntaxe n'inclut pas toutes les possibilités de la macro calmar2. Pour davantage de détails, cf. Le Guennec et Sautory (2005). On portera une attention particulière à leurs notations, qui diffèrent de celles adoptées ici.

Dans cet exemple, la table table_entree contient $Y, \widetilde{X}_2, Z$ et les inverses des probabilités de tirage inv_prob (i.e., $1/P(k \in S)$). La table table_marge contient les totaux des variables $\widetilde{X}_2$ et Z. La macro calmar2 crée la table table_sortie qui contient les mêmes variables que table_entree, ainsi que les poids W issus du calage généralisé.

3.3 Comparaison des estimateurs pondérés et non pondérés

L'objet de cette partie est de montrer qu'il est souvent possible de tester la compatibilité des hypothèses retenues sur la sélection et celles du modèle que l'on cherche à estimer. On suppose pour ce faire que $\widetilde{X} \subset X$ et que H0 ou H0' est vérifiée.

3.3.1 Modèles paramétriques

Considérons ici, comme dans l'exemple 2, un modèle paramétrique $F_{Y|X,\eta}$ avec $\theta = f(\eta)$. Les estimateurs pondérés et non pondérés convergent vers le même paramètre. De plus, si θ est estimé par maximum de vraisemblance sans les pondérations, il est asymptotiquement efficace, c'est-à-dire de variance asymptotique minimale.

Dans ce genre de situations, on peut, à l'aide d'un test d'Hausman, tester la compatibilité du processus de sélection (i.e. l'hypothèse H0 ou H0') et du modèle paramétrique retenu $F_{Y|X,\eta}$. Ainsi, dans l'exemple 2, il se peut ou que l'hypothèse sur la sélection soit fausse, ou que la forme fonctionnelle retenue $P(Y = 1|X) = F(X'\theta)$ ne soit pas correcte. Le test d'Hausman est basé sur la différence des estimateurs pondéré et non pondéré. Sous l'hypothèse nulle que H0 ou H0' et le modèle paramétrique sont compatibles, ils doivent en effet être proches l'un de l'autre alors que sinon, ils ne convergeront pas a priori vers la même valeur. Si on note $\widehat{\theta}_W$, l'estimateur calculé avec utilisation des poids et $\widehat{\theta}$ celui sans utilisation des poids, la statistique du test d'Hausman est la suivante :

$$T_H = \left(\widehat{\theta}_W - \widehat{\theta}\right)' \left[\widehat{V}\left(\widehat{\theta}_W\right) - \widehat{V}\left(\widehat{\theta}\right)\right]^{-1} \left(\widehat{\theta}_W - \widehat{\theta}\right).$$

Sous l'hypothèse nulle, on a, lorsque la taille de l'échantillon tend vers l'infini, on a $T_H \xrightarrow{L} \chi^2(r)$ avec

$$r = \mathrm{rg}\left(V\left(\widehat{\theta}_W\right) - V\left(\widehat{\theta}\right)\right).$$

De plus, T_H tend en général vers $+\infty$ sous l'hypothèse alternative. Ainsi, on rejettera au niveau α la compatibilité des hypothèses si $T_H > \chi^2_r(1-\alpha)$, où $\chi^2_r(1-\alpha)$ est le quantile d'ordre $1-\alpha$ d'un χ^2 à r degrés de liberté.

Dans le cas où l'hypothèse est rejetée, trois interprétations sont possibles. Soit on maintient

l'hypothèse que le processus de sélection vérifie H0 ou H0', ce qui revient à rejeter la validité du modèle. Soit on maintient l'hypothèse que le modèle est valide, ce qui conduit à rejeter H0 ou H0'. On peut évidemment également rejeter les deux hypothèses. Mais dans tous les cas, on ne peut maintenir simultanément les deux hypothèses.

3.3.2 Modèles linéaires

La discussion précédente portait sur les modèles paramétriques, ce qui exclut l'exemple important des modèles linéaires puisque le modèle est semi-paramétrique. Cependant le test est également valable dans ce cas car l'estimateur des MCO non pondéré est également optimal dans la classe des estimateurs linéaires sans biais sous l'hypothèse H0 ou H0', d'après le théorème de Gauss-Markov. La statistique T_H définie ci-dessus converge donc là-aussi, sous l'hypothèse nulle, vers un χ^2 à r degrés de liberté.

Si l'implémentation du test nécessite de façon générale le recours à la programmation matricielle (via la PROC IML sous SAS), il n'en est rien pour le modèle linéaire. Au lieu de faire les deux regressions séparées et de calculer "à la main" la différence des variances des deux estimateurs, on peut remarquer qu'une solution équivalente est d'estimer le système d'équation suivant :

$$\begin{cases} Y = X'\theta_1 + \varepsilon_1 \\ \sqrt{W}Y = \sqrt{W}X'\theta_2 + \varepsilon_2 \end{cases}$$

Dans ce cadre, le test d'Hausman revient au test de $\theta_1 = \theta_2$. Sous SAS, on peut alors utiliser la PROC SYSLIN pour estimer les deux modèles et implémenter le test [16], en utilisant la syntaxe suivante :

```
data a;
    set a;
    sqrtw=sqrt(w);            /* Racine carré des poids */
    xp=sqrt(w)*x;
    yp=sqrt(w)*y;
run;
```

16. Dans la proc syslin, on estime $V\left(\widehat{\theta}_W - \widehat{\theta}|X\right)$ sans utiliser le fait que, quand $\widehat{\theta}$ est efficace sous l'hypothèse nulle, on a :

$$V\left(\widehat{\theta}_W - \widehat{\theta}|X\right) = V\left(\widehat{\theta}_W|X\right) - V\left(\widehat{\theta}|X\right).$$

Cependant, l'estimateur de la variance reste convergent et le test est donc valide.

```
proc syslin data=a;
    model y=x;                /* Première équation du système ci-dessus */
    model yp=sqrtw xp/noint; /* Deuxième équation, sans la constante   */
                              /* (elle est captée par sqrtqw)           */
    stest y.x=yp.xp;          /* Test d'Hausman  */
run;
```

Hyp. sur la sélection	$\widetilde{X} \subset X$?	θ dépend de...	Est. pond. convergent ? (si oui, efficace ?)	Est. non pond. convergent ? (si oui, efficace ?)	Test possible entre les deux ?
$(Y, X) \perp\!\!\!\perp D \vert \widetilde{X}$	oui	$F_{Y\vert X}$	oui (non)	oui (oui)	oui
$(Y, X) \perp\!\!\!\perp D \vert \widetilde{X}$	non	$F_{Y\vert X}$ ou $F_{X,Y}$	oui (non)	non	non
$(Y, X) \perp\!\!\!\perp D \vert \widetilde{X}$	n'importe	$F_{X,Y}$	oui (non)	non	non
$Y \perp\!\!\!\perp D \vert X$	oui	$F_{Y\vert X}$	oui (non)	oui (oui)	oui
$Y \perp\!\!\!\perp D \vert X$	non	$F_{Y\vert X}$	non	oui (oui)	non
$Y \perp\!\!\!\perp D \vert X$	n'importe	$F_{X,Y}$	non	non	non
Non ignorable	n'importe	$F_{Y\vert X}$ ou $F_{X,Y}$	non*	non**	non

* : sauf en calculant les poids par calage généralisé, et sous l'hypothèse (3.4).

** : sauf en utilisant une procédure à la Heckman, et sous l'hypothèse (3.3).

TABLE 1: Propriétés des estimateurs pondérés et non pondérés

3.4 Récapitulatif

Le tableau de la page précédente présente, en fonction du mécanisme de sélection et de la nature du paramètre θ, si les estimateurs pondérés et non pondérés sont convergents ou non. Dans le cas où ils le sont, il précise lequel est asymptotiquement efficace. A l'exception d'un seul cas, les estimateurs pondérés seront convergents à chaque fois que les estimateurs non pondérés le sont. Il existe en revanche des situations importantes où les estimateurs pondérés seront convergents alors que les estimateurs non pondérés ne le seront pas.

4 Calcul de précision avec des poids

Il s'agit ici de savoir si la précision calculée par les logiciels d'estimateurs pondérés est "correcte", i.e. si les estimateurs des écarts-types correspondants sont convergents ou non. Comme précisé précédemment, on néglige ici le plan de sondage précis de l'enquête en supposant que les indicatrices d'appartenance à l'échantillon des répondants $(D_1, ..., D_N)$ sont i.i.d [17]. La règle simple (mais inconfortable !) à retenir est la suivante :

Même si les logiciels de statistique prévoient en général l'inclusion de poids (via par exemple l'instruction weight sous SAS), les estimateurs de variance calculés sont non convergents en général.

Pour obtenir des estimateurs convergents des écarts-types sous les hypothèses précédentes, il est possible d'utiliser l'algorithme de bootstrap de la page suivante, dont la validité est établie en annexe. Cette procédure est très proche du bootstrap classique, mais s'en distingue par l'aléa sur la taille de l'échantillon. Intuitivement, cet aléa provient du fait que la taille de l'échantillon des répondants est aléatoire.

17. Cette approximation n'est pas correcte lorsque le sondage consiste à tirer des grappes d'individus, et non des individus. Dans ce cas, il faut tenir compte des corrélations au sein des grappes, à l'aide par exemple de l'option Cluster sous Stata.

Algorithme de bootstrap tenant compte des pondérations.
Pour $b = 1$ à B :

1. Tirer $n_b \sim \text{Binomiale}(N, n/N)$, où n la taille de l'échantillon des répondants ;
2. Tirer à probabilités égales et avec remise un échantillon de taille n_b issu de l'échantillon initial. On peut utiliser pour cela la commande suivante sous SAS (ici on échantillonne dans a et $n_b = 2500$) :
   `proc surveyselect data=a method=urs sampsize=2500 out=boot ;`
   `run ;`
   On note U_i^b le nombre de fois où l'individu i a été tiré dans l'échantillon bootstrap.
3. Estimer les poids $W_i^b = 1/P(D_i = 1|\widetilde{X}_i)$, par un modèle de non-réponse et/ou un calage identique à celui effectué sur l'échantillon initial mais en utilisant les pondérations U_i^b (ou en construisant une table ayant U_i^b observations pour l'individu i de la table initiale) ;
4. Estimer le paramètre θ avec les poids $W_i^b U_i^b$. On note $\widehat{\theta}_b$ l'estimateur obtenu.

Fin.

On peut ensuite estimer (par exemple) la variance de $\widehat{\theta}$ par

$$\widehat{V} = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta} - \widehat{\theta}_b\right)^2.$$

Même si la procédure ci-dessus est simple, elle peut s'avérer coûteuse en temps lorsque l'échantillon est de grande taille [18]. Pour certains exemples simples, comme ceux considérés ci-dessous, il est possible d'obtenir plus simplement des estimateurs des écarts-types, si l'on est prêt à négliger la variance due à l'estimation des pondérations (i.e., on assimile $\widehat{P}(D = 1|\widetilde{X})$ à $P(D = 1|\widetilde{X})$).

Exemple 1 (suite).
Le modèle linéaire constitue un exemple important. On repart du modèle linéaire (3.1) sous l'hypothèse $E(X'\varepsilon) = 0$. On suppose que le processus de sélection satisfait H0. Alors on peut montrer (cf. annexe) que la matrice de variance de l'estimateur correspond à la matrice de White pour le modèle pondéré. Pour obtenir cette matrice sous SAS, il suffit

18. Pour accélérer la procédure, on pourra supprimer la phase 3 de l'algorithme précédent, et utiliser pour tous les échantillons bootstrap les poids W_i au lieu des poids W_i^b. Nous postulons que cette omission n'est pas trop problématique dans le sens qu'elle conduit en général à surestimer les variances des estimateurs.

d'utiliser l'option ACOV dans la proc REG :

```
proc reg data=a;
model y=x/acov;
weight w;
run;quit;
```

Exemple 2 (suite).
Supposons que l'on souhaite estimer un modèle logit ou probit en pondérant, par exemple parce que $\widetilde{X} \not\subset X$. Dans ce cas, l'estimateur $\widehat{\theta^W}$ peut être obtenu sous SAS aussi bien avec la PROC LOGISTIC que la PROC SURVEYLOGISTIC. En revanche, l'estimateur de variance de $\widehat{\theta^W}$ proposé par la PROC LOGISTIC n'est pas convergent (cf. annexe pour davantage de détails). On veillera donc à utiliser la PROC SURVEYLOGISTIC pour estimer la précision des estimateurs [19].

```
proc surveylogistic data=a;
model y (descending)= x; /* L'option link=probit permet d'estimer un probit */
weight w;
run;
```

19. L'intérêt d'utiliser la PROC LOGISTIC lorsque l'on pondère est qu'elle possède des options qui n'existent pas dans la PROC SURVEYLOGISTIC. Par exemple, il n'y a pas d'option permettant de récupérer directement l'estimation de $P(Y = 1|X)$ pour chaque individu dans la PROC SURVEYLOGISTIC.

Feuille de route

Cette page résume les différentes étapes décrites dans la présente note. On se met ici dans la peau d'un chargé d'étude ayant à estimer un modèle donné sur un fichier pondéré. Il connait donc la (les) variable(s) endogènes Y et les variables exogènes X ainsi que le modèle économétrique qu'il souhaite estimer.

1. Première question : quelles sont les variables $\widetilde{X}$ utilisées pour calculer les pondérations ? Pour répondre à cette question consulter la documentation de la source ou discuter avec le service producteur.
2. Deuxième question : quelle est la nature de la sélection ? En d'autres termes, est-on sous l'hypothèse $H0$, sous l'hypothèse $H0'$ ou dans un cas de non réponse non ignorable ? Si $H0$ ou $H0'$, aller en 3. Si non réponse non-ignorable déterminer si on peut utiliser un modèle de sélection ou une méthode de calage généralisé. Dans le premier cas on utilisera un estimateur non pondéré, dans le deuxième cas on utilisera un estimateur pondéré. Si la non-réponse est non-ignorable et que ni le modèle de sélection ni la méthode de calage généralisé ne sont utilisables, la théorie d'une estimation convergente reste à construire !
3. Troisième question : faut-il pondérer ou non ? Pour cela on se référera au tableau 1 p. 13. Pour identifier la ligne du tableau à considérer, il s'agit de savoir si $\widetilde{X} \subset X$ et si θ dépend seulement de la loi conditionnelle $F_{Y|X}$ ou de la loi jointe $F_{X,Y}$.
4. Calculer les estimateurs choisis à l'étape 3.
5. Estimer la précision des paramètres estimés à l'étape précédente. S'il s'agit d'un estimateur non pondéré, les procédures usuelles donnent directement des estimateurs corrects de précision. S'il s'agit d'un estimateur pondéré les précisions sont en général mal estimées par les procédures usuelles. Dans ce dernier cas, on peut soit utiliser des procédures *ad hoc* (cf. exemples 1 et 2 p. 15 et 16), soit utiliser la méthode de bootstrap présentée p. 15.
6. Dernière question : les hypothèses précédentes sont-elles raisonnables ? Si les estimateurs pondérés et non pondérés convergent tous les deux, il est possible de tester la compatibilité des hypothèses économétriques du modèle avec les hypothèses concernant la nature du processus de sélection. En cas de rejet de la compatibilité, il faut soit changer de modèle, soit revenir sur les hypothèses $H0$ et/ou $H0'$.

Annexe : preuves et précisions techniques

Résultats de la partie 3

Résultats 1 et 2

Montrons d'abord que l'estimateur pondéré de $F_{X,Y}$ est convergent. Sous H0, il tend vers [20] :

$$\begin{aligned} E\left[WD\mathbb{1}\{Y \leq y, X \leq x\}\right] &= E\left[WE[D|X,Y,\widetilde{X}]\mathbb{1}\{Y \leq y, X \leq x\}\right] \\ &= E\left[WE[D|\widetilde{X}]\mathbb{1}\{Y \leq y, X \leq x\}\right] \\ &= E\left[\mathbb{1}\{Y \leq y, X \leq x\}\right]. \end{aligned}$$

La deuxième égalité provient de H0, la troisième de la définition des poids. Ainsi, l'estimateur pondéré de θ sera convergent, que θ dépende de $F_{Y|X}$ seulement ou de la loi jointe de (X, Y).

Pour montrer que l'estimation non pondérée est valide et "efficace", il suffit de constater que H0 et $\widetilde{X} \subset X$ impliquent H0'. En effet, pour toute fonction g,

$$E\left[g(Y)|D=1,X\right] = \frac{E\left[Dg(Y)|X\right]}{P(D=1|X)} = \frac{E\left[E[D|X,Y,\widetilde{X}]g(Y)|X\right]}{P(D=1|\widetilde{X})} = E\left[g(Y)|X\right].$$

On peut alors utiliser le résultat 3 ci-dessous.

Résultat 3

On a, par l'hypothèse H0', $F_{Y|X,D=1} = F_{Y|X}$. Par conséquent, l'inférence menée sans pondération sur les répondants sera valide. De plus, les estimateurs non pondérés qui sont asymptotiquement efficaces en l'absence de sélection le seront aussi ici.

Validité de la méthode de bootstrap proposée

Sous les hypothèses retenues, l'échantillon $(Y_i, D_i, W_i)_{i\in U}$ est i.i.d. Le bootstrap standard effectué sur U est donc valide en général [21]. Ce bootstrap revient à tirer des poids $(U_i^b)_{i\in U}$

20. On suppose ici que les conditions de régularité assurant la convergence des contreparties empiriques sont vérifiées. Pour des détails sur cette question, le lecteur pourra se référer par exemple à van der Vaart (1998).

21. Il existe des contre-exemples. Ainsi, la distribution bootstrapée de l'estimateur standard de θ dans un modèle $U[0,\theta]$ est non convergente. De même, Abadie et Imbens (2008) montrent la non convergence du bootstrap dans les procédures d'appariement par le plus proche voisin.

vérifiant :

$$(U_1^b, ..., U_N^b) \sim \mathcal{M}(N, 1/N, ..., 1/N),$$

où $\mathcal{M}$ désigne une loi multinomiale. Cependant, il est inutile de tirer tous ces poids car seuls ceux correspondants aux individus répondants seront utilisés. Notons $n_b = \sum_{i \in \mathcal{R}} U_i^b$. On obtient facilement :

$$n_b \sim \text{Bin}(N, n/N).$$

Notons $\mathcal{R} = \{r_1, ..., r_n\}$. Conditionnellement à n_b, on a, par symétrie :

$$(U_{r_1}^b, ..., U_{r_n}^b) \sim \mathcal{M}(n_b, 1/n, ..., 1/n).$$

L'algorithme proposé est donc convergent.

Variance de l'estimateur pondéré dans le modèle linéaire

Dans le modèle (3.1), sous les hypothèses $E(X'\varepsilon) = 0$, $E(|X'X|) < \infty$ et $E(|W^2X'X\varepsilon^2|) < \infty$, on a le résultat asymptotique suivant :

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \longrightarrow \mathcal{N}(0, V)$$

La variance asymptotique de l'estimateur pondéré est

$$V = \left[E\left(WX'X|D=1\right)\right]^{-1} E\left(W^2X'\varepsilon^2X|D=1\right)\left[E\left(WX'X|D=1\right)\right]^{-1}.$$

Cette variance peut être estimé sur l'échantillon par les contreparties empiriques suivantes :

$$\widehat{V} = \left(\frac{1}{N}\sum W_iX_i'X_i\right)^{-1}\left(\frac{1}{N}\sum W_i^2\widehat{\varepsilon}_i^2X_i'X_i\right)\left(\frac{1}{N}\sum W_iX_i'X_i\right)^{-1}$$

Il s'agit de la matrice de White pour le modèle pondéré. On peut donc utiliser l'estimateur robuste de la matrice de variance-covariance, tel que calculé par SAS via l'option ACOV.

Variance de l'estimateur pondéré dans les modèles logit ou probit

On peut montrer que l'estimateur pondéré a comme matrice de variance covariance

$$\begin{aligned} V_0 \;=\; & E\left(\frac{F'(X'\theta)^2}{F(X'\theta)\left(1-F(X'\theta)\right)}XX'\right)^{-1} E\left(W\frac{F'(X'\theta)^2}{F(X'\theta)\left(1-F(X'\theta)\right)}XX'\right) \\ & E\left(\frac{F'(X'\theta)^2}{F(X'\theta)\left(1-F(X'\theta)\right)}XX'\right)^{-1} \end{aligned}$$

La PROC SURVEYLOGISTIC permet d'estimer V_0 de manière convergente [22] par

$$\widehat{V_0} = N\left(\sum_i W_i \frac{F'(X_i'\theta)^2}{F(X_i'\theta)\left(1-F(X_i'\theta)\right)} X_i X_i'\right)^{-1} \left(\sum_i W_i^2 \frac{F'(X_i'\theta)^2\left(Y_i - F(X_i'\theta)\right)^2}{F^2(X_i'\theta)\left(1-F(X_i'\theta)\right)^2} X_i X_i'\right) \left(\sum_i W_i \frac{F'(X_i'\theta)^2}{F(X_i'\theta)\left(1-F(X_i'\theta)\right)} X_i X_i'\right)^{-1}.$$

Il s'agit ni plus ni moins que la formule utilisée par les sondeurs pour calculer la variance sous un plan de sondage poissonien dont les probabilités d'ordre 1 sont les $\frac{1}{W_i}$.

Notons que l'estimateur de la variance asymptotique estimée avec la PROC LOGISTIC s'écrit

$$\widehat{V_1} = \left(\frac{1}{N}\sum_i W_i \frac{F'(X_i'\theta)^2}{F(X_i'\theta)\left(1-F(X_i'\theta)\right)} X_i X_i'\right)^{-1}$$

On peut remarquer que la variance diminue avec une augmentation des poids et N constant, ce qui est pathologique dans le cadre dans lequel nous sommes ! Cela conduit donc certains utilisateurs à normaliser les poids de telle manière que $\sum_i W_i = N$, dans ce cas

$$\widehat{V_1} = \left(\frac{1}{\sum_i W_i}\sum_i W_i \frac{F'(X_i'\theta)^2}{F(X_i'\theta)\left(1-F(X_i'\theta)\right)} X_i X_i'\right)^{-1}.$$

Dans ce dernier cas, l'estimateur de la variance tend vers

$$\lim \widehat{V_1} = E\left(\frac{F'(X'\theta)^2}{F(X'\theta)\left(1-F(X'\theta)\right)} XX'\right)^{-1} \neq V_0.$$

En d'autres termes, l'estimateur de la variance proposé par la PROC LOGISTIC n'est jamais convergent, quels que soient les poids utilisés.

22. Le terme central de l'estimateur de la PROC SURVEYLOGISTIC peut sembler étrange, mais c'est un estimateur convergent de $E\left(W\frac{F'(X'\theta)^2}{F(X'\theta)(1-F(X'\theta))}XX'\right)$ car sous les hypothèses du modèle, $E\left((Y_i - F(X_i'\theta))^2|X_i\right) = F(X_i'\theta)\left(1-F(X_i'\theta)\right)$.

Références

Abadie, A. et Imbens, G. W. (2008), 'On the failure of the bootstrap for matching estimators', *Econometrica* **76**, 1537–1557.

Deville, J. C. (2002), La correction de la non-réponse par calage généralisé, *in* 'Actes des Journées de Méthodologie Statistique 2002', INSEE, pp. 4–20.

Deville, J. C. et Sarndal, C. E. (1992), 'Calibration estimators in survey sampling', *Journal of the American Statistical Association* **87**, 376–382.

Imbens, G. W. (1992), 'An efficient method of moments estimator for discrete choice models with choice-based sampling', *Econometrica* **60**, 1187–1214.

Le Guennec, J. et Sautory, O. (2005), La macro calmar2 : redressement d'un échantillon par calage sur marges. Document INSEE.

Lerman, S. et Manski, C. F. (1977), 'The estimation of choice probabilities from choice based samples', *Econometrica* **45**, 1977–1988.

Sautory, O. (1993), La macro calmar : redressement d'un échantillon par calage sur marges. Document F9310, DSDS, INSEE.

Tillé, Y. (2001), *Théorie des sondages : Échantillonnage et estimation en populations finies : cours et exercices*, Dunod.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics.

Wooldridge, J. (2002), *Econometrics of Cross Section and Panel Data*, MIT Press.

Wooldridge, J. (2007), 'Inverse probability weighted estimation for general missing data problems', *Journal of Econometrics* **141**, 1281–1301.