Econométrie et Données d'Enquête: les effets de l'imputation de la non-réponse partielle sur l'estimation des paramètres d'un modèle économétrique [1]

C.Charreaux[2], C. Favre-Martinoz [3], H.Harle [4], R.Le Saout[5], P.-A. Robert[6]

[2] *Ecole Nationale de la Statistique et de l'Administration Economique, camille.charreaux@ensae-paristech.fr*

[3] *Direction de la Méthodologie et de la Coordination Statistique et Internationale, INSEE, 18 Boulevard Adolphe Pinard, 75014 Paris, cyril.favre-martinoz@insee.fr*

[4] *Ecole Nationale de la Statistique et de l'Administration Economique, honorine.harle@ensae-paristech.fr*

[5] *Direction de la Méthodologie et de la Coordination Statistique et Internationale, INSEE, 18 Boulevard Adolphe Pinard, 75014 Paris, ronan.le.saout@ensae.fr*

[6] *Ecole Nationale de la Statistique et de l'Administration Economique, pierre.antoine.robert@ensae-paristech.fr*

Résumé. Lors d'une enquête, l'imputation simple est la solution privilégiée pour corriger la non-réponse partielle et estimer de manière efficace des statistiques fonctions d'une seule variable aléatoire. En pratique, les données d'enquête sont utilisées pour conduire des analyses multivariées sans qu'il soit toujours possible de distinguer les données imputées. Tout se passe comme si l'information était parfaite, ce qui n'est en réalité pas le cas. Cette communication étudie les effets de l'imputation (simple) sur les analyses multivariées à l'aide de simulations, en liant théorie économétrique et des sondages. Si l'imputation simple est efficace pour estimer un paramètre de population finie comme une moyenne ou un total, les paramètres d'un modèle économétrique sont généralement estimés de façon biaisée même en cas d'exogénéité du processus de non-réponse.

Mots-clés. Econométrie, imputation, endogénéité.

[1]Nous remercions l'équipe pédagogique de l'ENSAE, ce travail de recherche ayant été initié lors d'un projet de statistique encadré par Cyril Favre-Martinoz et Ronan Le Saout. Cet article ne reflète pas l'opinion de l'INSEE et n'engage que ses auteurs.

1 Introduction

Lors d'un sondage, les unités statistiques (individus, ménages, entreprises...) renseignent des valeurs pour un ensemble de variables, par exemple l'âge, le revenu, la situation familiale, etc. Cependant, certaines unités ne répondent pas à toutes les variables pour lesquelles on les interroge. Ce phénomène est appelé la non-réponse. Plusieurs raisons peuvent expliquer cette absence de réponse. Il se peut que les questions abordées dans l'enquête soient sensibles pour l'individu interrogé (par exemple, les questions sur le revenu). Il est également possible que l'individu sélectionné ne comprenne pas la question ou que les réponses apportées soient incohérentes et par conséquent invalidées, ou que l'individu sélectionné abandonne au cours de l'enquête (enquête en plusieurs visites, carnet, panel). L'information recueillie par le sondeur est alors soit inexistante pour un individu, et on parle alors de non-réponse totale, soit incomplète et on parle alors de non-réponse partielle. La question est alors de savoir comment utiliser des données contenant de la non-réponse.

L'inférence basée sur des données contenant de la non-réponse donne en général des estimations biaisées pour des statistiques telles que des totaux ou des moyennes, sauf dans le cas d'une non-réponse complètement aléatoire (MCAR, pour Missing Completely At Random). Dans le cas d'une non-réponse supposée conditionnellement aléatoire (MAR pour Missing At Random) i.e. où la non-réponse est fonction de caractéristiques observables mais non de la caractéristique d'intérêt, ou non ignorable (MNAR pour Missing Not At Random), i.e. où la non-réponse dépend de la variable d'intérêt. Par exemple, on pourrait imaginer que la non-réponse sur le revenu est plus élevée pour les hauts revenus (MNAR Missing Not at random). Dans ce cas, les estimateurs sont potentiellement biaisés. Pour corriger l'éventuel biais induit par la non-réponse et améliorer la précision des estimateurs, des méthodes spécifiques ont été définies selon les hypothèses retenues pour modéliser le mécanisme de non-réponse. Nous nous focalisons ici sur l'étude de mécanismes de non-réponse ignorables (MCAR ou MAR), la prise en compte de la non-réponse non ignorable renvoyant à un cadre technique différent basé sur les modèles de sélection ou le calage généralisé (Davezies et D'Haultfoeuille 2009). Nous nous concentrons de plus sur le cas de la non réponse-partielle, mais une grande partie de notre analyse peut être étendue à la non réponse totale. Trois principales techniques sont envisageables pour tenir compte de la non-réponse partielle (Haziza 2005) : l'omission des non-répondants, la repondération et l'imputation. La dernière technique est généralement plus avantageuse. Cela vient du fait que l'imputation induit l'utilisation d'un jeu de données complet (contrairement aux deux autres techniques), l'existence de poids de sondage uniques

et l'utilisation d'informations auxiliaires sur les non-répondants. En pratique, l'imputation est donc la méthode la plus utilisée. Il s'agit alors de substituer chaque valeur manquante par une valeur *ad-hoc*, de manière déterministe ou stochastique.

La pertinence des méthodes d'imputation est traitée dans la littérature du point de vue des sondages, i.e. son effet sur le biais et la variance de totaux, de moyennes ou de statistiques descriptives plus complexes (qui sont des paramètres de population finie). En revanche, l'utilisation par les économètres des données contenant de la non-réponse préalablement imputées ne fait pas l'objet d'étude suffisamment approfondie (Little 1992 définit un cadre théorique mais qui ne s'appuie pas sur les méthodes d'imputation traditionnelles). La théorie économétrique suppose en effet une information complète au sens où l'information sur la présence ou non de données manquantes est connue. Plusieurs stratégies sont alors envisageables (Raghunathan 2004), par exemple une imputation par variables dites cachées (algorithme EM), une imputation multiple où plusieurs valeurs sont imputées à une même observation manquante ou des méthodes d'identification partielle (Manski 2005). Or, les instituts nationaux de statistique fournissent parfois aux acteurs de l'économie, mais aussi aux chercheurs des jeux de données complets, souvent sans préciser si les données ont été imputées. Manski (2015) dans un article prônant une meilleure communication de l'incertitude des statistiques publiques souligne ainsi l'absence quasi systématique d'informations sur les méthodes d'imputation dans les bases de données et statistiques américaines.

Les méthodes d'imputation construites pour inférer sur des paramètres de population finie comme des totaux ou des moyennes ne sont pas nécessairement adaptées pour inférer sur les paramètres de modèles économétriques. Il est déjà connu en Théorie des Sondages que l'imputation marginale de données manquantes, qui consiste à imputer les variables séparément conduit généralement à une estimation biaisée pour un paramètre de population finie bivariée tel que le coefficient de corrélation. Pour traiter ce problème des méthodes d'imputation basées sur des procédures de type Shao & Wang (2002) peuvent être mises en place pour permettent la préservation du lien entre plusieurs variables d'intérêt. Des analyses en économie (Heckman et LaFontaine 2006; Bollinger et Hirsch 2013) ont montré des exemples d'analyse biaisée et des premières propositions de correction. De manière intuitive, les principales méthodes d'imputation s'appuient sur des hypothèses sur la loi univariée associée au mécanisme de non-réponse (i.e. variable par variable), alors qu'un modèle économétrique s'appuie sur des hypothèses concernant le terme d'erreur conditionnellement aux variables explicatives (i.e. une loi jointe). Il n'y a donc aucune raison que ces hypothèses soient cohérentes dans le cas général. Par exemple dans le cas d'une non-réponse supposée MAR,

l'estimateur de la moyenne sera biaisé sans traitement de la non-réponse, ce ne sera pas le cas d'un modèle économétrique qui inclurait en plus de cette variable les caractéristiques qui permettent d'expliquer la non-réponse (pour des raisons comparables au choix de pondérer ou non le modèle, Davezies et D'Haultfoeuille 2009).

L'objectif de cet article n'est donc pas de proposer une nouvelle méthode de correction de la non-réponse mais d'étudier l'effet de ces méthodes ex-post lorsqu'une seule base de données complète est à la disposition du chargé d'étude ou du chercheur. Des simulations permettent d'illustrer qu'une méthode efficace pour des statistiques descriptives ne l'est pas pour un modèle économétrique simple. La suite de l'article est organisée comme suit. La partie 2 présente le cadre d'analyse ainsi que les principales méthodes d'imputation envisagées. La partie 3 illustre à travers des simulations l'effet de ces méthodes sur un modèle linéaire. Enfin, la partie 4 résume les premiers résultats issus de ces travaux et les perspectives.

2 Cadre théorique de l'étude

Dans cette partie, nous détaillons le modèle économétrique linéaire (que l'on appelle plus communément modèle "de superpopulation" en théorie des sondages) assez simple considéré dans la suite, ainsi que les méthodes d'imputations envisagées pour mettre en évidence leurs effets sur l'estimation des paramètres du modèle économétrique.

2.1 Modèle économétrique

On considère quatre variables aléatoires continues : X, Z, Z' et ε et le modèle économétrique sans constante suivant :

$$Y = \beta \cdot X + \gamma \cdot Z' + \varepsilon,$$

avec $\beta = 1$ et $\gamma = 0$ ou 1.

On suppose de plus que le terme d'erreur ε est indépendant des autres variables aléatoires afin de ne pas créer d'endogénéité. Ce terme d'erreur peut suivre une loi normale $\mathcal{N}(0,1)$ ou une autre loi par exemple du khi-deux $\chi^2(1)$, ce qui permet d'analyser les propriétés à distance finie des estimateurs. Les variables aléatoires X, Z et Z' sont supposées suivre des lois normales multivariées. Les variables aléatoires X et Z sont supposées corrélées entre elles mais non corrélées avec Z'. La non-réponse partielle est générée pour les variables X et Y à l'aide d'un mécanisme de non-réponse fonction des variables Z et/ou Z'.

Plusieurs cas vont ensuite être étudiés, selon les hypothèses retenues sur γ, les variables expliquant la non-réponse et les taux de non-réponse.

Cas 1. Exogénéité.

1.1 La non-réponse est fonction de Z uniquement. Cette variable n'étant pas corrélée avec le terme d'erreur $\omega = \gamma \cdot Z' + \varepsilon$ (que le paramètre γ soit supposé nul ou non, il y aura une plus grande variabilité du modèle lorsqu'il est non nul), l'application du modèle économétrique sur les seules observations (X, Y) sans données manquantes sera sans biais. Le fait que X et Z soient corrélés engendre un intérêt pour estimer des statistiques descriptives telles que la moyenne de X ou Y. Par contre l'effet sur le modèle économétrique est ambigu. On peut néanmoins noter que si seule de la non-réponse sur Y est observée, aucun biais ne devrait être constaté (résultat classique d'un modèle à erreur de mesure).

1.2 Le paramètre γ est supposé nul et la non-réponse fonction de Z et Z'. La variable Z' étant néanmoins inobservée, l'imputation n'est effectuée qu'en fonction de Z et est donc imparfaite. Cela revient à une mauvaise spécification du mécanisme de non-réponse.

Cas 2. Endogénéité. Le paramètre γ est supposé non nul et la non-réponse fonction de Z et Z' pour X uniquement (mais uniquement de Z pour Y, sinon la non-réponse est non ignorable). Cette situation crée de la sélection et donc de l'endogénéité. L'estimation du modèle économétrique sur les seules observations (X, Y) sans données manquantes sera biaisée. La question est alors de savoir si les méthodes d'imputation amplifient ou diminuent ce biais.

2.2 Méthodes d'imputation utilisées pour traiter la non-réponse partielle

Les méthodes couramment utilisées pour traiter la non-réponse partielle sont les suivantes (entre parenthèses, nous indiquons les abréviations utilisées par la suite): - imputation par la moyenne (Moyenne)
- imputation par Hot-Deck (HotDeck)
- imputation par Hot-Deck stratifié (HotDeckStrat)
- imputation par plus proche voisin (KNN)
- imputation par la régression (Regression)

Nous comparons dans la suite ces cinq méthodes classiques au cas où l'on supprime purement et simplement les unités non-répondantes appelé " Available Case".

Ces méthodes présentent un inconvénient majeur, dans la mesure où l'imputation est effectuée indépendamment sur chacun des variables présentant de la non-réponse partielle. Ces techniques d'imputation marginale détruisent la corrélation pré-existante entre les variables. D'autres méthodes, comme l'imputation multiple, l'imputation via un algorithme EM ou par maximum de vraisemblance, davantage utilisées dans un contexte économétrique pourraient également être envisagées, mais sont dans un premier temps volontairement écartées de l'étude. L'objectif ici est d'évaluer au moins de façon empirique l'impact de ces méthodes sur les paramètres d'un modèle économétrique.

3 Simulation du biais dans un modèle linéaire

3.1 Protocole de simulation

Pour illustrer les biais éventuels de ces méthodes d'imputation sur un modèle économétrique, nous menons un exercice de simulations résumé par le graphique suivant. Nous faisons une enquête dans une population de 5000 individus pour deux variables d'intérêt X et Y, cette dernière étant générée à partir du modèle $Y = \beta \cdot X + \gamma \cdot Z' + \varepsilon$ avec Z' une variable auxiliaire inobservée et ε un terme d'erreur (étape 1). Les variables aléatoires X, Z et $Z^{'}$ ont été générées selon une loi normale multivariée d'espérance $(0, 2, 2)^\top$ et de matrice de variance-covariance $\Sigma = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.
Dans cette population, nous utilisons un plan de sondage aléatoire simple sans remise (étape 2) de taille $n = 200$ ou $n = 500$ parmi $N = 5000$ (soit un taux de sondage de 4% ou 10%). Suite à la collecte des données, la non-réponse partielle est observée pour les variables X et Y (étape 3). Cette non-réponse est imputée par différentes méthodes, sous l'hypothèse d'un mécanisme de non-réponse fonction d'une variable auxiliaire parfaitement observée Z mais également de la variable auxiliaire inobservée Z'. L'objectif de nos simulations est d'observer

l'effet de la méthode d'imputation sur l'estimation du paramètre β et de sa variance.

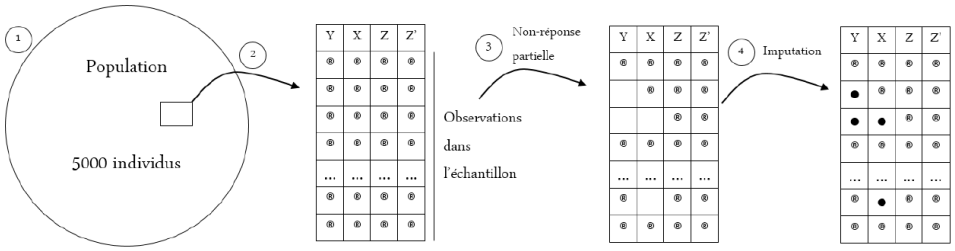

Figure 1: Protocole de simulation

Trois sources d'aléa sont ici présentes: le processus générateur des données (dit PGD), le plan de sondage et l'imputation des données. Le plan de sondage étant ici un sondage aléatoire simple sans remise. Le choix de ce plan de sondage n'est pas neutre, car le modèle économétrie qui tient au niveau de la population, tient également au niveau de l'échantillon. Cette étape n'engendre aucun biais dans l'estimation des paramètres du modèle. A un niveau désagrégé, les estimateurs pourraient être peu efficaces (i.e. avec une variance élevée) ce qui pourrait être résolu par la définition d'un plan de sondage stratifié par exemple. Cette question de l'efficacité du plan de sondage n'est pas reliée à celle des effets des méthodes d'imputation. Nous n'aborderons donc plus par la suite le choix du plan de sondage et resterons dans ce cadre simplifié. Un calcul complet de la variance des paramètres nécessiterait néanmoins de tenir compte de cette étape.

3.2 Impact de l'imputation sur un paramètre de population finie

Nous avons observé l'effet des méthodes d'imputation sur les moyennes estimées de Y et X. Les différences entre les moyennes de la variable X avant et après imputation sont présentées dans la Figure 2 pour le cas "Available Case" et les cinq méthodes d'imputation partielles

détaillées dans le paragraphe 2.2. Ces résultats ont été obtenus à l'aide 1000 simulations.

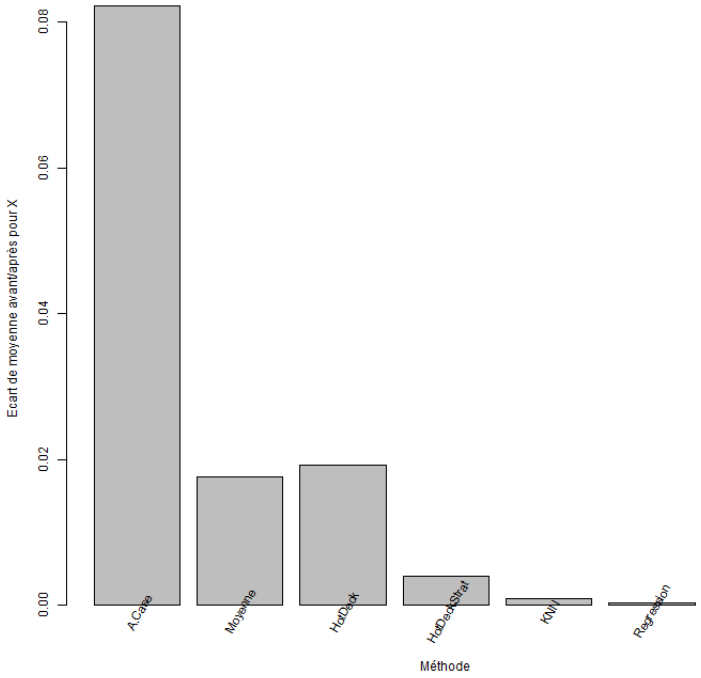

Figure 2: Effet de l'imputation sur la moyenne de X

L'utilisation des méthodes d'imputation est justifiée, puisque l'on constate que l'estimation sur les répondants seulement conduit au plus fort écart, et ce de manière claire: la différence par rapport aux cas où les méthodes d'imputation sont utilisées est importante. Il est important de noter que les écarts entre les différentes méthodes d'imputation ne sont pas représentatifs de la précision de ces méthodes. En particulier, ces écarts ne permettent pas d'évaluer et de comparer ces méthodes, il s'agit ici simplement de confirmer la nécessité d'utiliser des méthodes d'imputation partielle pour estimer des paramètres de population finie.

3.3 Étude du biais pour l'estimation des paramètres du modèle économétrique

Les méthodes d'imputation couramment utilisées dans le cadre de la Théorie des sondages atténuent le biais dans l'estimation des paramètres de population finie telles les moyennes et les totaux. Cependant leur utilisation peut poser problème pour l'estimation de paramètres issus du modèle économétrique présenté à la section 2.1. Ces méthodes détruisent la corréla-

tion entre les variables, afin de se concentrer sur les distributions marginales. L'estimation dans le cas d'un modèle économétrique fait intervenir les distributions conjointes des variables présentes dans le modèle.

Le graphique ci-dessous permet d'illustrer le fonctionnement de l'imputation pour deux méthodes : l'imputation par la moyenne et l'imputation par les proches voisins. Les répondants sont représentés en gris, les non-répondants avant imputation sont les cercles blanc, et les valeurs imputées correspondent aux croix rouges. L'objectif est de montrer à travers ce graphique les perturbations induites dans la distribution conjointe après imputation des variables X et Y. Ces perturbations se traduisent par une estimation biaisée de la corrélation en population finie entre ces deux variables, qui engendrent une estimation biaisée au niveau du modèle économétrique.

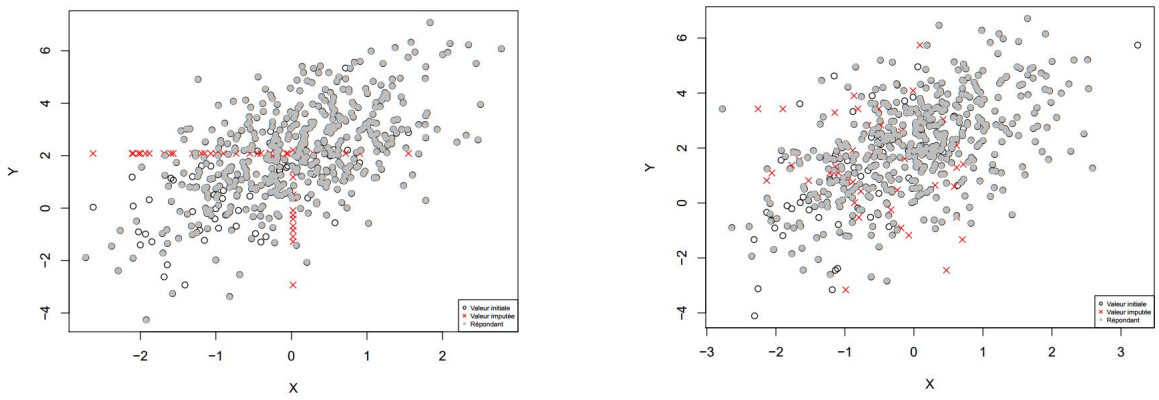

Figure 3: Effet de l'imputation par la moyenne (gauche) ou des plus proches voisins (droite) sur les valeurs de Y et X

Nous constatons que les valeurs imputées sont réparties de manière à former une croix : si X est inconnu, il sera imputé par la moyenne des répondants, quelle que soit la valeur des autres variables de l'individu, et de même si c'est Y qui est manquante. Cette méthode d'imputation modifie donc totalement la corrélation entre les variables. Il en va de même pour la méthode des plus proches voisins, pour laquelle des points bien en dehors du nuage initial apparaissent après imputation, modifiant ainsi la distribution conjointe de X et Y sur les données complétées.

A partir de 5000 simulations Monte-Carlo, nous avons estimé le biais et la variance associé

à l'estimateur des MCO calculé sur les répondants uniquement ou sur le jeu de données complétées à l'aide d'une des cinq méthodes d'imputation citées dans la partie 2.2. Les résultats des simulations sont présentés dans le Tableau 2.

Méthode	Biais pour β(%)			100 x Var($\hat{\beta}$)		
Cas	1.1	1.2	2	1.1	1.2	2
A.Case	-0.32	0.45	-1.19	1.59	0.70	1.18
Moyenne	-15.32	-12.95	-17.76	1.42	0.71	0.98
HotDeck	-28.54	-24.76	-21.53	1.58	0.91	1.29
HotDeckStrat	-21.80	-19.26	-15.71	1.55	0.88	1.29
KNN	-18.16	-16.35	-13.77	1.74	0.87	1.36
Regression	-10.41	-9.08	-11.83	1.43	0.68	1.05

Table 1: Biais et variance estimé pour β, n=200, 5000 simulations

Le premier constat est l'importance du biais estimé, qui peut atteindre une moyenne de plus 25%(ex: HotDeck). Le choix de la méthode d'imputation et son signalement n'est donc pas anodin pour l'analyse économétrique de données présentant de la non-réponse partielle.

D'autre part, l'augmentation de la taille de l'échantillon a un effet sur la variance du paramètre estimé : en utilisant non plus 200 mais 500 individus, soit un passage de 2% à 10% de la taille de la population, on constate une division par plus de 2 de la variabilité de $\hat{\beta}$ (Table 4). Le biais moyen estimé quant à lui conserve le même ordre de grandeur malgré l'échantillon plus grand.

Méthode	Biais pour β (%)			100 x Var($\hat{\beta}$)		
Cas	1.1	1.2	2	1.1	1.2	2
A.Case	-0.02	0.00	-1.38	0.52	0.27	0.51
Moyenne	-14.98	-13.41	-17.75	0.46	0.27	0.46
HotDeck	-28.15	-25.49	-21.70	0.58	0.33	0.58
HotDeckStrat	-21.31	-19.50	-15.80	0.57	0.33	0.57
KNN	-17.49	-16.94	-14.02	0.60	0.32	0.59
Regression	-9.82	-9.40	-11.87	0.46	0.27	0.47

Table 2: Biais et variance estimé pour β dans les trois cas, n=500, 5000 simulations

Quel que soit le cas étudié, nous constatons que la meilleure méthode pour limiter le biais sur β reste l'estimation par Available Case **dans le cas d'un modèle linéaire entre Y et X**, c'est-à-dire estimer le coefficient sur les répondants, sans imputation.

Nous avons donc constaté sous l'hypothèse d'un modèle linéaire que si les méthodes d'imputation présentent un intérêt fort pour accomplir leur objectif premier, c'est-à-dire améliorer l'estimation de statistiques descriptives telles que les moyennes, elles ont cependant un impact non négligeable sur l'évaluation économétrique, puisque contrairement à l'utilisation des répondants seule, elles biaisent le coefficient à hauteur de 15 à 25% pour les méthodes usuellement utilisées.

3.4 Extension aux variables qualitatives

Pour les variables qualitatives, le cadre d'analyse est différent. Prenons un exemple simplifié où on estime le modèle économétrique $Y_k = \alpha + \beta \cdot \mathbf{1}_{k \in \text{Gpe Traité}} + \varepsilon_k$ avec Y_k le salaire perçu par un individu k et $\mathbf{1}_{k \in \text{Gpe Traité}}$, une indicatrice valant 1 pour les individus ayant suivi un programme de formation. L'estimateur des Moindres Carrés Ordinaires (MCO) non pondéré vaut alors $\hat{\beta}^{MCO} = \overline{Y}_{\text{Gpe Traités}} - \overline{Y}_{\text{Gpe Non Traités}}$ i.e. la différence des moyennes simples de salaires entre les individus traités et non traités.

De manière plus générale, on peut interpréter le paramètre d'une variable qualitative (lorsqu'il n'y a pas de variable continue dans le modèle) comme une différence de moyenne de la variable endogène sur des populations spécifiques. On retrouve alors le cadre des statistiques usuellement étudiées en Théorie des Sondages. L'utilisation conjointe de variables continues et qualitatives est néanmoins plus difficile à interpréter.

3.5 Correction du biais selon l'information disponible

Deux cas principaux doivent être envisagés, selon que l'information sur les observations imputées est disponible dans la base de données ou non. Si les drapeaux d'imputation sont disponibles, il est possible de revenir aux données bruts. La question peut se poser alors de ne pas tenir compte des imputations effectuées et de mener une analyse spécifique au sujet traité à l'aide de méthodes de sélection, d'identification partielle, d'un algorithme EM ou d'imputations multiples. Ce choix reste bien sûr coûteux par rapport à l'utilisation directe de la base de données diffusée. Dans le cas contraire, lorsque l'information sur les imputations individuelles n'est pas accessible ou parce qu'une méthode spécifique serait trop coûteuse, il est nécessaire d'envisager une correction approchée selon l'information disponible: taux de non-réponse (éventuellement par strates), disponibilité d'informations auxiliaires exhaustives, voire aucune information. Cette contrainte pratique de non-disponibilité de l'information n'a été abordée à notre connaissance que par Chen et Shao (1996).

4 Conclusion

L'ensemble des travaux présenté dans cet article mettent empiriquement en évidence l'introduction de biais importants via l'utilisation des méthodes d'imputation classiquement utilisées en Théorie des Sondages pour l'estimation de paramètre d'un modèle économétrique. Une étude théorique des biais et une application sur un jeu de données réelles seront envisagés par la suite et permettront de compléter ces premières observations.

Bibliographie

Bollinger, C. R., et B. T. Hirsch. (2013) "Is Earnings Nonresponse Ignorable?" *The Review of Economics and Statistics*, 95(2), 407-416.

Chen, Y., et J. Shao. (1996) "Inference with Complex Survey Data Imputed by Hot Deck When Nonrespondents are Nonidentifiable." SSC Annual Meeting, Proceedings of the Survey Methods Section.

D'Haultfoeuille, X., et L. Davezies. (2009) "Faut-il pondérer ?... Ou l'éternelle question de l'économètre confronté à des données d'enquête." Document de travail de l'INSEE, G2009/06.

Haziza, D. (2005) "Inférence en présence d'imputation simple dans les enquêtes: un survol." *Journal de la société statistique de Paris*, 146(4), 69-118.

Heckman, J. J., et P. A. LaFontaine. (2006) "Bias Corrected Estimates of Ged Returns." NBER Working Paper n°12018.

Little, Roderick J.A. (1992) "Regression With Missing X's: A Review." *Journal of the American Statistical Association*, 87(420), 1227-1237.

Manski, C. F. (2005) "Partial Identification with Missing Data: Concepts and Findings." *International Journal of Approximate Reasoning*, 39(2-3), 151-165.

Manski, C. F. (2015) "Communicating Uncertainty in Official Economic Statistics: An Appraisal Fifty Years after Morgenstern." *Journal of Economic Literature*, 53(3), 631-53.

Raghunathan, Trivellore E. (2004) "What Do We Do With Missing Data? Some Options For Analysis of Incomplete Data. " *Annu. Rev. Public Health*, 25, 99-117.

Shao, J., et H. Wang. (2002) "Sample Correlation Coefficients Based on Survey Data Under Regression Imputation." *Journal of the American Statistical Association*, 97(458), 544-552.