Comment utiliser des données d'enquête ?

Guide illustré sur l'économétrie des données échantillonnées *

C.Favre-Martinoz †, M.Guillerm ‡, R. Le Saout §

Résumé

Cet article s'appuie sur les travaux d'un groupe de lecture sur l'économétrie des données d'enquête mis en place au sein du Département de la Méthodologie de l'Insee sur la période 2015-2016. Les données d'enquête sont largement utilisées pour mener des études économiques. Les liens entre la micro-économétrie et la théorie des sondages sont pourtant mal connus ou mal compris, ces deux pans de la statistique ayant des objectifs différents : inférence des paramètres d'un modèle vs estimation de statistiques descriptives. Les chargés d'études font souvent abstraction du fait qu'ils utilisent des données d'enquête. Ils ne savent pas toujours comment tenir compte des traitements d'enquête (plan de sondage et pondération, repondération liée à la non-réponse ou au calage). Nous conduisons ici une revue de littérature sans se limiter à la théorie économétrique. Les contraintes tant en termes de logiciels que de contenu des bases de données sont ainsi soulignées à travers des exemples basés sur l'enquête Patrimoine conduite par l'INSEE. La question de la pondération ou non des modèles économétriques a été largement abordée dans la littérature. Il n'est pas nécessaire de pondérer lorsque les variables ayant servi à définir le plan de sondage sont incluses dans le modèle. Dans les autres cas, une approche prudente reste de comparer les estimations des modèles pondérés et non pondérés. Pour le calcul de variance tenant compte du plan de sondage, deux types d'aléa doivent être pris en compte, l'aléa du tirage de l'échantillon et celui du modèle économétrique. Le calcul précis de cette variance nécessite ainsi des informations détaillées sur la méthodologie d'enquête et n'est que rarement effectué. Dans de nombreux cas, cette variance peut néanmoins être approchée. Les méthodes d'imputation utilisées pour corriger la non-réponse ont des effets ambigus sur les modèles économétriques, ce qui reste encore un sujet à explorer. Le peu de prise en compte de la méthodologie d'enquête peut ainsi provenir d'une absence de solutions théoriques identifiées mais également d'un manque de communication sur les traitements mis en oeuvre lors d'une enquête statistique. La diffusion d'informations détaillées sur la construction des données apparaît donc nécessaire pour permettre une véritable économétrie des données d'enquête.

Mots clés : poids de sondage, enquête statistique, sélection.
Codes JEL : C18, C50, C83.

*Ce travail a été initié à travers un groupe de lecture au sein de l'Insee, coordonné par Marine Guillerm et Ronan Le Saout. Il a été complété par un projet d'étudiants sur les effets de méthodes d'imputation encadré par Cyril Favre-Martinoz et Ronan Le Saout. Nous tenons à remercier l'ensemble des participants à ce groupe, notamment Pascal Ardilly, Thomas Balcone, Guillaume Chauvet, Martin Chevalier, Laurent Davezies, Xavier D'Haultfoueille, Pauline Givord, Thomas Merly-Alpa, Raphaël Lardeux, Eric Lesage, Trong-Hien Pham, Tiaray Razafindranovona, Olivier Sautory, et Benjamin Vignolles. Nous remercions également Rosalinda Coppoletta et Aline Ferrante pour leur aide sur l'utilisation de l'enquête Patrimoine. Cet article ne reflète pas les opinions de l'Insee ni de la Dares.

†CREST-INSEE. Email : cyril.favre-martinoz@insee.fr
‡DARES. Email : marine.guillerm@travail.gouv.fr
§CREST-ENSAI.Email : ronan.le.saout@ensae.fr

1 Introduction

De nombreux types de données sont utilisés pour conduire une étude économique : données agrégées ou synthétiques (indices et comptabilité nationale), données administratives (fichiers fiscaux ou déclarations de salaires), ou de manière plus récente des données massives (données de téléphonie mobile, de prix issus d'internet...). Chaque type de données crée bien sûr des contraintes techniques particulières. En microéconométrie, l'usage de données d'enquête, construites à partir de la théorie des sondages, reste néanmoins une norme. Or la conduite d'une enquête fait appel à des techniques mathématiques particulières issues de la théorie des sondages pour définir un plan de sondage et donc des poids de sondage (i.e. définir qui interroger) mais également corriger de la non-réponse à une enquête. Ces traitements ne tiennent généralement pas compte de l'usage économétrique ultérieur de ces données. De même, les chargés d'études et les chercheurs font souvent abstraction du fait qu'ils utilisent des données d'enquête, les liens entre la micro-économétrie et la théorie des sondages restant mal connus ou mal compris. L'information disponible sur la méthodologie d'enquête dans les bases de diffusion est de plus pauvre.

L'objectif de cette revue de littérature est triple. Tout d'abord, ce document vise à informer des avancées théoriques sur l'impact des traitements d'enquête sur les modèles économétriques (biais et précision) mais également des points en suspens. Cette prise en compte s'améliore, par la diffusion de documents de travail (Davezies et D'Haultfoeuille 2012 sur les poids de sondage par exemple) et de procédures orientées "sondages" dans les principaux logiciels statistiques. Mais la compréhension des différences d'approches entre la théorie des sondages et l'économétrie reste encore mal connue. Deuxièmement, il vise à présenter les procédures disponibles dans les principaux logiciels statistiques et leurs utilisations avec les informations limitées disponibles dans les bases de diffusion. Pour ce faire, l'enquête Patrimoine sera mobilisée. En dernier lieu, il y a un objectif de plus long terme d'identifier les informations complémentaires qui pourraient être ajoutées aux bases de diffusion (strates de tirage et de post-stratification, poids de tirage...). L'objectif n'est donc pas ici de donner des solutions applicables à toute situation mais d'alerter sur la nécessité de réfléchir à la construction des données lorsqu'on conduit une étude économique.

Dans certaines situations, il est indispensable de pondérer une analyse économétrique effectuée à l'aide de données d'enquête. Supposons qu'on cherche à évaluer l'effet d'une formation professionnelle sur le retour à l'emploi ou le salaire perçu. On ne dispose que de données d'enquête (un échantillon s), chaque individu k enquêté ayant une probabilité d'inclusion π_k et donc un poids de sondage $w_k = 1/\pi_k$.

On estime alors le modèle économétrique $Y_k = \alpha + \beta \cdot \mathbf{1}_{k \in \text{Gpe Traité}} + \varepsilon_k$ avec Y_k le salaire perçu par l'individu k et $\mathbf{1}_{k \in \text{Gpe Traité}}$, l'indicatrice valant 1 pour les individus ayant suivi la formation.

L'estimateur des Moindres Carrés Ordinaires (MCO) non pondéré vaut

$$\hat{\beta}^{MCO} = \overline{Y}_{\text{Gpe Traités}} - \overline{Y}_{\text{Gpe Non Traités}}$$

i.e. la différence des moyennes simples de salaires entre les individus traités et non traités.

A l'aide d'un estimateur de Hájek de ces moyennes, on obtiendrait

$$\hat{\beta}^{H} = \left(\sum_{j \in \text{Gpe Traités}} 1/\pi_j \right)^{-1} \sum_{j \in \text{Gpe Traités}} Y_j/\pi_j - \left(\sum_{k \in \text{Gpe Non Traités}} 1/\pi_k \right)^{-1} \sum_{k \in \text{Gpe Non Traités}} Y_k/\pi_k.$$

Il n'y a aucune raison que ces deux estimateurs soient égaux et, en général, $\hat{\beta}^{MCO} \neq \hat{\beta}^{H}$. Mieux vaudrait alors pondérer. C'est le cas pour toutes statistiques descriptives calculées à partir de données d'enquête. Dans le cas général d'un modèle économétrique, les choses se complexifient. L'objet de ce document sera d'aborder ces cas plus complexes et l'effet de la (non) pondération sur les estimateurs et leur variance.

Deux types d'approche pour l'estimation d'un paramètre avec des données d'enquêtes s'affrontent. Le statisticien-sondeur peut utiliser une approche fondée sur le plan de sondage, l'aléa est alors dans le tirage de l'échantillon, les valeurs observées sur la population étant considérées fixes (approche dite *Design-Based*). Le statisticien-économètre peut utiliser une approche à l'aide d'un modèle stochastique. Dans ce cadre, les données sont supposées être générées de manière i.i.d., les valeurs observées sur la population (et l'échantillon) étant considérées comme la réalisation de variables aléatoires issues d'un modèle de "super-population" (approche dite *Model-Based*) [1]. Les deux approches sont néanmoins cohérentes pour le calcul des estimateurs. Little (2004) souligne que l'estimateur de Horvitz-Thompson (HT) peut ainsi être vu comme l'estimateur MCO du modèle linéaire : $y_i = \beta \cdot \pi_i + \pi_i \varepsilon_i$ et ε_i i.i.d. de loi $\mathcal{N}\left(0, \sigma^2\right)$. $\hat{\beta} = \hat{t}_{HT}/n$, où $\hat{t}_{HT}$ est l'estimateur de HT du total $\sum_{i \in U} Y_i$. L'estimateur de HT du total ($\hat{t}_{HT} = \sum_{i \in s} y_i/\pi_i$) est un "bon" estimateur si le modèle décrit bien la population, i.e. $y_i/\pi_i \sim \mathcal{N}\left(\beta, \sigma^2\right)$. Binder et Roberts (2003) et Binder (2011) introduisent également une approche unifiée, dite approche *Model-Design based.* Il y a alors un mécanisme de sélection des données en deux phases. Les valeurs observées sur la population sont la réalisation de variables aléatoires issues d'un modèle de super-population, valeurs observées sur lesquelles s'applique ensuite un sondage. Pour les praticiens en sondages, l'intérêt des modèles est de débloquer des situations inextricables pour le calcul de variance complexe, par exemple pour l'estimateur d'un ratio. Les principaux domaines de la théorie des sondages qui font également intervenir les modèles sont le traitement de la non-réponse et l'estimation sur petits domaines. Enfin, les modèles sont invoqués comme justification théorique des sondages empiriques (quotas). Les 2 approches ne sont néanmoins pas cohérentes pour l'estimation des variances. Il n'y a alors pas de meilleur choix possible. Un économètre ne désire pas forcément obtenir une variance nulle sur une population exhaustive. Mais il ne désire pas non plus effectuer une hypothèse forte sur la relation entre les poids de sondages et la variance des observations. L'objet de ce document ne sera pas de trancher le débat académique sur les avantages et inconvénients de ces approches mais de présenter les solutions s'offrant aux utilisateurs.

L'organisation de la suite du document est la suivante. La deuxième partie présente, à destination d'un lecteur non spécialiste des enquêtes statistiques, les différentes étapes d'une enquête statistique. La troisième partie effectue une revue de littérature commentée sur la convergence des estimateurs et la prise en compte ou non de la pondération dans les modèles économétriques, le calcul de variance adéquat tenant compte du plan de sondage et les effets de l'imputation des données sur les modèles économétriques (illustrés à l'aide de simulations). La quatrième partie détaille un cas pratique à partir de l'enquête Patrimoine. La dernière partie conclut.

1. Särndal (2010) dresse un panorama plus large de l'emploi de modèles en sondages

2 Qu'est-ce qu'une enquête statistique ?

Une enquête statistique, c'est plusieurs étapes avant l'obtention de la base de diffusion. Des questions à la fois pratiques (coûts et délais) et théoriques (théorie des sondages) interviennent et peuvent engendrer des erreurs, se traduisant par un biais et/ou une baisse de la précision. Ce document se focalise sur les erreurs quantifiables à l'aide de la théorie des sondages, mais d'autres biais existent liés par exemple au mode de collecte (par voie postale, internet ou par réponse directe à un enquêteur), ou à la manière de rédiger ou de poser les questions. On pourra se référer à Razafindranovona (2015) pour une présentation détaillée de ces autres types d'erreurs et de la notion d'erreur d'enquête totale.

La première phase d'une enquête est celle de la conception, à travers la définition d'un questionnaire et la détermination du champ de l'enquête (population à interroger). Pour qu'une mesure soit juste, il faut que les concepts exposés dans le questionnaire soient ainsi clairs et compréhensibles de la même manière par les enquêtés. On parle également de biais de couverture lorsqu'une partie de la population n'est pas interrogée à tort.

Vient ensuite la construction du plan de sondage, i.e. la définition des unités à interroger et de leur nombre. Ce plan peut être très simple, par exemple un plan aléatoire simple (dit SAS) où les unités (individus, ménages ou entreprises) sont tirées aléatoirement dans la population. Chaque unité a alors le même poids. Ce type de plan est néanmoins peu efficace pour décrire des populations particulières (les ménages d'une région française, les entreprises de plus de 250 salariés par exemple). Les enquêtes « entreprise » sont ainsi habituellement stratifiées par taille (nombre de salariés) et secteurs d'activité (sondage dit stratifié où un poids de sondage différent par strate est défini). Les enquêtes « ménage » peuvent être stratifiées (par exemple au niveau région et âge de la population), ou tirées par grappes (par exemple pour une enquête sur les élèves du secondaire, on tirera des établissements, puis des classes dans lesquelles tous les élèves seront ou non interrogés). Plusieurs degrés (strates ou grappes imbriquées) peuvent être pris en compte. Il existe également des plans à probabilités inégales (plan de sondage systématique, ou Poissonien). Des contraintes interviennent également ; équilibrage, partage des poids, appariement avec une autre enquête. In fine le choix du plan de sondage sera fonction de critères de précisions et de diffusion ex-ante et de contraintes de coût (nombre d'unités enquêtées). On peut résumer cette étape comme une sélection des individus en fonction de caractéristiques Z_1, qui définissent un poids de sondage $\mathbb{P}(Z_1)$ égal ici à l'inverse de la probabilité d'inclusion.

La collecte des informations (mode de collecte, formation des enquêteurs, délais) arbitre entre la qualité des données, et des contraintes de coût et de délais. Par exemple une enquête via Internet est beaucoup moins coûteuse qu'une enquête en face à face mais le comportement des répondants n'est pas le même. Si les délais sont très serrés, il faudra accepter un taux de réponse plus faible. Des erreurs lors de la saisie et de la codification sont également possibles. Il reste délicat de quantifier les biais de cette étape. Une base de données brute est alors disponible, tous les enquêtés n'ont pas répondu (non-réponse totale), d'autres n'ont répondu que partiellement (non-réponse partielle) ou de manière incohérente. Des unités sont également exclues car devenues hors champs (par exemple une entreprise ayant initialement 15 salariés mais n'en ayant plus que 5).

Pour rendre la base exploitable, des retraitements sont donc effectués. Les poids de sondage sont modifiés pour tenir compte de la non-réponse totale (procédure dite de repondération). En pratique, des classes de réponse homogène sont définis à l'aide d'un modèle logistique et de variables Z_2. Les poids de sondage

$\mathbb{P}(Z_1, Z_2)$ sont ainsi ajustés par classes. Au lieu de cette phase de repondération, les données peuvent être imputées, l'ensemble des réponses d'un répondant sont alors dupliqués. Pour la non-réponse partielle, la solution la plus commune est d'imputer (de manière déterministe ou stochastique) les valeurs manquantes à l'aide de variables auxiliaires. L'imputation s'appuie sur l'amélioration (en termes de biais et de variance) de variables descriptives univariées (par exemple le total ou la moyenne). On ne tient généralement pas compte de la corrélation entre les variables dans ces imputations. Dans l'idéal, l'information sur le caractère imputé ou non d'une variable est conservée à l'aide de variables dites « drapeau ». La base de données ne présente alors plus de données manquantes.

Avant la diffusion de la base de données et des chiffres agrégés, un calage sur marges est réalisé pour s'assurer que les chiffres diffusés sont cohérents avec d'autres sources (par exemple du recensement de la population). Les poids de sondage $\mathbb{P}(Z_1, Z_2, Z_3)$ sont ainsi modifiés à l'aide des marges des variables Z_3. Pour des questions de secret statistique, les données individuelles sont parfois bruitées de telle sorte que les chiffres agrégés soient les mêmes mais pas les données individuelles.

A la différence d'une étude économique qui cherche à répondre à une question précise, l'ensemble de ces retraitements vise en premier lieu à rendre l'information économique primaire (i.e. les statistiques descriptives) la plus pertinente et précise possibles. Ils ne tiennent en particulier pas compte des corrélations entre variables. On comprend donc que les objectifs n'étant pas les mêmes, la simple application des techniques économétriques sans réflexion sur les modalités de construction des données peut engendrer des résultats incohérents.

La notion d'aléa de la théorie des sondages et de la théorie économétrique n'est enfin pas la même. Pour plus de détails sur la théorie des sondages, on pourra notamment se référer à Ardilly (2006) ou Tillé (2001).

3 Revue de littérature commentée

3.1 La convergence des estimateurs - Pondérer ou non le modèle économétrique

La différence d'approche est perceptible à travers les logiciels statistiques qui proposent de pondérer une régression linéaire principalement selon 2 approches.

3.1.1 Approche économétrique : le modèle de population

Ce n'est initialement pas un problème de sondages mais d'hétéroscédasticité. Le cadre est celui d'un modèle de population $\bar{Y}_g = \beta \cdot \bar{X}_g + \bar{\varepsilon}_g$ avec $\bar{\varepsilon}_g = \sigma^2/n_g$ et n_g le nombre d'individus de chaque groupe g.

La matrice de variance est alors connue $V(\bar{\varepsilon}_g) = W^{-1} \cdot \sigma^2$ avec $W = \text{Diag}(n_g)$. On peut alors calculer l'estimateur des moindres carrés généralisés, la forme de l'hétéroscédasticité étant connue.

L'estimateur efficace est obtenu en sphérisant le modèle initial en le multipliant par $W^{1/2}$. On applique alors les moindres carrés généralisés (MCG) sur le modèle "sphérisé" $\bar{Y}_g* = \beta \cdot \bar{X}_g * + \bar{\varepsilon}_g*$ avec $Z* = W^{1/2}Z$.

L'estimateur est alors $\hat{\beta} = (X'WX)^{-1}(X'WY)$ de variance estimée $\hat{V}\left(\hat{\beta}\right) = (X'WX)^{-1}\hat{\sigma^2}$, sous hypothèse d'homoscédasticité.

Pondérer un modèle économétrique est donc une hypothèse sur la variance des observations et non sur le tirage des individus mais...

3.1.2 Modèle pondéré type sondages

Si on pouvait observer l'ensemble des individus, les estimateurs du modèle $Y = \beta \cdot X + \varepsilon$ serait $\hat{\beta} = (X'X)^{-1} X'Y$. Pondérer dans une perspective "sondages" revient à estimer ces quantités à l'aide d'estimateurs d'Horvitz-Thompson :

$$\hat{\beta}_s = \left(X_s' W X_s\right)^{-1} \left(X_s' W Y_s\right)$$

On retrouve l'estimateur précédent ! Par contre, il n'y a pas d'hypothèse sur la distribution des résidus. La variance de cet estimateur correspond à la variance empirique sur tous les échantillons possibles :

$$V(\hat{\beta}) = \sum_s p(s) \left[\hat{\beta}_s - E(\hat{\beta})\right]^2$$

Avec $E(\hat{\beta}) = \sum_s p(s)\hat{\beta}_s$ et $p(s)$ la probabilité de tirer l'échantillon. On l'estime par

$$\hat{V}\left(\hat{\beta}\right) = \left(X_s' W X_s\right)^{-1} G^{\text{Sondages}} \left(X_s' W X_s\right)^{-1}$$

avec G^{Sondages} qui dépend du plan de sondage avec notamment un terme multiplicatif $(1-f)$, f étant le taux de sondage (qui peut être différent par strates, grappes...).

3.1.3 Cas général

D'Haultfœuille et Davezies (2012) visent à réconcilier les deux approches, économétrie et sondage. Ils modélisent le sondage (à travers le tirage mais également les traitements postérieurs d'enquête) comme un problème de sélection. Un modèle de superpopulation est associé aux observations, ce qui est nécessaire pour prendre en compte l'hétérogénéité d'individus semblables (selon des caractéristiques observables) mais qui ne font pas in fine les mêmes choix. Leurs principales conclusions sont qu'il est souvent préférable de pondérer même pour des modèles économétriques et que le chargé d'études doit connaître les variables jouant sur la probabilité de tirage, la non-réponse et le calage. Le choix de pondérer ou non une analyse économétrique est étudié d'un point de vue théorique (quelques hypothèses permettent de trancher) et par un test statistique du type Hausman qui compare les estimateurs non pondéré et pondéré.

Le cadre général est celui d'un plan de sondage poissonien. Les observations individuelles $\left(D, \tilde{X}, X, Y\right)_i$ sont alors supposées i.i.d. avec D l'indicatrice de réponse (i.e. le fait d'être tiré et de répondre), $\tilde{X}$ les variables expliquant la réponse finale (i.e. ayant servi à définir le plan de sondage, le modèle de non-réponse et le calage), X et Y les variables explicatives et expliquée du modèle économétrique. Les poids sont modélisés par $W_i = \mathbb{P}\left(D_i = 1/\tilde{X}_i\right)$, dont on dispose d'estimateurs convergents (i.e. les traitements post-collecte ne font pas d'erreurs systématiques).

La question de pondérer ou non repose sur la validité ou non de certaines hypothèses (cf supra). Le schéma 1 résume les différents cas possibles et l'estimateur qu'il est alors préférable d'utiliser.

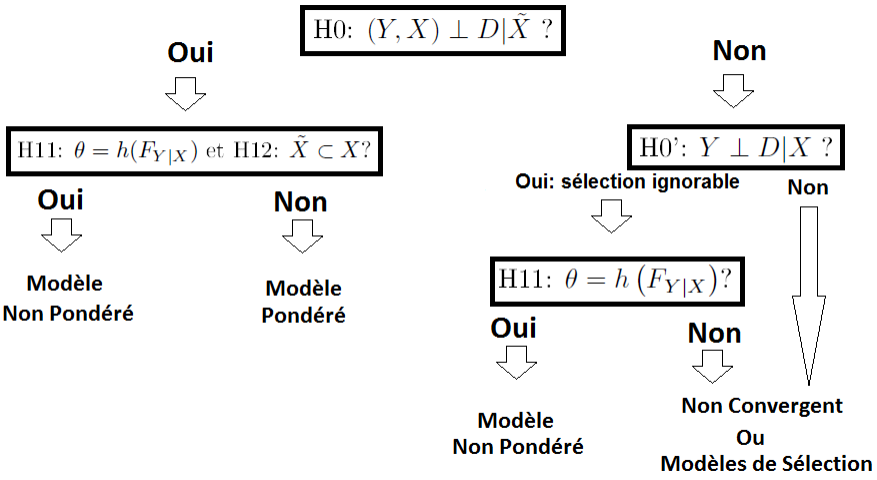

FIGURE 1 – Approche Davezies-D'Haultfœuille (2012) pour le choix de pondérer ou non un modèle

L'hypothèse notée H0 consiste à supposer que les variables utilisées pour définir le plan de sondage et corriger la non-réponse sont les facteurs pertinents de la sélection. Formellement, **H0 :** $(Y, X) \perp D|\tilde{X}$. La probabilité de tirage (à $\tilde{X}$ fixé) est indépendante de la probabilité de réponse à l'enquête et à (X, Y). La probabilité de réponse à l'enquête (à $\tilde{X}$ fixé) est indépendante de (X, Y).

Par exemple si $\tilde{X}$ inclut le type de ménage et l'âge de la personne de référence, Y = salaire et X = diplôme, on suppose que la non-réponse est indépendante du salaire et du diplôme à type de ménage et âge de la personne de référence fixée. Cette hypothèse n'est pas toujours vérifiée, par exemple la non-réponse peut être non ignorable, i.e. inclure d'autres facteurs que $\tilde{X}$. Les traitements d'enquête sont en effet effectués pour l'ensemble des variables de la base de données ; il n'y a pas de poids différencié par variables. Selon la question, le modèle de non-réponse adopté pourra être inadapté aux variables (X, Y) entrant en jeu.

Sous l'hypothèse H0, pondérer ou non repose sur la validité de deux autres hypothèses (H11) et (H12). Quand l'hypothèse H0 n'est pas vérifiée, on peut travailler avec une hypothèse alternative, H0' : $D|X \perp Y$. Sous cette hypothèse, la sélection, conditionnellement à X, ne dépend pas de Y.

H11 : Le paramètre θ qu'on cherche à estimer dépend uniquement de $F_{Y|X}$ (fonction de répartition de Y/X).

Par exemple, cette hypothèse est vérifiée pour toutes les méthodes qui définissent le paramètre θ comme solution annulant des moments conditionnels. C'est donc vrai pour les paramètres du modèle linéaire, pour les estimateurs du maximum de vraisemblance (logit, probit...) ou non-paramétriques. Ce n'est pas vrai par exemple pour les effets marginaux d'un modèle logit, qui dépendent de la loi des X ($\partial \mathbb{E}(Y/X)/\partial x_k = \partial \mathbb{P}(Y = 1/X)/\partial x_k = F'(X'\theta)\theta_k$ avec k l'indice d'une variable explicative).

H12 : $\tilde{X} \subset X$**.**

Cela signifie que le modèle inclut l'ensemble des variables expliquant le fait de répondre ou non à l'enquête. Elle se vérifie à l'aide de la documentation d'enquête. En pratique cette hypothèse est rarement vérifiée. Pour

l'enquête "patrimoine," le tirage est établi à un niveau géographique fin qui n'est pas disponible dans la base de diffusion pour des questions de confidentialité. De même, certaines variables de stratification d'un plan de sondage peuvent être exclues de l'analyse économétrique pour des raisons d'endogénéité. Enfin, Y peut intervenir dans l'échantillonnage. Quand on étudie une maladie, on sélectionne des sujets sains et des sujets malades pour avoir suffisamment de sujets malades même si l'occurrence de la maladie est faible.

Sous l'hypothèse H0, lorsque les deux hypothèses H11 et H12 sont vérifiées, il est préférable de ne pas pondérer. Dans ce cas en effet, les estimateurs pondérés et non pondérés sont tous deux convergents mais l'estimateur non pondéré est plus précis. Si H0 n'est pas vérifiée mais que H0' l'est, lorsque l'hypothèse H11 est vérifiée, il est préférable de ne pas pondérer. Dans les autres cas, H0' et/ou H11 non vérifiées, on ne peut pas en général obtenir des estimateurs convergents, que ce soit en pondérant ou non. D'autres méthodes (Heckman, calage généralisé) doivent être mobilisées.

Sous certaines hypothèses, il est préférable de choisir l'estimateur non pondéré (cf schéma 1), car il est convergent et efficace (i.e. de variance minimale). L'estimateur pondéré reste convergent mais est moins précis. Si les hypothèses ne sont pas vérifiées, l'estimateur non pondéré peut ne pas être convergent contrairement à l'estimateur pondéré. Pour confirmer les hypothèses et le choix du modèle non pondéré, les auteurs proposent de mettre en œuvre un test d'Hausman. L'idée du test d'Hausman est de comparer un estimateur convergent sous l'hypothèse nulle et l'hypothèse alternative (ici l'estimateur pondéré) et un estimateur convergent et efficace sous l'hypothèse nulle mais non convergent sous l'hypothèse alternative (ici l'estimateur non pondéré). La statistique de test s'appuie donc sur la différence des estimateurs pondérés et non pondérés (et des termes de variance), qui converge asymptotiquement vers une loi du χ^2. Si le choix de l'estimateur non pondéré est valide, la différence entre les estimateurs devrait être faible. Si l'hypothèse nulle est rejetée, soit les hypothèses H0 ou H0' ne sont pas vérifiées (et qui ne peuvent être testées), soit le modèle est mal spécifié.

Solon, Haider et Wooldridge (2014) présentent trois principales motivations à pondérer un modèle économétrique. Ils visent à mettre en avant, à l'aide d'exemples, que le choix de ne pas pondérer est parfois préférable. Sans proposer trop de formalisme, sa conclusion pratique est de toujours présenter les résultats des modèles pondérés et non pondérés. Le fait de constater de fortes divergences doit amener à s'interroger en premier lieu sur la spécification du modèle et la présence d'effets hétérogènes.

La première justification est la correction de l'hétéroscédasticité. Ce type de pondération n'a pas trait aux données d'enquête mais au fait d'utiliser un modèle de population avec des données agrégées. Des données au niveau des États américains peuvent correspondre à l'agrégation de comportements individuels : le taux de divorce dans un État est la moyenne des indicatrices au niveau individuel du fait d'avoir divorcé ou non. Les États ne comportant pas tous le même nombre d'individus, on peut suspecter que le modèle de population est hétéroscédastique, ce dont il faut tenir compte pour obtenir des estimateurs convergents de la précision des estimateurs. Les modèles sont alors pondérés avec des poids proportionnels à la taille de l'État. Ceci permet théoriquement d'obtenir des estimateurs plus précis. Mais cette correction n'est néanmoins valide que si les données individuelles peuvent être considérées i.i.d. Cette hypothèse peut être remise en cause si des effets sont spécifiques à chaque État (et sont donc partagés par les individus d'un même État). Dans ce cas, pondérer empire le problème d'hétéroscédasticité et on peut observer que l'estimateur non pondéré est plus précis. En pratique, il convient donc d'analyser la forme de l'hétéroscédasticité à l'aide de tests avant de la corriger. Il convient aussi de privilégier des estimations robustes à l'hétéroscédasticité (type matrice de

White).

La deuxième justification est celle également traitée dans l'article de Xavier D'Haultfœuille et Laurent Davezies, à savoir la sélection endogène des observations à travers le processus d'enquête (plan de sondage et mécanisme de la non-réponse). On retrouve les principales conclusions de Davezies et D'Haultfœuille, sans formalisme mathématique. Un sondage exogène est défini par l'indépendance du terme d'erreur et du poids de sondage ($\varepsilon \perp W$ ou, $Y/X \perp D$). Il n'y a pas besoin de pondérer le modèle dans ce cas. Comparer, d'un point de vue statistique, modèle pondéré et non pondéré est délicat notamment en présence d'hétéroscédasticité dans le modèle global (i.e. même sans données d'enquête). C'est pourquoi une approche "visuelle" est privilégiée pour comparer les deux estimateurs. Cette conclusion peut apparaitre contradictoire avec la mise en place d'un test d'Hausman, proposée par Davezies et D'Haultfoeuille. Le test d'Hausman fait néanmoins intervenir le calcul de la précision des estimateurs, pondéré et non pondéré. Sous l'hypothèse nulle, l'estimateur doit en particulier être efficace, ce qui suppose des observations i.i.d. et un terme d'erreur homoscédastique. Or, comme souligné précédemment, le calcul de la précision de l'estimateur non pondéré ne peut faire abstraction de la dépendance entre les observations introduite par le plan de sondage. Les hypothèses du test d'Hausman sont donc très fortes. Ce test apparaît donc comme une aide complémentaire à la décision, en complément d'une approche "visuelle".

La troisième justification pourrait être d'estimer de manière convergente des effets (notamment de traitements orientés évaluation des politiques publiques) en présence de comportements hétérogènes. Mais les auteurs montrent alors que la pondération ne peut pas tout. S'il y a des populations hétérogènes (avec des effets mais également des variances différentes), aucun de deux estimateurs, pondéré et non pondéré, ne convergent. Pondérer ne préserve donc en rien des problèmes de mauvaise spécification. Lorsque les estimateurs pondérés et non pondérés sont différents, distinguer ce qui relève d'une sélection endogène ou d'une mauvaise spécification du modèle n'est pas évident. Étudier l'hétérogénéité des effets (à travers l'inclusion de termes croisés) est nécessaire.

L'utilisation de poids normalisés (i.e. dont la somme correspond à la taille de l'échantillon) est une pratique classique lors d'estimation de modèle non linéaire pour éviter de surestimer la précision des estimateurs. Il n'y a pas de règle théorique mais ce choix apparait par la pratique moins néfaste que de ne pas pondérer. La prise en compte de la pondération est par ailleurs plus complexe avec des modèles non linéaires.

3.2 Pour conduire des tests statistiques, encore faut-il des variances justes

Le Guennec (2005) a détaillé le calcul de variance proposé par les logiciels avec une approche sondages. Les procédures orientées "sondages" (proc surveyreg dans SAS par exemple) ne tiennent pas compte par défaut de la correction de la variance par le facteur $(1 - f)$. Pour que la variance soit nulle en observant une population de manière exhaustive, il faut le demander explicitement. Par ailleurs, il est souvent possible avec une pondération "économétrique" de corriger certaines formes de corrélations inter-individuelles (en cohérence avec le plan de sondage), en définissant par exemple des clusters. Imaginons par exemple qu'on dispose de données obtenues à l'aide d'un plan de sondage stratifié, plusieurs solutions sont envisageables, par exemple 1) calculer l'estimateur de la variance orientée "sondages" en supprimant la correction par le facteur $1 - f$ (ce qui est l'option par défaut de la Proc Surveyreg dans SAS), 2) pondérer avec une perspective économétrique

(à l'aide de la proc reg dans SAS) et corriger l'hétéroscédasticité à l'aide de clusters par strate, ou 3) ne pas pondérer mais corriger l'hétéroscédasticité à l'aide de clusters par strate. Le choix sera en pratique fonction de la question et des données descriptives de l'enquête disponibles dans la base de diffusion (cf. application).

Ces différentes solutions restent imparfaites, car elles négligent certains aléas, celui du tirage ou celui du modèle de super-population. Graubard et Korn (2002) s'intéressent à l'estimation de la précision d'un estimateur d'un paramètre de superpopulation. Lorsque les données sont des données d'enquête, il est fréquent de ne pas tenir compte du plan de sondage et de faire comme si les données étaient directement générées selon un modèle de superpopulation. Les auteurs montrent que cela conduit à sous-estimer la variance des estimateurs. En se plaçant dans le cadre d'un sondage aléatoire simple stratifié sans remise, ils identifient trois composantes de cette variance : (1) une composante associée à l'aléa dû au plan de sondage, (2) une composante liée à l'aléa dû au modèle et (3) une composante liée à la variabilité inter-strates du modèle (et de la taille des strates). Sous l'hypothèse d'un taux de sondage faible et en l'absence de correction de population finie, les procédures logiciels orientées sondages (proc surveyreg dans SAS) identifient les deux premières composantes mais négligent la troisième. La démarche proposée par Graubard et Korn (2002) demeure complexe, à l'instar du calcul analytique de variance en théorie des sondages. Face à cette complexité, la question de l'estimation de certains termes par bootstrap se pose. Mais dans le cas général le bootstrap ne fonctionne pas quand on ne dispose pas de formule analytique de variance. C'est un constat d'échec. De plus, quand le taux de sondage est plus élevé, dès 0,3/0,5, les conditions asymptotiques ne sont pas remplies.

3.3 Données bruitées ou imputées, quelles conséquences ?

La gestion de la non-réponse partielle est abordée de manière très différente par les chargés d'études, selon l'information dont ils disposent. Un code "Valeur Manquante" est parfois disponible. Pour les variables qualitatives, une modalité spécifique peut être incluse dans le modèle. Pour les variables quantitatives, des modèles de sélection (Heckmann, Tobit) peuvent être utilisés.

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. Charreaux *et al.* (2016) étudient 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. Se restreindre aux seules observations sans données manquantes peut être un choix plus pertinent que l'utilisation de données imputées, ce qui est illustré par les graphiques suivants.

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.

Abrevaya et Donald (2017) montrent les biais de sélection associés aux choix classiques de traitement des données manquantes : 1) Supprimer les variables explicatives avec des données manquantes (biais de

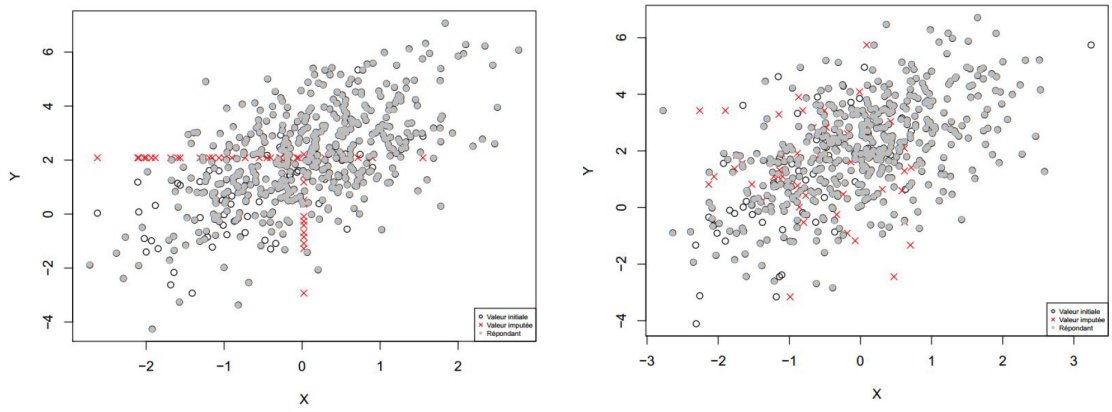

FIGURE 2 – Effet de l'imputation par la moyenne (gauche) ou des plus proches voisins (droite) sur les valeurs de Y et X

variable omise), 2) Ajouter une indicatrice de non-réponse pour les valeurs manquantes dans le modèle (biais de mauvais contrôle), 3) Imputer les données manquantes, par exemple à l'aide d'un modèle linéaire sur les seules observations complètement observées. Lorsque l'hypothèse (très forte) d'indépendance complète de la non-réponse (MCAR, Missing Completly At Random) peut être faite, supprimer les observations présentant des données manquantes reste valide, seule une perte de précision est dans ce cas observée. Les auteurs proposent une méthode de moments généralisés permettant de tenir compte dans l'estimation des observations présentant des données manquantes.

Au-delà des données d'enquête, la recherche en économie appliquée s'appuie sur des données individuelles administratives. Pour des raisons de confidentialité, ces données peuvent être bruitées (en tenant compte des corrélations entre les variables) et qualifiées de données synthétiques. Des recherches récentes (https ://www2.vrdc.cornell.edu) étudient la validité des études économiques utilisant de telles données.

3.4 Sélection de domaines d'études

Il est courant dans une étude économique de sélectionner seulement une partie de l'échantillon. Pour étudier un modèle d'offre de travail des femmes, on restreindra ainsi l'échantillon issu de l'enquête emploi aux femmes mariées conjointes de salarié. Cette sélection tient à la question et non au plan de sondage, ce qui peut engendrer des problèmes méthodologiques. L'effectif de ces sous-populations est alors aléatoire. D'un point de vue logiciel, SAS propose par exemple une option DOMAIN qui permet d'effectuer une analyse par domaine. La documentation de SAS précise ainsi "It is common practice to compute statistics for domains. The formation of these domains might be unrelated to the sample design. Therefore, the sample sizes for the domains are random variables. Use a DOMAIN statement to incorporate this variability into the variance estimation. Use the DOMAIN statement on the entire data set to perform a domain analysis. Creating a new data set from a single domain and analyzing that with PROC SURVEYREG yields inappropriate estimates of variance.". Ce problème de variance est surtout présent pour des échantillons de petite taille et donc de petits domaines. Pour des domaines de grande taille, la modification de variance induite par la variabilité de la taille du domaine est marginale. De même, plusieurs enquêtes sont parfois appariées. La notion de domaine ou de pondération n'est alors pas claire, lorsque les plans de sondage sont différents.

4 Un cas pratique : l'enquête Patrimoine

On propose d'illustrer les deux approches "économétrique" et "sondage" à partir des données de l'enquête Patrimoine 2010. Cette partie insiste sur les aspects pratiques : possibilités logiciels, estimations à partir des informations limitées à la disposition des chargés d'études.

4.1 Présentation de l'enquête

L'enquête Patrimoine est menée tous les six ans pour mesurer les actifs immobiliers, financiers et professionnels des ménages. À l'image des autres enquêtes ménages menées à l'Insee, les données mises à disposition des chargés d'études ont connu différents "traitements". Le manuel d'utilisation des données en donne un descriptif (succinct) :

1. Le plan de sondage
   Le plan de sondage diffère selon qu'on se trouve en France métropolitaine ou dans les DOM [2]. En France métropolitaine, le plan de sondage est à deux degrés. Dans un premier temps, des zones d'action enquêteur (ZAE) sont tirées. Ensuite, des adresses sont tirées dans les fichiers fiscaux en particulier de la taxe d'habitation. L'échantillon est finement stratifié. Pour tenir compte de l'extrême concentration du patrimoine et atteindre des populations relativement rares, certaines strates ont été fortement surreprésentées. Il en résulte une extrême variabilité des poids de tirage.
   Dans les DOM, le sondage est équilibré sur la microrégion. Plus précisément, il y a six strates pour l'échantillon "standard" (agriculteurs (sauf à La Réunion) ; hauts indépendants ; cadres ; revenus du patrimoine ; âgés ; reste) et 4 strates pour l'échantillon "hauts patrimoines" (riches urbains ; patrimoines à dominante mobilière ; patrimoines à dominante immobilière ; patrimoines plus faibles). L'enquête connaît également une extension "agriculteurs" en métropole.
2. Les traitements post-enquête
   (a) Pour corriger de la non-réponse totale, les données sont repondérées. La probabilité de répondre à l'enquête est modélisée par quatre logits non pondérés à partir des variables disponibles dans la base de sondage. Ces modèles permettent de construire des groupes de réponse homogènes. Au sein de ces groupes, le poids des répondants est multiplié par l'inverse du taux de réponse de l'ensemble.
   (b) Ensuite le calage sur marges est réalisé séparément pour la métropole et les DOM. Il repose en métropole sur la pyramide des âges (sexe × âge au niveau individu, et âge de la personne de référence), la catégorie socio-professionnelle et le diplôme de la personne de référence, le patrimoine net pour les "hauts patrimoines", les revenus d'activité et les revenus du patrimoine, la tranche d'unité urbaine, la ZEAT et le type de ménage.

2. À La Réunion, le plan de sondage est similaire à celui de la France métropolitaine, hormis l'absence d'une strate pour les agriculteurs. Dans les Antilles (Guadeloupe et Martinique), le sondage est tiré dans les Enquêtes Annuelles de Recensement et équilibré sur la microrégion. La stratification adoptée pour cet échantillonnage s'inspire de celle adoptée pour l'échantillon "standard" de la métropole et de la Réunion. Concernant le calage sur marges, pour la Réunion, il repose sur la pyramide des âges (sexe × âge au niveau individu), le nombre de ménages, le taux de propriétaires du logement principal, la catégorie socio-professionnelle de la personne de référence, les revenus d'activité et les revenus du patrimoine, le lieu de naissance de la personne de référence, le nombre de pièces du logement, la micro-région. Pour les Antilles, on utilise l'âge et la catégorie socio-professionnelle de la personne de référence, le type de bâti, le type de logement (individuel ou collectif), le nombre total de logements, le nombre de ménages propriétaires de leur résidence principale, le nombre d'individus par tranche d'âge.

(c) Imputation des données manquantes

Certaines données manquantes sont imputées. C'est le cas des variables de détention (binaire). La méthode utilisée est un hot-deck aléatoire équilibré, par strate. Le nombre de telles données imputées est faible.

Pour les variables de montant, l'information est collectée sous forme de fourchettes pour les actifs immobiliers et professionnels (le ménage évalue un montant minimum et maximum). Pour les actifs financiers, il était proposé au ménage une échelle de montants dont une modalité "montant en clair". Lorsque l'information obtenue est sous forme de fourchette ou d'échelle, les montants sont ensuite imputés de manière stochastique par un modèle économétrique auquel on ajoute des résidus simulés sous contrainte de respect des tranches initiales et de plafonds réglementaires (modèles sur données censurées).

La base a donc subi de nombreux traitements. Le chargé d'études ne dispose cependant que d'une information partielle. Les poids de tirage ne sont pas disponibles dans la base de données. On ne connait pas précisément les strates de tirage, de repondération ou de calage. Le premier degré du plan de sondage consiste à tirer des ZAE. Mais cette information, trop fine, n'est pas disponible dans la base.
Il est possible de repérer les ménages pour lesquels une imputation a été réalisée. À chaque variable de la table ménage est associée une variable avec le suffixe "_drap" dans la table Drap_men. On repère par ces variables les ménages pour lesquels une imputation a été réalisée : "0" (sans objet), "1" (réponse), "-1" (ne sait pas), "-2" (refus de répondre). Prendre en compte les traitements dans les estimations apparait donc compromis. On peut cependant se demander quelle est la meilleure stratégie pour le chargé d'études dans cette situation d'information incomplète.

4.2 Deux modèles

Nous proposons ici l'estimation de deux modèles, l'un linéaire, l'autre non linéaire, sur les données de l'enquête Patrimoine. Trois stratégies d'estimation sont proposées : deux estimations selon l'approche "économétrique" : sans tenir compte des poids, en tenant compte des poids, et une estimation selon l'approche "sondage" en essayant de tenir compte au mieux du plan de sondage.

4.2.1 Les procédures dans les logiciels statistiques

Nous décrivons ici brièvement les procédures disponibles sous les trois logiciels SAS, R et STATA pour mener les estimations. Lorsque les poids sont utilisés pour corriger de l'hétéroscédasticité dans l'estimation de la précision des estimateurs (i.e. que nous sommes en présence d'un modèle de population), il convient d'utiliser l'option WEIGHT sous SAS, AWEIGHT sous STATA, et WEIGHTS sous R (avec les procédures de régression adaptées).

Pour tenir compte du plan de sondage, il existe par contre des procédures adaptées.

Sous SAS, c'est la procédure SURVEYREG avec l'instruction WEIGHT pour les poids de sondage. Les instructions STRATA et CLUSTER permettent respectivement de définir les strates et les grappes. Le Guennec (2005) donne un descriptif détaillé de cette procédure et de l'équivalent pour les modèles non linéaires SURVEYLOGISTIC.

Sous STATA, l'option PWEIGHT permet d'indiquer que les poids correspondent à des poids de sondage (à probabilités inégales) dans les procédures classiques. Pour définir des plans de sondage plus complexes, on utilise les instructions SVYSET et SVY. La procédure se passe en deux temps. SVYSET définit le plan de sondage. SVY procède à l'estimation du modèle en tenant compte du plan de sondage défini.

Sous R, le package SURVEY permet de conduire des estimations orientées sondages. Comme sous STATA, on définit d'abord le plan de sondage par l'instruction SVYDESIGN. Ensuite, l'instruction SVYGLM procède à l'estimation du modèle en tenant compte du plan de sondage.

4.2.2 Les estimations sur les données de l'enquête Patrimoine 2010

Modèle linéaire

On estime d'abord un modèle linéaire. La variable d'intérêt est le logarithme du patrimoine brut. Trois modèles sont proposés : un modèle linéaire, sans tenir compte de la pondération ; un modèle linéaire pondéré ; un modèle linéaire tenant compte (imparfaitement) du plan de sondage. L'information à la disposition du chargé d'études étant limitée, on fait comme si les poids étaient des poids de sondage et que le plan de sondage était à probabilités inégales. On ne tient donc pas compte des strates. On ne tient pas compte non plus des imputations et repondérations.

Variable	Modèle 1	Modèle 2	Modèle 3
Age	$0,16^{***}$	$0,14^{***}$	$0,14^{***}$
	(0,0078)	(0,0077)	(0,011)
Age^2	$-0,0012^{***}$	$-0,00098^{***}$	$-0,00098^{***}$
	(0,00007)	(0,00007)	(0,000096)
CSP			
Ouvrier spécialisé	$4,46^{***}$	$4,64^{***}$	$4,64^{***}$
	(0,22)	(0,20)	(0,28)
Ouvrier qualifié	$5,57^{***}$	$5,86^{***}$	$5,86^{***}$
	(0,21)	(0,20)	(0,27)
Technicien	$6,60^{***}$	$6,88^{***}$	$6,88^{***}$
	(0,22)	(0,21)	(0,29)
Personnel de catégorie B	$6,79^{***}$	$7,13^{***}$	$7,13^{***}$
	(0,23)	(0,22)	(0,29)
Agent de maîtrise	$6,83^{***}$	$7,26^{***}$	$7,26^{***}$
	(0,23)	(0,22)	(0,29)
Personnel de catégorie A	$7,53^{***}$	$7,71^{***}$	$7,71^{***}$
	(0,22)	(0,22)	(0,29)
Ingénieur, cadre	$7,79^{***}$	$7,95^{***}$	$7,95^{***}$
	(0,22)	(0,21)	(0,28)
Personnel de catégorie C, D	$5,77^{***}$	$6,00^{***}$	$6,00^{***}$
	(0,22)	(0,21)	(0,29)
Employé	$5,49^{***}$	$5,60^{***}$	$5,60^{***}$
	(0,21)	(0,19)	(0,28)
Directeur général	$8,30^{***}$	$7,95^{***}$	$7,95^{***}$
	(0,26)	(0,30)	(0,40)

Le modèle 2 donne des estimateurs plus précis que le modèle 1. Les estimateurs des modèles 2 et 3 sont les mêmes, comme attendu. Les différentes estimations apparaissent proches. La significativité d'une variable ne change pas d'une estimation à une autre.

Modèle non linéaire

On estime un deuxième modèle, un modèle non linéaire dont la variable d'intérêt est "le ménage est propriétaire de sa résidence principale" ou non. De même trois stratégies d'estimation sont mises en œuvre : selon une approche économétrique un modèle logistique non pondéré et un modèle logistique pondéré [3], et selon une approche sondage une estimation en tenant compte du plan de sondage. De même que précédemment, on suppose que les poids correspondent à des poids de tirage d'un sondage à probabilités inégales.

3. Les poids ont été normalisés pour ne pas surestimer la précision. Il s'agit d'une transformation utilisée classiquement.

Variable	Proc logistic		Proc logistic pondérée		Proc surveylogistic	
			Poids normalisés			
	Coeff.	OR	Coeff.	OR	Coeff.	OR
Constante	$0,61^{***}$ (0,13)		$0,59^{***}$ (0,11)		$0,59^{***}$ (0,15)	
CSP						
Ouvrier spécialisé	$-1,71^{***}$ (0,14)	0,181	$-1,86^{***}$ (0,12)	0,16	$-1,86^{***}$ (0,16)	0,16
Ouvrier qualifié	$-1,00^{***}$ (0,12)	0,37	$-1,10^{***}$ (0,11)	0,33	$-1,10^{***}$ (0,14)	0,33
Technicien	$-0,14$ (0,15)	0,87	$-0,27^{**}$ (0,12)	0,76	$-0,27$ (0,17)	0,76
Personnel de catégorie B	$-0,22$ (0,15)	0,80	$-0,36^{***}$ (0,13)	0,70	$-0,36^{**}$ (0,17)	0,70
Agent de maîtrise	ref.		ref.		ref.	
Personnel de catégorie A	$0,62^{***}$ (0,15)	1,85	$0,40^{***}$ (0,14)	1,50	$0,40^{**}$ (0,18)	1,50
Ingénieur, cadre	$0,70^{***}$ (0,13)	2,02	$0,37^{***}$ (0,12)	1,45	$0,37^{**}$ (0,16)	1,45
Personnel de catégorie C, D	$-0,79^{***}$ (0,14)	0,45	$-0,94^{***}$ (0,12)	0,39	$-0,94^{***}$ (0,16)	0,39
Employé	$-1,05^{***}$ (0,12)	0,35	$-1,22^{***}$ (0,11)	0,30	$-1,22^{***}$ (0,14)	0,30
Directeur général	$0,90^{***}$ (0,26)	2,46	0,21 (0,26)	1,23	0,21 (0,40)	1,23
Taille de l'UU						
Commune rurale	$1,53^{***}$ (0,08)	4,63	$1,49^{***}$ (0,072)	4,42	$1,49^{***}$ (0,10)	4,42
UU de moins de 5 000 hbts	$0,88^{***}$ (0,12)	2,41	$0,84^{***}$ (0,11)	2,32	$0,84^{***}$ (0,14)	2,32
UU de 5 000 à 9 999 hbts	$0,67^{***}$ (0,12)	1,95	$0,55^{***}$ (0,11)	1,74	$0,55^{***}$ (0,15)	1,74
UU de 10 000 à 19 999 hbts	$0,49^{***}$ (0,11)	1,63	$0,51^{***}$ (0,10)	1,66	$0,51^{***}$ (0,14)	1,66
UU de 20 000 à 49 999 hbts	0,11 (0,10)	1,11	0,065 (0,095)	1,067	0,065 (0,13)	1,07
UU de 50 000 à 99 999 hbts	$-0,051$ (0,10)	0,95	$-0,062$ (0,095)	0,94	$-0,062$ (0,11)	0,94
UU de 100 000 à 199 999 hbts	0,12 (0,11)	1,13	0,11 (0,096)	1,12	0,11 (0,13)	1,12
UU de 200 000 à 1 999 999 hbts	ref.		ref.		ref.	
UU de Paris	$-0,31^{***}$ (0,079)	0,74	$-0,42^{***}$ (0,071)	0,65	$-0,42^{***}$ (0,094)	0,65
Age						
Moins de 30 ans	$-2,01^{***}$ (0,13)	0,13	$-1,98^{***}$ (0,099)	0,14	$-1,98^{***}$ (0,16)	0,14
30-40 ans	$-0,59^{***}$ (0,085)	0,56	$-0,54^{***}$ (0,072)	0,58	$-0,54^{***}$ (0,097)	0,58
40-50 ans	ref.		ref.		ref.	
50-60 ans	$0,58^{***}$ (0,082)	1,78	$0,48^{***}$ (0,074)	1,62	$0,48^{***}$ (0,092)	1,62
Plus de 60 ans	$0,87^{***}$ (0,070)	2,38	$0,80^{***}$ (0,064)	2,22	$0,80^{***}$ (0,083)	2,22

Ce modèle fait apparaître des différences dans les estimations par l'un ou l'autre des modèles. En particulier, le coefficient associé à la modalité "directeur général" est significatif au seuil 1% dans le modèle 1, et non significatif dans les modèles 2 et 3.

4.3 Pondérer ou ne pas pondérer

La première application est l'explication (descriptive) du patrimoine des ménages (en log) par leur âge et leur CSP. C'est donc un modèle linéaire. Les différences constatées pour les estimateurs, pondérés et non pondérés, sont faibles. L'hypothèse H11 est respectée mais pas l'hypothèse H12 (certaines variables de stratification ne sont dans tous les cas pas incluses dans la base de diffusion). Sous l'hypothèse H0, au vu de cette analyse, l'estimateur pondéré paraît préférable. Mais on pourrait aussi faire l'hypothèse que H0 n'est pas respectée (les hauts patrimoines répondent moins, et certaines CSP ont été surpondérés) mais H0' $(Y \perp D|X)$ l'est, i.e. qu'à CSP et âge fixés, les comportements de réponse sont indépendants de la valeur du patrimoine. Cette dernière hypothèse parait plus réaliste et plausible. En cohérence avec le fait d'observer des estimateurs proches, l'estimateur non pondéré pourrait alors être choisi. Les résultats doivent néanmoins amener à questionner le respect de l'hypothèse H0 ou H0'. Les comportements de réponse pourraient en effet toujours dépendre de la valeur du patrimoine. Les deux modèles seraient dans ce cas non convergents.

La deuxième application est l'explication (descriptive) du fait d'être propriétaire de sa résidence principale par l'âge, la CSP et la taille de l'unité urbaine. C'est donc un modèle non linéaire type logistique. La variable à expliquer et les variables explicatives sont moins sujettes à une censure des réponses des enquêtés. Les hypothèses H0 et H0' apparaissent plausibles. Des différences sont néanmoins constatées entre les estimateurs, pondérés et non pondérés, du fait de la dégradation de la précision des estimateurs pondérés. Il apparait possible de choisir également l'estimateur non pondéré.

5 Conclusions

Ces exemples illustrent le fait que la prise en compte des traitements d'enquête dans les modèles économétriques est difficile à deux titres. Tout d'abord, la théorie n'est pas encore complètement établie. En second, lieu, à ces difficultés théoriques s'ajoutent des difficultés pratiques. L'information sur les traitements opérés sur les données est en général incomplète. La question est donc plutôt : que peut-on faire au mieux avec cette information limitée ?

L'utilisation des procédures orientées sondages apparaît ainsi comme la solution la plus robuste. Mais lorsque les informations sur la construction des données d'enquête sont limitées, une approche mixte (et qui correspond à la pratique économétrique) pourrait donc être d'utiliser des estimateurs pondérés et de spécifier des termes de variance approchant au maximum la structure du plan de sondage. L'approche modèle permet également de comprendre pourquoi la variance n'est pas nulle, même pour un recensement.

Trois grandes règles peuvent être formulées en conclusion :

— La qualité d'une étude économique dépend en premier lieu de la qualité des données utilisées. De ce point de vue, étudier la manière dont les données ont été construites peut permettre de comprendre des résultats aberrants et corriger si nécessaire la méthodologie mise en oeuvre ;

— Il convient donc d'adopter une approche prudente avec des données d'enquête, pour tester la robustesse des analyses. Après étude de l'information d'enquête, il convient de comparer estimateurs pondérés et non pondérés, et de calculer les variances selon différentes stratégies si toute l'information n'est pas connue. De fortes divergences doivent amener à réfléchir à la mise en oeuvre de traitements spécifiques, par exemple recalculer des poids adaptés.

— A plus long terme, l'information contenue dans les bases de diffusion devrait contenir des variables détaillant la construction des données (poids initiaux, strates de tirage, indicatrices indiquant les observations imputées par exemple). En l'absence de ces informations, les solutions théoriques ne peuvent être mises en oeuvre.

6 Bibliographie

Abrevaya, J. et S. G. Donald. [2017] « A GMM Approach For Dealing With Missing Data On Regressors », *The Review of Economics and Statistics*, 99(4), 657-662.

Ardilly, Pascal. [2006] « Les Techniques de Sondage », Editions TECHNIP.

Binder, David A. [2011] « Estimating model parameters from a complex survey under a model-design randomization framework », *Pakistan Journal of Statistics*, 27 (4), p. 371-390.

Binder, David A. et Georgia R. Roberts. [2003] « Design-based and model-based methods for estimating model parametres », dans Chambers, R.L., et C.J. Skinner (dir.), *Analysis of Survey Data*, Wiley, Chapitre 3.

Charreaux, C., C. Favre-Martinoz, H. Harle, R. Le Saout et P.-A. Robert. [2016] « Econométrie et Données d'Enquête : les effets de l'imputation de la non-réponse partielle sur l'estimation des paramètres d'un modèle économétrique », Colloque francophone sur les sondages.

Davezies, Laurent et Xavier D'Haultfoeuille. [2012] « Faut-il pondérer ? Ou l'éternelle question de l'économètre confronté à des données d'enquête », *Acte des Journées de Méthodologie Statistique*, INSEE.

Graubard, Barry I. et Edward L. Korn. [2002] « Inference for superpopulation parameters using sample surveys », *Statistical science*, 17 (1), p. 73-96.

Little, Roderick J. [2004] « To model or not to model ? Competing modes of inference for finite population sampling », J*ournal of the American Statistical Association*, 99 (466), p. 546-556.

Le guennec, Josiane. [2005] « La régression sur échantillon avec SAS », *Acte des Journées de Méthodologie Statistique*, INSEE.

Razafindranovona, Tiaray. [2015] « La Collecte Multimode et le Paradigme de l'Erreur d'Enquête Totale », Document de travail INSEE n°M2015/01.

Särndal, Carl-Erik. [2010] « Models in survey sampling » dans Carlson, M., H. Nyquist et M. Villani (dir.), *Official Statistics - Methodology and Applications in Honour of Daniel Thorburn*, p. 15-28.

Solon, Gary, Steven J. Haider et Jeffrey M. Wooldridge. [2014] « What are we weighting for ? », *NBER Working Paper*, 18859.

Tillé, Yves. [2001] « Théorie des Sondages - Échantillonnage et Estimation en Populations Finies », Dunod.