« What are we weighting for? »
G. Solon, S. J. Haider, J. Wooldridge

Groupe de lecture « Économétrie des données d'enquêtes »

Martin Chevalier (DMCSI) – 15 décembre 2014

Démarche générale de l'article

Orienter les praticiens dans la conduite de leur études économétriques.

Ne cherche pas à donner une réponse en toute généralité : selon les **objectifs** de l'étude et les données analysées, il est opportun ou non de pondérer.

Fait suite à plusieurs travaux de Jeffrey Wooldridge sur la question, notamment :

- (1999) « Asymptotic properties of weighted M-estimators for variable probability samples », *Econometrica*, Vol. 67, No. 6, p. 1385-1406 ;
- (2002) « Inverse probability weighted M-estimators for sample selection, attrition and stratification », *Portugese Economic Journal*, Vol. 1, No. 2, p. 117-139.

Objectifs envisagés

Objectif 1 : Estimer une statistique (éventuellement complexe) sur une population à partir des données d'un échantillon.

Objectif 2 : Rechercher l'inférence dans un modèle économétrique

- Objectif 2.1 : Améliorer la précision des estimateurs en présence d'hétéroscédasticité ;
- Objectif 2.2 : Obtenir des estimations convergentes dans le cas d'un échantillonnage non-indépendant de la variable d'intérêt ;
- Objectif 2.3 : Estimer de façon convergente un effet marginal quand le modèle est mal spécifié.

Pondérer ou ne pas pondérer ?

Estimer une statistique sur une population à partir des données d'un échantillon

Dans tous les cas, **il faut pondérer** par l'inverse des probabilités d'inclusion dans la logique de l'estimateur d'Horvitz-Thompson.

Exemples : estimations de totaux, moyennes, quantiles ou quantités plus complexes (coefficients de Gini, etc.).

La précision de ces estimations peut être calculée :

- de façon analytique : en utilisant au besoin des techniques de linéarisation ;
- par réplication : à condition de s'assurer que ces méthodes conduisent à des estimateurs convergents sous le plan de sondage.

Pondérer ou ne pas pondérer ?

Estimer une statistique sur une population à partir des données d'un échantillon

Un cas particulier : une statistique construite à partir d'un modèle de régression.

Cette logique reste valide dans le cas d'une statistique construite à partir d'un modèle de régression.

Exemple : impact d'une caractéristique (couleur de la peau, sexe) ou d'un « traitement » toutes choses égales par ailleurs.

Là encore, la précision peut être obtenue :

- de façon analytique : dès lors que l'estimateur peut être exprimé comme une fonction d'un ou plusieurs totaux sur l'échantillon, il est possible de construire une linéarisée sur laquelle calculer la précision (Ardilly, 2006) ;
- par réplication.

Pondérer ou ne pas pondérer ?

Améliorer les estimateurs en cas d'hétéroscédasticité

La décision de pondérer ou non (et par quoi) dépend de la forme que prend l'hétéroscédasticité.

À partir de travaux portant sur l'impact de la législation sur le taux de divorce, les auteurs soulignent les difficultés liées à l'utilisation de la **pondération par la taille d'une unité dans le cas de données micro-agrégées**.

Exemples : caractéristiques d'élèves agrégées au niveau de leur classe, caractéristiques de salariés agrégées au niveau de leur entreprise, etc.

Pondérer ou ne pas pondérer ?

Améliorer les estimateurs en cas d'hétéroscédasticité

L'utilisation de la pondération par la taille de l'unité n'a de sens que si les termes d'erreur *individuels* sont indépendants et identiquement distribués.

Dickens (1990) souligne que cette hypothèse est rarement vérifiée en pratique : bien souvent les caractéristiques des individus d'une même unité sont corrélées (*clustered*).

De façon pratique :

- toujours chercher à analyser la forme de l'hétéroscédasticité avant de chercher à la corriger (test de Breusch-Pagan, de White, etc.) ;
- privilégier les estimations robustes à l'hétéroscédasticité de la matrice de variance-covariance (White, 1980)

Pondérer ou ne pas pondérer ?

Obtenir des estimations convergentes quand l'échantillonnage n'est pas indépendant de la variable d'intérêt

La décision de pondérer ou non par l'inverse de la probabilité d'inclusion dépend du **mécanisme de sélection** et des **variables incluses dans le modèle**.

L'exemple envisagé ici est un cas d'échantillonnage endogène dans une étude sur les modes de transport (Manski, Lerman, 1977), mais le cadre est généralisable.

Il envisage l'échantillonnage comme un mode de sélection sur des variables **observables**, **au moins au moment de la construction des pondérations** (Wooldridge, 1999, 2002).

Il se distingue ainsi du cadre d'Heckman (1979) où :

1. la variable déterminant la sélection est inobservable ;
2. de l'information auxiliaire est disponible pour les répondants et les non-répondants au moment de l'estimation.

Pondérer ou ne pas pondérer ?

Obtenir des estimations convergentes quand l'échantillonnage n'est pas indépendant de la variable d'intérêt

Formellement (Wooldridge, 2002), on définit le **M-estimateur pondéré par l'inverse de la probabilité de sélection** $\hat{\theta}_w$ comme la solution de :

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} s_i \frac{1}{p_i} q(X_i, \theta)$$

avec s_i l'indicatrice d'appartenance à l'échantillon, p_i la probabilité de sélection et q la fonction objectif associée au modèle.

Sous l'hypothèse que p_i **est connue pour tout individu de l'échantillon** (et d'autres hypothèses de régularité), Wooldridge montre que :

$$\hat{\theta}_w \xrightarrow{N \to \infty} \theta_0$$

Pondérer ou ne pas pondérer ?

Obtenir des estimations convergentes quand l'échantillonnage n'est pas indépendant de la variable d'intérêt

Quand l'**ensemble des variables déterminant la sélection** dans l'échantillon **sont incluses dans le modèle**, **il n'est pas nécessaire de pondérer** pour obtenir des estimateurs convergents. Par ailleurs, ils conservent leur propriété d'optimalité.

En revanche, quand **toutes les variables déterminant la sélection** dans l'échantillon **ne sont pas incluses dans le modèle**, **il est nécessaire** de pondérer pour obtenir des estimateurs convergents. Ceux-ci ne sont pas optimaux.

On retrouve les conclusions du document de travail de Laurent Davezies et Xavier d'Haultfoeuille (2009).

Pondérer ou ne pas pondérer ?

Estimer de façon convergente un effet marginal quand le modèle est mal spécifié

Dans tous les cas, il n'est pas possible d'estimer correctement un effet marginal dans ce cadre.

Quand un effet d'interaction est négligé dans la modélisation, l'utilisation de la pondération ne permet pas de garantir la convergence des estimations.

La déformation de l'échantillon par la pondération ne préserve donc pas des problèmes de **mauvaise spécification** (on raisonne sous le modèle).

Il convient d'**étudier les sources d'hétérogénéité** dans les données pour adapter les modèles en conséquence.

Éléments de conclusion et de discussion

Feuille de route

Pour estimer des statistiques (y compris issues de régressions) **sans chercher l'inférence** :

- pondérer par l'inverse des probabilités d'inclusion ;
- calculer la précision de la statistique en tenant uniquement compte du plan de sondage (analytique ou par réplication).

Pour mener à bien des **tests sur les paramètres de régression** :

- si toutes les variables décrivant le plan de sondage sont dans la base de données, ne pas pondérer et les intégrer comme variables de contrôle (convergent et optimal) ;
- dans le cas contraire (le plus fréquent en raison des contraintes de diffusion), utiliser un modèle pondéré (convergent).

Éléments de conclusion et de discussion

Principaux apports de l'article

Il clarifie l'utilisation de la pondération dans plusieurs cas pratiques et établit une **distinction claire entre estimation d'une quantité seule** (toujours pondérer) **et recherche de l'inférence** (pondérer selon les situations).

En proposant d'utiliser l'inverse de la probabilité de réponse pour rendre compte de la sélection opérée par le plan de sondage, il **réconcilie traitement économétrique du biais de sélection** (Heckman, 1979) **et estimateur d'Horvitz-Thompson**.

Directement tourné vers la pratique, il donne des **conseils d'ordre général** : toujours utiliser la matrice de variance-covariance de White, analyser l'hétéroscédasticité et les sources d'hétérogénéité, faire figurer les estimations non-pondérées et pondérées.

Éléments de conclusion et de discussion

Quelques limites

En pratique, on s'intéresse souvent **à la fois à la significativité** (inférence) **et à la valeur d'un coefficient de régression** : en suivant les auteurs, il pourrait être nécessaire dans le premier cas de ne pas pondérer et dans le second de pondérer, ce qui n'est pas naturel.

Quand on pondère une régression par l'inverse de la probabilité de sélection, souvent les tests sont menés **comme si l'échantillon était de la taille de la population**. Ce cadre ne permettrait-il pas de **justifier théoriquement l'utilisation d'une pondération proportionnelle à l'inverse de la probabilité de sélection mais valant en moyenne 1** ?

Bibliographie complémentaire

Ardilly P. (2006), *Les techniques de sondage*, Technip, Paris, 675 p.

Davezies L., D'Hautfoeille X. (2009), « Faut-il pondérer ? » , *Document de travail de la DESE*, 23 p.

Heckman J. (1979), « Sample Selection Bias as a Specification Error », *Econometrica*, Vol. 47, No. 1, p. 153-161