Comment utiliser les données d'enquête?

Guide illustré sur l'économétrie des données échantillonnées

Favre-Martinoz, Guillerm, Le Saout

JMS INSEE

Juin 2018

Introduction

- Les données d'enquête, construites à partir de la théorie des sondages, sont largement utilisées pour des études économiques avec des méthodes économétriques.
- Des liens économétrie/sondages mal connus ou mal compris.
- Peu de prise en compte des traitements d'enquête (plan de sondage, repondération et calage, imputation) en économétrie.
- Peu d'informations sur les traitements d'enquête accessibles aux chargés d'études dans les bases de diffusion.
- Des logiciels qui intègrent le traitement des données d'enquête.

Introduction

- Travail initié par un groupe de lecture, complété d'applications.
- Trois objectifs :
  - Quel est l'impact des traitements d'enquête sur les modèles économétriques ?
  - Que peut faire le chargé d'études avec les informations limitées dont il dispose (méthode de calcul des poids finaux, strates du plan de sondage...) ?
  - Quelles informations pourraient être ajoutées aux bases de diffusion pour améliorer les traitements économétriques ?
- L'exemple de l'enquête Patrimoine pour illustrer ces questions.

Sommaire

Les grandes étapes d'une enquête

- Conception
- Plan de sondage
- Collecte
- Retraitements : repondération, imputations
- Calage sur marges

Objectif : Rendre les statistiques descriptives les plus pertinentes et précises possibles

Théorie des sondages et économétrie

- L'approche « sondages » classique : la population est finie, l'aléa est dans le tirage, il n'y a pas de modèle ;
  On souhaite estimer le paramètre de population finie.
- L'approche « économétrique » classique : il y a une « super-population » générée par un modèle probabiliste ;
  On souhaite estimer un paramètre du modèle de super-population.
- Des débats académiques sur les avantages/inconvénients des 2 approches.

Pourquoi pondérer ?

- On dispose de données d'enquête, chaque individu k a une probabilité d'inclusion π_k d'un échantillon s.
- On cherche à mesurer l'effet d'un traitement $\mathbf{1}_{k \in \text{Gpe Traité}}$ sur une variable d'intérêt Y_k

$$Y_k = \alpha + \beta \cdot \mathbf{1}_{k \in \text{Gpe Traité}} + \varepsilon_k$$

- L'estimateur des MCO non pondéré, $\hat{\beta}^{MCO} = \overline{Y}_{\text{Gpe Traité\&s}} - \overline{Y}_{\text{Gpe Non Traité\&s}}$.
- L'estimateur d'Hájek, $\hat{\beta}^H = \left(\sum_j 1/\pi_j \right)^{-1} \sum_{j \in \text{Gpe Traité\&s}} Y_j/\pi_j - \left(\sum_k 1/\pi_k \right)^{-1} \sum_{k \in \text{Gpe Non Traité\&s}} Y_k/\pi_k$.
- En général $\hat{\beta}^{MCO} \neq \hat{\beta}^H$... et mieux vaut pondérer.

La pondération en économétrie

- Ce n'est initialement pas un problème de sondage mais d'hétéroscédasticité.
- Le cadre est celui d'un modèle de population $\bar{Y}_g = \beta \cdot \bar{X}_g + \bar{\varepsilon}_g$ avec $\bar{\varepsilon}_g = \frac{\sigma^2}{n_g}$ et n_g le nombre d'individus de chaque groupe g.
- La matrice de variance est alors connue, $V(\bar{\varepsilon}) = W^{-1}\sigma^2$ avec $W = \text{Diag}(n_g)$.
- On peut appliquer les moindres carrés généralisés sur le modèle « sphérisé » $W^{1/2} \cdot \bar{Y}_g = W^{1/2} \cdot \beta \cdot \bar{X}_g + W^{1/2} \cdot \bar{\varepsilon}_g$.
- L'estimateur est $\hat{\beta} = \left(X'WX\right)^{-1}\left(X'WY\right)$ de variance $\hat{V}\left(\hat{\beta}\right) = \left(X'WX\right)^{-1}\widehat{\sigma^2}$ (sous hypothèse d'homoscédasticité).
- Pondérer un modèle économétrique (avec les options pond des logiciels) est donc une hypothèse sur la variance des observations et non sur le tirage des individus mais...

La pondération en tenant compte du plan de sondage

- Si on pouvait observer l'ensemble des individus, les estimateurs du modèle $Y = \beta \cdot X + \varepsilon$ seraient

$$\hat{\beta} = (X'X)^{-1} X'Y$$

$$\hat{V}\left(\hat{\beta}\right) = (X'X)^{-1} \hat{V}\left(X'\hat{\varepsilon}\right) (X'X)^{-1}$$

- Pondérer avec 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 !

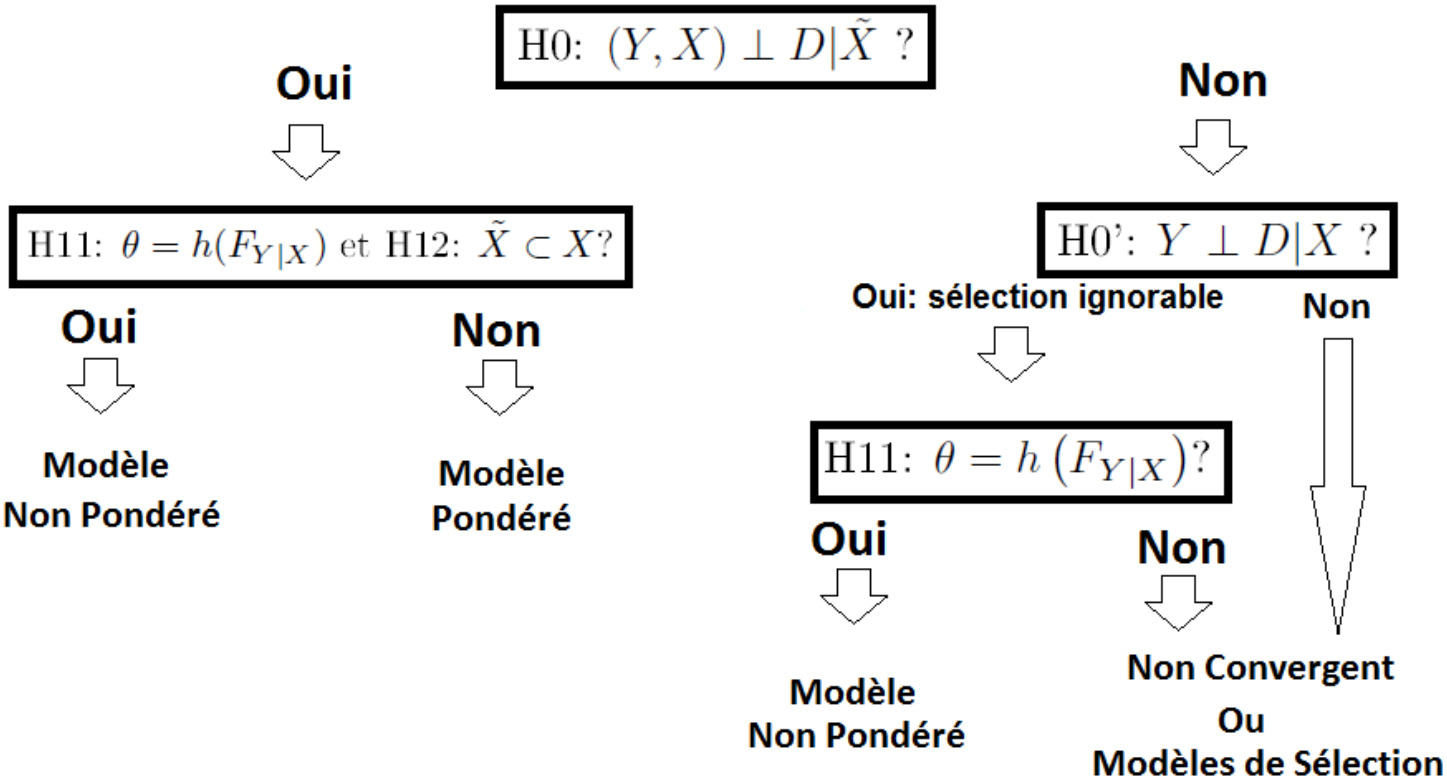
La pondération en tenant compte du plan de sondage

- 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 $E(\hat{\beta}) = \sum_s p(s)\hat{\beta}_s$
  On l'estime par :

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

G^{Sondages} 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...).

Un cadre général : Davezies-D'Haultfoeuille (2009)



Des variances justes ?

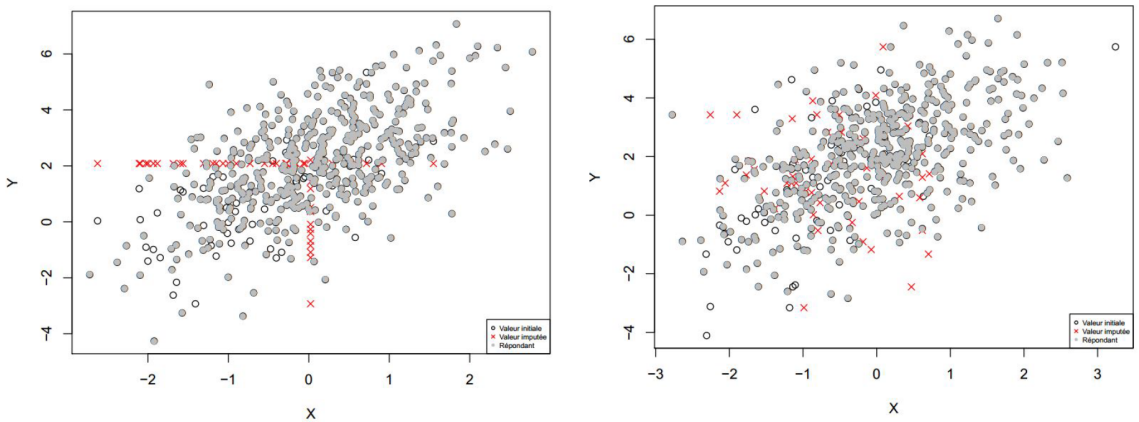
- Problème de l'approche sondages : si c'est un recensement $f = 1$, la variance est nulle.
- Lorsque les données sont des données d'enquête, il est donc fréquent de ne pas tenir compte du plan de sondage et de faire comme si les données étaient directement générées selon un modèle de superpopulation.
- Avec une régression pondérée, il est possible de «robustifier » la variance en définissant des clusters et en autorisant certaines formes d'hétéroscédasticité

$$\hat{V}\left(\hat{\beta}\right) = \left(X^{'}WX\right)^{-1} G^{\text{Econométrie}} \left(X^{'}WX\right)^{-1}$$

- Graubard et Korn (2002) : cela conduit à sous-estimer la variance des estimateurs, la variance totale dépend des aléas sondages, modèle et de la variabilité inter-strates de la taille des strates.

Les effets de l'imputation de la non-réponse partielle

- Les méthodes d'imputation type stochastique (hot-deck) ou déterministes engendrent-elles des erreurs de mesure d'un point de vue économétrique ?
- Charreaux et al. (2016) : simulations pour étudier les effets de l'imputation (simple) sur les analyses multivariées ;
- Les paramètres d'un modèle économétrique sont généralement estimés de façon biaisée même en cas d'exogénéité du processus de non-réponse.



Sommaire

Description de l'enquête

- Mesure des actifs immobiliers, financiers et professionnels des ménages tous les 6 ans ;
- Plan de sondage (France métropolitaine)
  - Tirage à deux degrés : 1) zones d'action enquêteur (ZAE) ; 2) stratification fine à partir des fichiers fiscaux de la taxe d'habitation.
  - Prise en compte de l'extrême concentration du patrimoine, sur-représentation des « agriculteurs »... une importante variabilité des poids de tirage.
- Repondération pour corriger de la non-réponse totale par score de propension à répondre
  - Modèle Logit pour expliquer la réponse et construire des strates de réponse homogène ;
  - Repondération au sein des classes par la moyenne des taux de réponse.

Description de l'enquête

- Calage sur marges (nombre de ménages, pyramide des âges, csp et diplôme de la personne de référence, tranche d'unité urbaine, ZEAT et type de ménage).
- Montants des actifs obtenus soit en clair, soit avec des cartes.
  - Les montants sont ensuite imputés (non-réponse partielle ou déclaration en tranches) de manière stochastique.
  - Modèle économétrique auquel on ajoute des résidus simulés sous contraintes de respect des tranches initiales.
- Faible imputation des variables qualitatives par hot-deck stratifié équilibré.

Contenu de la base de diffusion

- Pas les poids de tirage. Pas les strates de tirage, de repondération ou de calage. $\Longrightarrow$Obligation d'approximer ces informations pour en tenir compte.
- La variable de pondération permet les exploitations locales : Métropole, Réunion et Antilles (ensemble Guadeloupe - Martinique - Guyane). Aucune autre exploitation régionale que celles précitées n'est possible.
- Pondérations spécifiques pour les modules secondaires.
- Pour repérer les ménages pour lesquels une imputation a été réalisée, il y a la variable _drap 0 (sans objet), 1 (réponse), -1 (ne sait pas), -2 (refus de répondre).
- Quand il y a une incohérence entre la variable et la variable _drap associée, c'est qu'une imputation a été réalisée.

Régression non pondérée

Sous SAS

```
PROC REG DATA = matable;
MODEL y = x1 x2;
RUN;QUIT;
```

Sous STATA

```
reg y x1 x2
```

Sous R

```
summary(lm(y ~ x1+x2,data=matable))
```

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{V}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$$

$$\text{avec } \hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n-p} \text{ et } \hat{u} = Y - X\hat{\beta}$$

Régression pondérée

Sous SAS

```
PROC REG DATA = matable;
MODEL y = x1 x2;
WEIGHT pond;
RUN;QUIT;
```

Sous STATA

```
reg y x1 x2 [aweight=pond]
```

Sous R

```
summary(lm(y ~ x1+x2,data=matable,weights=POND))
```

$$\hat{\beta} = (X'WX)^{-1}X'WY \text{ avec } W = Diag(w_k)$$
$$\hat{V}(\hat{\beta}) = \hat{\sigma}^2(X'WX)^{-1}$$

En tenant compte du plan de sondage

Sous SAS

```
PROC SURVEYREG DATA = matable;
WEIGHT pond;
MODEL y = x1 x2;
RUN;
```

`STRATA` pour définir les strates
`CLUSTER` pour définir les grappes.

Sous STATA

```
reg y x1 x2 [pweight=pond]
```

Ou pour définir des plans de sondage plus complexes (grappes, clusters) :

```
svyset [pweight=pond]
svy: reg y x1 x2
```

Sous R

```
library(survey)
mondesign<-svydesign(ids=~0,strata=NULL,
weights=~POND,data=matable)
summary(svyglm(y ~ x1+x2,design=mondesign))
```

`ids` pour définir les grappes, `STRATA` pour définir les grappes.

- On obtient le même $\hat{\beta}$ qu'avec une PROC REG pondérée.
- L'estimateur de la variance est de la forme :

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

G dépend du plan de sondage. Voir Le Guennec (2005).

Enquête Patrimoine

Variable d'intérêt : Le log du patrimoine brut

- Modèle 1 : régression non pondérée
- Modèle 2 : régression pondérée
- Modèle 3 : Avec la procédure SURVEYREG.
  - Le sondage est stratifié par ZAE notamment, mais l'information n'est pas disponible dans la base.
  - On définit dans la procédure un sondage à probabilités inégales.
  - On ne tient pas compte d'autres traitements : le calage, les imputations.

Enquête Patrimoine

Variable d'intérêt : Le log du patrimoine brut

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)
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)

Quelques remarques sur les logiciels

- Les estimateurs peuvent être trés différents. Une approche pragmatique est nécessaire.
- La multiplication de la variance par $(1 - f)$ est optionnelle. Une seule perspective sondages ne suffit pas.
- Plusieurs méthodes pour le calcul de variance dans Stata : taylor linearized, bootstrap, jackknife, balanced repeated replicate, successive difference replicate.

Sommaire

- Prise en compte des traitements d'enquête dans les modèles économétriques difficile :
  - La théorie n'est pas encore complètement établie ;
  - L'information sur les traitements opérés sur les données est en général incomplète.
- L'utilisation des procédures orientées sondages apparaît comme la solution la plus robuste.
- Trois grandes règles :
  - étudier la manière dont les données ont été construites ;
  - adopter une approche prudente (comparer estimateurs pondérés et non pondérés, calculer les variances selon différentes stratégies).
  - ajouter dans les bases de diffusion des variables détaillant la construction des données (poids initiaux, strates de tirage, indicatrices indiquant les observations imputées par exemple)