Econométrie des données d'enquête

Groupe de lecture - Réunion 1

Marine GUILLERM et Ronan LE SAOUT

INSEE, division MAEE

24 Novembre 2014

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

- A travers la lecture de textes, des objectifs pratiques :
  - 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 (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.
- Objectif de cette première réunion : valider les thématiques, les textes et l'organisation.

Sommaire

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'Horvitz-Thompson,
$\hat{\beta}^{HT} = \sum_{j \in \text{Gpe Traité\&s}} Y_j/\pi_j - \sum_{k \in \text{Gpe Non Traité\&s}} Y_k/\pi_k$.
- En général $\hat{\beta}^{MCO} \neq \hat{\beta}^{HT}$... et mieux vaut pondérer.
- Comment définir le «en général » ?

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 $Y = \beta \cdot X + \varepsilon$ serait

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

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

- 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...).

Pourquoi prendre en compte le plan de sondage ?

- Problème : si c'est un recensement $f = 1$, la variance est nulle. Est-ce vraiment ce que veut l'économètre ? Pourquoi dans ce cas l'économètre calcule une précision ?
- Avec une régression pondérée, il est possible de «robustifier » la variance en définissant des clusters et en autorisant certaines formes d'hétéroscédasticité

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

Il n'y a pas de terme multiplicatif $(1 - f)$. N'est-ce pas préférable pour l'économètre (qui ne connaît pas le détail du plan de sondage) ?

Théorie des sondages et Econométrie

- L'approche « sondages » classique : la population est finie, l'aléa est dans le tirage, il n'y a pas de modèle ;
- L'approche « économétrique » classique : il y a une « super-population » générée par un modèle probabiliste ;
- Une approche « modèle » qui permet de rapprocher ces 2 paradigmes ?

L'estimation sur un domaine

- Il est courant dans une étude économique de sélectionner seulement une partie de l'échantillon

Par exemple, les femmes mariées conjoints de salarié pour étudier un modèle d'offre de travail.

- Cette sélection tient à la question et pas au plan de sondage.
- L'effectif de ces sous-populations dans l'échantillon est aléatoire (idem pour notre exemple de groupe traité).
- Quelles implications pour l'économètre ? Quelles relations avec l'estimation sur domaine ?
- Quel sens a l'appariement de 2 échantillons d'enquête avec des plans de sondage différents (mais qui ont par exemple été coordonnés) ?

Les traitements de la non-réponse

- La repondération pour tenir compte de la non-réponse totale est-il une manière efficace de traiter la sélection des répondants ? Quid de la différence entre poids de tirage et poids finaux ?
- Le calage peut-il être perçu comme des équations de moments économétriques ?
- Les méthodes d'imputation type stochastique (hot-deck) ou déterministes engendrent-elles des erreurs de mesure d'un point de vue économétrique ?
- Comment tenir compte de ce qui n'est pas observé en économétrie (identification partielle) ?

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 containtes 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.

Sommaire

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)

Enquête Patrimoine - Modèle non linéaire (1/2)

Variable	Proc logistic		Proc logistic pondérée Poids normalisés		Proc surveylogistic	
	Coeff.	OR	Coeff.	OR	Coeff.	OR
Constante	$0,6099^{***}$ (0,1274)		$0,5901^{***}$ (0,1107)		$0,5901^{***}$ (0,1516)	
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

Enquête Patrimoine - Modèle non linéaire (2/2)

Variable	Proc logistic		Proc logistic pondérée Poids normalisés		Proc surveylogistic	
	Coeff.	OR	Coeff.	OR	Coeff.	OR
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

Quelques remarques sur les logiciels

- Les estimateurs peuvent être très différents. Quels critères statistiques pour le choix du modèle, pondéré ou non pondéré ?
- La multiplication de la variance par $(1 - f)$ est optionnelle. Est-ce là introduire une hypothèse de « sur-population » avec une perspective « sondages » ?
- Analyse possible par domaine dans SAS et Stata.
- Plusieurs méthodes pour le calcul de variance dans Stata : taylor linearized, bootstrap, jackknife, balanced repeated replicate, successive difference replicate.
- Stata permet de tenir compte de la poststratification mais que faire quand les strates de tirage ou de poststratification ne sont que partiellement connues ?