Groupe de lecture "Econométrie des données d'enquête"

Compte-rendu de la première réunion, 24 novenbre 2014: objectifs et organisation

Suivi par Marine Guillerm et Ronan Le Saout

Les données d'enquête, construites à partir de la théorie des sondages, sont largement utilisées pour des études économiques avec des méthodes économétriques. Les liens entre l'économétrie et les sondages sont pourtant mal connus ou mal compris. Les chargés d'études et les chercheurs 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, répondération liée à la non-réponse ou au calage). L'information disponible sur ces traitements dans les bases de diffusion sont de plus pauvres. Cette prise en compte s'améliore, par la diffusion de documents de travail (Davezies et D'Haultfoeuille sur les poids de sondage par exemple) et de procédures orientées "sondages" dans les principaux logiciels statistiques.

La division des méthodes appliquées de l'économétrie et de l'évaluation a donc proposé de mettre en place un groupe de lecture sur l'économétrie des données d'enquête qui se réunira une fois par mois. La présentation des articles sera suivie de discussions techniques et sur l'inférence associée dans les principaux logiciels statistiques. Cette discussion sera aussi l'occasion d'étudier comment appliquer ces méthodes sur les données d'enquête de l'INSEE. Les objectifs sont donc tournés vers la pratique à l'aide des logiciels, plus que sur la présentation de méthodes économétriques théoriques. Trois objectifs peuvent ainsi être identifiés:

- La diffusion de la connaissance sur l'impact des traitements d'enquête sur les modèles économétriques (biais et précision), et les différences d'approches entre la théorie des sondages et l'économétrie;
- Proposer des solutions pratiques à l'aide des principaux logiciels statistiques avec les informations limitées disponibles dans les bases de diffusion;
- Identifier, à plus long terme, les informations complémentaires qui pourraient être ajoutées aux bases de diffusion (strates de tirage et de post-stratification, poids de tirage...).

Le groupe a validé l'organisation générale des prochaines séances (cf.annexe). La prochaine session aura lieu le lundi 15 décembre de 15h30 à 17h30 salle 18-119 (MK2), sur la prise en compte de la pondération en économétrie. Les documents du groupe (textes, compte-rendu) seront mis en ligne sur le site Intranet de la division des méthodes appliquées de l'économétrie et de l'évaluation.

1 Survol des enjeux théoriques

1.1 Pourquoi pondérer une analyse économétrique ?

Pour justifier la nécessité de pondérer une analyse économétrique effectuée à l'aide de données d'enquête, un court exemple illustratif a été donné.

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

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

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

A l'aide d'un estimateur Horvitz-Thompson de ces moyennes, on obtiendrait

$$\hat{\beta}^{HT} = \sum_{j \in \text{Gpe Traité\&s}} Y_j/\pi_j - \sum_{k \in \text{Gpe Non Traité\&s}} Y_k/\pi_k.$$

Il n'y a aucune raison que ces deux estimateurs soient égaux et, en général, $\hat{\beta}^{MCO} \neq \hat{\beta}^{HT}$. Mieux vaudrait alors pondérer. Mais au-delà du simple constat "visuel", comment définir statistiquement que l'estimateur pondéré est préférable à l'estimateur non pondéré ?

1.2 Deux approches différentes de la pondération: économétrie vs sondages

Les logiciels statistiques proposent de pondérer une régression linéaire principalement selon 2 approches.

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...

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:

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

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...).

Des différences d'approche

Un débat s'est engagé parmi les membres du groupe sur ces 2 approches différentes. D'un côté, la théorie des sondages "classique" considère qu'il n'y a pas de modèle statistique et, en particulier, pas d'hypothèses sur les résidus. L'aléa est dans le tirage, à l'intérieur d'une population de taille finie. De l'autre, l'économétrie "classique" considère que les données sont générées par un modèle probabiliste (DGP, Data Generating Process) et qu'il existe donc toujours une "super-population".

De manière plus récente, la théorie des sondages intègre la notion de modèle ("model based approach"). De même, l'économétrie, avec l'émergence de données massives, réfléchit à l'inférence sur population exhaustive, à l'image de l'article récent de Abadie et al. (2014). Les 2 approches sont cohérentes pour le calcul des estimateurs, mais pas 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. Tout est donc question de compromis et de comprendre ce que signifient ces différentes variances.

Dans une perspective logicielle, on constate en effet que 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. Ces différentes approches sont-elles cohérentes asymptotiquement? Dans le cas où peu d'éléments sur le plan de sondage et les traitements d'enquête sont connus, l'une est-elle préférable à l'autre ? Il s'agit de bien réfléchir à ce que revêt la notion de variance des estimateurs selon ces différentes approches.

On pourra noter que Stata est plus intuitif sur la gestion de la pondération. Les principales procédures économétriques intègrent en effet plusieurs options de poids : aweight pour l'appoche économétrique, pweight pour l'approche sondages, fweight pour des poids de duplication des observations et iweight pour des cas spécifiques. Seule la prise en compte de plan de sondage complexe fait appel à une procédure spécifique (svy). Au contraire, SAS et R proposent avec les principales procédures économétriques une unique option de pondération, qui correspond à l'approche économétrique. L'appel à l'approche "sondages" s'effectue par des procédures spécifiques.

La prise en compte de la pondération est par ailleurs plus complexe avec des modèles non linéaires. Dans le cas où les observations sont corrélées (ce qui est le cas avec des données d'enquête), l'estimateur du maximum de vraisemblance n'est en effet plus convergent au contraire de l'estimateur des MCO.

1.3 Autres questions associées aux données d'enquête

Dans les exemples précédents, les poids de tirage et les poids finaux étaient considérés comme identiques. Ils n'étaient issus que du plan de sondage. Or ces poids sont rarement identiques pour 2 raisons.

En premier lieu, il y a de la non-réponse totale qui entraîne une répondération. Ces techniques correspondent à une prise en compte particulière de la propension des individus à répondre et donc à des hypothèses implicites sur le traitement de la sélection. En second lieu, un calage sur marges est souvent effectué pour assurer une cohérence avec des chiffres agrégés issus de sources exhaustives (par ex. le recensement), ce qui peut s'analyser d'un point de vue économétrique comme des équations de moments.

La théorie des sondages fait donc des hypothèses économétriques, qu'il convient de mettre en avant. L'implication de ces traitements sur l'identification des paramètres et de leur variance pourra également être étudiée. D'un point de vue logiciel, Stata propose par exemple la prise en compte de la poststratification.

D'autres questions ont trait à la pratique des chargés d'études, concernant la sélection de leur échantillon et la gestion de la non-réponse partielle.

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 conjoints de salarié. Cette sélection tient à la question et non au plan de sondage. 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.

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 inluse dans le modèle. Pour les variables quantitatives, des modèles de sélection (Heckmann, Tobit) peuvent être utilisés. Mais, bien souvent, la non-réponse partielle a été imputée de manière stochastique (hot-deck) ou déterministe. A noter que la non-réponse totale peut également être imputée par hot-deck. L'implication de ces imputations (erreurs de mesure?) sur l'identification des paramètres et leur variance pourra également être étudiée.

2 Illustration à l'aide de l'enquête patrimoine

On propose d'illustrer les deux approches "économétrique" et "sondage" définies dans la première partie 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.

2.1 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[1]. 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 $\times$ 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.

   (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

[1] 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 $\times$ 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.

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.

2.2 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" définie dans la partie précédente: 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.

2.2.1 Les procédures sous les logiciels statistiques

Nous décrivons ici brièvement les procédures disponibles sous les trois logiciels SAS, R et STATA pour mener les estimations.

1. **Sans tenir compte des poids:**

   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))
   ```

2. **En tenant compte des poids:**

   Sous SAS:

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

   Sous STATA:

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

   `aweight` indique qu'on a un modèle de population. Les poids sont utilisés pour corriger de l'hétéroscédasticité dans l'estimation de la précision des estimateurs.

   Sous R:

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

3. **En tenant compte du plan de sondage**

   Sous SAS:

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

Les instructions `STRATA` et `CLUSTER` permettent respectivement de définir les strates et les grappes. Guennec (2005) donne un descriptif détaillé de la procédure `SURVEYREG` et de l'équivalent pour les modèles non linéaires `SURVEYLOGISTIC`.

Sous STATA:
Pour un sondage à probabilités inégales:

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

`pweight` permet d'indiquer que les poids correspondent à des poids de sondage.

Pour définir des plans de sondage plus complexes, on utilise `svyset` et `svy`:

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

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

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

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.

2.2.2 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:
Modèle 1: Un modèle linéaire, sans tenir compte de la pondération.
Modèle 2: Un modèle linéaire pondéré.
Modèle 3: 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 (voir partie 1). Les différentes estimations apparaissent proches. La significativité d'une variable ne change pas d'une estimation à un 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[2], et selon une approche sondage une estimation en tenant compte du plan de sondage. Cette dernière estimation est réalisée avec la procédure SAS `SURVEYLOGISTIC`. De même que précédemment, on suppose que les poids correspondent à des poids de tirage d'un sondage à probabilités inégales.

[2]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 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
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.

Ces exemples illustrent le fait que la prise en compte des traitements dans les estimations est difficile à deux titres. En théorie, on ne sait pas bien comment estimer notre modèle. Une approche économétrique ou une approche sondage? et si on opte pour une approche économétrique: pondérer ou pas? D'autres alternatives sont peut-être encore possibles. 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?
Il serait intéressant de compléter ces résultats par des estimations permettant de mieux tenir compte du plan de sondage. La section Patrimoine a proposé de nous apporter de l'information supplémentaire. Des tests sur des données simulées pourraient également venir compléter ultérieurement l'analyse.

References

Abadie, A., S. Athey, G. W. Imbens et J. M. Wooldridge. 2014, «Finite Population Causal Standard Errors», NBER Working Papers 20325, National Bureau of Economic Research, Inc.

Guennec, J. L. 2005, «La régression sur échantillon avec sas», Acte des JMS, INSEE.

Annexe 1 : Organisation prévisionnelle des prochaines séances

Séance 2 : Lundi 15 Décembre, 15h30/17h30, Salle 18-119 (MK2)

La (non) prise en compte des poids de sondage sur les estimateurs

Davezies, Laurent, et Xavier D'Haultf œuille. (2009) "Faut-il pondérer ? Ou l'éternelle question de l'économètre confronté à des données de sondage." Document de travail n° G2009/06, Direction des Études et Synthèses Économiques, INSEE.

Solon, Gary, Steven J. Haider, et Jeffrey Wooldridge. (2014) "What are we weighting for ?" NBER Working Paper n°18859.

Séance 3 : Lundi 19 Janvier, 15h30/17h30, Salle 1139

La modélisation économétrique en théorie des sondages

Chapitre 3 "Design-based and Model-based Methods for Estimating Model Parametres." Chambers, R.L., et C.J. Skinner. (2003) "Analysis of Survey Data" Wiley.

Little, R. (2004) "To model or not to model ? Competing modes of inference for finite population sampling", JASA

Séance 4 : Lundi 9 Février, 15h30/17h30, Salle 933

La (non) prise en compte du plan de sondage sur la variance

Bhattacharya, Debopam. (2005) "Asymptotic inference from multi-stage samples." Journal of Econometrics, 126(1), 145-171.

Séance 5 : Lundi 16 Mars, 15h30/17h30, Salle 933

La (non) prise en compte de la non-réponse totale

D'Haultfœuille, Xavier. (2010) "A New Instrumental Method for Dealing with Endogenous Selection." Journal of Econometrics, 154(1), 1-15

Chapitre 1 de Manski, Charles F. (2003) "Partial identification of probability distributions." Springer Series in Statistics.

Séance 6 : Lundi 13 Avril, 15h30/17h30, Salle 933

La (non) prise en compte de la non-réponse partielle

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

Von Hippel, P. T. (2007) "Regression With Missing Y's: An Improved Strategy for Analysing Multiply Imputed Data." Sociological Methodology, 37, 83–117.

Chen, Qingxia, Joseph G. Ibrahim, Ming-Hui Chen, et Pralay Senchaudhuri, « Theory and inference for regression models with missing responses and covariates. » Journal of Multivariate Analysis, 99(6), 1302-1331.

Séance 7 : Lundi 18 Mai, 15h30/17h30, Salle 933
A définir

Séance 8 : Lundi 15 Juin, 15h30/17h30, Salle 933
A définir

Par exemple :

Bhattacharya, Debopam. (2008) "Inference in panel data models under attrition caused by unobservables ." Journal of Econometrics, 144(2), 430-446.

Behaghel, Luc, Bruno Crépon, Marc Gurgand, et Thomas Le Barbanchon. (2012) "Please Call Again: Correcting Non-Response Bias in Treatment Effect Models" IZA DP No. 6751.

Davezies, Laurent, et Xavier D'Haultfœuille. (2014) "Endogenous attrition in panels" Mimeo.

Raghunathan, Trivellore E. (2004) "What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data." Annual Review of Public Health, 25, 99-117.

Souche, Stéphanie. (2014) "Pourquoi ne veut-on pas répondre à une question sur son revenu? L'aide des modèles censurés pour les enquêtes transport." Communication JMA.

Extrait de Mallinckrodt, Craig. (2013) "Preventing and treating missing data in longitudinal clinical trials: a practical guide" Cambridge (GBR) ; West Nyack, N.Y. : CUP. Cambridge University Press.

Extrait de Chambers, R.L., et C.J. Skinner. (2003) "Analysis of Survey Data" Wiley.

Extrait de Skinner, C.J., Holt, et Smith. (1999) "Analysis of complex surveys" Wiley.

Smith, TMF. (1987) "To weight or not to weight, that is the question", dans Bayesian statistics, 3 (Valencia, 1987), 437–451, Oxford Univ. Press, New York.

Nathan, G. et D. Holt (1980) "The effect of Survey Design on Regression Analysis", Journal of the Royal Statistical Society. Series B (Methodological), Vol. 42, No. 3, pp. 377-386.

Pfeffermann , D. (1993) "The role of Sampling weights when modeling survey data" International Statistical Review