Groupe de lecture "Econométrie des données d'enquête"

Compte-rendu de la deuxième réunion, 15 décembre 2014
La (non) prise en compte des poids de sondage sur les estimateurs

Suivi par Marine Guillerm et Ronan Le Saout

Cette deuxième séance du groupe de lecture avait pour thème: La (non) prise en compte des poids de sondage sur les estimateurs. Deux textes autour de ce thème ont été présentés. Xavier D'haultfœuille et Laurent Davezies ont présenté leur document de travail (Actes des JMS 2012) "Faut-il pondérer? Ou l'éternelle question de l'économètre confronté à des données de sondage". Martin Chevalier a présenté le document de travail du NBER "What are we weighting for?" de Solon, Haider et Wooldridge (2014).

L'utilisation des poids de sondage pour le calcul des statistiques descriptives fait consensus. Cette session avait pour objectif d'étudier la convergence des estimateurs d'un modèle de régression, selon que les observations sont pondérées ou non. Ces deux articles traitent également du calcul des variances (de ces estimateurs) associées à des données d'enquête. Cette question sera abordée de nouveau en session 4 du groupe de lecture, le 9 février 2015. Les présentations n'ont donc traitées que de manière succincte ce point.

1 Faut-il pondérer? Ou l'éternelle question de l'économètre confronté à des données de sondage

Le document de Xavier D'Haultfœuille et Laurent Davezies vise 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. D'un point de vue pratique, 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.

1.1 Faut-il ou non pondérer pour obtenir des estimateurs convergents?

Le cadre général est celui d'un plan de sondage poissonnien. 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).

1.1.1 Les différentes hypothèses

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.

La première hypothèse notée H0 doit obligatoirement être vérifiée. Dans le cas contraire, aucun des deux estimateurs (pondéré et non pondéré) n'est convergent, à moins de travailler avec une hypothèse alternative

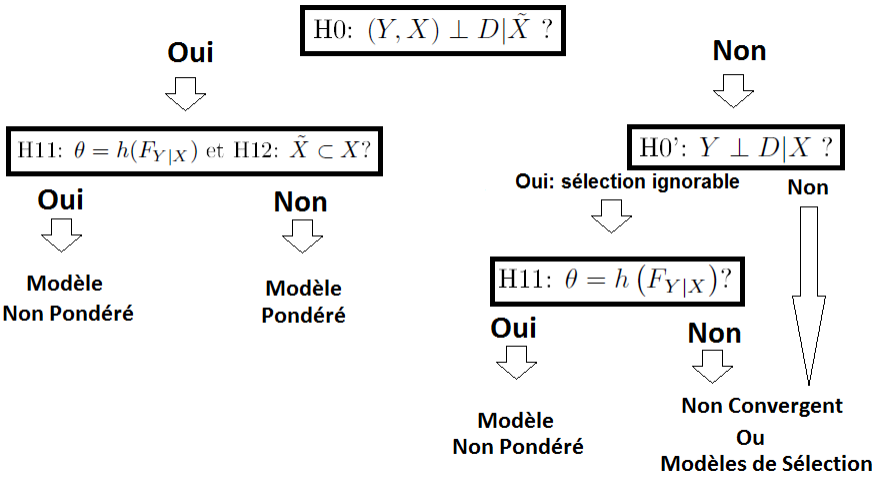

Schéma 1: Approche Davezies-D'Haultfœuille pour le choix de pondérer ou non un modèle

(cf supra). Cette hypothèse 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 paramètreθ 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.

1.1.2 Quand pondérer?

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.

Des approches mixtes sont possibles, par exemple pour estimer les effets marginaux d'un modèle Logit. Les estimateurs peuvent être obtenus à l'aide d'un modèle non pondéré si les hypothèses sont vérifiées. Mais pour les effets marginaux, l'hypothèse H11 n'est pas vérifiée. Ce calcul sera alors pondéré.

1.1.3 Quelques remarques

Lors des discussions, la question de la validité en pratique de l'hypothèse d'un plan de sondage poissonnien et du caractère i.i.d. des observations s'est posée. Les plans de sondage pour les enquêtes menées à l'Insee ne sont en effet en général pas compatibles avec l'hypothèse de sondage poissonnien. Il s'agit pourtant du cadre général sur lequel repose la présentation. L'impact sur les estimations se pose donc. Cette hypothèse est classique en économétrie mais peut être relâchée, sans que cela ait une incidence sur la convergence des estimateurs. Il serait possible de tenir compte de la corrélation entre D_i et D_j pour $i \neq j$. Ceci a par contre un impact sur le calcul de la précision des estimateurs, qui doit en tenir compte.

On peut vouloir tenir compte du plan de sondage pour estimer correctement la précision des estimateurs dans un plan de sondage complexe qui aboutit *in fine* à des probabilités d'inclusion égales (on a alors $\tilde{X} = \emptyset$ et l'estimateur pondéré et non pondéré sont identiques).

1.2 Tester le choix de l'estimateur, pondéré ou non

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, il est alors possible 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, non abordés ici), 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.

2 What are we weighting for?

L'article de Solon, Haider et Wooldridge (2014) présente trois principales motivations à pondérer un modèle économétrique. Il vise 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 (cf. session 1). 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 pondéré et non pondéré 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.

Lors des discussions, l'utilisation de poids normalisés (i.e. dont la somme correspond à la taille de l'échantillon) a été abordée, 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 c'est moins faux que de ne pas pondérer. Idéalement, il faudrait utiliser les procédures adaptées (SURVEYREG et SURVEYLOGISTIC sous SAS par exemple).

3 Application pratique

Nous reprenons dans cette partie l'exemple sur les données de l'enquête patrimoine présenté lors de la session 1.

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 mais H0' ($Y \perp D|X$) l'est. L'estimateur non pondéré serait alors efficace. Nous mettons donc en œuvre un test d'Hausman (sous Stata).

```
*Modele lineaire, test d'Hausman
*Estimateurs consistant sous H0 et H1 : Regression ponderee
. quietly reg logpatribrut agepr agepr2 classifpr_02-classifpr_10
if classifpr!="" [aweight=pond]
. estimates store est_H0_H1
*Estimateurs efficaces uniquement sous H0 : Regression non ponderee
. quietly reg logpatribrut agepr agepr2 classifpr_02-classifpr_10
if classifpr!=""
. estimates store est_H0
*Test
. hausman est_H0_H1 est_H0, constant

Test: Ho: difference in coefficients not systematic
                chi2(11) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                                 =      258.85
            Prob>chi2 =      0.0000
(V_b-V_B is not positive definite)
```

Le test rejette fortement l'hypothèse nulle (mais les conditions techniques du test ne semblent pas respectées), ce qui conforterait le choix du modèle pondéré. Les résultats doivent néanmoins amener à questionner le respect de l'hypothèse H0 ou H0'. Les comportements de réponse pourraient en effet dépendre 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. Les différences constatées pour les estimateurs, pondérés et non pondérés, peuvent être importantes.

De la même manière que précédemment, on met en œuvre un test d'Hausman. Le test rejette fortement l'hypothèse nulle. La validité de l'hypothèse H0 paraît ici réaliste. Le modèle non pondéré serait donc choisi.

```
*Modele logistique, test d'Hausman
*Estimateurs consistant sous H0 et H1 : Regression ponderee
. quietly xi: glm proprietaire i.cat_agepr i.classifpr i.tu [aweight=pond],
link(logit)
. estimates store est_H0_H1
*Estimateurs efficaces uniquement sous H0 : Regression non ponderee
. quietly xi: glm proprietaire i.cat_agepr i.classifpr i.tu, link(logit)
. estimates store est_H0
*Test
. hausman est_H0_H1 est_H0, constant

Test: Ho: difference in coefficients not systematic
                chi2(22) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                                 =      327.52
            Prob>chi2 =      0.0000
(V_b-V_B is not positive definite)
```

Ces applications ont été effectuées sous Stata. Il est à noter que pour le modèle convergent sous l'hypothèse nulle et alternative, seuls les poids économétriques (aweight) peuvent être utilisés. Il n'est pas possible d'utiliser les poids de sondage (pweight) ou de corriger la matrice de variance-covariance à l'aide de clusters. Ce sont donc des limites très restrictives du test. Les estimateurs de variance utilisés dans le test ne sont pas convergents et peuvent amener à rejeter trop facilement l'hypothèse nulle. Il n'est pas possible de pondérer un modèle logit sous Stata à l'aide de poids économétriques. On estime donc un modèle GLM avec un lien Logit.

Les codes SAS sont fournis en partie dans le document de travail de Xavier D'Haultfœuille et Laurent Davezies.