Faut-il pondérer ?

...Ou l'éternelle question de l'économètre confronté à des données d'enquête

Laurent Davezies et Xavier D'Haultfœuille

CREST

JMS 2012

Plan de la présentation

Le ... entre deux chaises

Considérons un chargé d'études souhaitant estimer un modèle à partir de données d'enquête... Quand le doute l'assaille : faut-il utiliser des poids ?

- Il revient à ses cours de sondage. Là le message est clair :
  - Les quantités mesurées sont supposées fixes et non aléatoires ;
  - Dès lors les poids sont indispensables pour obtenir des estimateurs sans biais (ou presque) du paramètre d'intérêt défini sur la population totale.
- Puis il relit son manuel d'économétrie préféré. Là aussi, c'est (à peu près) clair :
  - Les quantités mesurées sont supposées aléatoires, et l'échantillonnage i.i.d. ;
  - Dès lors la question des poids est totalement ignorée, le sous-entendu étant qu'il n'est pas nécessaire de les utiliser.
- Que croire ? ?

L'objet de ce papier

- Réconcilier les deux approches en utilisant un modèle de superpopulation et en modélisant le sondage comme un problème de sélection ;
- Montrer qu'il est souvent préférable de pondérer même lorsqu'on fait des modèles ;
- Insister sur l'importance pour le chargé d'études de connaître les variables jouant sur la probabilité de tirage et la non-réponse ;
- Développer un test simultané du processus de sélection et du modèle considéré ;
- Évoquer l'inférence en présence de poids.

Comment réconcilier les deux approches ?

- On suppose que la population totale est un échantillon de taille N issue d'un modèle statistique. Par exemple, $(y_1, ..., y_N)$ est la réalisation des v.a. $(Y_1, ..., Y_N)$.
- On observe non pas cet "échantillon" mais seulement l'échantillon final des répondants de l'enquête. On note D_i l'indicatrice de réponse, si bien que $D_i = S_i \times R_i$ où $S_i = \mathbb{1}\{i \text{ appartient à l'échantillon initial}\}$ et $R_i = \mathbb{1}\{i \text{ répond à l'enquête}\}$.
- On note $\widetilde{X}$ les variables jouant sur D et $W_i = 1/P(D_i = 1|\widetilde{X}_i)$ le poids de i. $\widetilde{X}$ inclut les variables utilisées pour fixer les poids de sondages initiaux et celles utilisées pour redresser de la non-réponse.

Comment réconcilier les deux approches ?

- 1ère hypothèse importante : On suppose $(D_1, ..., D_N)$ i.i.d :
  - Cela revient à considérer le tirage comme poissonnien. Ok pour la 2ème phase de non-réponse, pas pour la 1ère (tirages de taille fixe), avec des unités primaires... Mais c'est une approximation courante (cf. Deville) pour calculer des écarts-types.
  - Il est souvent possible d'inclure la présence de grappes dans le calcul des écarts-types (cf. l'option cluster sous Stata).
  - Cette hypothèse n'est pas incompatible avec des probabilités de tirage/de réponse inégales.
- 2ème hypothèse importante : il est possible d'avoir un estimateur convergent des poids. Cela suppose que l'on connaisse le "bon" modèle de non-réponse.

Plan de la présentation

Plan de la présentation

Notations et hypothèses retenues

- On note Y la variable expliquée, X les variables explicatives. A priori $X \neq \widetilde{X}$. On note F_Y (resp. $F_{X,Y}$, $F_{Y|X}$) la fonction de répartition de Y (resp. de (X, Y), de Y sachant X). Le paramètre θ est fonction de $F_{X,Y}$.
- On étudiera en particulier la situation où l'hypothèse suivante est vérifiée :

$$H0.\ (Y, X) \perp\!\!\!\perp D|\widetilde{X}.$$

H0 stipule que $\widetilde{X}$ capte correctement les facteurs de sélection. Par exemple si $\widetilde{X}$ 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 PR fixée.

Premier résultat sous H0.

- Supposons par ailleurs H1 = (H11, H12), avec

$$H11. \quad \theta \text{ dépend seulement de } F_{Y|X}$$
$$H12. \quad \widetilde{X} \subset X.$$

- H11 est une hypothèse sur le modèle statistique. H12 peut directement se vérifier, en consultant la liste des variables utilisées pour le tirage et le calage/modèle de non-réponse !
- **Si H0 et H1 sont vérifiées, alors on peut pondérer, mais l'estimateur non pondéré est plus précis.**
- Intuition : dans ce cadre $Y \perp\!\!\!\perp D|X$. La sélection est "ignorable", et l'estimateur non pondéré évite le problème de dispersion des poids.

Premier résultat sous H0.

- H11 est vérifiée si le modèle économétrique considéré est vrai :
  - Exemple 1 : modèle linéaire $Y = X'\theta + \varepsilon$ avec $E(\varepsilon|X) = 0$. Alors $\theta = \partial E(Y|X = x)/\partial x$.
  - Exemple 2 : modèles binaires types logit/probit $Y = \mathbb{1}\{X'\theta + \varepsilon \geq 0\}$ avec $\varepsilon \perp\!\!\!\perp X$ de loi F connue. Alors $\theta = \partial F^{-1}(E(Y|X = x))/\partial x$.
- H12 par contre est restrictive et n'est a priori pas vérifiée en pratique.

  Exemple : équations de salaire. Ces équations incluent rarement la tranche d'unité urbaine, qui est pourtant souvent utilisée dans $\widetilde{X}$.

Deuxième résultat sous H0.

- **Si H0 est vérifiée mais pas H1, il faut pondérer** : les estimateurs pondérés seront convergent, contrairement en général aux estimateurs non pondérés.
- H11 n'est pas satisfaite dans les cas suivants :
  - Statistiques simples : par exemple $E(Y)$. L'espérance dépend de $F_{Y|X}$ mais aussi de F_X car $E(Y) = E(E(Y|X))$.
  - Exemple 1 (suite) : Si l'on suppose seulement $E(X'\varepsilon) = 0$ plutôt que $E(\varepsilon|X) = 0$, alors θ dépend de $F_{Y,X}$ et pas seulement de $F_{Y|X}$.

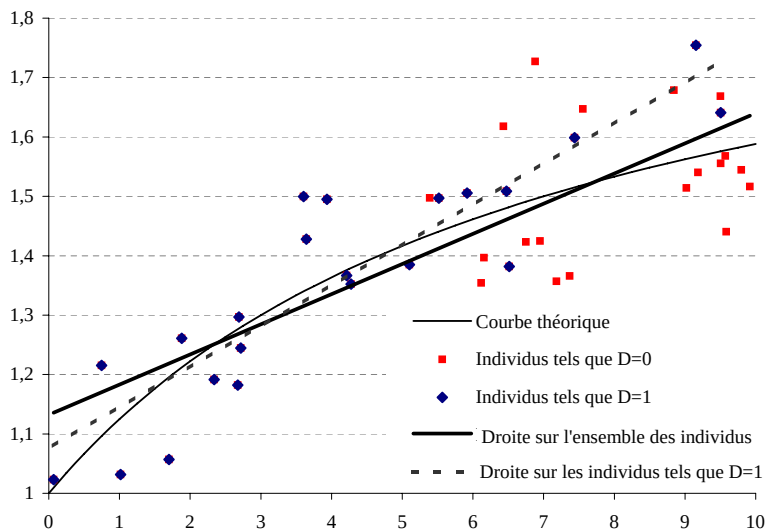

FIGURE: estimation des rendements de l'éducation.

Deuxième résultat sous H0.

- H11 n'est pas satisfaite dans les cas suivants (suite) :
  - Exemple 2 (suite) : plutôt que θ, on s'intéresse souvent dans ces modèles à l'effet marginal moyen

$$\widetilde{\theta}_k = E\left[\frac{\partial E(Y|X)}{\partial x_k}\right].$$

    Cet effet dépend de $F_{Y|X}$ mais aussi de F_X car $\widetilde{\theta}_k = \theta_k E[F'(X'\theta)]$.
- Comme vu précédemment, H12 est restrictive.
  Autre contre-exemple : le "choice-based sampling" (cf. Lerman et Manski, 1977), i.e. lorsque Y intervient dans l'échantillonnage. Par exemple si l'on s'intéresse à la probabilité d'être cadre en utilisant une enquête où les cadres sont surreprésentés.

Résultat quand H0 n'est pas vérifiée.

Dans ce cas, D dépend de (X, Y) même conditionnellement à $\widetilde{X}$. On peut identifier deux situations :

1. La sélection dépend "uniquement" de X, si bien que $D \perp\!\!\!\perp Y|X$: sélection ignorable.
2. La sélection est liée à Y même conditionnellement à $(X, \widetilde{X})$. Situation délicate (non-réponse "non ignorable").

Exemple : équations de salaire. La 1ère situation se présente si la non-réponse dépend du niveau d'éducation $(= X)$ mais pas directement de la tranche d'unité urbaine $(= \widetilde{X})$. La 2ème situation correspond à une non-réponse fonction aussi du salaire.

Résultat quand H0 n'est pas vérifiée.

- Si la sélection est ignorable :
  - si H11 est vérifiée, on peut estimer de façon convergente θ **sans pondérer**.
  - sinon on ne pourra en général pas estimer de façon convergente θ.
- Si la sélection est non-ignorable, il faut des hypothèses supplémentaires !
  Par exemple supposer qu'une variable Z joue sur Y mais pas directement sur D : $Z \perp\!\!\!\perp D | Y, \widetilde{X}$. On peut alors faire du calage généralisé avec calmar2 (cf. Deville, 2002, Le Guennec et Sautory, 2005). Il faut pondérer dans ce cas !

Plan de la présentation

L'idée de départ.

- Supposons qu'on estime un modèle avec et sans poids et que les résultats soient très différents. Qu'est-ce que cela veut dire ?
- Si H0, H11 et H12 sont vérifiées, les deux estimateurs sont convergents donc devraient être "proches". S'ils sont "très différents", l'une des hypothèses est fausse.
- Si l'on maintient H12 ($\widetilde{X} \subset X$), ou bien H0 n'est pas vérifiée (i.e., $\widetilde{X}$ ne capte pas tous les facteurs pertinents de la non-réponse), ou bien θ ne dépend pas que de $F_{Y|X}$ (i.e., le modèle est faux : cf. l'exemple du modèle linéaire).

Description du test

- Notons $\widehat{\theta}$ (resp. $\widehat{\theta}_W$) l'estimateur non-pondéré (resp. pondéré), et $\widehat{V}$ (resp. $\widehat{V}_W$) un estimateur de sa variance.
- Si H0 et H12 sont vraie, l'estimateur $\widehat{\theta}$ est asymptotiquement efficace. On peut alors faire un test d'Hausman, en s'appuyant sur la statistique

$$T_H = \left(\widehat{\theta}_W - \widehat{\theta}\right)' \left[\widehat{V}\left(\widehat{\theta}_W\right) - \widehat{V}\left(\widehat{\theta}\right)\right]^{-1} \left(\widehat{\theta}_W - \widehat{\theta}\right).$$

- On rejette l'hypothèse jointe (H0, H12) si $T_H > \chi^2_r(1-\alpha)$, où $\chi^2_r(1-\alpha)$ est le quantile d'ordre $1-\alpha$ d'un χ^2 à r degrés de liberté.
- Si l'on rejette le test et que l'on croit à H0, on rejette alors H12 et dans ce cas il faut obligatoirement pondérer.

Plan de la présentation

Le point de départ

- La plupart des logiciels statistiques (SAS, Stata, R...) laissent la possibilité d'inclure des poids dans leurs procédures.
- Cependant, les écarts-types des estimateurs calculés avec ces poids ne sont pas nécessairement corrects.
- Sous SAS, les écarts-types fournis ne sont corrects que dans les procédures de préfixe "survey" (surveymeans, surveyreg, surveylogistic).
- Sous Stata il y a plusieurs possibilités pour inclure les poids. L'option sweight conduit à des écarts-types corrects.

Une procédure de bootstrap

- En l'absence de calcul préprogrammé correct des écarts-types, on peut utiliser un algorithme de bootstrap.
- Dans notre modélisation initiale, la population totale est un échantillon i.i.d.
- On pourrait donc appliquer un bootstrap "standard" sur cette population, en tirant dans celle-ci avec remise un échantillon de taille N.
- On ne considérerait ensuite que les individus tels que $D_i = 1$ dans l'échantillon bootstrap final.
- En fait, on peut de façon équivalente tirer directement dans l'échantillon des répondants, pourvu que la taille de l'échantillon soit aléatoire.

Une procédure de bootstrap

On obtient alors l'algorithme de bootstrap suivant :

Pour $b = 1$ à B :

1. Tirer $n_b \sim$ Binomiale$(N, n/N)$, où n la taille de l'échantillon des répondants ;
2. Tirer à probabilités égales et avec remise un échantillon de taille n_b issu de l'échantillon initial. On peut utiliser pour cela la commande suivante sous SAS (ici on échantillonne dans a et $n_b = 2500$) :
```
proc surveyselect data=a method=urs sampsize=2500 out=boot;
run;
```
On note U_i^b le nombre de fois où l'individu i a été tiré dans l'échantillon bootstrap.
3. Estimer les poids $W_i^b = 1/P(D_i = 1|\widetilde{X}_i)$, par un modèle de non-réponse et/ou un calage identique à celui effectué sur l'échantillon initial mais en utilisant les pondérations U_i^b (ou en construisant une table ayant U_i^b observations pour l'individu i de la table initiale) ;
4. Estimer le paramètre θ avec les poids $W_i^b U_i^b$. On note $\widehat{\theta}_b$ l'estimateur obtenu.

Fin.

On peut ensuite estimer (par exemple) la variance de $\widehat{\theta}$ par

$$\widehat{V} = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta} - \widehat{\theta}_b\right)^2.$$

Plan de la présentation

Conclusion

- Si les facteurs de sélection ont été correctement pris en compte, il est plus prudent de pondérer car les estimateurs correspondant seront toujours convergents (même si parfois ils sont moins précis).
- Les estimateurs non-pondérés seront plus précis si le modèle est vrai et $\widetilde{X} \subset X$. Cette dernière condition souligne l'importance de se renseigner sur la liste des variables utilisées dans le calage / le modèle de non-réponse.
- Il faut être prudent dans le calcul des estimateurs pondérés.

Aparté

- On a supposé pouvoir estimer de manière convergente les poids

$$W_i = P(D_i = 1|\widetilde{X}_i) = \pi_i \times P(R_i = 1|\widetilde{X}_i),$$

où $\pi_i = P(i \in S)$ est le poids de tirage.

- Pour obtenir W_i une pratique courante est :
  - de faire un modèle de non-réponse en utilisant des variables $\widetilde{X}_{1i}$ disponibles sur les répondants et non-répondants ;
  - de faire un calage sur des variables $\widetilde{X}_{2i}$ disponibles sur les répondants seuls mais dont les totaux sont connus.
- Cependant cette pratique ne permet pas en général de corriger correctement la non-réponse si celle-ci dépend à la fois de $\widetilde{X}_{1i}$ et de $\widetilde{X}_{2i}$.

Aparté

- Supposons que le vrai modèle soit logistique :

$$P(D=1|\widetilde{X}) = \Lambda(\widetilde{X}'\beta), \quad \text{avec } \Lambda(x) = \frac{1}{1+e^{-x}}.$$

- Il est alors possible d'estimer de manière convergente β à l'aide d'un calage "modifié". En effet :

$$\begin{aligned} E\left[\frac{D\widetilde{X}}{\Lambda(\widetilde{X}'\beta)}\right] &= E\left[D\widetilde{X}\left(1+e^{-\widetilde{X}'\beta}\right)\right] = E(\widetilde{X}) \\ \Longleftrightarrow E\left[D\widetilde{X}e^{-\widetilde{X}'\beta}\right] &= E(\widetilde{X}) - E(D\widetilde{X}) \end{aligned}$$

- On retrouve donc une équation de calage classique par le raking ratio *mais* on cale sur des marges différentes !
  - Pour les variables $\widetilde{X}_{1i}$ disponibles sur les non-répondants on cale sur *la moyenne des non-répondants* ;
  - Pour les variables $\widetilde{X}_{2i}$ on cale sur *la moyenne auxiliaire* $E(\widetilde{X})$ *à laquelle on soustrait la moyenne non-pondérée des répondants*.
- N.B. : il faut enfin ajouter 1 aux poids obtenus ainsi par calmar !