Faut-il pondérer ?

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

Laurent Davezies et Xavier D'Haultfœuille

CREST

Groupe de lecture

Plan de la présentation

Le calcul 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 cette présentation

- 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.

Pourquoi deux approches ?

En simplifiant à l'extrême, la théorie des sondages se concentre sur la mesure de certaines grandeurs dans la population alors que seul un échantillon aléatoire est observé.

- Le sondeur reste souvent agnostique sur les relations entre variables étudiées.
- Les paramètres d'intérêt (totaux, moyenne, mesure de liaisons entre variables, distribution...) sont donnés. Par exemple, l'INSEE doit produire une description des revenus salariaux.
- L'aléa est « instrumental » (i.e. mis en œuvre par le sondeur).
- La notion d'écart-type est basée sur la dispersion des estimations obtenues en répétant les tirages sous le même plan de sondage. Donc pas d'écart-type si recensement.
- L'inférence est difficile à fonder dans ce cadre (à distance finie, pourquoi $\pm 1.96\sigma$? Quelle asymptotique ?). L'extrapolation de la mesure sur une autre population n'appartient pas à ce registre.

Pourquoi deux approches ?

En simplifiant à l'extrême : L'économétrie (structurelle) se pose la question de l'articulation entre théorie économique et évaluation de paramètres intervenant dans cette théorie.

- Ex : les revenus salariaux dépendent (entre autres) de la productivité, qui dépend (entre autres) des études suivies par les individus. Rendement des diplômes : quelle est la nature (accumulation de capital humain vs signaling) et l'intensité de cette dépendance ?
- Le rendement des diplômes ne peut se réduire uniquement à une question de mesure. Il faut une théorie car les individus choisissent leurs études.
- Par ailleurs, il faut une modélisation de l'hétérogénéité car deux individus « observationnellement semblables » ne font pas les mêmes choix d'étude et n'ont pas les mêmes salaires.
- ...

Pourquoi deux approches ?

En simplifiant à l'extrême : L'économétrie (structurelle) se pose la question de l'articulation entre théorie économique et évaluation de paramètres intervenant dans cette théorie.

- ...
- Le meilleur moyen qu'on a trouvé pour prendre en compte cette hétérogénéité, c'est d'introduire la notion de variables aléatoires.
- L'aléa est épistémique, i.e. vient palier un défaut de connaissance. Introduction d'hypothèses (fortes ?) sur la structure des observations.
- Dans ce cadre le biais se défini comme l'écart moyen au sens de l'aléa épistémique entre le paramètre défini *in abstracto* par la théorie et l'estimateur. Idem pour l'écart-type. L'inférence est fondée sur une théorie asymptotique.

Cette séparation est-elle si claire ?

- Non, car la non-réponse en sondage nécessite une modélisation, car extrapolation sur les non-répondants. Dans les deux cas : contrôle de la sélection, dans l'échantillon ou dans le « traitement ».
- En pratique : les estimations obtenues peuvent être sensibles au choix de pondérer ou non.
- Pour « faire des choses qui se tiennent »(A. Desrosières), les chargés d'étude, les chercheurs en sciences sociales ou les méthodologues doivent mobiliser des arguments appartenant aux deux registres.

Cette séparation est-elle si claire ?

Pendant longtemps :

- tensions entre (certains) économètres soutenant qu'il est inutile voire néfaste de pondérer les régressions et des sondeurs/concepteurs d'enquêtes soutenant l'inverse.
- Argument couramment entendu à l'Ensae et au Crest dans les années 2000 : « Quand on modélise des comportements, cela n'a rien à voir avec le plan de sondage donc on ne pondère pas. » (Justification ? ?)

Depuis le milieu des années 90, renouvellement de l'économétrie :

- Modèle causal de Rubin : reformulation de la question de l'endogénéité et de l'hétérogénéité en économétrie.
- Développement des expériences randomisées, les économètres appliqués se sont mis à produire des données sur échantillons.
- Evolution qui a rendu plus claire la question de l'usage des pondérations pour les économètres appliqués, qui sont aujourd'hui nettement moins réticents à l'usage des pondérations.

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 $\widetilde{X}_i$ les variables jouant sur S_i et R_i (donc sur D_i) et $W_i = 1/\mathbb{P}(D_i = 1|\widetilde{X}_i)$ le poids de i. $\widetilde{X}_i$ inclut les variables utilisées pour fixer les poids de sondages initiaux et celles utilisées pour redresser de la non-réponse.
- On note Y_i la variable expliquée, X_i les variables explicatives (avec a priori $X_i \neq \widetilde{X}_i$).
- N.B. : comme on supposera les variables i.i.d., on omettra l'indice i par la suite.

Comment réconcilier les deux approches ?

- 1ère hypothèse : $((D_1, \widetilde{X}_1, X_1, Y_1), ..., (D_N, \widetilde{X}_N, X_N, Y_N))$ i.i.d.
  - 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 : on dispose d'un estimateur des poids $\widehat{W}$ convergent. Cela suppose que l'on connaisse le vrai modèle de non-réponse. ▸ détails
- 3ème hypothèse : condition de support $\mathbb{P}(D = 1|\widetilde{X}) > 0$. Sinon certains paramètres ne sont pas identifiables. En particulier, la probabilité de sélection vérifie $\mathbb{P}(S = 1|\widetilde{X}) > 0$.

Plan de la présentation

Plan de la présentation

Plan de la présentation

Une hypothèse cruciale

- On considérera principalement les situations 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.

- Comme $\widetilde{X}$ contient les facteurs de sélection dans l'échantillon initial (=variables qui interviennent dans le plan de sondage), on a :

$$(R, Y, X) \perp\!\!\!\perp S|\widetilde{X}$$

- Donc H0 signifie que la non-réponse est indépendante de (Y, X), à $\widetilde{X}$ fixé :

$$H0 :\ (Y, X) \perp\!\!\!\perp D|\widetilde{X} \Longleftrightarrow (Y, X) \perp\!\!\!\perp R|\widetilde{X}$$

- 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 personne de référence fixée.

Une hypothèse supplémentaire

- On considérera par ailleurs H1 = (H11, H12), avec

  $H11.$ θ dépend seulement de $F_{Y|X}$ (fonction de répartition de $Y|X$)

  $H12.$ $\widetilde{X} \subset X$.

- H11 est une hypothèse sur le modèle statistique. En pratique, θ est défini par des moments conditionnels :

$$\mathbb{E}(m(Y, X, \theta)|X) = 0.$$

  - Modèle linéaire : $m(Y, X, \theta) = Y - X\theta$
  - Maximum de vraisemblance : $m(Y, X, \theta) = \frac{\partial \ln f(Y|X,\theta)}{\partial \theta}$
  - Régression non-paramétrique : $m(Y, X, \theta) = Y - \theta(X)$

- H12 peut directement se vérifier, en consultant la liste des variables utilisées pour le tirage et le calage/modèle de non-réponse !
- On étudie dans la suite les situations où $H1$ est vérifiée comme celles où elle ne l'est pas.

Questions posées ou non ici

- On s'intéresse ici à la question de la convergence des estimateurs pondérés et non pondérés.
- Dans ce cadre, il s'agit de savoir si les conditions de moment théoriques sous-jacentes à l'estimation sont bien vérifiées.
- On ignore dans cette section les questions d'inférence :
  - Comme $\widehat{W}$ est supposé convergent, on peut supposer, pour étudier la convergence des estimateurs de θ, que $\widehat{W} = W$;
  - On ne s'intéresse pas pour le moment à l'estimation de la variance asymptotique des estimateurs de θ.

Plan de la présentation

Résultat principal.

- **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.
- C'est la formalisation de l'argument habituel des économètres.
- Mais on verra dans la suite que H1 est souvent violé...

Un petit détour par les GMM

- Oublions pour le moment la sélection, i.e. qu'on observe (Y, X) seulement quand $D = 1$.
- H11 est vérifiée si le modèle économétrique considéré est vrai.
- Si θ est un vecteur défini par des conditions de moments conditionnels, on a en général plus d'équations que d'inconnues (car le nb. de valeurs de $X >$ taille de θ).
- θ doit annuler $\mathbb{E}(g(X)m(Y, X, \theta))$ pour tout g. Problème : quel g choisir pour prendre les contreparties empiriques ?
- Le choix optimal, i.e. permettant d'atteindre la borne d'efficacité, vérifie :

$$g^\star(X) = \mathbb{E}\left(\frac{\partial m}{\partial \theta'}(Y, X, \theta)|X\right) \mathbb{V}(m(Y, X, \theta)|X)^{-1}.$$

On se ramène au cas où on a autant d'équations que d'inconnues.

Un petit détour par les GMM

- Exemple 1 : modèle linéaire $Y = X\theta + \varepsilon$ avec $\mathbb{E}(\varepsilon|X) = 0$. Alors $\theta = \partial\mathbb{E}(Y|X = x)/\partial x$, $m(Y, X, \theta) = Y - X\theta$ et

$$\frac{\partial m}{\partial \theta'}(Y, X, \theta) = -X'.$$

  Sous l'homoscédasticité, $g^\star(X) = -X/\sigma^2$ et on obtient la version théorique des CPO des MCO : $\mathbb{E}(X'(Y - X\theta)) = 0$.

- Exemple 2 : Maximum de vraisemblance $Y|X \sim f(Y|X, \theta)$. $m(Y, X, \theta) = \partial \ln f(Y|X, \theta)/\partial\theta$ est le score du modèle et

$$g^\star(X) = \mathbb{E}\left(\frac{\partial^2 \ln f(Y|X, \theta)}{\partial\theta\partial\theta'}\right)\mathbb{V}(m(Y, X, \theta)|X)^{-1} = -Id.$$

  Dans ce cas on utilise simplement $\mathbb{E}(m(Y, X, \theta)) = 0$.

Retour à notre cadre

- Ces conditions de moments sont a priori inutiles ici car les seuls moments théoriques que l'on va pouvoir exploiter sont de la forme $\mathbb{E}(f(Y,X)|D=1)$.
- Mais sous les hypothèses H0 et H1, $Y \perp\!\!\!\perp D|X$ donc

$$\mathbb{E}(m(Y,X,\theta)|D=1,X)=0.$$

- On peut donc appliquer la théorie précédente telle quelle. L'estimateur efficace de θ est basé sur la contrepartie empirique de

$$\mathbb{E}\left[g^{\star}(X)m(Y,X,\theta)|D=1\right]=0.$$

- L'estimateur non pondéré est donc convergent et asymptotiquement efficace.

Retour à notre cadre

- L'estimateur pondéré est quant à lui basé sur la contrepartie empirique de

$$\mathbb{E}\left[Wg^{\star}(X)m(Y,X,\theta)|D=1\right]=0. \tag{1}$$

- On utilise donc une autre fonction $g(X)=Wg^{\star}(X)$, différente de la fonction optimale $g^{\star}(X)$. L'estimateur n'est donc pas efficace...
- ... Mais il est convergent ! Car (1) est vérifiée :

$$\mathbb{E}\left[Wg^{\star}(X)m(Y,X,\theta)|D=1,X\right]=Wg^{\star}(X)\mathbb{E}\left[m(Y,X,\theta)|X\right]=0,$$

et on réintègre ensuite sur X (sachant $D=1$).

Plan de la présentation

Résultat principal

- **Si H0 est vérifiée mais pas H1, il faut pondérer** : les estimateurs pondérés seront convergent, contrairement en général aux estimateurs non pondérés.
- Intuitions de la non-convergence des estimateurs non-pondérés :

  - Si H11 n'est pas vérifiée, θ dépend de F_X et $F_X \neq F_{X|D=1}$. Ignorer la sélection revient à surpondérer certains individus, ce qui biaise en général l'estimation de θ.

  - Si H12 n'est pas vérifiée, ne pas pondérer revient à supposer que $F_{Y|D=1,X} = F_{Y|X}$ mais cette égalité ne tient plus si $\widetilde{X}$ et Y sont corrélées, à X fixé.

Exemples où H11 n'est pas satisfaite

- Statistiques simples : par exemple $\mathbb{E}(Y)$. L'espérance dépend de $F_{Y|X}$ mais aussi de F_X car $\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X))$.
- Considérons un modèle binaire paramétrique (logit, probit...) :

$$Y = \mathbb{1}\{X\beta + \varepsilon \geq 0\},\ -\varepsilon \perp\!\!\!\perp X \text{ de fdr } F \text{ supposée connue.}$$

- β vérifie bien H11 ici. Mais on est souvent plus intéressé par l'effet marginal moyen de de X_k que par β_k lui-même :

$$\theta = \beta_k \mathbb{E}\left[F'(X\beta)\right]$$

  Cet effet dépend de $F_{Y|X}$ à travers β, mais aussi de F_X.
- Dans ce cas, on peut estimer β sans poids si H12 est vérifiée, mais θ doit être estimée avec des poids !

Exemples où H11 n'est pas satisfaite

- Effet moyen du traitement : $T(= X_1)$ traitement binaire, $Y(0), Y(1)$ outcome potentiels, $Y = Y(0) + T(Y(1) - Y(0))$ outcome observé et l'on suppose

$$T \perp\!\!\!\perp (Y(0), Y(1)) | X_2$$

  avec $X = (X_1, X_2)$.

- Dans ce cas on a :

$$\theta = \mathbb{E}\left[Y(1) - Y(0)\right] = \mathbb{E}\left[\mathbb{E}(Y|T = 1, X_2) - \mathbb{E}(Y|T = 0, X_2)\right].$$

- Si $\mathbb{E}(Y|T, X_2) = \mathbb{E}(Y|X)$ ne dépend que de $F_{Y|X}$ et peut s'estimer non-paramétriquement sans pondération, θ dépend de F_X et doit donc être estimé avec les poids !

Conséquence sur les estimateurs pondérés et non pondérés

- Dans tous les exemples cités, θ est définie par une condition de moments non conditionnelle :

$$\mathbb{E}(m(Y, X, \theta)) = 0,$$

où m est une fonction connue ou estimée des observations et du paramètre.

- Or sous H0 et H12 seuls, on a généralement :

$$\mathbb{E}(m(Y, X, \theta)|D = 1) \neq \mathbb{E}(m(Y, X, \theta)) = 0.$$

Donc l'estimateur non pondéré n'est pas convergent en général.

Conséquence sur les estimateurs pondérés et non pondérés

- Par contre l'estimateur pondéré est convergent :

$$\begin{aligned}
\mathbb{E}(Wm(Y,X,\theta)|D=1) &= \frac{\mathbb{E}[WDm(Y,X,\theta)]}{\mathbb{P}(D=1)} \\
&= \frac{\mathbb{E}[Wm(Y,X,\theta)\mathbb{E}(D|X,Y,\widetilde{X})]}{\mathbb{P}(D=1)} \\
&= \frac{\mathbb{E}[Wm(Y,X,\theta)\mathbb{E}(D|\widetilde{X})]}{\mathbb{P}(D=1)} \\
&= \frac{\mathbb{E}[Wm(Y,X,\theta)/W]}{\mathbb{P}(D=1)} \\
&= 0.
\end{aligned}$$

Exemples où H12 n'est pas satisfaite

- Il n'y a pas de raison que les variables jouant sur le tirage ou la non-réponse soient des variables explicatives pertinentes du modèle théorique.
- Exemple : équations de salaire : Ces équations n'incluent en général ni la tranche d'unité urbaine, ni la taille du ménage, qui sont pourtant souvent utilisées dans $\widetilde{X}$. Une des raisons : ces variables sont a priori endogènes !
- 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.

Exemples où H12 n'est pas satisfaite

- Variables instrumentales :

$$Y_1 = Y_2\theta_1 + X_1\theta_2 + \varepsilon,\ \mathbb{E}(\varepsilon|X) = 0$$

avec $\theta = (\theta_1, \theta_2)$, $Y = (Y_1, Y_2)$ et $X = (X_1, X_2)$.

- Sous la condition de rang habituelle, θ est alors défini par les moments :

$$\mathbb{E}\left[Y_1 - Y_2\theta_1 - X_1\theta_2|X\right] = 0.$$

Donc H11 est vérifiée, mais H12 ne l'est pas si Y_2 et $\widetilde{X}$ ont des composantes communes.

- Exemple : effet de la fertilité (Y_2 =nombre d'enfants) sur la participation au marché du travail (cf. par ex. Angrist et Evans, 1998). Y_2, a priori endogène, peut aussi être un élément de $\widetilde{X}$.

Pourquoi faut-il pondérer si H12 (mais pas H11) est violée ?

- En pondérant, on annule les contreparties empiriques de $\mathbb{E}(Wg^\star m|D=1)$. Or :

  $$\mathbb{E}(Wg^\star m|D=1) = \mathbb{E}(WDg^\star m)/\mathbb{P}(D=1) = \mathbb{E}(g^\star m)/\mathbb{P}(D=1) = 0.$$

  L'estimateur est donc convergent.
- En ne pondérant pas, on annule les contreparties empiriques de $\mathbb{E}(g^\star m|D=1)$. Or

  $$\mathbb{E}(g^\star m|D=1) = \mathbb{E}(g^\star m/W)/\mathbb{P}(D=1),$$

  qui n'a pas de raison d'être nul ! Estimateur **non convergent**.

Pourquoi faut-il pondérer si H12 (mais pas H11) est violée ?

- Notons que l'estimateur pondéré n'est pas efficace en général.
- L'estimateur efficace est celui qui annule les contreparties empiriques de $\mathbb{E}(Wg^{\star\star}m|D=1)$, avec $g^{\star\star}$ la fonction optimale correspondant aux moments conditionnels $\mathbb{E}(Wm|D=1,X)=0$.
- Cet estimateur n'a pas de forme simple en général. Par exemple, pour le modèle linéaire :

$$g^{\star\star}(X) = -X'\mathbb{E}(W|X)/\mathbb{E}(W^2|X).$$

  Il faut donc estimer en première étape $\mathbb{E}(W|X)/\mathbb{E}(W^2|X)$, construire de nouveaux poids $W\widehat{\mathbb{E}}(W|X)/\widehat{\mathbb{E}}(W^2|X)$, puis faire les MCO avec ces poids.
- N.B. : quand H12 est vérifiée, on a $Wg^{\star\star} = g^{\star}$: la non-pondération peut alors se voir comme le choix optimal de moments conditionnels pondérés, où les poids disparaissent comme par magie !

Plan de la présentation

H0 n'est plus 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 dite 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 ou de la taille du ménage $(= \widetilde{X})$. 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.
  - Mais c'est une situation peu crédible en général.
- Si la sélection est non-ignorable, il faut des hypothèses supplémentaires :
  - Approche classique (en économétrie) : supposer qu'on a un instrument Z jouant sur D mais pas sur Y. On utilise alors par exemple un modèle de sélection généralisée.
  - Calage généralisé (cf. Deville, 2002, Le Guennec et Sautory, 2005) : 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`. 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.

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 H1 sont vraies, 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 $T_H > \chi^2_{\dim(\theta)}(1-\alpha)$, où $\chi^2_{\dim(\theta)}(1-\alpha)$ est le quantile d'ordre $1-\alpha$ d'un χ^2 à $\dim(\theta)$ degrés de liberté.
- Si l'on rejette le test, que l'on maintient H12 et que l'on croit à H0, alors on rejette H11. 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 $W_i^b = 1/\mathbb{P}(D_i = 1|\widetilde{X}_i)$, 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 l'on s'intéresse à un paramètre d'un modèle conditionnel et si $\widetilde{X} \subset X$.
- Cette dernière condition est cependant rarement vérifiée en pratique. Elle souligne aussi 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.

Remarques sur l'estimation des poids

- Si l'on suppose, comme on le fait par la suite, que $R \perp\!\!\!\perp S|\widetilde{X}$, alors

$$1/W = \mathbb{P}(R=1, S=1|\widetilde{X}) = \mathbb{P}(S=1|\widetilde{X})\mathbb{P}(R=1|\widetilde{X}),$$

  où $\mathbb{P}(S=1|\widetilde{X})$ est la probabilité de tirage qui est connue.
- Estimer W revient donc à estimer $\mathbb{P}(R=1|\widetilde{X})$.
- Une pratique courante est :
  - de faire un modèle de non-réponse en utilisant des variables $\widetilde{X}_1$ disponibles sur les répondants et non-répondants ;
  - de faire un calage sur des variables $\widetilde{X}_2$ disponibles sur les répondants seuls mais dont les totaux sont connus.
- Cette pratique ne permet pas en général de corriger correctement la non-réponse si celle-ci dépend à la fois de $\widetilde{X}_1$ et de $\widetilde{X}_2$.

Remarques sur l'estimation des poids

- Supposons que le vrai modèle soit logistique :

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

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

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

- On retrouve donc une équation de calage classique par le raking ratio *mais* on cale sur des marges différentes !
  - Pour les variables $\widetilde{X}_1$ disponibles sur les non-répondants on cale sur *la moyenne des non-répondants* ;
  - Pour les variables $\widetilde{X}_2$ on cale sur *la moyenne auxiliaire* $\mathbb{E}(\widetilde{X})$ *à laquelle on soustrait la moyenne non-pondérée des répondants*.
- N.B. : il faut enfin ajouter 1 aux poids obtenus ainsi par calmar !

Retour