Groupe de lecture "Econométrie des données d'enquête"

Compte-rendu de la 4^{e}réunion, 9 février 2015
La (non) prise en compte du plan de sondage sur la variance

Suivi par Marine Guillerm et Ronan Le Saout

Cette quatrième séance du groupe de lecture "Sondage et économétrie" avait pour thème: la (non) prise en compte du plan de sondage sur la variance. Elle s'est articulée autour de l'article de Graubard et Korn (2002) "Inference for superpopulation parameters using sample surveys" proposé et présenté par Guillaume Chauvet et Cyril Favre-Martinoz (CREST-Ensai).

1 Présentation de l'article de Graubard et Korn (2002)

On s'intéresse ici à l'estimation de la précision d'un estimateur d'un paramètre de superpopulation. Lorsque les données sont des données d'enquête, il est fréquent de ne pas tenir compte du plan de sondage et de faire comme si les données étaient directement générées selon un modèle de superpopulation. Les auteurs montrent que cela conduit à sous-estimer la variance des estimateurs.

1.1 Modèle sans cluster

On se place dans le cadre d'un sondage aléatoire simple stratifié sans remise. Les données dans la population sont d'abord générées selon un modèle à deux niveaux: une indicatrice de strate est tout d'abord générée, puis la variable d'intérêt est générée conditionnellement à la strate. Dans chaque strate h de taille K_h, on sélectionne k_h individus. On cherche à estimer la précision de l'estimateur sans biais classique $\bar{y} = \sum_{h=1}^{H} \frac{K_h}{K} \bar{y}_h$ de $\bar{Y} = \sum_{h=1}^{L} \frac{K_h}{K} \bar{Y}_h$.
La variance de cet estimateur a trois composantes: (1) une composante associée à l'aléa dû au plan de sondage, (2) une composante liée à l'aléa de modèle sur la variable Y et (3) une composante liée à la variabilité de modèle de la taille des strates qui correspond à la variabilité inter-strates.
L'estimateur habituel d'un sondage stratifié sans remise $\hat{var}_{wo}(\bar{y}) = \sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{K_h - k_h}{K_h} \frac{s_h^2}{k_h}$ estime sans biais la première composante. L'estimateur de la variance $\hat{var}_{wr}(\bar{y}) = \sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{s_h^2}{k_h}$ qui ne tient pas compte du caractère sans remise du tirage estime sans biais les deux premières composantes. Il sous-estime donc la vraie variance et ceci d'autant plus que la fraction de sondage augmente et que la variance inter-strates est grande par rapport à la variance intra. Graubard et Korn propose un estimateur non biaisé de la troisième composante:

$$\frac{\hat{\Delta}_{betw,y}}{K} = \frac{K}{K-1} \sum_{h=1}^{L} \frac{K_h}{K} (\bar{y}_h - \bar{y})^2 - \sum_{h=1}^{L} \frac{K_h(K-K_h)}{K(K-1)} \frac{s_h^2}{k_h}$$

1.2 Modèle avec clusters

Les données dans la population U sont générés selon le modèle suivant:

$$F : \begin{pmatrix} M_{ij} \\ T_{ij} \end{pmatrix} \sim_{iid} \mathcal{L}\left(\begin{pmatrix} \alpha_i \\ \tau_i \end{pmatrix}, \begin{pmatrix} \sigma_{11i} & \sigma_{12i} \\ \sigma_{12i} & \sigma_{22i} \end{pmatrix} \right)$$
$$(\alpha_i, \tau_i, \sigma_{11i}, \sigma_{22i}, \sigma_{12i}, N_i, Z_i, \eta_i) \sim_{iid} F$$

M_{ij} est le nombre d'unités tertiaires dans l'unité secondaire j de l'unité primaire i, T_{ij} le total de la variable d'intérêt dans l'US j de l'unité primaire i.

On cherche à estimer la variance de l'estimateur $\bar{y}$ du paramètre de superpopulation μ. L'estimateur $\bar{y}$ se présente sous la forme d'un ratio, avec au numérateur un estimateur du total et au dénominateur un estimateur de la taille de la population. On utilise donc une technique de linéarisation pour estimer sa variance qu'on approxime par la somme de deux composantes: (1) une composante liée au plan de sondage et (2) une composante liée à la variabilité de Y. On connait des estimateurs sans biais de ces composantes. La première composante est estimé par l'estimateur de variance sans remise, asymptotiquement sans biais. L'estimateur de la deuxième composante nécessite la connaissance des probabilités d'inclusion d'ordre 2 des unités primaires. Pour se libérer de cette contrainte, Graubard et Korn proposent un estimateur en partant de l'estimateur de variance avec remise pour la première composante.

Ces estimateurs de variance sont asymptotiquement sans biais. Une des conditions requises est un taux de sondage faible pour les unités primaires dans chaque strate. Le fait d'avoir une fraction de sondage faible au premier degré ne garantit pas l'obtention d'un estimateur de variance "avec remise" approximativement sans biais. Le biais de l'estimateur de variance "avec remise" peut être élevé si une proportion même petite de grosses unités primaires est échantillonnée, dans une strate avec une fraction de sondage importante.

2 Discussion

1. La démarche des auteurs est originale, l'inférence repose sur le plan et le modèle. En général, on fait l'hypothèse d'un plan de sondage non informatif et on bascule vers l'économétrie. Or le plan peut être informatif, par exemple quand l'allocation dépend des tailles des strates.

2. La question de l'estimation de certains termes par bootstrap se pose. Mais Guillaume Chauvet fait remarquer qu'en général le bootstrap ne fonctionne pas quand on n'a pas de formule analytique de variance. C'est un constat d'échec.

3. Quand le taux de sondage est plus élevé, dès 0,3/0,5, les conditions asymptotiques ne sont pas remplies. On a donc un problème.

4. La procédure surveymeans fait l'hypothèse d'un tirage avec remise, dès le premier degré, avec la possibilité d'inclure une correction de population finie. Sans correction de population finie, dans le modèle sans cluster, on néglige la troisième composante.

5. Les auteurs fournissent des estimateurs non biaisés de la variance, mais sont-ils plus ou moins précis? Peut-être moins précis mais on préfère sur-estimer que sous-estimer.

6. Le succès de l'article? L'article fait l'objet d'un chapitre du livre "Analysis of health survey". Une recherche sur google scholar montre que l'article a été cité 51 fois.

3 Application sur les données de l'enquête Patrimoine

La section Patrimoine de la division "Revenus et patrimoine des ménages" nous a fourni pour chaque ménage enquêté dans l'enquête Patrimoine, la strate de tirage "anonymisée". Nous pouvons ainsi repérer les ménages appartenant à la même strate, sans pouvoir les repérer géographiquement. Une application sera effectuée à partir de ces données.