Groupe de lecture "Econométrie des données d'enquête"

Compte-rendu de la troisième réunion, 19 janvier 2015
La modélisation en théorie des sondages

Suivi par Marine Guillerm et Ronan Le Saout

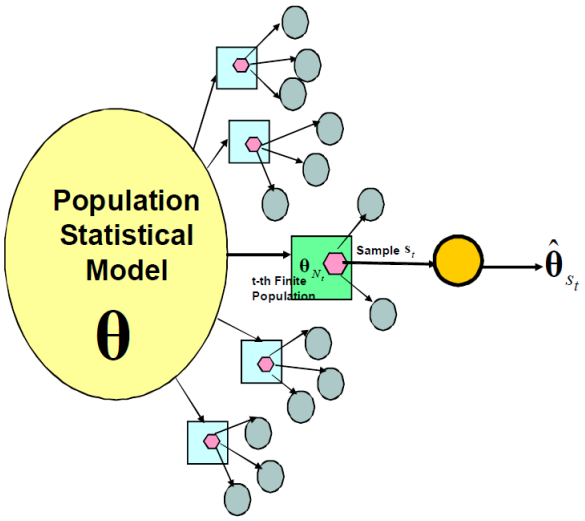
Cette troisième séance du groupe de lecture avait pour thème: La modélisation en théorie des sondages. Deux textes autour de ce thème ont été présentés. Tiaray Razafindranovona a présenté l'article (Binder et Roberts, 2003) "Design-based and model-based methods for estimating model parameters". Marine Guillerm a présenté l'article "To model or not to model? Competing modes of inference for finite population sampling" de Little (2004).

Cette session avait pour objectif d'approfondir les débats sur les différences d'approche entre économétrie et sondages. Les deux premières sessions insistaient sur l'utilisation des poids et de la connaissance du plan de sondage pour améliorer si besoin les modèles économétriques mis en œuvre. L'orientation des deux articles de la session 3 est inversée, et se concentre sur l'apport de la modélisation en théorie des sondages. La liaison avec la mise en œuvre des méthodes économétriques a été abordée pendant la discussion sur les textes.

1 Design-based and model-based methods for estimating model parameters

Deux types d'approche pour l'estimation d'un paramètre avec des données d'enquêtes s'affrontent. Le statisticien-sondeur peut utiliser une approche fondée sur le plan de sondage, l'aléa est alors dans le tirage de l'échantillon, les valeurs observées sur la population étant considérées fixes (approche dite *Design-Based*). Le statisticien-économètre peut utiliser une approche à l'aide d'un modèle stochastique. Dans ce cadre, les données sont supposées être générées de manière i.i.d., les valeurs observées sur la population (et l'échantillon) étant considérées comme la réalisation de variables aléatoires (approche dite *Model-Based*). Peut-on réconcilier ces deux approches, l'une des deux approches domine-t-elle l'autre pour estimer un paramètre? Ces questions font par ailleurs l'objet de controverses et de débats scientifiques, toujours d'actualité. Särndal (2010) dresse un panorama plus large de l'emploi de modèles en sondages, et donne deux définitions à la réconciliation des approches: les rendre compatibles et les faire converger, mais également revenir à des relations amicales!

L'intérêt de l'article est d'introduire une approche unifiée, dite approche *Model-Design based.* Il y a alors un mécanisme de sélection des données en deux phases, comme illustré dans le graphique ci-dessous (Binder 2011).



Les valeurs observées sur la population sont la réalisation de variables aléatoires issues d'un modèle de super-population. La convergence asymptotique signifie alors que le nombre d'individus de la population tend vers l'infini, qu'on note $\xi-$convergence, conditionnellement au processus de tirage à l'aide du plan de sondage (les variables aléatoires I_t égales à 1 pour les individus tirés et 0 sinon). Binder et Roberts (2003) font l'hypothèse que le modèle est du type $f_U(y;\theta)$ avec θ paramètre d'intérêt et U la population. A noter que l'approche de Binder (2011) est moins restrictive et introduit un modèle conditionnel à des variables explicatives X (ce qui est plus cohérent avec la modélisation économétrique), $f_U(y/x;\theta)$.

Les valeurs observées sur l'échantillon sont la réalisation d'un plan de sondage (les traitements d'enquête sont négligés). La convergence asymptotique signifie alors que le nombre d'individus tirés dans la population supposée fixe par le plan de sondage tend vers l'infini, qu'on note $p-$convergence, conditionnellement au processus générateur des données Y_t.

On définit alors la $\xi p-$convergence, qui tient compte du modèle et du plan de sondage. Au lieu d'une approche conditionnelle (par rapport au processus de tirage ou des observations), c'est donc une approche non conditionnelle. On tient compte de la loi jointe des processus (I_t, Y_t).

Dans une approche *Design-Based*, l'estimateur d'intérêt est $\hat{b} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i$ avec s l'échantillon et π_i la probabilité d'inclusion. Dans une approche *Model-Based*, l'estimateur d'intérêt est $\hat{\beta} = \frac{1}{n}\sum_{i\in s} y_i$. Si le vrai modèle est bien celui qui a été spécifié, les deux approches sont convergentes, l'estimateur $\hat{\beta}$ étant plus efficace. En théorie, $b = \beta = \bar{y}_U$, i.e. la moyenne sur la population totale. Dans le cas général (hétérogénéité des comportements rendant le modèle faux par exemple), $\mathbb{E}_p\left(\hat{\beta}\right)$ ne converge pas vers b ou β. Par contre, $\mathbb{E}_\xi\left(\hat{b}\right)$ converge toujours vers b et β.

Dans le cadre de l'approche *Model-Design based*, on définit la variance totale pour $\hat{b}$.

$$\mathbb{V}_{\xi p}\left(\hat{b}\right) = \mathbb{V}_\xi\left(\mathbb{E}_p\left(\hat{b}\right)\right) + \mathbb{E}_\xi\left(\mathbb{V}_p\left(\hat{b}\right)\right)$$

Sous l'hypothèse que $\mathbb{V}_\xi\left(\mathbb{E}_p\left(\hat{b}\right)\right) = \mathbb{V}_\xi\left(\bar{y}_U\right) = O\left(N^{-1}\right)$ (hypothèse moins restrictive que le caractère i.i.d. des observations) et qu'on dispose d'un estimateur $\widehat{\mathbb{V}_p}\left(\hat{b}\right)$ convergent de la variance $\mathbb{V}_p\left(\hat{b}\right)$, la variance totale peut être estimée par $\widehat{\mathbb{V}_p}\left(\hat{b}\right)$. L'approche *Design based* est donc robuste au sens où elle fournit

asymptotiquement des estimateurs et des variances convergentes. Si le modèle est vrai, l'estimateur *Design based* peut souffrir d'un problème d'efficacité par rapport à l'estimateur *model-based*. Ce problème d'efficacité reste mineur par rapport au gain de robustesse tant que la taille de l'échantillon est grande et que le taux de sondage est faible. L'approche est ensuite étendue à des statistiques plus complexes (et aux modèles de régression dans Binder (2011)).

L'article se conclut par l'étude des conséquences de choisir une approche *Model-Based* lorsque le modèle a été mal spécifié. On retrouve ici les exemples de la session 2. Les estimateurs peuvent être biaisés et non convergents, si le plan de sondage est non ignorable *(choice based sample*, i.e. le plan de sondage est déterminé en fonction de Y) ou de populations hétérogènes. Dans le cas où le terme de variance est mal spécifié dans le modèle, les estimateurs de variance seront également non convergents.

2 To model or not to model? Competing modes of inference for finite population sampling

L'article étend l'approche intégré *Model-Design based* dans un cadre bayésien et avec des exemples plus complexes. Les avantages et faiblesses de l'approche *Design-Based* sont rappelés. Elle permet d'obtenir des estimateurs fiables pour des échantillons de grande taille sous de faibles hypothèses. Par contre, l'inférence n'est valide que dans un cadre asymptotique d'une part et n'est plus adaptée quand le plan de sondage est perturbé, par exemples avec de la non réponse ou des erreurs de mesure d'autre part. Il n'y a pas de théorie pour construire des estimateurs optimaux, i.e. il n'existe pas de plan de sondage et d'estimateurs, qui quelque soit les valeurs de Y, fournissent un estimateur de variance minimum. Enfin, les calculs de variance sont compliqués pour les plans de sondage complexes.

Les deux approches sont néanmoins cohérentes. L'approche intégrée classique (i.e. fréquentiste) est déjà présente dans les estimateurs classiques issus de la théorie des sondages. L'approche *model-assisted design-based* (ou Model-Design based dans le précédent article) décrit l'usage d'outils orientés modèles en théorie des sondages soit pour améliorer le plan de sondage ou un estimateur. Mais l'inférence repose toujours sur le plan de sondage. Le texte donne plusieurs exemples. L'estimateur de Horvitz-Thompson (HT) peut ainsi être vu comme l'estimateur OLS du modèle linéaire: $y_i = \beta \cdot \pi_i + \pi_i \varepsilon_i$ et ε_i i.i.d. de loi $\mathcal{N}\left(0, \sigma^2\right)$. On montre en effet que $\hat{\beta} = \hat{t}_{HT}/n$, où $\hat{t}_{HT}$ est l'estimateur de HT du total $\sum_{i \in U} Y_i$. On en déduit que l'estimateur de HT du total ($\hat{t}_{HT} = \sum_{i \in s} y_i/\pi_i$) est un "bon" estimateur si le modèle décrit bien la population, i.e. $y_i/\pi_i \sim \mathcal{N}\left(\beta, \sigma^2\right)$.
Dans le célèbre exemple de Basu (1971) relatif à l'estimation du poids total de 50 éléphants, $\pi_i = 99/100$ pour l'éléphant "moyen" Sambo et $\pi_i = 1/4900$ pour chacun des 49 autres. Ce qui conduit à un estimateur certes sans biais mais aberrant et avec une variance très élevée en raison d'un choix de plan de sondage qui n'est pas du tout pertinent. Les modèles interviennent également en théorie des sondages lorsque l'on souhaite inclure de l'information auxiliaire. On procède classiquement à un ajustement linéaire entre la variable d'intérêt Y et les variables auxiliaires disponibles. La prédiction $\hat{y}_i = x_i\hat{\beta}$ est ensuite utilisée dans l'expression de l'estimateur. Enfin, la pseudo-vraisemblance permet de construire des estimateurs. Dans une proche modèle classique, l'estimateur du maximum de vraisemblance correspond à la valeur du paramètre qui annule le score. L'approche *model-assisted* reprend cette équation en estimant le score par HT.
Cette approche permet certes de construire des estimateurs et/ou des plans de sondage mais ne résout pas les problèmes de fond de l'approche sondage. Elle ne dicte pas en particulier la construction d'estimateurs de variance minimale.

L'article présente enfin l'approche bayésienne dans l'approche model-design based. L'idée est toujours de définir un modèle paramétrique pour Y (éventuellement conditionnellement à des variables explicatives X) associé à une distribution *a priori* pour les paramètres de cette loi. Cela permet d'inclure ou non de l'information *a priori* sur le paramètre d'intérêt. Par exemple, on peut supposer que les observations y_i sont générées i.i.d. par une loi normale $\mathcal{N}\left(\mu, \sigma^2\right)$ munie d'un *a priori* non informatif sur les paramètres $\left(\mu, \log \sigma^2\right)$. La formule de Bayes permet alors de déduire la loi *a posteriori* c'est-à-dire la loi de $Y|Y_{inc}$ où

Y_{inc} est la partie sélectionnée de la population. De cette loi *a posteriori*, on déduit celle de $Q(Y)|Y_{inc}$ dont découle l'estimation de la quantité d'intérêt $Q(Y)$.

Comme illustré dans l'article, cette approche ne résout pas le problème de mauvaise spécification du modèle générateur des données. Par exemple négliger les strates dans la spécification du modèle conduit à un estimateur biaisé du point de vue de l'approche *design-based*. Par contre, avec un modèle cohérent avec le plan de sondage et un a priori non informatif, les estimateurs de l'approche sondage et de l'approche modèle sont cohérents (pour les exemples qui illustrent l'article). Cette approche bayésienne présente l'avantage par rapport à l'approche sondage d'améliorer l'inférence pour des petits échantillons. Cela permet de plus d'améliorer le calcul des variances.

3 Echanges autour des textes

Pour les praticiens en sondages, l'intérêt des modèles est de débloquer des situations inextricables pour le calcul de variance complexe, par exemple pour l'estimateur d'un ratio. Les principaux domaines de la théorie des sondages qui font également intervenir les modèles sont le traitement de la non-réponse et l'estimation sur petits domaines. Enfin, les modèles sont invoqués comme justification théorique des sondages empiriques (quotas).

Les articles mettent en avant qu'un mauvais modèle conduit à de mauvais estimateurs. Cette conclusion n'est pas incohérente avec les présentations des deux premières sessions. Il apparaît en effet que les estimateurs pondérés sont plus robustes, même pour un modèle économétrique. On peut néanmoins regretter que l'analyse proposée par Binder et Roberts (2003) soit non conditionnelle pour le modèle de super-population. La définition du modèle est très restrictive, avec des observations non conditionnelles i.i.d. Or dans un modèle économétrique, on conditionne le caractère i.i.d. à des variables observables et explicatives X. Dans l'article de Binder et Roberts (2003), ce conditionnement par des observables est pris en compte pour l'approche sondage (via le plan de sondage et donc des variables $\tilde{X}$) mais pas pour l'approche modèle. Cela peut donc être perçu comme une simplification de l'approche modèle. Little (2004) au contraire envisage un modèle conditionnel, mais restreint l'analyse au cadre Bayésien. La lecture de Binder (2011), qui introduit des modèles conditionnels dans une approche intégrée classique et avec des exemples de modèles économétriques, pourrait compléter ces deux articles.

En économétrie, on peut également prendre en compte l'hétérogénéité de plusieurs façons. Des indicatrices de sous-populations peuvent être introduites dans le modèle (ce qui a été présenté à la session 2). Des modèles différents peuvent être définis pour les sous-populations (par exemple pour la discrimination salariale). Des phénomènes de sélection peuvent également être introduits dans le modèle (par exemple, on corrigera le salaire en tenant compte du fait qu'il n'est observé que pour les personnes en emploi). Un plan de sondage est général et non spécifique à une question (les poids sont identiques pour toutes les variables). Un modèle économétrique traite au contraire d'un point précis. Les articles pourraient donc être prolongés, en étudiant les estimateurs d'un modèle économétrique et les cas de deux approches de l'hétérogénéité différentes (voire d'un modèle de sélection), l'une via le plan de sondage, l'autre par le modèle. Ce n'est évidemment pas l'objet des articles, qui utilisent la modélisation comme un input dans une perspective sondages. De plus, un chargé d'études n'a souvent que peu d'informations sur la conduite de l'enquête et par défaut, il utilise donc une hypothèse de caractère i.i.d. des observations (ou du moins des variables qui ne tiennent compte que très partiellement du plan de sondage).

L'utilisation des procédures orientées sondages (PROC SURVEYREG par exemple) apparaît donc comme la solution la plus robuste. Little (2004) mentionne néanmoins dans sa conclusion "I prefer estimates of precision to be based on the Bayesian posterior distribution for a carefully specified model, but other methods of precision estimation that trade efficiency for robustness, such as replication methods and the "sandwich" estimator, have some appeal in the production survey setting, where sample sizes are large and detailed model assessment is not practical.". Une approche mixte (et qui correspond à la pratique économétrique) pourrait donc être d'utiliser des estimateurs pondérés et de spécifier des termes de variance approchant au maximum

la structure du plan de sondage. L'approche modèle permet également de comprendre pourquoi la variance n'est pas nulle, même pour un recensement. Le calcul de la variance sera étudié en session 4.

Les articles étudient le cas d'une mauvaise spécification du modèle. Lors de la session 2, pour étudier la prise en compte de la pondération dans un modèle économétrique, il y avait une hypothèse que les poids utilisés étaient convergents. Cette hypothèse pose la question inverse à celle des articles. Existe-il des plans de sondage mal spécifiés et des cas de poids non convergents ? Pour la phase de tirage, les poids sont supposés connus. Un plan de sondage peut être mal spécifié (par exemple dans le cas des éléphants de Basu (1971)) mais les estimateurs resteront asymptotiquement convergents, mais peu efficaces. Il n'y a pas de plans de sondage faux (sauf à supposer une erreur grossière avec des probabilités d'inclusion nulles pour certains individus), seulement des plans de sondage inefficaces. Par contre, le traitement de la non-réponse peut être mal effectué mais c'est alors encore un problème de modèle.

L'intérêt de l'approche Bayésienne a été questionné, notamment pour le calcul des variances. L'approche demeure complexe et d'une application non immédiate. Elle vise à améliorer ou corriger des cas où 1) l'estimateur d'Horvitz-Thompson ne serait pas raisonnable (d'où la nécessité d'une approche modèle) et 2) l'approche intégrée classique (i.e. fréquentiste) serait non efficace de part, par exemple, la présence d'observations ou de poids aberrants. Ce qui pose plutôt la question de la définition en amont du plan de sondage.

A noter que l'inversion des conditionnements ($\mathbb{E}_p\left(\mathbb{E}_\xi\left(h\left(Y\right)\right)\right) = \mathbb{E}_\xi\left(\mathbb{E}_p\left(h\left(Y\right)\right)\right)$) n'est possible que si le plan de sondage n'est pas informatif, i.e. la loi de Y ne dépend pas des unités tirées.

Bibliographie complémentaire

Binder, David A. (2011) "Estimating Model Parameters from a Complex Survey under a Model-Design Randomization Framework" Pakistan Journal of Statistics, 27(4), 371-390.

Särndal, Carl-Erik. (2010) "Models in Survey Sampling" in Official Statistics in Honour of Daniel Thorburn, 15-27.