*S*éminaire de *M*éthodologie *S*tatistique

mardi 2 juillet 2013
14h-17h, Insee - Malakoff 1 - salle 1245

Le traitement des unités influentes dans les enquêtes

Estimation robuste en présence d'unités influentes : une approche unifiée

David Haziza - *Université de Montréal et Crest-Ensai*

Les unités influentes peuvent avoir un effet important sur la qualité des estimations. Nous distinguons les valeurs résultant d'erreurs de mesure (erreurs grossières, erreurs d'unité de mesure) des vraies valeurs faisant partie de la population à l'étude, appelées par la suite valeurs influentes. Les premières sont habituellement identifiées à l'étape de vérification et sont corrigées, soit manuellement, soit par imputation. Dans cette présentation, nous nous attacherons uniquement au traitement des valeurs influentes. Le problème des valeurs influentes survient particulièrement dans les enquêtes auprès des entreprises qui collectent des variables économiques dont les distributions sont fortement asymétriques. De plus, des unités exhibant des poids extrêmes et certaines erreurs de la base de sondage sont propices à l'apparition d'unités influentes dans l'échantillon. Les valeurs influentes sont problématiques car elles mènent généralement à des estimateurs instables (c'est-à-dire des estimateurs ayant une grande variance). Il est possible de minimiser leur effet au moyen d'un plan de sondage approprié. Par exemple, il est de coutume d'utiliser un plan stratifié comportant une ou plusieurs strates exhaustives. Cependant, il est généralement impossible d'éliminer complètement le problème des unités influentes à l'étape du plan de sondage. Il est donc souhaitable de développer des méthodes d'estimation robustes à la présence d'unités influentes. Dans cette présentation, nous tenterons de répondre aux trois questions suivantes :

(1) Qu'est-ce qu'une valeur influente dans le contexte des enquêtes ?
(2) Comment quantifier l'influence d'une unité sur un estimateur ?
(3) Commet réduire l'impact des unités ayant une grande influence à l'étape de l'estimation?

Les réponses aux deux dernières questions s'appuient sur le concept de biais conditionnel d'une unité, qui est une mesure d'influence prenant en compte le plan de sondage.
En pratique, des estimations sont requises non seulement au niveau de la population mais également au niveau de sous-populations appelées domaines. Nous discuterons également de l'estimation robuste pour des domaines et présenterons une méthode s'apparentant au calage permettant d'assurer la cohérence entres les estimations robustes obtenues sur des domaines et l'estimation robuste sur la population totale.

Cette présentation repose sur des travaux réalisés en collaboration avec Jean-François Beaumont, Cyril Favre Martinoz et Anne Ruiz-Gazen.

.../...

La gestion des unités influentes dans l'ESA par winsorisation

Emmanuel Gros - *Département des méthodes statistiques, Insee*
Fabien Guggemos - *Département de l'emploi et des revenus d'activité, Insee*

L'enquête sectorielle annuelle, ou ESA, permet de recueillir les valeurs d'un grand nombre de variables économiques d'intérêt auprès des entreprises françaises, notamment la ventilation du chiffre d'affaires par branche d'activité. Les unités constituant l'échantillon de l'ESA sont sélectionnées selon un plan de sondage stratifié, les strates étant définies par croisement des codes d'activité principale exercée, de tranche d'effectif et de région.
Cependant, un mauvais classement sectoriel au lancement de l'enquête ou la hausse des effectifs d'une entreprise sur l'année en cours sont alors autant de facteurs générant, dans les strates non exhaustives, des points qui s'avèrent à la fois atypiques et non aberrants : atypiques au sens où ils sont situés dans la queue de distribution de la strate à laquelle ils appartiennent, non aberrants car leur valeur est certifiée et ne résulte pas d'une erreur de mesure.
Ces points atypiques, présents dans l'échantillon de l'ESA, engendrent une forte variance et par conséquent une grande instabilité des estimateurs considérés. Pour réduire cette dernière, une procédure de winsorisation est mise en œuvre dans l'ESA. Pour une variable d'intérêt donnée, cette procédure consiste à définir dans un premier temps un jeu de seuils par strate permettant d'identifier les unités atypiques, à savoir les unités dont la valeur dépasse le seuil de leur strate. Puis ces unités atypiques sont traitées, en rognant alternativement le poids de sondage ou la valeur de la variable d'intérêt considérée, en proportion de l'importance de leur caractère atypique, de façon à réduire leur influence dans les estimations.
Cette présentation détaillera la technique de winsorisation retenue ainsi que la procédure mise en œuvre pour déterminer les seuils par strate, et dressera un bilan de l'application de cette méthode dans le processus de production de l'ESA au cours des dernières campagnes.

Traitement des valeurs influentes dans l'enquête TIC-entreprises et expérimentations en cours

Elvire Demoly - *Département des synthèses sectorielles, Insee*

L'enquête européenne sur les technologies de l'information et de la communication et le commerce électronique (enquête TIC) est une enquête annuelle auprès des entreprises d'au moins 10 personnes, qui vise à connaître le niveau d'informatisation et la diffusion des TIC dans ces entreprises, en particulier l'importance du commerce électronique. Comme dans beaucoup d'enquêtes auprès d'entreprises, une des difficultés est l'estimation de données quantitatives, par exemple le montant total des ventes par internet, en présence d'unités qui ont un effet important sur la qualité des estimations, dites unités influentes.
Nous présenterons les différents traitements appliqués à l'enquête TIC visant à réduire l'effet de ces unités influentes, comme le tirage dans des strates exhaustives, la troncature des poids à 1 ou la winsorisation, ainsi que des travaux en cours sur ce sujet (mise en cohérence d'estimations robustes par domaine, amélioration de la méthode actuelle de winsorisation).

Traitement des unités influentes dans les enquêtes en présence de non-réponse totale

Cyril Favre Martinoz - *Crest-Ensai*

Dans la première présentation, un cadre unifié pour le traitement des unités influentes a été exposé. Ce cadre supposait toutefois l'absence d'erreurs non dues à l'échantillonnage. En pratique, les estimations produites par les organismes statistiques sont sujettes à plusieurs erreurs (erreurs de non-réponse, erreurs de couverture et erreurs de mesure). Dans cette présentation, nous généralisons les résultats en présence de non-réponse totale. Dans ce contexte, il est de coutume de former des groupes de réponse homogènes et d'ajuster le poids des individus répondants par l'inverse du taux de réponse observé dans ces groupes. Les estimateurs résultants ne sont toutefois pas robustes à la présence d'unités influentes. En présence de non-réponse totale, une unité influente peut avoir un effet important sur l'erreur due à l'échantillonnage et/ou sur l'erreur de non-réponse. Nous généralisons le concept de biais conditionnel au cas d'une non-réponse totale, ce qui permet de quantifier l'influence d'une unité sur l'erreur totale. Un estimateur robuste est construit au moyen des biais conditionnels estimés. Les résultats d'une étude par simulation seront présentés.