Un survol des méthodes d'estimation en présence de valeurs influentes dans les enquêtes

Cyril Favre-Martinoz

9^e Colloque Francophone sur les sondages
Gatineau

13 octobre 2016

Unité aberrante vs. unité influente

- **Définition :** une valeur aberrante est une donnée générée selon un modèle différent de celui qui tient pour la majorité des données.
- Une valeur aberrante
  - peut être une observation légitime (non entachée d'erreur); C'est notamment le cas lorsque le modèle considéré est un modèle de mélange.
  - Le jeu de données peut être contaminé par des erreurs de mesure.
- L'erreur d'unité de mesure est un exemple classique d'erreurs de mesure.
- Ce type d'erreur est le plus souvent détecté et corrigé au moment de l'étape d'apurement des données.

Détection des valeurs aberrantes

- Dans le cadre des enquêtes, détecter les valeurs aberrantes = détecter les erreurs de mesure.
- En l'absence d'erreur, il est inutile de procéder à cette étape de détection
- Si une unité est identifiée comme suspecte :
  - on réalise un suivi de cette unité (rappel);
  - on considère sa valeur comme manquante et on l'impute;
  - on peut la corriger manuellement;
  - on la garde telle quelle.
- En pratique, il existe un grand nombre de méthodes permettant de détecter les valeurs aberrantes.

Unité influente

- Une unité influente fait partie intégrante de la population finie.
- Il s'agit d'une observation légitime.
- Problème induit par les unités influentes : elles rendent les estimateurs classiques très instables $\rightarrow$ variances élevées.
- Dans quelles situations, les estimateurs ponctuels sont instables ?
  - Quand les poids de sondage sont très peu corrélés aux variables d'intérêt **et** les poids de sondage sont très dispersés $\rightarrow$ fréquent dans les enquêtes "Ménage".
  - Quand la distribution de la variable d'intérêt est très asymétrique et/ou quand il y a des erreurs dans la base de sondage (i.e., mauvaises classifications) $\rightarrow$ problème des sauteurs de strate $\rightarrow$ fréquent dans les enquêtes "Entreprise".
- Si un estimateur a une variance importante, il est fort probable qu'il y ait une ou plusieurs unités influentes dans l'échantillon

Comment se prémunir contre les unités influentes à l'étape du plan de sondage ?

- Idéalement, on souhaiterait éliminer le problème des valeurs influentes dès le plan de sondage.
- Dans le cas des enquêtes "Ménage", cela n'est pas toujours possible car les variables utilisées pour construire le plan (le plus souvent de l'information géographique) sont souvent très peu corrélées aux variables d'intérêt.
- Dans le cas des enquêtes "Entreprise", on peut se prémunir en construisant une ou plusieurs strates exhaustives $\rightarrow$ Les unités appartenant à ces strates ont une probabilité d'inclusion égale à 1 $\rightarrow$ Elles n'ont plus d'influence sur l'estimateur.
- Même avec un "bon plan", le problème des valeurs influentes n'est jamais complètement réglé :
  - on s'intéresse à beaucoup de variables d'intérêt et on dispose que d'un nombre limité de variables auxiliaires.
  - Problème des sauteurs de strate.

Notation

- Soit U une population finie de taille N.
- Objectif : estimer le total, $t_y = \sum_{i \in U} y_i$.
- Un échantillon s est sélectionné selon un plan de sondage $p(.)$ de probabilités d'inclusion $\pi_1, \ldots, \pi_N$.
- On considère un estimateur linéaire de t_y de la forme

$$\widehat{t}_y = \sum_{i \in s} w_i y_i,$$

où w_i est le poids de sondage de l'unité i.

Les estimateurs linéaires standards

- En l'absence d'erreur d'échantillonnage, le système de pondération classique $\{d_i = \pi_i^{-1}; i \in s\}$ conduit à l'estimateur Horvitz-Thompson :

$$\widehat{t}_{HT} = \sum_{i \in s} d_i y_i.$$

- Propriété de l'estimateur Horvitz-Thompson :
  - Sans biais sous le plan pour t_y : $E_p(\widehat{t}_{HT}) = t_y$.
  - Pour un plan de taille fixe : si $y_i \propto \pi_i$ alors $\widehat{t}_{HT} = t_y$ quel que soit s, on a :

$$V_p(\widehat{t}_{HT}) = 0.$$

- On s'attend à ce que $\widehat{t}_{HT}$ soit très efficace si y est approximativement proportionnelle à π.

Les estimateurs classiques

- Un estimateur alternatif : l'estimateur de type Hájek

$$\widehat{t}_{HA} = \frac{N}{\widehat{N}_{HT}} \widehat{t}_{HT}$$

avec $\widehat{N}_{HT} = \sum_{i \in s} d_i$.

- Les propriétés de l'estimateur de type Hájek :
  - Asymptotiquement sans biais pour t_y : $E_p(\widehat{t}_{HA}) \approx t_y$ si la taille d'échantillon n est suffisamment grande.
  - Si $y_i = \beta$, $\beta \in \mathbb{R}$ alors $\widehat{t}_{HA} = t_y$ quel que soit s, on a :

$$MSE_p(\widehat{t}_{HA}) = 0.$$

- On s'attend à ce que l'estimateur de type Hájek soit efficace si y_i et π_i ne sont pas liées.

Variabilité importante des poids

- Une variabilité importante des poids est souvent associée à une variance importante de l'estimateur.
- Souvent vérifiée, mais pas toujours vraie.
- Contre-exemple :
  - Pour $\widehat{t}_{HT}$ si $y_i \propto \pi_i$.
  - Pour $\widehat{t}_{HA}$ si $y_i = \beta$.
- Combinaison de deux facteurs qui rend les estimateurs instables :
  - une forte variation des poids de sondage
  - une corrélation faible entre les poids de sondage et la variable d'intérêt.
- Exemples classiques : Rao (1966) et Basu (1971).
- Dans une enquête avec une multitude de variables d'intérêt, il n'est pas rare d'observer des poids dispersés non corrélés à une ou plusieurs variables d'intérêt.$\rightarrow$ Cas des enquêtes environnementales.

Variabilité importante des poids

- Une variabilité importante des poids de sondage peut provenir
  - d'un sondage à probabilités inégales;
  - des corrections pour la non-réponse;
  - des ajustements liés au calage.
- A l'étape de la correction de la non-réponse, l'utilisation de classe de repondération permet de se prémunir contre des variations extrêmes des facteurs de correction.
- A l'étape de calage, il est possible de borner le facteur de variation de poids via l'utilisation d'une fonction de distance adaptée $\rightarrow$ permet de se prémunir contre une variation extrême des poids.

Que peut-on faire pour agir sur les valeurs influentes à l'étape d'estimation ?

- Un nombre important de techniques existe dans la littérature :
  - méthodes de troncature des poids ;
  - méthodes de troncature des valeurs ;
  - méthodes de lissage des poids;
- La plupart du temps les méthodes de troncature des poids et troncature des valeurs sont équivalentes.
- Différentes sur le papier mais elles poursuivent le même objectif :
  - modifier le poids ou la valeur de la variable d'intérêt afin de produire un estimateur dont l'erreur quadratique moyenne est plus faible que l'estimateur classique (Horvitz-Thompson, Hájek, calage).
  - Cette réduction de l'erreur quadratique moyenne se fait au détriment de l'introduction d'un léger biais.
  - Le traitement des valeurs influentes : compromis entre biais et variance.

Que peut-on faire pour agir sur les valeurs influentes à l'étape d'estimation ?

- Toutes les méthodes de troncature des poids ou des observations reposent sur la détermination d'un seuil au-delà duquel une unité est considérée comme influente.
- Le choix de ce seuil est crucial. Un mauvais choix pourrait engendrer un estimateur avec une erreur quadratique moyenne plus élevée que l'estimateur classique.
- Le seuil doit être fonction de la taille d'échantillon n de telle sorte que :

$$\lim_{n \to \infty} c_n = \infty.$$

- Dans ce cas, l'estimateur robuste résultant sera consistant sous le plan, car il convergera vers l'estimateur classique, qui est consistant.

Que peut-on faire pour agir sur les valeurs influentes à l'étape d'estimation ?

- En pratique, le seuil c est souvent choisi indépendamment de la taille de l'échantillon; par exemple, il est courant d'utiliser une méthode de détermination des valeurs aberrantes sur la distribution des poids ou des observations. Dans ce cas, l'estimateur robuste résultant n'est pas consistant.
- Il existe au moins deux critères pour choisir le seuil c de telle sorte que l'estimateur robuste soit consistant sous le plan :
  - Déterminer le seuil c qui minimise l'erreur quadratique moyenne de l'estimateur robuste.
  - Déterminer le seuil c qui minimise la plus grande influence (déterminée éventuellement par le biais conditionnel) dans l'échantillon.

Troncature des poids

- Un nombre important de procédures de troncature des poids sont utilisées en pratique.
- Idée : les unités avec un poids élevé w_i sont influentes et nuisibles.
- Les unités avec un poids extrêmes sont considérées comme influentes.
- Une méthode populaire consiste à
  (1) choisir une valeur $w_0 = c\overline{w}$, où $\overline{w} = n^{-1} \sum_{i \in s} w_i$;
  (2) définir un poids tronqué $\widetilde{w}_i$ qui vérifie

$$\widetilde{w}_i = \begin{cases} w_0 & \text{si} \quad w_i \geq w_0 \\ \gamma w_i & \text{si} \quad w_i < w_0 \end{cases}$$

où γ est facteur d'ajustement qui assure que

$$\sum_{i \in s} \widetilde{w}_i = \sum_{i \in s} w_i.$$

- L'étape 2 est répétée jusqu'à ce que les poids $\widetilde{w}_i$ soient plus petits w_0.

Troncature des poids : une méthode populaire

- L'estimateur de t_y après troncature est donné par

$$\widehat{t}_{trim}(w_0) = \sum_{i \in s} \widetilde{w}_i y_i.$$

- Si $c = \infty$, alors $\widetilde{w}_i = w_i$ quel que soit i et $\widehat{t}_{trim}(w_0) = \widehat{t}_{HT}$.
- Si $c = 1$, l'estimateur $\widehat{t}_{trim}(w_0)$ est équivalent à l'estimateur non pondérée :

$$\widehat{t}_{trim}(w_0) = \frac{N}{n} \sum_{i \in s} y_i.$$

- Habituellement, la valeur de c est fixée *a priori* (e.g., $3 \leq c \leq 6$)
- On peut aussi remplacer $\overline{w}$ par la médiane des poids w_i, dans ce cas $w_0 = cMed(w_i)$.

Méthodes basées uniquement sur la distribution des poids

- Potter (1988, 1990) propose le premier traitement formel de troncature des poids.
- Modélisation de la distribution des poids : on suppose que la fonction réciproque de la distribution normalisée des poids suit une loi Beta.

$$f(w_i) = \frac{n\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(1/(nw_i))^{\alpha-1}(1-1/(nw_i))^{\beta-1}$$

- Les poids appartenant à la queue supérieure de la distribution, vérifiant par exemple $1 - F(w_i) < 0,01$ sont tronqués à w_0 tel que $1 - F(w_0) = 0,01$.
- Avantage : cette méthode ne dépend pas de la variable d'intérêt.
- Inconvénient : l'estimateur robuste sera efficace pour certaines variables d'intérêt et inefficace pour d'autres.

Méthodes tenant compte de la variable d'intérêt

- Potter (1990) propose de choisir le seuil w_0 qui minimise une estimation de l'erreur quadratique moyenne de l'estimateur tronqué $\widehat{t}_{trim}(w_0)$:

$$\widehat{MSE}(\widehat{t}_{trim}) = \widehat{V}(\widehat{t}_{trim}) + \left(\widehat{t}_{trim} - \widehat{t}_{HT}\right)^2 - \widehat{V}(\widehat{t}_{trim} - \widehat{t}_{HT}).$$

- L'estimateur tronqué optimal : $\widehat{t}_{trim}(w_0^{opt})$.
- Contrairement aux méthodes précédentes, celle-ci tient compte de la variable d'intérêt y.
- Ce seuil optimal w_0^{opt} pour une certaine variable d'intérêt peut être très loin du seuil optimal pour une autre variable d'intérêt.
- En général, une minimisation de l'erreur quadratique moyenne estimée est relativement complexe.

Winsorisation

- Winsorisation standard : Soit $\widetilde{y}_i$ la valeur de y associée à l'unité i après winsorisation. On a :

$$\widetilde{y}_i = \begin{cases} y_i & \text{si } w_i y_i \leq c \\ \frac{c}{w_i} & \text{si } w_i y_i > c \end{cases}$$

- L'estimateur par winsorisation standard est donné par :

$$\widehat{t}_{stand}(c) = \sum_{i \in s} w_i \widetilde{y}_i = \sum_{i \in s} \widetilde{w}_i y_i,$$

où

$$\widetilde{w}_i = w_i \frac{min\left(y_i, \frac{c}{w_i}\right)}{y_i}.$$

- Si $min\left(y_i, \frac{c}{w_i}\right) = y_i$ (i.e., l'unité i n'est pas influente), alors $\widetilde{w}_i = w_i$.
- Pour une unité influente : $\widetilde{w}_i < w_i$.
- Problème : $\widetilde{w}_i$ peut être plus petit que 1.

Winsorisation

- Winsorisation type Dalen-Tambay: Dalen (1987) et Tambay (1988) propose une winsorisation du type :

$$\widetilde{y}_i = \begin{cases} y_i & \text{si } w_i y_i \leq c \\ \frac{c}{w_i} + \frac{1}{w_i}(y_i - \frac{c}{w_i}) & \text{si } w_i y_i > c \end{cases}$$

- L'estimateur winsorisé de Dalen-Tembay est donné par

$$\widehat{t}_{DT}(c) = \sum_{i \in s} w_i \widetilde{y}_i = \sum_{i \in s} \widetilde{w}_i y_i,$$

où

$$\widetilde{w}_i = 1 + (w_i - 1)\frac{min\left(y_i, \frac{c}{w_i}\right)}{y_i} \geq 1$$

.

- Si $min\left(y_i, \frac{c}{w_i}\right) = y_i$, alors $\widetilde{w}_i = w_i$.
- Pour une unité influente : $\widetilde{w}_i < w_i$.
- Propriété sympathique : $\widetilde{w}_i$ est nécessairement supérieur à 1.

Qu'est ce qu'une unité influente ?

- Jusqu'ici, une unité avec un poids w_i élevé (méthodes de Potter) ou une valeur pondérée $w_i y_i$ élevée (estimateur winsorisé) était considérée comme influente.
- Cependant, ces mesures de l'influence ne prennent pas en compte :
  - le plan de sondage ;
  - l'estimateur utilisée pour estimer t_y;
  - la forme du paramètre à estimer (total, ratio, etc);
- Une mesure alternative de l'influence ?
  - le biais conditionnel d'une unité ;

Mesurer l'influence : le biais conditionnel

- Il a été introduit par Moreno-Rebollo, Munoz-Reyes and Munoz-Pichardo (1999), puis repris par Beaumont, Haziza and Ruiz Gazen (2013) pour constuire des estimateurs robustes.
- Le biais conditionnel d'une unité échantillonnée i pour l'estimateur $\widehat{t}_{HT}$:

$$\begin{aligned} B_{1i}^{HT} &= E_p(\widehat{t}_{HT} - t_y | I_i = 1) \\ &= \sum_{j \in U} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_j, \end{aligned}$$

où $\pi_{ij} = P(\{i \ \in s\} \cap \{j \in s\})$ désigne la probabilité d'inclusion jointe des unités i et j dans l'échantillon.

- De façon similaire, on peut définir le biais conditionnel d'une unité non échantillonnée :

$$B_{0i} = E_p(\widehat{\theta} - \theta | I_i = 0).$$

Remarques

- Le biais conditionnel dépend en général des probabilités d'inclusion d'ordre deux, $\pi_{ij} \mapsto$ prend en compte le plan de sondage.
- Le biais conditionnel est généralement inconnu $\rightarrow$ il doit être estimé.

- L'influence d'une unité dépend de la valeur de la variable d'intérêt y et du plan de sondage.
- Si $\pi_i = 1$ alors $B_{1i}^{HT} = 0$.

Les cas particuliers

- Sondage aléatoire simple sans remise : $\pi_i = n/N$ pour tout i et $\pi_{ij} = n(n-1)/N(N-1)$ quel que soit $i \neq j$.

$$B_{1i}^{HT} = \frac{N}{(N-1)}\left(\frac{N}{n} - 1\right)(y_i - \overline{Y}),$$

où $\overline{Y} = Y/N$ est la moyenne dans la population.

- Plan poissonien : $\pi_{ij} = \pi_i \pi_j$, $i \neq j$

$$B_{1i}^{HT} = (d_i - 1)y_i = (d_i - 1)(y_i - 0)$$

Lien entre le biais conditionnel et l'erreur d'échantillonnage

- Pour un plan poissonien, l'erreur d'échantillonnage peut se décomposer de la façon suivante :

$$\widehat{t}_{HT} - t_y = \sum_{i \in s} B_{1i}^{HT} + \sum_{i \in U-s} B_{0i}^{HT}.$$

- Le biais conditionnel de l'unité i (échantillonnée ou non) est une mesure de sa contribution à l'erreur d'échantillonnage.
- Pour un sondage aléatoire simple sans remise et les plans à grande entropie :

$$\widehat{t}_{HT} - t_y \approx \sum_{i \in s} B_{1i}^{HT} + \sum_{i \in U-s} B_{0i}^{HT}.$$

si la taille de la population N est grande.

Lien entre le biais conditionnel et la variance d'échantillonnage de $\widehat{t}_{HT}$

- Pour un plan de sondage quelconque, la variance d'échantillonnage de $\widehat{t}_{HT}$ est donnée par :

$$\begin{aligned} V_p\left(\widehat{t}_{HT}\right) &= \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)y_i y_j \\ &= \sum_{i\in U} B_{1i}^{HT} y_i \end{aligned}$$

- Maintenant que l'on est capable d'identifier les unités influentes, on souhaite construire des estimateurs robustes.

Méthode robuste basée sur le biais conditionnel

- Proposée par Beaumont, Haziza et Ruiz-Gazen (2013).
- Il considère un estimateur robuste de $\widehat{t}_{HT}$ de la forme :

$$\widehat{t}_{BHR}(c) \quad = \quad \widehat{t}_{HT} - \sum_{i \in s} \widehat{B}_{1i}^{HT} + \sum_{i \in s} \psi_c \left(\widehat{B}_{1i}^{HT} \right)$$

où $\psi_c(\cdot)$ est la fonction de Huber donnée par :

$$\psi_c(x) = \begin{cases} c & \text{si } x > c \\ x & \text{si} |x| \leq c \\ -c & \text{si } x < -c \end{cases}$$

- Rôle de la fonction ψ : réduire l'influence des unités ayant une grande influence.

Méthode robuste basée sur le biais conditionnel

- La constante d'ajustement c peut être choisie de telle sorte qu'elle minimise l'erreur quadratique estimée de $\widehat{t}_{BHR}(c) \mapsto$ complexe sans hypothèses simplificatrices
- Un choix alternatif pour c : Choisir le seuil c qui minimise $\max_{i \in s} \left\{ |\hat{B}_{1i}^{R}| \right\}$:

$$\widehat{t}_{BHR}(c_{opt}) = \widehat{t}_{HT} - \frac{1}{2}(\widehat{B}_{min}^{HT} + \widehat{B}_{max}^{HT}).$$

- Facile à implémenter.
- $\widehat{t}_{BHR}(c_{opt})$ est consistant sous le plan pour t_y.

Méthode robuste basée sur le biais conditionnel

- L'estimateur BHR peut se réécrire de la façon suivante :

$$\widehat{t}_{BHR}(c) = \sum_{i \in s} d_i \widetilde{y}_i$$

où

$$\widetilde{y}_i = y_i - \frac{\alpha_i}{d_i} \hat{B}_{1i}^{HT}, \qquad \alpha_i = 1 - \psi_c(\hat{B}_{1i}^{HT}) / \hat{B}_{1i}^{HT}.$$

- Mais également sous la forme :

$$\widehat{t}_{BHR}(c) = \sum_{i \in s} \widetilde{d}_i y_i,$$

où

$$\widetilde{d}_i = d_i - \frac{\alpha_i}{y_i} \widehat{B}_{1i}^{HT}, \qquad \alpha_i = 1 - \psi_c(\widehat{B}_{1i}^{HT}) / \widehat{B}_{1i}^{HT}.$$

Bibliographie I

- Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *To appear in Biometrika*.
- Clark, R.G. (1995). Winsorization methods in sample surveys. Masters thesis, Department of Statistics, Australian National University.
- Dalén, J. (1987). Practical estimators of a population total which reduce the impact of large observations. R and D Report. Statistics Sweden.
- Kokic, P.N., and Bell, P.A. (1994). Optimal Winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics*, 10, 419–435.
- Moreno-Rebollo, J.L., Muñoz-Reyez, A.M. and Muñoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling: conditional bias. *Biometrika*, 86, 923–968.

Bibliographie II

- Moreno-Rebollo, J.L., Muñoz-Reyez, A.M., Jimenez-Gamero, M.D. and Muñoz-Pichardo, J. (2002). Influence diagnostics in survey sampling: estimating the conditional bias. *Metrika*, 55, 209–214.
- Rivest, L.-P. (1994). Statistical properties of Winsorized means for skewed distributions. *Biometrika*, 81, 373–383.
- Rivest, L.-P. and Hidiroglou, M. (2004). Outlier treatment for disaggregated estimates. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, Virginia, 4248–4256.
- Rivest, L.-P. and Hurtubise, D. (1995). On Searls Winsorized means for skewed populations. *Survey Methodology*, 21, 119–129.
- Tambay, J.-L. (1988). An integrated approach for the treatment of outliers in sub-annual surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, Virginia, 229–234.

Merci de votre attention !