Section 3: Utilisation de variables auxiliaires

Pierre Duchesne

August 1, 2017

Quelques notations

- Variable d'intérêt: y.
- Variable auxiliaire: x.
- L'objectif est le même qu'avant, à savoir estimer $t_y = \sum_U y_k$ ou encore la moyenne $\bar{y}_U = N^{-1} \sum_U y_k$.
- Jusqu'à maintenant l'estimateur considéré est l'estimateur Horvitz-Thompson $\hat{t}_{y\pi} = \sum_s y_k/\pi_k$.

Base de sondage

La base de sondage la plus simple ne consiste que des étiquettes des unités. On considère une base de sondage qui présume la connaissance d'informations sur les unités. La base de sondage prend la forme suivante.

Étiquette	Valeur de x
1	x_1
2	x_2
3	x_3
$\ldots$	$\ldots$
k	x_k
$\ldots$	$\ldots$
N	x_N

Information auxiliaire

- Les valeurs x_k sont supposées connues pour l'ensemble de la population, *avant de faire le sondage*.
- Question: Peut-on améliorer nos estimations à l'aide de l'information auxiliaire?
- Autrement formulé, au lieu de considérer l'estimateur d'Horvitz-Thompson, peut-on considérer un autre estimateur, qui incorporerait d'une manière ou d'une autre les x_k?

Estimation du total t_y

- On sélectionne s dans la base de sondage selon un plan d'échantillonnage $p(\cdot)$.
- On déduit les probabilités d'inclusion d'ordre un: $\pi_k = P(s \ni k)$.
- On observe alors $\{y_k | k \in s\}$.
- On se propose de remplacer $\hat{t}_{y\pi} = \sum_s y_k/\pi_k$ par une autre méthode d'estimation.
- Autres choix possibles que l'on se propose d'étudier:
  1. Estimateur par le ratio;
  2. Estimation par la régression.

Estimateur par le ratio

- L'estimateur est défini comme suit:

$$\hat{t}_{yra} = \sum_U x_k \frac{\sum_s y_k/\pi_k}{\sum_s x_k/\pi_k}$$

- Le total de l'information auxiliaire $t_x = \sum_U x_k$ est connu.
- La quantité

$$\frac{\sum_s y_k/\pi_k}{\sum_s x_k/\pi_k}$$

est calculable sur s.

Estimateur par la régression

L'estimateur peut s'écrire des deux façons suivantes:

$$\begin{aligned}\hat{t}_{yreg} &= \sum_s y_k/\pi_k + \hat{B}_s\left(\sum_U x_k - \sum_s x_k/\pi_k\right), \\ &= t_x\hat{B}_s + \sum_s (y_k - \hat{B}_s x_k)/\pi_k.\end{aligned}$$

L'estimateur de la pente de régression de y_k sur x_k est:

$$\hat{B}_s = \frac{\sum_s (y_k - \tilde{y}_s)(x_k - \tilde{x}_s)/\pi_k}{\sum_s (x_k - \tilde{x}_s)^2/\pi_k},$$

avec

$$\tilde{y}_s = \frac{\sum_s y_k/\pi_k}{\sum_s 1/\pi_k}, \ \tilde{x}_s = \frac{\sum_s x_k/\pi_k}{\sum_s 1/\pi_k}.$$

Remarque importante sur la disponibilité de l'information auxiliaire

- On a présumé que $x_1, x_2, \ldots, x_N$ sont connues.
- Est-ce que c'est absolument nécessaire de connaître tous les x_k pour l'ensemble de la population?
- Si on regarde attentivement les formules pour l'estimateur par le ratio et pour l'estimateur pour la régression, on remarque que:
  1. Nous avons en fait besoin de connaître le total t_x de la variable x. Cette information pourrait provenir d'un recensement, de fichiers administratifs, etc.
  2. Les x_k et y_k disponibles pour les k dans l'échantillon s, qui seront disponibles lors de la collecte des données (mise en oeuvre du sondage).

Disponibilité de l'information auxiliaire: premier cas

Ainsi deux situations peuvent se produire. Dans le premier cas:

- On connaît $x-1, x_2, \ldots, x_N$ avant le sondage car l'on dispose d'une base de sondage complète.
- On sélectionne s dans $U = \{1, 2, \ldots, N\}$.
- Observer les y_k, $k \in s$. Ainsi les données de sondage sont:

$$\{y_k, k \in s\}, \ \{x_k, k \in U\}.$$

Disponibilité de l'information auxiliaire: second cas

Dans le second cas, on pourrait se retrouver dans ce contexte:

- On connaît t_x, le total de l'information auxiliaire avant le sondage (mais on ne connaît pas individuellement $x_1, x_2, \ldots, x_N$ avant le sondage).
- On sélectionne s dans $U = \{1, 2, \ldots, N\}$.
- Lors de la collecte des données, observer $\{(x_k, y_k), k \in s\}$.
- Ainsi les données du sondage sont:

$$\{(x_k, y_k), k \in s\}, \text{ et } t_x.$$

Estimateurs Horvitz-Thompson, par le ratio, et par la régression pour le plan SI, $\pi_k = n/N$

- Estimateur Horvitz-Thompson: $\hat{t}_{y\pi} = N\bar{y}_s$.
- Estimateur par le ratio:

$$\hat{t}_{yra} = N\bar{x}_U \frac{\bar{y}_s}{\bar{x}_s}.$$

- Estimateur par la régression:

$$\hat{t}_{yreg} = N\left\{\bar{y}_s + \hat{B}_s(\bar{x}_U - \bar{x}_s)\right\},$$

avec

$$\hat{B}_s = \frac{\sum_s (y_k - \bar{y}_s)(x_k - \bar{x}_s)}{\sum_s (x_k - \bar{x}_s)^2},$$

Coefficient de corrélation

- Comme on va le voir, les estimateurs par le ratio et la régression trouvent leur justification à l'aide de modèles.
- Quand le modèle sera satisfaisant, ces estimateurs seront d'autant plus performants (efficaces, avec typiquement de faibles variances).
- Le lien linéaire entre y et x est mesuré par le coefficient de corrélation.
- Le coefficient de corrélation est quantité utile afin de comparer les estimateurs.

Coefficient de corrélation (suite)

Le coefficient de corrélation de population, comme un total de population ou moyenne de population, est un paramètre calculé sur U:

$$r_{xyU} = \frac{S_{xyU}}{S_{xU} S_{yU}},$$

avec

$$S_{xyU} = \frac{1}{N-1} \sum_U (x_k - \bar{x}_U)(y_k - \bar{y}_U)$$

et

$$S_{xU} = \sqrt{\frac{1}{N-1} \sum_U (x_k - \bar{x}_U)^2}$$

En invoquant l'inégalité de Cauchy-Schwartz, on trouve que $|r_{xyU}| \leq 1$.