

L'estimation d'un total en présence d'information auxiliaire

Mohammed El Haj Tirari

Crest - Ensai

Plan

- Introduction et notations
- Classe d'estimateurs du total
- Borne inférieure pour la variance des estimateurs de cette classe
- Échantillonnage équilibré et l'estimateur optimal du total
- Cas de l'estimateur par la régression généralisée.
- Conclusion

Introduction

- Plusieurs statistiques peuvent s'écrire comme des fonctions de totaux
- Plusieurs estimateurs du total ont été proposés
- L'objet principal de ce travail est de montrer l'intérêt d'utiliser un échantillonnage équilibré pour construire des estimateurs du total d'une population

Notations

- Population $U = \{1, \dots, k, \dots, N\}$
- Variable d'intérêt $\mathbf{y}_U = (y_1, \dots, y_k, \dots, y_N)'$
- Total de la population $t_y = \sum_{k \in U} y_k$
- Un échantillon s de la population U
- Un plan de sondage $p(s)$ de probabilités d'inclusion π_k et π_{kl}

Information auxiliaire

On dispose des valeurs de J variables auxiliaires

$$\mathbf{X}_1, \dots, \mathbf{X}_j, \dots, \mathbf{X}_J$$

$$\mathbf{x}_k = \left(x_{1k}, \dots, x_{jk}, \dots, x_{Jk} \right)', \quad k \in U$$

$$\mathbf{t}_x = \sum_{k \in U} \mathbf{x}_k$$

Approche modèle

Pour l'approche modèle, on suppose que $y_1, \dots, y_k, \dots, y_N$ sont des réalisations de N variables aléatoires dont la distribution conjointe ξ est définie par le modèle de régression :

$$\begin{cases} \mathbf{y}_U = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ E_{\xi}(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \text{et} \quad \text{var}_{\xi}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{V} \end{cases}$$

où

$$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_k, \dots, \mathbf{x}_N)' \quad ; \quad \boldsymbol{\beta} = (\beta_1, \dots, \beta_j, \dots, \beta_J)'$$

$$\mathbf{V} = \text{diag}(v_1^2, \dots, v_k^2, \dots, v_N^2)$$

Classe des estimateurs \hat{t}_{yR}

Considérons la classe des estimateurs linéaires définis par :

$$\hat{t}_{yR} = \sum_{k \in S} r_k y_k + \left(\sum_{k \in U} \mathbf{x}_k - \sum_{k \in S} r_k \mathbf{x}_k \right)' \hat{\boldsymbol{\beta}}_R$$

où

$$\hat{\boldsymbol{\beta}}_R = \left(\sum_{k \in S} \frac{r_k}{v_k^2} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in S} \frac{r_k}{v_k^2} \mathbf{x}_k y_k$$

avec $r_1, \dots, r_k, \dots, r_N$ connus et strictement positifs.

Classe des estimateurs \hat{t}_{yR}

Exemples

- Quand $r_k = 1$ pour tout $k \in U$, on a :

$$\hat{t}_{yR} = \hat{t}_{yBLUE} = \sum_{k \in s} y_k + \left(\sum_{k \in U} \mathbf{x}_k - \sum_{k \in s} \mathbf{x}_k \right)' \hat{\boldsymbol{\beta}}_{BLUE}$$

- Quand $r_k = \pi_k^{-1}$ pour tout $k \in U$, on a :

$$\hat{t}_{yR} = \hat{t}_{ygreg} = \sum_{k \in s} \frac{y_k}{\pi_k} + \left(\sum_{k \in U} \mathbf{x}_k - \sum_{k \in s} \frac{\mathbf{x}_k}{\pi_k} \right)' \hat{\boldsymbol{\beta}}_{greg}$$

Problème

- La précision de l'estimateur \hat{t}_{yR} dépend en partie de la qualité du modèle de régression.
- Existe-t-il des plans de sondage qui peuvent assurer la robustesse de \hat{t}_{yR} quand le modèle de régression ξ est incorrect ?

Solution

L'échantillonnage équilibré

Problème

Objectif

Chercher l'estimateur \hat{t}_{yR} qui met à profit le modèle de régression ξ sans être trop dépendant d'une éventuelle mal spécification de ce dernier.

Critère de choix

$$E_p E_\xi \left(\hat{t}_{yR} - t_y \right)^2$$

Estimateur \hat{t}_{yR} optimal

Résultat

Sous le modèle de régression :

$$\mathbf{y}_U = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, E_{\xi}(\boldsymbol{\varepsilon}) = \mathbf{0} \text{ et } \text{var}_{\xi}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{V}$$

si les vecteurs $\mathbf{V}\mathbf{1}_U$ et $\mathbf{V}^{1/2}\mathbf{1}_U \in C(\mathbf{X})$

où $\mathbf{V} = \text{diag}\{v_1^2, \dots, v_N^2\}$; $\mathbf{V}^{1/2} = \text{diag}\{v_1, \dots, v_N\}$

alors

$$E_p E_{\xi} \left(\hat{t}_{yR} - t_y \right)^2 \geq \left[\frac{a}{n} \left(\sum_{k \in U} v_k \right)^2 - \sum_{k \in U} v_k^2 \right] \sigma^2$$

$$\text{où } a = \frac{\min_{k \in U} r_k}{\max_{k \in U} r_k}.$$

Estimateur \hat{t}_{yR} optimal

De plus, la borne inférieure est atteinte ssi l'échantillon s vérifie la contrainte d'équilibrage suivante :

$$h \sum_{k \in s} \frac{\mathbf{x}_k}{v_k r_k^{-1}} = \sum_{k \in U} \mathbf{x}_k$$

$$\text{où } h = \sqrt{\frac{a}{n \sum_{k \in U} r_k^2 \pi_k} \sum_{k \in U} v_k}.$$

Dans ce cas, on a

$$\hat{t}_{yR} = \sum_{k \in s} w_{ks} y_k$$

où

$$w_{ks} = h \frac{r_k}{v_k} = \left[\sqrt{\left(\frac{a}{n \sum_{k \in U} r_k^2 \pi_k} \right) \sum_{k \in U} v_k} \right] \frac{r_k}{v_k}.$$

Remarques

- Une condition suffisante pour que la contrainte d'équilibrage soit satisfaite :

$$CV_v^2 \leq 1 - \frac{n}{aN}$$

où CV_v est le coefficient de variation de $v_1, \dots, v_k, \dots, v_N$.

- Cette condition est toujours satisfaite quand les séries v_1, \dots, v_N et r_1, \dots, r_N ne sont pas trop dispersées.
- La contrainte d'équilibrage est toujours satisfaite si, pour tout $k \in U$,
 $r_k = r$ et $v_k = cx_k^\alpha$; ($\alpha \in \square$)

L'Estimateur par la régression généralisée (*GREG*)

Sous la contrainte d'équilibrage, l'estimateur GREG optimal est donné par :

$$\hat{t}_{ygreg} = \sum_{k \in s} w_{ks} y_k$$

où

$$w_{ks} = h(v_k \pi_k)^{-1} = \left[\sqrt{\left(a/n \sum_{k \in U} \pi_k^{-1} \right) \sum_{k \in U} v_k} \right] (v_k \pi_k)^{-1}.$$

$$\text{et } a = \frac{\min_{k \in U} \pi_k}{\max_{k \in U} \pi_k}.$$

L'Estimateur par la régression généralisée (*GREG*)

- La contrainte d'équilibrage devient :

$$h \sum_{k \in S} \frac{\mathbf{X}_k}{v_k \pi_k} = \sum_{k \in U} \mathbf{X}_k$$

$$\text{où } h = \sqrt{\frac{a}{n \sum_{k \in U} \pi_k^{-1} \sum_{k \in U} v_k}}$$

- Cette contrainte est toujours satisfaite quand les séries v_1, \dots, v_N et π_1, \dots, π_N ne sont pas trop dispersées.
- La contrainte d'équilibrage est toujours satisfaite si, pour tout $k \in U$,

$$\pi_k = \frac{n}{N} \text{ et } v_k = cx_k^\alpha; \quad (\alpha \in \square)$$

Conclusion

Le fait de soumettre l'échantillon à la contrainte d'équilibrage permet de :

- Construire un estimateur de total \hat{t}_{yR} optimal sous le plan de sondage et le modèle de régression.
- Garantir une protection contre un éventuel biais engendré par l'addition d'une ou de plusieurs variables auxiliaires dans le modèle de régression ξ .