The Regression Estimator of Total Domain and Two-Stage Sampling Design

Getka-Wilczynska Elzbieta
Warsaw Schol of Economics,
,Institute of Econometrics,Division of Mathematical Statistics
Al. Niepodleg³ oœci 162
02-554 Warsaw, Poland
get-wil@sgh.waw.pl

1. Introduction

Let finite population U of size N be divided into D non-overlapping and nonempty subsets U_d (domains) of known size $N_d > 0$, $d = 1,...,D$. Assume that the ordered sample s of size $n(s)$ is drawn from the whole population U under a two-stage sampling design with stratification of primary sampling units, such that $P(k \in s) = \pi_k > 0$ for all k and $P(k,l \in s) = \pi_{kl} > 0$ for all $k \neq l;\ k,l = 1,...,N$. We suppose that the total values of an auxiliary variables X_j, $j = 1,..., p$ of the domain are available. We are interested in estimating the total domain value $Y_d = \sum_{k=1}^{N_d} Y_k$, where $d = 1,...,D$ on the random number observation from the subset (domain) of the finite population in the sample drawn from the whole population by two-stage sampling design. The variable of interest Y takes value y_k for k-th unit of U.

2. The regression estimators

Under the above assumption generalised difference estimator of domain total Y_d can be written as $\hat{t}_{diff} = \sum_{k \in s_d} \frac{y_k - E_k}{\pi_k} + \sum_{k \in U_d} E_k$, where E_k, $k = 1,...,N$ are constant.

When $E_k = c_j x_{jk}$, $k = 1,...,N;\ j = 1,...,p$ we have (1) $\hat{t}_{diff}^{(p)} = \hat{t}_{yd} + \sum_{j=1}^{p} c_j \left(X_{dj} - \hat{t}_{xdj}\right)$, where $\hat{t}_{yd}$ and $\hat{t}_{xdj}$ denotes the unbiased Horvitz-Thompson estimator of total domain Y_d and X_{dj}, $j = 1,..., p$ respectively.

The generalised regression estimator are developed from the formula (1) by linear regression approximation of the coefficient c_j, $j = 1,..., p$ only on the random number observation from the domain in the sample drawn from the whole population by two-stage sampling design. By using model assisted approach (Särndal, Swensson and Wretman (1992)) we derive the ratio and the regression estimators for the cases of the introduced generalised regression model. One-step regression model for the domain generated the coefficients $\hat{c}_1 = \frac{\hat{t}_{yd}}{\hat{t}_{Nd}}$ and $\hat{c}_2 = \frac{\hat{t}_{yd}}{\hat{t}_{xd}}$ for the additional value $x_k = 1$ and x_k for the k-th element of the domain. Then the regression estimators are the ratio estimators of domain total: $\hat{t}_{1d} = \frac{\hat{t}_{yd}}{\hat{t}_{Nd}} N_d$, $\hat{t}_{2d} = \frac{\hat{t}_{yd}}{\hat{t}_{xd}} X_d$. From the two-step regression model we derived coefficient

$\hat{c}_3 = \left[\frac{\hat{t}_{yd}}{\hat{t}_{Nd}} - \frac{Cov(\hat{t}_{yd}, \hat{t}_{xd})}{Var(\hat{t}_{xd})} \frac{\hat{t}_{xd}}{\hat{t}_{Nd}},\ \frac{Cov(\hat{t}_{yd}, \hat{t}_{xd})}{Var(\hat{t}_{xd})} \right]^T$ and the regression estimator of domain total

can be written as $\hat{t}_{3d} = N_d \left[\frac{\hat{t}_{yd}}{\hat{t}_{Nd}} + \hat{B}_d \left(\frac{X_d}{N_d} - \frac{\hat{t}_{xd}}{\hat{t}_{Nd}} \right) \right]$. Analogously, from one-step and from two-step regression model for the population are derived the regression estimators of the population total calculated on the basis on the sample drawn from the whole population.

Result. Under the conditions stated above expectation, MSE and the estimator of MSE of the statistic $\hat{t}_{3d} = N_d \left[\frac{\hat{t}_{yd}}{\hat{t}_{Nd}} + \hat{B}_d \left(\frac{X_d}{N_d} - \frac{\hat{t}_{xd}}{\hat{t}_{Nd}} \right) \right]$ are given by:

$$E(\hat{t}_{5d}) = Y_d - \frac{1}{N_d}\left[Cov(\hat{t}_{yd}, \hat{t}_{Nd}) - B_d\, Cov(\hat{t}_{xd}, \hat{t}_{Nd}) - (\bar{Y}_d - B_d \bar{X}_d) V(\hat{t}_{Nd})\right] + O(n^{-2})$$

$$MSE(\hat{t}_{5d}) = V(\hat{t}_{yd}) - 2 B_d\, Cov(\hat{t}_{yd}, \hat{t}_{xd}) + B_d^2 V(\hat{t}_{Nd}) -$$

$$2(\bar{Y}_d - B_d \bar{X}_d)\left[Cov(\hat{t}_{yd}, \hat{t}_{Nd}) - B_d\, Cov(\hat{t}_{xd}, \hat{t}_{Nd}) - \frac{1}{2}(\bar{Y}_d - B_d \bar{X}_d) V(\hat{t}_{Nd})\right] + O(n^{-2})$$

$$\hat{MSE}(\hat{t}_{5d}) = \hat{V}(\hat{t}_{yd}) - 2 \hat{B}_d\, \hat{Cov}(\hat{t}_{yd}, \hat{t}_{xd}) + \hat{B}_d^2 \hat{V}(\hat{t}_{Nd}) -$$

$$2\left(\frac{\hat{t}_{yd}}{\hat{t}_{Nd}} - \hat{B}_d \frac{\hat{t}_{xd}}{\hat{t}_{Nd}} \right)\left[\hat{Cov}(\hat{t}_{yd}, \hat{t}_{Nd}) - \hat{B}_d\, \hat{Cov}(\hat{t}_{xd}, \hat{t}_{Nd}) - \frac{1}{2}\left(\frac{\hat{t}_{yd}}{\hat{t}_{Nd}} - \hat{B}_d \frac{\hat{t}_{xd}}{\hat{t}_{Nd}} \right) \hat{V}(\hat{t}_{Nd})\right]$$

where $\hat{B}_d = \frac{Cov(\hat{t}_{yd}, \hat{t}_{xd})}{V(\hat{t}_{xd})}, B_d = \frac{Cov(Y_d, X_d)}{V(X_d)}$.

Approximate expressions for the expectation and the mean square error of the statistics of domain total $\hat{t}_{1d}$, $\hat{t}_{2d}$ (Getka-Wilczynska (2000)), $\hat{t}_{3d}$ (or of population total) and the ratio of total, the covariance, the regression coefficient are developed by using the Taylor linearization technique. For the two-stage sampling design the first and the second conditional moments and the variance estimator of the Horvitz-Thompson estimator are derived in (Getka-Wilczynska (2000)).

REFERENCES

Getka-Wilczynska, E. (2000). Estimation of total domain in finite population. Statistics in Transition v.4, no 4, 711-728. Warsaw.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). Model assisted survey sampling.

Springer - Verlag New York.

RESUME

On a considéré les propriétés des estimateurs de la valeur globale de la caractéristique examinée pour les domaines de la population finie. Les estimateurs sont calculés pour le nombre aléatoire des observations du sous-ensemble dans l'échantillon trié de toute la population selon le schéma de tirage effectué à deux étapes.

Après avoir défini l'estimateur généralisé de la différence et l'estimateur généralisé de regression, on a introduit indépendamment un modèle généralisé de regression pour la population et pour les sous-ensembles de la population. En appliquant une approche adjointe au modèle, on a obtenu pour les cas détaillés du modèle de regression, les estimateurs de quotient et de regression (les fonctions non-linéaires des estimateurs Horvitz-Thompson).

A partir de la méthode de la linéarisation de Taylor on a désigné les erreurs moyennes quadratiques et on a donné les estimateurs des erreurs moyennes quadratiques. Les conditions dans lesquelles les estimateurs construits ne sont pas chargés, la valeur de charge et les mesures diverses de l'efficacité des estimateurs sont analysés à partir d'une simulation.