ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 8, 2004

Multivariate finite population inference under the assumption of linear pattern in the population

KRISTIINA RAJALEID

ABSTRACT. In this paper a multivariate form of regression estimator for finite population inference is studied. Using Taylor linearization the asymptotic covariance matrix of the estimator is derived. Properties of the estimator are examined. The results are illustrated by a simulation study.

1. Introduction

The regression estimator allows to use auxiliary information to increase the precision of estimates. The estimator is approximately unbiased and has smaller variance than the classical Horvitz–Thompson estimator.

Usually there is more than one study variable under interest in a sample survey. However, population totals are estimated one at a time, without considering their joint behaviour. In the current paper a multivariate form of regression estimator is introduced. Covariance matrix of the Taylor linearized form of the estimator is derived and properties of the covariance matrix are studied.

2. Population under study

Let us have a finite population consisting of N elements. Let there be p study variables in the population. We write the population data as an $N \times p$ matrix

$$\mathbf{Y} = y_{(ij)},$$

where each row represents a population element with its p variable values and each column represents a variable measured on N population elements. The $N \times k$ matrix of auxiliary variables is written in the same way, $\mathbf{X} = x_{(ij)}$.

Received October 25, 2003.
2000 *Mathematics Subject Classification.* 62D05.
Key words and phrases. Finite population inference, regression estimator.
The research was supported by Estonian Science Foundation Grant 5523.

Values of study variables are known only for those population elements which are sampled. Values of auxiliary variables are known for each element in the population (or at least the population totals are known). Auxiliary information is used to increase the quality of estimates. It can be used both before and after sampling (for example, to calculate inclusion probabilities proportional to a size variable and later calibrate by auxiliary variables).

3. Superpopulation model

We assume that the finite population is a realization of a superpopulation model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where the dimensions of the matrices $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are $k \times p$ and $N \times p$ accordingly. For example, the population data may be generated by a biological, chemical or economical process.

We assume that the expected values of errors are equal to zero and the population units are independent, but the study variables are dependent.

According to the model the expected value of the ith variable (ith column of $\mathbf{Y}$) is

$$\mathrm{E}(Y_i) = \mathbf{X}\beta_i,$$

where β_i is the ith column of matrix $\boldsymbol{\beta}$. The expected value of an element of $\mathbf{Y}$ is

$$\mathrm{E}(y_{mn}) = \sum_{j=1}^{k} x_{nj}\beta_{jm}.$$

4. Sampling

Sampling is performed by a random vector $\mathbf{I} = (I_1, I_2, \cdots, I_N)^{\mathrm{T}}$ called sampling vector. Behaviour of $\mathbf{I}$ depends on the sampling design used. Elements of $\mathbf{I}$ – inclusion indicators – show the number of times a population element is sampled. For without-replacement sampling schemes the possible values of inclusion indicators are 0 and 1, for with-replacement schemes inclusion indicators may take values from 0 to n, where n is the sample size. At the estimation stage the expanded sampling vector

$$\breve{\mathbf{I}} = \left(\frac{I_1}{E(I_1)}, \frac{I_2}{E(I_2)}, \cdots, \frac{I_N}{E(I_N)}\right)^{\mathrm{T}}$$

is used. For without-replacement sampling schemes $\mathrm{E}(I_i)$ equals inclusion probability, $\mathrm{E}(I_i) = \Pr(I_i = 1) = \pi_i$. In general it is the expected sampling count of unit i. Note that the expectation

$$\mathrm{E}(\breve{\mathbf{I}}) = \mathbf{1}, \tag{2}$$

is the vector of ones. About the vector approach in sampling theory see e.g. Traat (2003).

5. Parameter estimation

Minimizing $(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ gives the least squares estimator for superpopulation parameter $\boldsymbol{\beta}$ on the finite population level. The estimator is

$$\mathbf{B} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}. \tag{3}$$

We denote

$$\mathbf{t}_{\mathbf{XX}} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$$

and

$$\mathbf{t}_{\mathbf{XY}} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}.$$

The sample estimators for $\mathbf{t}_{\mathbf{XX}}$ and $\mathbf{t}_{\mathbf{XY}}$ are

$$\hat{\mathbf{t}}_{\mathbf{XX}} = \mathbf{X}^{\mathrm{T}}\breve{\mathbf{I}}_{\mathrm{diag}}\mathbf{X}$$

and

$$\hat{\mathbf{t}}_{\mathbf{XY}} = \mathbf{X}^{\mathrm{T}}\breve{\mathbf{I}}_{\mathrm{diag}}\mathbf{Y}.$$

Due to (2) the estimators $\hat{\mathbf{t}}_{\mathbf{XX}}$ and $\hat{\mathbf{t}}_{\mathbf{XY}}$ are unbiased with respect to the sampling design. By combining them we get an estimator for $\mathbf{B}$:

$$\hat{\mathbf{B}} = \hat{\mathbf{t}}_{\mathbf{XX}}^{-1}\hat{\mathbf{t}}_{\mathbf{XY}}. \tag{4}$$

Estimator $\hat{\mathbf{B}}$ is not unbiased anymore because it is not linear in $\breve{\mathbf{I}}$ (the expectation of the inverse of $\breve{\mathbf{I}}$ is not equal to the inverse of its expectation).

6. Estimators of population totals

Usually the aim of a survey is the estimation of population totals of the study variables or a function (average, ratio) of the population totals. Written in matrix form the vector of population totals is

$$\mathbf{t}_{\mathbf{Y}} = \mathbf{Y}^{\mathrm{T}}\mathbf{1}. \tag{5}$$

As in reality only part of the values of study variables is known, one cannot find the totals but has to estimate them using the values of sampled elements. Well-known Horvitz–Thompson estimator of population totals is

$$\hat{\mathbf{t}}_{\mathbf{Y}} = \mathbf{Y}^{\mathrm{T}}\breve{\mathbf{I}}. \tag{6}$$

Assuming that the linear pattern in the population,

$$\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta},$$

holds and having estimated the parameter $\boldsymbol{\beta}$, it is possible to find predicted values for the whole population:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}. \tag{7}$$

Residuals $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}$ are known for sampled elements. Using the relationship

$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{R}$$

we get another estimator for the totals of study variables:

$$\hat{\mathbf{t}}_{\mathbf{Y}} = \hat{\mathbf{Y}}^{\mathrm{T}}\mathbf{1} + \mathbf{R}^{\mathrm{T}}\check{\mathbf{I}}. \tag{8}$$

The estimator of the population total of a variable (an element in $\hat{\mathbf{t}}_{\mathbf{Y}}$) is equal to the sum of predicted values, and the sample residuals multiplied by the elements of the expanded sampling vector. Estimator (8) is called generalized regression estimator of $\mathbf{t}_{\mathbf{Y}}$ (for one-dimensional case see Särndal et al. (1992), p. 246). Denoting it by $\hat{\mathbf{t}}_{\mathbf{Y},\mathbf{reg}}$ and replacing terms in (8) we get

$$\hat{\mathbf{t}}_{\mathbf{Y},\mathbf{reg}} = (\mathbf{X}\hat{\mathbf{B}})^{\mathrm{T}}\mathbf{1} + (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^{\mathrm{T}}\check{\mathbf{I}}. \tag{9}$$

Random terms in this expression are $\check{\mathbf{I}}$ and $\hat{\mathbf{B}}$ (it includes $\check{\mathbf{I}}$).

7. Variance of regression estimator

In order to study variance of the regression estimator we use Taylor linearization method. We write the estimator (9) as

$$\hat{\mathbf{t}}^{\mathrm{T}}_{\mathbf{Y},\mathbf{reg}} = \hat{\mathbf{t}}^{\mathrm{T}}_{\mathbf{Y}} + \hat{\mathbf{t}}^{\mathrm{T}}_{\mathbf{XY}}\hat{\mathbf{t}}_{\mathbf{XX}}(\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}}), \tag{10}$$

where

$$\begin{aligned}
\mathbf{t}_{\mathbf{X}} &= \mathbf{X}^{\mathrm{T}}\mathbf{1},\\
\hat{\mathbf{t}}_{\mathbf{X}} &= \mathbf{X}^{\mathrm{T}}\check{\mathbf{I}},\\
\hat{\mathbf{t}}_{\mathbf{Y}} &= \mathbf{Y}^{\mathrm{T}}\check{\mathbf{I}}.
\end{aligned}$$

We see that it is a function of four random arguments $\hat{\mathbf{t}}_{\mathbf{Y},\mathbf{reg}} = f(\hat{\mathbf{t}}_{\mathbf{X}}, \hat{\mathbf{t}}_{\mathbf{Y}}, \hat{\mathbf{t}}_{\mathbf{XX}}, \hat{\mathbf{t}}_{\mathbf{XY}})$. We expand (10) into a Taylor series at the point of expected values of the arguments, using the following rules (Kollo (1991), pp. 66–67):

$$\frac{d\mathbf{X}^{\mathrm{T}}}{d\mathbf{X}} = \mathbf{I}_{\mathrm{q,p}},$$

$$\frac{d\mathbf{X}^{-1}}{d\mathbf{X}} = -\mathbf{X}^{-1} \otimes (\mathbf{X}^{\mathrm{T}})^{-1},$$

$$\frac{d\mathbf{AYB}}{d\mathbf{X}} = (\mathbf{B}^{\mathrm{T}} \otimes \mathbf{A})\frac{d\mathbf{Y}}{d\mathbf{X}},$$

$$\frac{d\mathbf{Y} \otimes \mathbf{Z}}{d\mathbf{X}} = (\mathbf{I}_{\mathrm{n}} \otimes \mathbf{I}_{\mathrm{m,s}} \otimes \mathbf{I}_{\mathrm{r}}) \left[(\mathbf{I}_{\mathrm{mn}} \otimes \mathrm{vec}\mathbf{Z})\frac{d\mathbf{Y}}{d\mathbf{X}} + (\mathrm{vec}\mathbf{Y} \otimes \mathbf{I}_{\mathrm{rs}})\frac{d\mathbf{Z}}{d\mathbf{X}} \right],$$

where $\mathbf{X}$ is a $p \times q$, $\mathbf{Y}$ is an $m \times n$ and $\mathbf{Z}$ is an $r \times s$ matrix.

Derivatives of function f are

$$\frac{df}{d\hat{\mathbf{t}}_{\mathbf{X}}} = -\hat{\mathbf{t}}_{\mathbf{XY}}^{\mathrm{T}}\mathbf{t}_{\mathbf{XX}}^{-\mathbf{1}},$$

$$\frac{df}{d\hat{\mathbf{t}}_{\mathbf{Y}}} = \mathbf{1}\ ,$$

$$\frac{df}{d\hat{\mathbf{t}}_{\mathbf{XX}}} = -\hat{\mathbf{t}}_{\mathbf{XY}}^{\mathrm{T}}\mathbf{t}_{\mathbf{XX}}^{-\mathbf{1}} \otimes (\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}})^{\mathrm{T}}\mathbf{t}_{\mathbf{XX}}^{-\mathbf{1}},$$

$$\frac{df}{d\hat{\mathbf{t}}_{\mathbf{XY}}} = (\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}})^{\mathrm{T}}\mathbf{t}_{\mathbf{XX}}^{-\mathbf{1}}.$$

Evaluating the derivatives at $(\hat{\mathbf{t}}_{\mathbf{X}}, \hat{\mathbf{t}}_{\mathbf{Y}}, \hat{\mathbf{t}}_{\mathbf{XX}}, \hat{\mathbf{t}}_{\mathbf{XY}}) = (\mathbf{t}_{\mathbf{X}}, \mathbf{t}_{\mathbf{Y}}, \mathbf{t}_{\mathbf{XX}}, \mathbf{t}_{\mathbf{XY}})$ we get the linear part of the Taylor series

$$\hat{\mathbf{t}}_{\mathbf{Y,lin}} = \hat{\mathbf{t}}_{\mathbf{Y}} + (\mathbf{t}_{\mathbf{X}} - \hat{\mathbf{t}}_{\mathbf{X}})^{\mathrm{T}}\mathbf{t}_{\mathbf{XX}}^{-\mathbf{1}}\hat{\mathbf{t}}_{\mathbf{XY}} \tag{11}$$

or

$$\hat{\mathbf{t}}_{\mathbf{Y,lin}} = (\mathbf{XB})^{\mathrm{T}}\mathbf{1} + (\mathbf{Y} - \mathbf{XB})^{\mathrm{T}}\breve{\mathbf{I}}. \tag{12}$$

For bigger samples $\hat{\mathbf{t}}_{\mathbf{Y,lin}}$ approximates well $\hat{\mathbf{t}}_{\mathbf{Y,reg}}$. In (12) only $\breve{\mathbf{I}}$ is random and it is easy to find the expectation and covariance matrix of it. One can see that (12) is unbiased, i.e. its expectation is equal to the vector of population totals. Covariance matrix of the linearized regression estimator is

$$\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{Y,lin}}) = (\mathbf{Y} - \mathbf{XB})^{\mathrm{T}}\breve{\mathbf{\Delta}}(\mathbf{Y} - \mathbf{XB}), \tag{13}$$

where

$$\breve{\mathbf{\Delta}} = \mathrm{Cov}(\breve{\mathbf{I}}) \tag{14}$$

is the covariance matrix of expanded sampling vector.

For comparison, the covariance matrix of Horvitz–Thompson estimator is

$$\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{Y},\pi}) = \mathbf{Y}^{\mathrm{T}}\breve{\mathbf{\Delta}}\mathbf{Y}. \tag{15}$$

Elements of matrix $\breve{\mathbf{\Delta}}$ are multiplied with population values in (15) but with population residuals in (13). Therefore, one can significantly decrease the variances and covariances of the estimators by using auxiliary information.

8. Simulation study

A simulation study was carried out to examine the behaviour of the introduced multivariate estimator. A population with $N = 1000$ elements was generated. The values of $k = 4$ auxiliary variables were generated independently, as well as the random errors. Superpopulation parameter $\mathbf{B}$ was given. Values of study variables were calculated according to model (1). There were $p = 3$ study variables and $k = 4$ auxiliary variables in the generated population. Auxiliary variables were correlated with study variables as shown in Table 1.

	Y_1	Y_2	Y_3
X_1	0.65	0.16	0.41
X_2	0.52	0.96	0.84
X_3	0.15	0.16	0.14
X_4	0.49	0.22	0.36

TABLE 1. Correlations between study variables and auxiliary variables

Correlation matrix of study variables was

$$\mathrm{Cor}(\mathbf{Y}) = \begin{pmatrix} 1 & 0.71 & 0.88 \\ 0.71 & 1 & 0.95 \\ 0.88 & 0.95 & 1 \end{pmatrix}.$$

Population totals of study variables were $t_1 = 79994$, $t_2 = 243330$ and $t_3 = 169160$.

1000 independent samples with size $n = 100$ were taken from the population. Two different sampling designs were used – simple random sampling and multinomial sampling. Simple random sampling is a without-replacement sampling design. For simple random sampling matrix (14) has $\frac{N-n}{n}$ on the diagonal and $-\frac{N-n}{n(N-1)}$ in other places. Multinomial sampling is a with-replacement design with unequal selection probabilities p_i, where selection count of population element i is described by a binomial random variable, $I_i \sim B(n, p_i)$. Hence (14) has $\frac{1-p_i}{np_i}$ on the diagonal and $-\frac{1}{n}$ in other places.

Population totals were estimated in two different ways – using Horvitz–Thompson estimator (6) and regression estimator (9). Altogether we considered four different situations (two sampling schemes combined with two estimation methods). For each situation we calculated the estimates of the totals and the covariance matrix of the estimates. Determinant of the covariance matrix (generalized variance) was used to characterize both the variances and covariances of the estimates with one figure.

	Y_1	Y_2	Y_3
SI+reg	-4	-2	0
SI+HT	-2	13	15
MN+reg	-3	-3	0
MN+HT	-5	-1	-1

TABLE 2. Bias of estimates

In Table 2 the biases of the estimates are shown (SI = simple random sampling, MN = multinomial sampling, HT = Horvitz–Thompson estimator, reg = regression estimator). As known, Horvitz–Thompson estimator

is unbiased and regression estimator is approximately unbiased. All four situations give good estimates with ignorable bias.

	SI			MN		
HT	340102	1049559	725973	26014	453	222
	1049559	6352908	3414640	453	4962	-2
	725973	3414640	2022949	222	-2	1813
reg	21817	-125	-168	23785	-247	156
	-125	4000	-92	-247	4578	-144
	-168	-92	1580	156	-144	1831

TABLE 3. Covariance matrices of the estimates

Table 3 shows covariance matrices of the estimates. As expected, the variances and covariances are the biggest for simple random sampling and Horvitz–Thompson estimator. Using auxiliary information decreased the variability and covariance of sample estimators.

In Table 4 the generalized variances of the estimates are given. The value is the biggest for simple random sampling coupled with Horvitz–Thompson estimator. In this case auxiliary information is used neither on the sampling stage nor on the estimation stage. Generalized variances for the other situations are smaller and of equal magnitude.

Sampling situation	Determinant of covariance matrix
SI+reg	$1.37 \cdot 10^{11}$
SI+HT	$3.22 \cdot 10^{16}$
MN+reg	$1.98 \cdot 10^{11}$
MN+HT	$2.33 \cdot 10^{11}$

TABLE 4. Generalized variance of estimates

We also calculated the correlation matrices of the estimates for all the four situations. For three of the four situations correlation matrices showed independence of estimators. For simple random sampling coupled with Horvitz–Thompson estimator correlation matrix coincided with correlation matrix of study variables what is theoretically the case. Positive correlations between estimates mean that using one and the same sample the population totals of different variables tend to be overestimated, or underestimated at the same time.

References

Kollo, T. (1991). *Matrix Derivative in Multivariate Statistics*, Tartu University Press, Tartu. (In Russian)

Traat, I. (2003). On the estimation of finite population covariance matrix. *Statistics in Transition* **6(1)**, 67–82.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag.

INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITY OF TARTU, J. LIIVI 2, 50409 TARTU, ESTONIA

E-mail address: rajaleid@ut.ee