SOME MODEL-BASED ESTIMATOR FOR A FINITE POPULATION TOTAL

D. Krapavickaitė
Institute of Mathematics and Informatics
Vilnius, LITHUANIA
e-mail: krapav@ktl.mii.lt

Abstract
A model-based estimator of the total of nonnegative study variable in finite population is investigated. The case of the variable acquiring significant number of zero values is studied. The tobit model for non-sampled values of such a variable is applied. Properties of proposed model-based estimator are studied analytically and by simulation.

1 Introduction
The variable of expenditure on environmental protection is quite irregular: some of the enterprises have high expenditure and some of them have none at all. Thus, distribution of such a variable is skewed with the peak at zero and has a high population variance. For this reason the well-known design-based Horvitz-Thompson estimator [2] of the total of such a variable in finite population has a high variance as well. We are looking for another way of estimating the total.

One of the ways is to use the model-based approach of a study variable. In this case the values of variables of the elements of finite population \( U \) are assumed to be generated according to some super-population model. The non-sampled values of the study variable are predicted by this model and used for the estimation of the total. If we can find a distribution model that closely resembles the distribution of the study variable, a model based-estimator may have a smaller mean square error than a design-based one. This way of estimation is given in Valliant et. al. [7].

2 Model-based estimator of a total
Let us denote by \( U = \{1, 2, \ldots, N\} \) the finite population, consisting of \( N \) elements, and by \( y \) the study variable with the values \( y_k, k = 1, \ldots, N \), defined for the elements of the population, correspondingly; the total of this variable is \( t_y = y_1 + \cdots + y_N \). Let \( x \) be an auxiliary variable, and its values \( x_k, k = 1, \ldots, N \), known for all the population elements. Let \( i, \ i \subset U, \) be a probability sample of size \( n \) and \( \bar{i} = U \setminus i \) be a subset of the non-sampled elements of the population. The values \( y_k \) are known only for the elements \( k \) that belong to the sample: \( k \in i \).

The parameter of interest is the population total \( t_y = \sum_{k=1}^{N} y_k \). It can be expressed as a sum \( t_y^{(i)} \) of the sampled elements and a sum \( t_y^{(\bar{i})} \) of the non-sampled ones:

\[
t_y = t_y^{(i)} + t_y^{(\bar{i})} = \sum_{k \in i} y_k + \sum_{k \in \bar{i}} y_k.
\]
Here the value of \( t_y^{(i)} \) is known from the sampled data. A super-population model for the values \( y_k, k \in \mathcal{U} \) has been built. This model is used to construct the estimators \( \hat{y}_k, k \in \hat{\mathcal{I}} \). Then \( t_y^{(i)} \) is estimated by 
\[
\tilde{t}_y^{(model)} = t_y^{(i)} + \tilde{t}_y^{(i)}.
\]
(1)
Tobit model for the variable \( y \) with zero and positive values is used.

3 Tobit model of a study variable

Let us denote by \( x_k = (1, x_k)' \), \( k = 1, 2, \ldots, N \), values of the known non-random auxiliary vector \( x \) and by \( z_k^* \), values of some unobserved random variable \( z^* \). Suppose the values \( z_k^*, k = 1, 2, \ldots, N \), can be written in the form
\[
z_k^* = x_k^* \beta + \varepsilon_k
\]
(2)
with the unknown regression model parameter \( \beta = (\beta_0, \beta_1)' \) and error terms \( \varepsilon_k \sim \mathcal{N}(0, \sigma^2) \), which are independent. Let us define a variable \( z \) with the values
\[
z_k = \begin{cases} 
z_k^*, & \text{if } z_k^* \geq 0, \\
0, & \text{if } z_k^* < 0, 
\end{cases}
\]
(3)
k = 1, \ldots, N. The model given by equations (2) and (3) and associating \( z_k \) with \( z_k^* \) is called a censored regression model or a tobit model of the variable \( z \) ([6], [1], [5]). It is proved there that
\[
E(z_k|x_k) = x_k^* \beta \phi\left(\frac{x_k^* \beta}{\sigma}\right) + \sigma \phi\left(\frac{x_k^* \beta}{\sigma}\right),
\]
(4)
k = 1, \ldots, N. Here \( \Phi(\cdot) \) is a standard normal distribution function, and \( \phi(\cdot) \) is its density function.

In the case of the real population, the tobit model is fitted to the variable \( z \) with the values obtained by the transformation \( z_k = \ln(y_k + 1), k = 1, \ldots, N \), rather than to the initial variable \( y \) (for example, expenditure of the enterprise to the environmental protection). We have proved in [3] that
\[
E(y_k|x_k) = e^{x_k^* \beta + \frac{z_k^2}{2}} \Phi\left(\frac{x_k^* \beta}{\sigma} + \sigma\right) - \Phi\left(\frac{x_k^* \beta}{\sigma}\right), \quad k = 1, \ldots, N.
\]
(5)
After estimation of the tobit model parameters \( \sigma, \beta = (\beta_0, \beta_1)' \) by the maximum likelihood method and obtaining the estimators \( \hat{\sigma}, \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)' \), the predictors \( \hat{z}_k \) and \( \hat{y}_k^{(tobit)} \) are obtained from (4) and (5):
\[
\hat{z}_k^{(tobit)} = x_k^* \hat{\beta} \Phi\left(\frac{x_k^* \hat{\beta}}{\hat{\sigma}}\right) + \hat{\sigma} \phi\left(\frac{x_k^* \hat{\beta}}{\hat{\sigma}}\right),
\]
(6)
\[
\hat{y}_k^{(tobit)} = e^{x_k^* \hat{\beta} + \frac{z_k^2}{2}} \Phi\left(\frac{x_k^* \hat{\beta}}{\hat{\sigma}} + \hat{\sigma}\right) - \Phi\left(\frac{x_k^* \hat{\beta}}{\hat{\sigma}}\right), \quad k = 1, \ldots, N.
\]
(7)
They are used to estimate the totals \( t_z = \sum_{k=1}^{N} z_k \) and \( t_y = \sum_{k=1}^{N} y_k \) by

\[
\hat{t}_z^{(tobit)} = \hat{t}_z^{(i)} + \hat{t}_z^{(ii)} = \sum_{k \in i} z_k + \sum_{k \in \bar{i}} \hat{z}_k^{(tobit)}, \tag{8}
\]

\[
\hat{t}_y^{(tobit)} = \hat{t}_y^{(i)} + \hat{t}_y^{(ii)} = \sum_{k \in i} y_k + \sum_{k \in \bar{i}} \hat{y}_k^{(tobit)}. \tag{9}
\]

These estimators are not unbiased. Some properties of such estimators are studied analytically, some of them are shown by simulation in [3], [4].

References


