Domain estimation for a censored study variable

Danutė Krapavickaitė
Vilnius Gediminas Technical University, Vilnius, Lithuania – danute.krapavickaite@vgtu.lt

Abstract

The estimator for domain totals of a study variable having many zero values is studied in the paper. Design-based model-assisted estimators are usually inefficient in such a situation because of a high variance of a study variable. Therefore, a censored regression model introduced by Tobin is suggested in this paper to be used in the usual domain estimation framework. It is suggested to estimate model parameters using Bayesian inference, and the model-based estimator of domain total is obtained.

Keywords: variable containing many zero values; unit-level model; Bayesian inference.

1. Introduction

Let us take a non-negative study variable defined in a finite population which contains many zero values. It is called a censored variable. For example, an enterprise may have high environmental protection expenses or it may have no such expenses at all; a household may have expenses for luxury goods or it may have no such expenses at all; the area under crop that is grown not so often on farms, for example, rape in Lithuania. Usually, variables in a sampling frame may be correlated with positive values of the study variable, but there are no auxiliary variables in a sampling frame indicating that the value of a study variable may be equal to zero. The design-based estimator of the population total for such a variable usually has a high variance; the model-assisted estimator does not have a significantly lower variance. The censored regression (or tobit) model and the two-line (or change-point regression) model have been applied to the model-based and model-assisted estimator to estimate the total in Krapavickaite (2011).

The aim of the current study is to use a unit-level model for domains for a study variable having many zero values with the application of the censored regression model, and to estimate it using Bayesian inference. The model-based estimator for domain totals is constructed afterwards.

2. Tobit model

The tobit model was introduced by Tobin (1958) and is widely used in econometrics. It is also called a censored regression model with censoring of the study variable at any level. We consider only censoring at zero. Suppose that a non-observable random variable Y_k^*, k=1,2,...,N, can be described by the linear regression model:

$$Y_k^* = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k,\ \varepsilon_k \sim N(0,\psi),\ k=1,2,\ldots,N,$$

with variance $\psi > 0$, auxiliary J-dimensional vectors $\mathbf{x}_k = \left(x_{k1},\ldots,x_{kJ}\right)'$, vector of parameters $\boldsymbol{\beta} = \left(\beta_1,\ldots,\beta_J\right)'$, and identically distributed errors ε_k. Errors ε_k, ε_l are independent for $k \neq l$, k, l =1,2,…,N. Let the observed variable

$$Y_k = \max(0, \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k) = \begin{cases} Y_k^*, & \text{if} \quad Y_k^* \geq 0, \\ 0, & \text{if} \quad Y_k^* < 0, \quad k = 1,2,\ldots,N. \end{cases} \qquad (1)$$

This is a tobit model for Y_k, it has been deeply studied by Amemiya (1984). An expression of the mean for Y_k under the model can be found in Greene (2003):

$$E(Y_k \mid \mathbf{x}_k) = \mathbf{x}'_k\boldsymbol{\beta}\Phi\left(\frac{\mathbf{x}'_k\boldsymbol{\beta}}{\psi}\right)+\sigma\varphi\left(\frac{\mathbf{x}'_k\boldsymbol{\beta}}{\psi}\right). \tag{2}$$

Here φ is a density and Φ is a distribution function of a standard normal random variable. The mean of the tobit model is shown in Fig. 1.

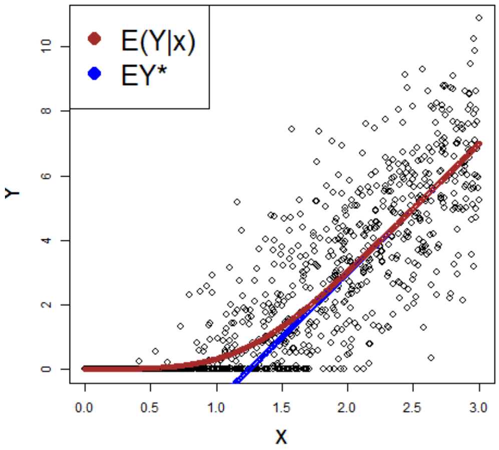

Fig. 1. Data containing many zero values and mean of the tobit model

The estimates of the parameters $\boldsymbol{\beta}$, ψ may be obtained by maximizing a logarithm of the likelihood function

$$\ln L(\boldsymbol{\beta},\psi \mid y) = \sum_{k:y_k=0} \ln\left(1-\Phi\left(\frac{\mathbf{x}'_k\boldsymbol{\beta}}{\psi}\right)\right) - \frac{1}{2}\sum_{k:y_k>0}\left(\ln 2\pi + \ln\sigma^2 + \left(\frac{y_k}{\psi}-\frac{\mathbf{x}'_k\boldsymbol{\beta}}{\psi}\right)^2\right)$$

for observed data y and known values $\mathbf{x}_k$ of the auxiliary vector. This function will be used in a traditional small area estimation framework.

3. Unit-level model for domains

We introduce a unit-level model for domains and then discuss the way of its estimation. The estimator suggested will be applied in the next section.

Let $U=\{1,2,\dots,N\}$ be a finite population, $s\subset U$ – probability sample, y – a study variable with values y_k, $k\in U$, having many zero values. Let our population consists of D domains $U=U_1\cup\dots\cup U_D$ of sizes $N_1,\dots,N_D$, $N_1+\dots+N_D=N$. The sample s is splitting into domains by portions $s=s_1\cup\dots\cup s_D$ of sizes $n_1,\dots,n_D$, $n_1+\dots+n_D=n$. We re-denote the study variable y_{dj}, j=1,..., N_d, and assume that a value $\mathbf{x}_{dj}$ of a unit-specific auxiliary vector $\mathbf{x}$ is available for each population element, j=1,..., N_D, d=1,...,D. The basic unit-level model is presented in Rao (2005). We modify it and assume the following unit-level model for a variable y:

$$y_{dj}=\mu_{dj}+v_{dj},\quad v_{dj}\sim N(0,\sigma^2),\ j=1,\dots,N_d,\ d=1,\dots,D; \tag{3}$$

$$\mu_{dj}=\max(0,\mathbf{x}'_{dj}\boldsymbol{\beta}+e_{dj}),\quad e_{dj}\sim N(0,\psi_d). \tag{4}$$

Here v_{dj} are independent, identically distributed (iid) random variables, e_{dj} are iid random variables, independent of v_d's, $\sigma > 0$, $\psi_d > 0$. This model allows some negative values for y because of random error in (3).

We introduce a model-based estimator for domain totals of a study variable *y*. First of all, model parameters $\boldsymbol{\beta}$, $\boldsymbol{\psi} = (\psi_1,...,\psi_D)'$, σ^2 have to be estimated. For this, we use Bayesian analysis.

Let us denote by f_{dj} density function of the observed domain data and the likelihood of the observed data *y* by

$$f(y\,|\,\boldsymbol{\beta},\boldsymbol{\psi},\sigma^2) = \bigcap_{d=1}^{D}\bigcap_{j=1}^{n_d} f_{dj}\left(y\,|\,\boldsymbol{\beta},\boldsymbol{\psi},\sigma^2\right).$$

Let us assume parameters $\boldsymbol{\beta},\boldsymbol{\psi},\sigma^2$ to be random. The un-normalized joined posterior density of these parameters may be expressed by a relation of proportionality

$$f(\boldsymbol{\beta},\boldsymbol{\psi},\sigma^2\,|\,y) \propto f\left(y\,|\,\boldsymbol{\beta},\boldsymbol{\psi}\right)\varphi_{0,\sigma}(e)\,f(\boldsymbol{\beta},\boldsymbol{\psi})\,f\left(\sigma^2\right),$$

and the logarithm of the joint posterior density

$$\ln f(\boldsymbol{\beta},\boldsymbol{\psi},\sigma^2\,|\,y) \propto \ln L(\boldsymbol{\beta},\boldsymbol{\psi}\,|\,y) + \ln\varphi_{0,\sigma}(e) + \ln f(\boldsymbol{\beta}) + \ln f(\boldsymbol{\psi}) + \ln f\left(\sigma^2\right),$$

with density functions *f* of the corresponding domain parameters and density $\varphi_{0,\sigma}$ of the normally distributed error *e*.

We assume a priori for estimation procedure that variances $\sigma^2, \psi_1,...,\psi_d$ are distributed "uniformly" on $(0,\infty)$, $\boldsymbol{\beta}$ components distributed normally with high variances *N*(0,1000).

An iterative procedure is used where the numerical approximation algorithm is maximizing the logarithm of the un-normalized joint posterior density building a Monte Carlo chain, and the estimated values $\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\psi}},\hat{\sigma}^2$ of the posterior density of $\boldsymbol{\beta},\boldsymbol{\psi},\sigma^2$ are obtained on each iteration. After the repetition of the iterative procedure *B* of times, the means $\hat{\hat{\boldsymbol{\beta}}},\hat{\hat{\boldsymbol{\psi}}},\hat{\hat{\sigma}}^2$ of the posterior distribution values $\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\psi}},\hat{\sigma}^2$ are taken as final estimates, and the unobserved values of a study variable are estimated by

$$\hat{y}_{dj} = \hat{E}y_{dj}\,\Big|_{\substack{\boldsymbol{\beta}=\hat{\hat{\boldsymbol{\beta}}}\\ \psi_d=\hat{\hat{\psi}}_d\\ \sigma=\hat{\hat{\sigma}}}} = \mathbf{x}'_j\hat{\hat{\boldsymbol{\beta}}}\Phi\left(\frac{\mathbf{x}'_j\hat{\hat{\boldsymbol{\beta}}}}{\hat{\hat{\psi}}_d}\right) + \hat{\hat{\psi}}_d\varphi\left(\frac{\mathbf{x}'_j\hat{\hat{\boldsymbol{\beta}}}}{\hat{\hat{\psi}}_d}\right). \qquad (5)$$

The model-based estimator of the domain total $t_d = \sum_{k\in U_d} y_k$ is then obtained by

$$\hat{t}_{\text{mod},d} = \sum_{j\in s_d} y_j + \sum_{j\in U_d\setminus s_d}\hat{y}_{dj}\,. \qquad (6)$$

This estimator will be applied in the next section.

For comparison, another model-based estimator of a population total may be used. The tobit model (1) is assumed for all population elements, $k\in U$. Its parameters are estimated by maximum likelihood method. y_{dj} are predicted by plug-in estimators of the parameters into (2) and $\hat{y}_{dj}$ are obtained. The estimator of total $\hat{t}_{ml.d}$ is then obtained by (6).

4. Simulation study

Data for a simulation study have been generated according to the tobit model.

Two domains U_1, U_2 of size $N_1 = N_2 = 500$ are simulated as population values of an auxiliary variable *x* by a uniform distribution in the interval (0,2). The values of a study variable *y* are simulated by (3), (4) with $\boldsymbol{\beta} = (\beta_1, \beta_2)' = (-2{,}3)'$, $\boldsymbol{\psi} = (\psi_1, \psi_2)' = (0.09{,}0.25)'$, and $\sigma^2 = 0.09$. Simple random samples s_1 and s_2 of size $n_1 = n_2 = 50$ are drawn independently from each domain. The parameters of the unit level model (3), (4) are estimated by Bayesian inference. $K = 10$ Monte Carlo chains of length *B*=7200 are simulated using a componentwise hit-and-run Metropolis iterative algorithm. The R package *LaplacesDemon* (Statistikat, 2014) has been used for this. The results obtained by (5), (6) are presented in Table 1.

For comparison, the model-based estimates $\hat{t}_{ml.d}$ and unbiased estimates of the total in the case of simple random sampling for domains $\hat{t}_{srs.d} = N_d / n_d \sum_{j \in s_d} y_{dj}$ are presented in Table 1.

The averages of K estimates and their accuracy measures are found by

$$\hat{\hat{t}} = \frac{1}{K}\sum_{i=1}^{K} \hat{t}_i\,, \quad Bias(\hat{t}) = \hat{\hat{t}} - t\,, \quad V\hat{a}r(\hat{t}) = \frac{1}{K}\sum_{i=1}^{K}\left(\hat{t}_i - \hat{\hat{t}}\right)^2, \quad rMSE(\hat{t}) = \sqrt{Bias^2(\hat{t}) + V\hat{a}r(\hat{t})}\,/\,\hat{\hat{t}}\,.$$

Table 1. Simulation results. Estimates for the domain totals

Do-main	*Total true*	$\hat{\hat{t}}_{srs}$	$\hat{\hat{t}}_{bayes}$	$\hat{\hat{t}}_{ml}$	*Bias srs*	*Bias Bayes*	*Bias ml*	*rMSE srs*	*rMSE Bayes*	*rMSE ml*
1	667	687	690	672	16	23	6	0.14	0.05	0.03
2	649	744	689	674	95	40	25	0.18	0.07	0.05
Popu-lation	1316	1431	1379	1268	111	63	51	0.11	0.04	0.03

5. Conclusions

- The estimates of σ very close to 0 show random term v_{dj} in (3) being insignificant.
- The model-based estimator using Bayesian inference seems to have lower variation than that of the design-based estimator for the total when the study variable contains many zero values.
- There is no essential difference in application of the one-level model and two-level model for the data studied. But accuracy of the estimates may depend on the data generated for simulation.

The model used here is the first approach to the small area estimation problem for the variable having many zero values. The idea presented in this paper will be developed further.

References

Amemiya T. (1984) Tobit Models: A Survey. *Journal of Econometrics*, v. 24, 3–61.

Greene W. H. (2003). *Econometric Analysis*. Prentice Hall, Upper Saddle River, 2003.

Krapavickaitė D. (2011). Some models for estimation of total of a study variable having many zero values, *Lithuanian Mathematical Journal*, v. 51(3), 370–384.

Rao J. N. K. (2003). *Small area estimation*, Hoboken, John Wiley & Sons.

Statisticat, LLC (2014). *LaplacesDemon: Complete Environment for Bayesian Inference*. R package version 14.04.05, URL http://www.bayesian-inference.com/software.

Tobin J. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica*, v. 26, 24–36.