Effect of using Tobit and Heckit models in regression estimation for data with many zeros

Tetiana Ianevych[1]and Veronika Serhiienko[2]

[1]Taras Shevchenko National University of Kyiv, e-mail: yata452@univ.kiev.ua
[2]Taras Shevchenko National University of Kyiv, e-mail: nichka_2009@ukr.net

Abstract

In the work we examine the effect of using the Tobit and Heckit models as assisting for the generalized regression estimator in order to improve it for data containing many zero values.
Keywords: Tobit model, Heckit model, regression estimation, excess of zeros in data.

1 Introduction

It is rather frequent situation when the economic data, especially microeconomic data, contain observations where some variable of interest is equal to zero for a number of the observations in the data set. Such data have excess of zero values and this can lead to a number of econometric problems when using Ordinary Least Squares (OLS) to estimate the unknown parameters of a regression model. We faced with this problem when start to work with Ukrainian capital expenditure survey.

One of the models that widely used in such situations is the Tobit model introduced by Tobin in 1958. It is developed for the censored data. For the data, suffering from big number of zeros but not caused by censoring, another models can be used – the Heckit model. We examined the usefulness and accuracy of these models utilizing general linear regression estimator (GREG). For this we use Monte Carlo simulation method measuring the efficiency with the Absolute Relative Bias and the Relative Root Mean Square Error.

2 Models for data with excess of zeros

The key decision facing any researcher working with a data set containing zeros is the choice of the appropriate model. The following summarizes the key elements of such a decision. Suppose that the variable of interest is y_i, and there are a large number of zero values for y in a given data set. The first step is to determine why the zeros are present in the data. There can be two alternatives:

(1) the zeros appear as a result of censoring or

(2) the zeros represent a decision that the researcher has no control over for some reason.

The first alternative usually corresponds to Tobit model whereas the second one – to Heckit model. Let us consider them in details.

2.1 Tobit Model

The Tobit Model was introduced by Tobin in 1958. The Tobit model, also called a censored regression model, is designed to estimate linear relationships between variables when there is either left- or right-censoring in the dependent variable. Formally, it can be written as

$$y = \begin{cases} y^*, if\ \ y^* > \tau \\ \tau_y\ , if\ y^* \leq \tau \end{cases}$$

where $y_i^* = X_i\beta + e_i$, $e_i \sim N(0, \sigma^2)$. The most common choice is $\tau = \tau_y = 0$.

The coefficient of such model can be calculated using the Maximum Likelihood method. In R there is a function "tobit()" in the *AER* package developed for this task. The Tobit Model allows different generalizations. For more information see Humphreys (2013).

2.2 Heckit Model

This type of model is appropriate when $y_i = 0$ because of the non-observable response. It means that knowledge $y_i = 0$ is uninformative in estimating the determinants of the level of y_k We can formulate it starting from the "participation" equation

$$z_i^* = \omega_i\gamma + u_i$$

$$z_i = \begin{cases} 1, if\ \ z_i^* > 0 \\ 0, if\ z_i^* \leq 0 \end{cases}$$

and continuing with "consumption" equation

$$y_i = \begin{cases} x_i\beta + e_i, if\ z_i^* > 0 \\ \quad 0\ , if\ z_i^* \leq 0 \end{cases}$$

with errors $u_i \sim N(0,1)$ and $e_i \sim N(0, \sigma^2)$, to be correlated in general case $corr(u_i, e_i) = \rho$. This specific terminology comes from Jones (1989) who investigated cigarette consumption.

The coefficient of the Heckit model can be calculated using the Maximum Likelihood method or 2 step method, developed by Heckman (1976). In R you can use the function "selection()" inside the *sampleSelection* package developed for this task.

3 Analysing the simulated data

So, we want to investigate the efficiency of the general regression estimator based on Tobit and Heckit models comparing to the classical Horvitz-Thompson and regression estimator based on classical linear model.

Our first simulated population U consists of N=1000 elements for which we produce the values of y_i as follows

$$y^* = -2.35 + 1.6578 \cdot x + e, \text{ where } e \sim N(0,1) \text{ and}$$

$$y = \begin{cases} y^*, & y^* > 0 \\ 0, & y^* \leq 0 \end{cases}.$$

The parameter of interest is the total $Y = \sum_{i \in U} y_i$. The underlying design is simple random sampling of size 100.

The main relative measure of efficiency for the estimators we used are:

the absolute relative bias

$$ARB = \left| \frac{1}{K} \sum_{k=1}^{K} \hat{Y}(s_k) - Y \right| / Y$$

and the relative root mean square error

$$RRMSE = \sqrt{\tfrac{1}{K} \textstyle\sum_{k=1}^{K} (\hat{Y}(s_k) - Y)^2} / Y.$$

Making K=1000 Monte-Carlo simulations we obtained the results given in the Table 1.

Table 1: Comparison of estimators

	Horvitz-Thompson estimator (%)	**GREG estimator LM assisted (%)**	**GREG estimator Tobit assisted (%)**
ARB	0.1394116	2.764075	0.185612
RRMSE	9.2875309	7.200544	7.073001

For the second population U consisting of N=1000 elements we simulated the values of y_i as $z_i^* = 1 + \omega_i + x_i + u_i + e_i$, where $u_i \sim N(0,1)$, $e_i \sim N(0,0.6)$,

$$z_i = \begin{cases} 1, z_i^* > 0 \\ 0, z_i^* \leq 0 \end{cases} \quad \text{and} \quad y_i = \begin{cases} 1 + x_i + u_i, & z_i^* > 0 \\ 0, & z_i^* \leq 0 \end{cases}.$$

After K=1000 Monte-Carlo simulations we obtained the following results.

Table 2: Comparison of estimators

	Horvitz-Thompson estimator (%)	**GREG estimator LM assisted (%)**	**GREG estimator Heckit assisted (%)**
ARB	0.8497651	14.86922	13.33495
RRMSE	27.8450915	23.30287	22.53748

Conclusion

As we can see, usage of GREG estimator leads to biased but better results with regards to the accuracy. The usage of the Tobit and Heckit-based estimators improve the quality of GREG estimator with regard to both bias and mean square error if the underlying processes of zero-values appearing corresponds well with estimator. If it does not correspond the improvement can be lost. And the main useful thing is that all theses GREG estimators can be used for the small area estimation.

References

Jones, A. M. (1989). A double-hurdle model of cigarette consumption. *Journal of Applied Econometrics.* Vol.**4**, No.1, 23-39.

Heckman J. (1979) Sample selection bias as a specification error. *Econometrica,* Vol.**7,** No. 1, 153-161.

Humpreys, B.R. (2013). *Dealing With Zeros in Economic Data.* University of Alberta, https://pdfs.semanticscholar.org/35c3/8229c8f7393acffc93b4a83120661df1f02c.pdf .

Tobin, J. (1958) Estimation of relationships for limited dependent variables. *Econometrica*, **26**, 24-36.