SOME APPROACHES IN ANALYZING THE DATA WITH EXCESS OF ZEROS

Tetiana Ianevych
Taras Shevchenko National University of Kyiv, Ukraine, yata452@univ.kiev.ua

I examine the different models and corresponding LM estimators for data containing many zero values and analyze their usefulness for designing sample survey of capital expenditure in Ukraine.

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.

In order to avoid underestimation of the capital expenditure value in the quarterly surveys the probabilistic sampling should be implemented into the investigation. Since the large and medium-sized enterprises are more valuable they are all observed every time. We survey by the means of the sample only the small and new enterprises. Thereby we focus on the investigation of small enterprises and perceive them as population.

I have made several attempts to incorporate into the designing and estimation process the additional information, e.g. the data on expenditure from the previous surveys, revenue, number of employee, etc. There were utilized the ordinary regression estimator and different robust estimators in order to obtain more accurate estimates not only for the estimate on the country level but also for different domain estimates. This led us to the problem of dealing with many zero values.

There are a number of econometric approaches to dealing with the problem of zeros. These approaches differ depending on the type of the data. If data is countable than statisticians often used such model as zero inflated Poisson or similar. Capital expenditure is not a count data, so we need to use the models that deal with semi-continuous data. It means that it has a continuous distribution except for a probability mass at 0. Good review paper devoted to all this models is a paper by Min&Agresti (2002).

So, we decided to compare the following estimators for the estimation of the capital expenditure in 2010: Horwitz-Tompson; GREG with and without log-transformation of independent variable which is capital expenditure in 2009; regression estimator based on Tobit model with capital expenditure in 2009 as a regressor; and a regression based on Heckit model capital expenditure in 2009 as a regressor in the outcome equation and identification variable on whether the enterprise had expenditure last year as a regressor for selection equation. We have made 1000 Monte Carlo simulation for all these estimators and calculated the ARB and RRMSE. The results of this simulation study are presented in the Table 1.

Table 1: Comparison of different estimators

	HT	GREG	Log-transform + GREG	Tobit	Heckit	Log-transfor + Heckit
ARB, %	0.58	10.18	0.72	24.18	4.09	0.41
RRMSE, %	41.70	36.02	37.77	51.18	40.88	41.41

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 fell short of expectations but still can be improved by changing or incorporating more independent variables. And the main useful thing is that all theses estimators except HT can be used for the construction of small area estimators.

References

Min, Yo., Agresti, (2002) A., Modeling nonnegative data with clumping at zero: A survey. JIRSS, Vol. 1, Nos. 1-2, 7-33.