Forum Statistika dan Komputasi : *Indonesian Journal of Statistics* available online at:
ISSN : 0853-8115 *journal.ipb.ac.id/index.php/statistika*
Observation

ALTERNATIVE SEMIPARAMETRIC ESTIMATION FOR NON-NORMALITY IN CENSORED REGRESSION MODEL WITH LARGE NUMBER OF ZERO OBSERVATION

(Case Study : LPG Demand On Household Sector)

Andres Purmalino[1], Asep Saefuddin[1], Hari Wijayanto[1]

Department of Statistics, Bogor Agricultural University, Indonesia

ABSTRACT

A large number of zero observation on the response variable in the socio-economic field are often found in household demand models. This will imply on the method to estimate parameters in the model used. Ordinary least square estimators of linear models to be biased and inconsistent. One model to overcome is using censored regression model is also know as tobit model. However, non-normality in the Tobit Estimators being inconsistent. Another alternative estimators is censor least absolute deviations (CLAD). CLAD estimator is consistent and asymptotically normal for a wide class of distribution. This study was to focus on the application of Tobit and Censored Least Absolute Deviations (CLAD) estimators for LPG demand. The data used is the LPG expenditure in rural areas in the provinces of West Java that the number zero observations is 39 percent of the sample. The result shows that CLAD and Tobit estimators are consistent estimators. But along with increasing the number of samples, the CLAD estimators performance is getting better than Tobit estimators.

Keywords : Zero observation, CLAD, Tobit, Consistent estimator, LPG demand

INTRODUCTION

Survey in socio-economic field are often conducted with a large sample. The collected data then be used by the consumer / researchers to arrange the models in the socio-economics field. In the field of socio-economics, ,zero expenditure often occurs in household expenditure. This is caused by the presence of households that do not consume certain types of goods and services, while other households consume in varying amounts.

The data structure where the response variable has a value of zero for most observations in the regression model will imply on the method to estimate parameters. The assumptions of normality and homoskedastistas required in the regression analysis tend not met. Disposing of zero observations in the analysis can throw existing information and do not reflect the actual situation.

One methods to overcome that are used is tobit models. However Tobit estimators is inconsisten if errors have non normality. As alternative, the method used was censored least absolute deviation (CLAD). This method is semiparametic estimators. The advantage of this method is a robust estimator in the presence of non-normality and heteroscedasticity.

MATERIALS AND METHODS

Data

The data used in this study were obtained from the SUSENAS (2013) and PODES (2011) by Central Board of Statistics. The data is about household LPG expenditure in rural areas in West Java. The LPG expenditure as the response variable, while explanatory variables consisted of 10 variables, namely expenditure household (x1), number of household members (x2) , generation y and z (x3), education level household head (x4), agricultural household (x5), the existence of forests (x6), distance of the town / village live to the district / city nearby (x7), and the presence of the agent / kerosene retail dealer (x9), the existence of wife works(x10). The number of zero observations on the response variable is about 37 percent.

The data used in this study were obtained from the SUSENAS (2013) and PODES (2011) by Central Board of Statistics. The data is about household LPG expenditure in rural areas in West Java. The LPG expenditure as the response variable, while explanatory variables consisted of 10 variables, namely expenditure household (x1), number of household members (x2) , generation y and z (x3), education level household head (x4), agricultural household (x5), the existence of forests (x6), distance of the town / village live to the district / city nearby (x7), existing of the kerosene agent / kerosene retail dealer (x8), existing of the LPG agent / LPG retail dealer (x9)the existence of wife works(x10). The number of zero observations on the response variable is about 39 percent.

Censored Regression

Censored regression model in standard tobit model was established by assuming there is a linear relationship between the response variable with the independent variables were expressed by the equation:

$$y_i^* = X_i\beta + u_i$$

$$y_i = y_i^*, \quad y_i^* > 0$$

$$y_i = 0, \quad y_i^* < 0$$

$$y_i^* \; is \; latent \; or \; unobservable \; dependent \; variable.$$

The density function of y_i is

$$f(y_i) = \left[\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{y_i - X_i\beta}{\sigma}\right)^2}\right]^{d_i} \left[\Phi\left(\frac{-X_i\beta}{\sigma}\right)\right]^{1-d_i}$$

$$= \left[\frac{1}{\sigma}\phi\left(\frac{y_i - X_i\beta}{\sigma}\right)\right]^{d_i} \left[1 - \Phi\left(\frac{X_i\beta}{\sigma}\right)\right]^{1-d_i}$$

Where d_i=1 if y>0 and d_i=0 if y=0, and $\phi(.)$ and $\Phi(.)$ be the density and distribution functions of the standard normal distribution. Censored regression parameter estimation is using Maximum Likelihood Estimation commonly called tobit estimators. The log likelihood function is

$$LL = \sum_{i=1}^{N}\left\{d_i\left(-ln\sigma + ln\phi\left(\frac{y_i - X_i\beta}{\sigma}\right)\right) + (1 - d_i)ln\left(1 - \Phi\left(\frac{X_i\beta}{\sigma}\right)\right)\right\}$$

Parameter estimation β and σ obtained using first and second derivatives of the log likelihood function using the Newton-Raphson iterations for nonlinear functions. Tobit estimators has been proven by Amemiya (1973) are consistent estimators and asymptotically normal.

Conditional Moment Test

It is well known that the OLS estimators in the standard normal regression model retains its consistency (although not its efficiency) when some of the basic assumption of normality are violation. In censored regression, violation assumptions of normality cause Tobit estimator becomes inconsistent (Long, 1997). One frequently used approach for testing normality in censored regression model is conditional moment test. The statistics conditional moment test is

$$\tau = \iota'\widehat{M}\left[\widehat{M}'\widehat{M} - \widehat{M}'\widehat{G}\left(\widehat{G}'\widehat{G}\right)^{-1}\widehat{G}'\widehat{M}\right]^{-1}\widehat{M}'\iota \sim \chi_{(r)}$$

where ι is an (Nx1) vector of ones, $\widehat{M}$ is the (Nxr) matrix of sample realization of the r moment restrictions, $\widehat{G}$ are the terms in gradient/score (Nxk) matrix of the log likelihood function. Newey (1985) noted that one benefit of τ is that could be easily calculated via an artificial regression with no constan,

$$1 = \widehat{M}\delta_1 + \widehat{G}\delta_2 + e$$

he noted that τ=N-SSR = N x R^2 , where N is the number of observation in the artifcial regression and SSR is the sum of squared residuals from this regression and R^2 is the coefficients of determination. For normality test, the null hypothesis is normality, reject if $\tau > \chi_{(2)}$.

Censored Least Absolute Deviation

Censored Least Absolute Deviation (CLAD) be alternative estimators if the assumptions of normality are violated on

Tobit estimators. Powell (1984) has developed semiparametric estimators to relax the strict assumptions which were needed to estimate censored regression models, which are defined as follow

$$\beta_{CLAD} = \underbrace{argmin}_{\beta_0\beta_1\ldots\beta_p} \sum_{i=1}^{n} |Y_i - max\{0, x_i\beta\}|$$

The estimators are called semiparametric estimation because its combine parametric and non parametric component. As the uncensored mean $x_i\beta$ is parameterized but the error distribution is not. CLAD estimator has the advantages of not sensitive to outlier data, is able to produce robust estimates. The estimation technique used in clad for the CLAD estimator is Buchinsky's (1991) iterative linear programming algorithm (ILPA).) The first step of the ILPA is to estimate a LAD for the full sample, then delete the observations for which the predicted value of the dependent variable is less than zero. LAD is estimated on the new sample, and again negative predicted values are dropped. More generally, observations are dropped if the predicted value is less than the censoring value when the left tail of the distribution is censored, or they are dropped if the predicted value is greater than the censoring value when the right tail of the distribution is censored. Buchinsky (1991) shows that if the process converges, then a local minimum is obtained. Convergence occurs when there are no negative predicted values in two consecutive iterations.

Consistent Estimators

It often happens that an estimator does not satisfy one or more of the desirable statistical properties in small samples. But as the sample size increases indefinitely, the estimator possesses several desirable statistical properties. These properties are known as asymptotic properties of the sample. One of the asymptotic properties of the sample is consistent estimators. Variable is said to be a consistent estimators if along with the increase in the number of samples, the variance of parameter estimators getting smaller near zero In this study a simulation done by resampling on the data with a sample of 300, 1000, 3000 and 5000 from 8035 sample. Each sample was repeated 30 times.

RESULTS AND DISCUSSION

The results of estimation coefficients for LPG demand in the Tobit and CLAD can be seen in the following chart :

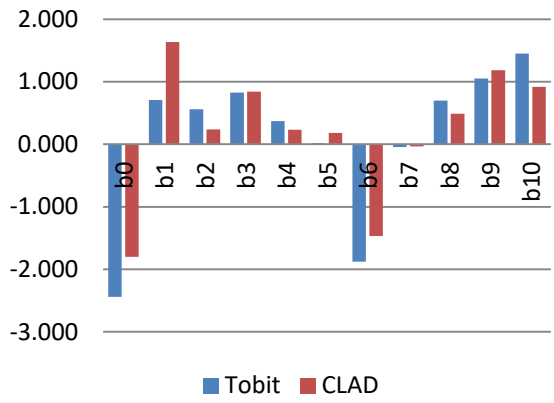
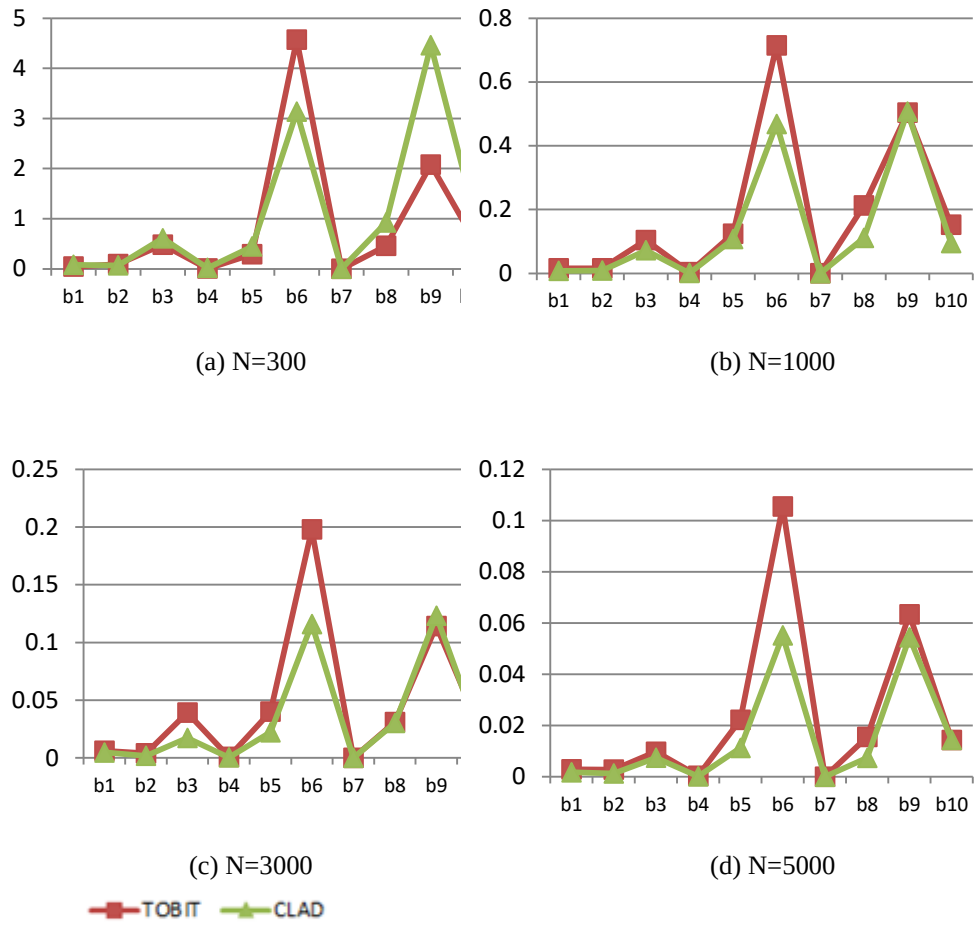


Chart 1 Coeffisient Value of Estimators TobitE and CLAD

Tobit and CLAD estimators showed coefficients estimation values are different. However, both methods showed similarity in relation with the independent variables to the response variable to see the positive and negative values of each coefficient.

Based on conditional moment test the censored regression does not meet the assumptions of normality. Statistic value of conditional moment test is 85.62, is higher than chi-square table at the 0.05 significance level, with 2 degree of freedom. So the null hypothesis, that error normally distributed is rejected. This result indicates that the Tobit estimators becomes inconsistent estimator.

The simulation results showing the method of Tobit and CLAD produce consistent estimators of parameters. The variance estimators of each method showed smaller value with the addition of the sample. However, along with increasing the number of samples, the CLAD estimator showed good performance than Tobit estimators, the chart below shows that.



(a) N=300 (b) N=1000

(c) N=3000 (d) N=5000

Graph 1. Variance of Tobit and CLAD estimators for 300, 1000, 3000, and 5000 Samples

Based on assumption and simulation test, model of LPG demand in the household sector in the rural areas of West Java is using CLAD estimators. Tobit estimators as parametric estimation not meet the assumption and can have an impact on the resulting estimator. While CLAD as semiparametric estimator does not require assumptions as parametric estimator. The simulation results also show CLAD have more accurate estimators along with a large number of samples. The equation LPG demand model is :

$$Y= -1.8X1 + 1.636X2 + 0.237X3 + 0.839X4 + 0.229X5 + 0.179X6 - 1.468X7 - 0.035X8 + 0.486X9 + 1.185X10$$

The regression coefficient is negative occurs at the variable distance (X6) and the forest area (X7), while the other variable is positive. This indicates that the decline in demand for LPG household sector in rural areas caused by distance and forest area.

CONCLUSION AND REMARKS

According to the result, we can conclude :

1. The model of LPG demand in the household sector in the rural areas of West Java using CLAD estimator is more consistent than MLE.
2. LPG demand decline occurred in households near forest and away from the capital of the district / city.
3. After all, necessary to add resampling and repetition in the simulation to obtain consistent estimators of parameters in order to obtain results in a more precise

REFERENCES

Amemiya, T. 1984. Tobit models: A survey. Journal of Econometrics 24: 3-61

Barnes, Shahidur, Husain. 2010. Policy Research Working Paper 5332. Energy

Access, Efficiency, and Poverty : The World Bank Development Research Group.

Cameron, Trivedi. 2005. Microeconometrics (Methods and Aplications). Cambridge University Press. United States Of America.

Deaton, A. 1988. The analysis of household surveys. John Hopkins University Press. London.

Draper, N R., Smith H. 1981. Applied regression analysis. Second Edition. John Wiley & Sons, Inc. New York.

Greene, H W. 2003. Econometric Analysis. Fifth Edition. New York University

Gujarati,N D. 2012. Dasar Dasar Ekonometrika. Edisi 5. Buku 2. Salemba Empat

Maddala, G S. 1983. Limited dependent and qualitative variables in econometrics. Cambridge University Press. New York.

Pranadji KD, Djamaludin DM, Kiftiah N. 2010. Analisis Perilaku Penggunaan LPG Pada Rumah Tangga Di Kota Bogor. Jurnal Ilmu Kel. & Kons., Agustus 2010, p:173-183. Vol.3, No. 2. ISSN : 1907-6037.

Tobin, J. 1958. Estimation of relationships for limitied dependent variables. Econometrica Vol 26:24-36

Virgantari, F. 2005. Perbandingan model Tobit, regresi terpotong dan regresi biasa pada data konsumsi rumah tangga. Bogor. Fakultas Matematika dan Ilmu Pengetahuan Alam. Institut Pertanian Bogor.