

# Analysis of Household Expenditures on Education in Semarang City, Indonesia Using Tobit Regression Model

Alan Prahutama<sup>1\*</sup>, Agus Rusgiyono<sup>1</sup>, Sudarno<sup>1</sup>, Moch. Abdul Mukid<sup>1</sup> and Tatik Widiharini<sup>1</sup>

<sup>1</sup>Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Tembalang, Semarang

**Abstract.** Data on household expenditure for education is censored data. One method used to analyze this censored data is Tobit regression. Tobit regression changes the response variable which was originally a nominal scale into a mixed scale measurement. Tobit regression can overcome the non-biased categorical regression model. This study aims to analyze the factors that influence household expenditure on education in Semarang city. The dependent variable used in this study is household expenditure for education. The independent variables used include the Education of the Head of the Household, Occupation of the Head of the household, number of household members, Number of Working Household Members, the proportion of household members who attend school  $\leq$  in Junior High School, Senior High School and College, food expenditure in households and regions. Based on the Tobit regression model obtained, the factors that influence household expenditure for education include the number of household members (ART), Number of Working Household Members, the proportion of household members who attend school  $\leq$  in Junior High School, Senior High School and College, food expenditure in households. Generate market research especially in customer satisfaction.

**Keywords:** Tobit Regression Model; Household expenditures on education.

## 1 Introduction

Education is a right that should be obtained by every Indonesian citizen. It is written in the constitutional foundation of the Indonesian state, namely the 1945 Constitution. The Indonesian government requires 9 years of education, namely Elementary and Secondary Schools by Government Regulation No. 47 of 2008. The existence of government assistance does not mean that all households do not pay for education at all. The amount of education costs incurred by these households is very variable. The difference in the number of costs incurred can be influenced by the characteristic differences between households [1].

Households that do not pay for education due to various reasons. The existence of government assistance in the form of BOS (School Operational Assistance) can be one reason for a household not to pay for education. Not all households have household members who are studying. It can be because there are indeed no household members who are in compulsory education or because of the economic conditions of the household that force household members not to study. The poverty factor is one of the things that is thought to have an effect on education expenditure in attractive households for further research [2].

Household expenditure for education is ratio scale data. In reality, some households do not pay for education at all. It means that there are education households whose household expenditure for education

is in the lower limit, which is worth 0. The existence of households whose expenditure is within this limit causes the data to be censored. To solve a regression problem that contains censored data, Tobit regression method is used. This method will reduce the bias of the censored data when compared to the classic linear regression model. Research with Tobit regression has been developed by several researchers. Research using Tobit regression including Jeryana, Kencana, and Gandhiadi [3] uses Tobit regression to model the consumption of liquid milk in the province of Bali. Austin [4] use the Tobit Model to identify the status of Health. The objectiveness of this paper is modelling the expenditures of education in Semarang city using Tobit's regression Model.

## 2 Tobit Regression Model

Tobit's regression model was introduced by James Tobin in 1958. Tobin [5] used data from the interview section of the 1952 and 1933 Consumer Financial Surveys as observations. Data as the dependent variable is a household expenditure in the United States to buy a car [6]. These variables are divided into two categories, where households that do not buy cars mean zero, while households that buy cars have a value of expenditure (an absolute value and not zero).

According to Greene [7] Tobit regression model can be defined as follow as:

\* Corresponding author: [alan.prahutama@gmail.com](mailto:alan.prahutama@gmail.com)

$$y_i = \begin{cases} y_i^* = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, & \text{for } y_i > c \\ c & , \text{ for } y_i \leq c \end{cases} \quad (1)$$

where  $y_i$  dependent variable and censored data, the value of  $y_i$  will be worth  $y_i^*$  if the value of  $y_i$  higher than constant of  $c$  and will be writhed  $c$  if the value of  $y_i$  less than constant of  $c$  or same as  $c$ . While  $\mathbf{x}_i^t$  is vectoa r of independent variables with  $\mathbf{x}_i^t = [1 \ X_{1i} \ X_{2i} \ \dots \ X_{pi}]$ ,  $\boldsymbol{\beta}$  is vector parameter of independent variables with the value of  $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \beta_2 \ \dots \ \beta_p]^t$ , and  $\varepsilon_i$  is error of  $y_i$  where compared based on estimation of  $y_i$  where the the value of  $\varepsilon_i$  is assumed normal distribution with zeroa mean and variance constant  $\sigma^2$  or can be written as  $\varepsilon_i \sim N(0, \sigma^2)$ . In Tobit regression model the value of  $c$  assumed 0 so in Eq. (1) can be written as [8]:

$$y_i = \begin{cases} y_i^* = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, & \text{for } y_i > 0 \\ 0 & , \text{ for } y_i \leq 0 \end{cases} \quad (2)$$

Parameter estimation of the Tobit model uses the Maximum Likelihood Estimator (MLE). By using MLE, the Likelihood function for the Tobit model with  $c$  as the sensor point is formulated as follows:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n_1} \Phi\left(\frac{c - \mathbf{x}_i^t \boldsymbol{\beta}}{\sigma}\right) \prod_{i=n_1+1}^{n_2} \phi\left(\frac{y_i - \mathbf{x}_i^t \boldsymbol{\beta}}{\sigma}\right) \quad (3)$$

The likelihood function in equation (3) consists of two parts, namely the function for censored data that runs from the 1<sup>st</sup> to the  $n_1$  data and the function for non-censored data that runs from  $n_1 + 1$  data to  $n_2$  data. In order to estimate the parameters of the Tobit Regression, equation (3) than in and down to the parameters to be searched is then equated with zero. However, because

the final result of the first derivative of the parameter is in the form of openform, it cannot be searched directly but must use Newton-Raphson iterations (NR). The NR iteration is written mathematically as follows [9]:

$$\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{(l)} - [\mathbf{H}^{(l)}]^{-1} \mathbf{q}^{(l)} \quad (4)$$

where,

$$\boldsymbol{\theta} = [\beta_0 \ \beta_1 \ \beta_2 \ \dots \ \beta_p]^t$$

$$\mathbf{q} = \left[ \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \beta_0} \quad \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \beta_1} \quad \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \beta_2} \quad \dots \quad \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \beta_p} \right]^t$$

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_0^2} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_0 \partial \beta_2} & \dots & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_0 \partial \beta_p} \\ \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_1^2} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_1 \partial \beta_2} & \dots & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_1 \partial \beta_p} \\ & & & \ddots & \\ \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_0 \partial \beta_p} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_1 \partial \beta_p} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_2 \partial \beta_p} & \dots & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_p^2} \end{bmatrix} \quad (2)$$

NR iteration will stop when the value  $\|\boldsymbol{\theta}^{(l+1)} - \boldsymbol{\theta}^{(l)}\| \leq \delta$ , where  $\delta$  is very sa mall value.

### 3 Methodology of Research

The type of data used in this study is secondary data taken from the 2016 Susenas Data issued by the Central Java Central Bureau of Statistics (BPS). The data taken is data on household expenditure for education (Y) in Semarang city. In additional environmental factors, as independent variables, used for analysis in household modeling for education include:

**Table 1.** Independent Variables

No	Variable	Scale	Category	No	Variable	Scale	Category
1	Education of the Head of the Household ( $x_1$ )	Ordinal	1. Un Graduate in Elementary School 2. Elementary School 3. Junior High School 4. Senior High School 5. Collage	6	The proportion of household members who are taking education in Senior High School ( $x_6$ )	Ratio	-
2	Occupation of the Head of the Household ( $x_2$ )	Nominal	1. No Working 2. Working	7	The proportion of household members who are taking education in University ( $x_7$ )	Ratio	-
3	Number of household members ( $x_3$ )	Ratio	-	8	Household Food Expenditures ( $x_8$ )	Ratio	-
4	Number of Working Household Members ( $x_4$ )	Ratio	-	9	Region ( $x_9$ )	Nominal	1. Village 2. City
5	The proportion of household members who are taking education in $\leq$ Senior High School ( $x_5$ )	Ratio	-				

## 4 Results and Discussions

In Tobit regression modeling, the education variable of the head of the household will be transformed into the old variable ratio scale education. Another absolute scale variable is the work status of the head of the household and the area transformed into a dummy variable. The transformed variables are used in Tobit regression modeling as follows:

**Table 2.** Transformation of Categorical Variable

Variable	Category	Transformation Variable
Education of the Head of the Household ( $x_1$ )	1. Un Graduate in Elementary School	0
	2. Elementary School	6
	3. Junior High School	9
	4. Senior High School	12
	5. Collage	16
Occupation of the Head of the Household ( $x_2$ )	No Working	0
	Working	1
Region ( $x_9$ )	Village	0
	City	1

Tobit regression modeling in this discussion is divided into two, namely individual Tobit modeling and multiple Tobit regression modeling. Individual Tobit regression modeling is done in each independent variable while multiple Tobit regression modeling is done in all independent variables.

### 4.1 Individual Tobit Modelling Regression

Modeling of factors that influence household expenditure for education begins with conducting individual modeling for each variable. These individual modeling results will be tested to see the effect of individual predictor variables on household expenditure on education.

The hypothesis used for this individual modeling is the partial test hypothesis as follows:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0, j = 1, 2, \dots, 9$$

**Table 3.** Modeling Tobit Regression Model Individually

Variables	Estimation of Parameter	P-value	Conclusion
Intercept	0.3456	0.0043	Not significant
Education of the Head of the Household ( $x_1$ )	0.25678	0.3231	
Intercept	-1.278	0.00056	Not significant
Occupation of the Head of the Household ( $x_2$ )	0.3567	0.265	

Intercept	0.7875	0.00871	significant
Number of household members ( $x_3$ )	0.669	0.0497	
Intercept	1.008	0.0065	significant
Number of Working Household Members ( $x_4$ )	0.979	0.0016	
Intercept	0.0567	0.0078	significant
The proportion of household members who are taking education in $\leq$ Junior High School ( $x_5$ )	1.8643	0.000	
Intercept	-2.087	0.0006	significant
The proportion of household members who are taking education in Senior High School ( $x_6$ )	3.045	0.0000	
Intercept	1.087	0.0006	significant
The proportion of household members who are taking education in College ( $x_7$ )	4.788	0.000	
Intercept	0.655	0.0067	significant
Household Food Expenditures ( $x_8$ )	1.0889	0.00564	
Intercept	0.076	0.234	Not significant
Region ( $x_9$ )	0.0156	0.876	

Table 3 shows the results of the estimated Tobit regression model individually for 9 independent variables. Based on the results of the analysis obtained by comparing to show the variables that are not significant include the length of education of the head of the household, the work of the head of the household, and the region. Based on individual modeling, the variable that has the most significant influence is the proportion of household members who attend school either elementary, junior high, high school or college. All variables have coefficients marked positive. This positive coefficient means that the higher the value of the variable, the higher the expenditure of households for education.

### 4.2 Multiple Tobit Regression Model

Multiple modeling in Tobit regression is done by modeling all independent variables simultaneously based on significant individual modeling. The parameter estimation results from this multiple modeling will be tested using simultaneous testing. Partial tests were carried out using the Wald test as in individual

modeling. While simultaneous testing is carried out with the following hypothesis:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_9 = 0$$

$H_1$ : At Least There is one of  $\beta_j \neq 0, j = 1, 2, \dots, 9$

The test statistic used is the G test statistic, with a G value of 267,453 compared to the value of Chi-Square table  $\chi^2_{0.05;6} = 12.59$ . Rejected  $H_0$  if  $G > \chi^2_{0.05;6}$ , then can be concluded that  $H_0$  is rejected, so the Tobit Regression model is significant. In addition to testing the significance of each parameter using the Wald statistical test. The results of the tests are explained in Table 4.

**Table 4.** Multiple Tobit Regression Model

Variable	Parameter Estimate	P-value	Conclusion
Intercept	2.4451	0.00235	Significant
Number of household members ( $x_3$ )	0.3067	0.0467	Significant
Number of Working Household Members ( $x_4$ )	-0.62404	0.00567	Significant
The proportion of household members who are taking education in $\leq$ Junior High School ( $x_5$ )	1.023	0.00034	Significant
The proportion of household members who are taking education in Senior High School ( $x_6$ )	0.9876	0.00021 3	Significant
The proportion of household members who are taking education in College ( $x_7$ )	2.6054	0.00043	Significant
Household Food Expenditures ( $x_8$ )	0.08765	0.0056	Significant

Table 4 shows the test parameters of the Tobit regression model partially. Based on the table all predictor variables are significant. Parameter coefficients are positive, meaning that the more the number of household members, the more the proportion of household members in Junior High School, Senior High School, and college, and the more expenditure to eat

then shows the greater household expenditure for education. While the parameter coefficient of the number of working household members is negative, which means that the more household members who work show that there is less household expenditure for education. It means that household members do not spend education costs while working.

Based on Table 4, it can be concluded that the proportion of household members who are currently in tertiary education gives the most significant contribution to household expenditure for education. While food expenditure provides the smallest contribution to household expenditure for education in Semarang city.

## 5 Conclusion

Based on the results and discussions that have been carried out, it can be concluded that Tobit regression modeling for household expenditure in education factors that influence include number of household members, Number of Working Household Members, Proportion of household members who are taking education in  $\leq$  Junior High School, Senior High School and College and Household Food Expenditures. The most significant contribution to the high cost of household expenditure for education is the Proportion of household members who are taking education in College.

## Acknowledgment

This Research Supported by Diponegoro University as a founding in "Usul Riset Madya 2018" Statistics Department Faculty of Sciences and Mathematics.

## References

1. P. Alberto, Journal of The Pacific Economic **16**, 233
2. T. Amemiya, Journal of Econometrics **24**, 61 (1984)
3. I. Jernaya, I. Kencana, G. Gandhiadi, E-Jurnal Matematika **3**, 75 (2014)
4. P. Austin, M. Escobar, J. Kopeck, Quality of Life Research **9**, 901 (2000)
5. J. Tobin, Econometrica **26**(1), 24-36 (1958)
6. L. Li, J. Simonoff, C. Tsai, Statistical Modelling **7**, 107 (2007)
7. W. Greene, *Censored Data and Truncated Distribution*, New York University : New York (2005)
8. Y. Adelekan, A. Omotayo, The Journal of Developing Area **51**318 (2017)
9. H. Guo, Y. Zhao, T. Niu, K. Tsui, Plos ONE **8**(1) (2017)