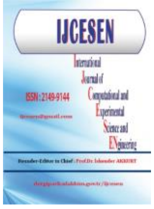
International Journal of Computational and Experimental Science and ENgineering (IJCESEN)
Vol. 6-No.1 (2020) pp. 63-74
http://dergipark.org.tr/en/pub/ijcesen



ISSN: 2149-9144

Research Article

Dependent Dummy Variable Models: An Application of Logit, Probit and Tobit Models on Survey Data

Öznur İŞÇİ GÜNERİ[1], Burcu DURMUŞ[2]*

[1]Muğla Sıtkı Koçman Üniversity, Faculty of Science, Department of Statistics, 48000, Muğla-Turkey
[2] Muğla Sıtkı Koçman Üniversity, Rectorate Unit, 48000, Muğla-Turkey

* **Corresponding Author**: burcudurmus@mu.edu.tr
ORCID: 0000-0002-0298-0802

Article Info:

DOI: 10.22399/ijcesen.666512
Received : 27 December 2019
Accepted : 17 March 2020

Keywords

Logit model
Probit model
Tobit model
Marginal effects
Information Criteria

Abstract:

In the current study, logit, probit and tobit models which are commonly used among dependent dummy variable models are included. These models are also known as limited dependent variable models in the literature. Surveys, which are widely used in the field of social sciences, are carried out with limited options due to their nature. Linear regression models cannot be used in statistical estimations since they do not provide assumptions in limited analysis. In this case, different regression models may be preferred. The main purpose of this study is to compare the Tobit model used in censored data and the binary logit and binary probit regression models derived from this model. For this purpose, analysis were conducted on survey data. Logit, probit and Tobit model coefficient estimates and marginal effects were calculated. In addition, AIC and BIC values were obtained from the model selection criteria for these 3 models.

1. Introduction

Categorical models whose dependent or explained variables are coded as “0” and “1” are called as two-ended or dummy dependent variable models. Models in which the dependent variable consists of answers having two categories such as whether the consumer buys a product or not, whether an individual participates in the labor market or not, yes-no, successful-unsuccessful, male-female can be given as examples. When qualitative variable models can take two such values, the first models that come to mind are linear probability model (LPM), logit and probit models. In the linear probability model which is one of these models the most obvious problem is that the estimated probability values fall outside the range of “0” and “1”. The problems encountered in linear probability models and detailed information about such models are explained by Gujarati [1] and Aldric and Nelson [2]. Logit and probit models are the most widely used models for estimating the functional relationship between dependent and independent variables in practice.

Logit and probit models are also among the generalized linear models (GLM) family. If the latent variable is unobserved or the dependent variable is binary, this model cannot be estimated using the normal least squares method (OLS). Instead, the maximum probability estimate is used which requires assumptions about the distribution of errors. Often, the choice is between the normal errors in the probit model and the logistic errors in the logit model [3].

If independent variables can be observed in regression models where the range of change of the dependent variable is limited in any way, the censored model is a possibility. However, if observations outside a certain range are completely lost, there is a discrete model. Limited dependent variables are generally divided into two groups: censored and truncated regression models. The Tobit model is also known as the censored regression model. When the dependent variable is censored, least squares estimates give biased results. Therefore, when censored is applied to the dependent variable, the Tobit model allows us to derive consistent and asymptotically efficient predictors.

Tobit model, which is known as models where the dependent variable has a lower or upper limit, was first used by Tobin to analyze household expenditures by working on durable consumer goods considering the fact that expenditure cannot be negative. Tobin's idea was to change the probability function to reflect the probability of unequal sampling for each observation depending on whether the latent dependent variable rises or falls below the specified threshold [4]. This standard Tobit model was later developed.

When it is known that the error terms for Tobit models are normally distributed, maximum similarity and other similarity-based processes yield consistent and asymptotically normally distributed estimators. However, when the assumed parametric form of the similarity function is incorrectly determined, the estimators become inconsistent [5].

In this study, the Tobit model used for censored data and two-option logit and probit models which are derived from censored dependent variable are emphasized. In the second part of the study, logit, probit and Tobit models are introduced; in the third section, the formulations of the marginal effects used in the coefficient interpretation of the models are shown and in the fourth section, the model selection criteria used in the application are given. In the fifth part of the study, model analyses were carried out based on the family income and expenditure survey data and the results of the analyses were explained at the end of the study.

2. Dependent Dummy Variable Models

In this study, logit, probit and tobit models are studied. These models are described below.

2.1. Logit Model

In the logit regression model, none of the assumptions (linear distribution of the dependent variable, withdrawal of independent variables from normal distribution, normal distribution of the error term and no relationship between error term values, etc.) involved in the linear regression analysis are not sought. Therefore, it provides researchers with considerable flexibility and has become a more preferred method. A general linear regression model can be written as expressed in Equation 1, where y_i is a dependent variable and x_i is an independent variable.

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i \qquad (1)$$

In the above model, α constant term and β are regression coefficients. This model can be predicted by classical OLS when the dependent variable is continuous. However, logit or probit regression methods are used in cases where the dependent variable is discrete [6].

The Logit model can be used to model the probability of a particular class or event with two states. Suppose that the unobservable or latent variable generated from the observed variable y_i between $-\infty$ and $+\infty$. is y_i^*. Values greater than y_i^* are considered $y_i = 1$ and values less than or equal to y_i^* are considered $y_i = 0$. The latent variable y_i^* is assumed to be linearly dependent on the observed x_i throughout the structural model. y_i^* is connected to the binary variable y_i observed by the measurement equation in Equation 2:

$$y_i = \begin{cases} 1, & y_i^* > \tau \\ 0, & y_i^* \leq \tau \end{cases} \qquad (2)$$

Where τ is the breakpoint or threshold value. If $y_i^* > \tau$ is $y_i = 1$ and $y_i^* \leq \tau$ it takes $y_i = 0$. The relationship between observed y_i and latent y_i^* is shown in Fig. 1.

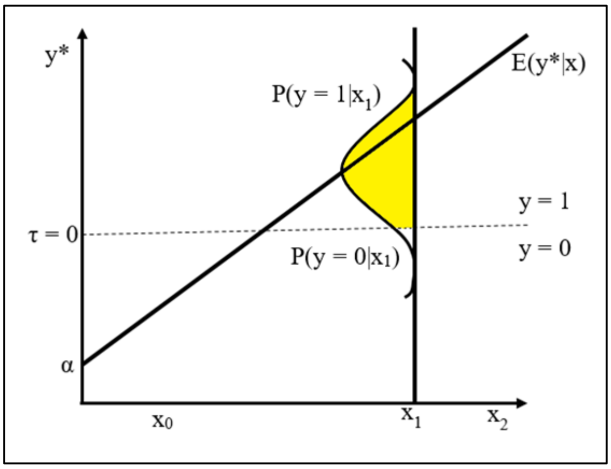


Figure 1. *Distribution of y_i^* according to $x_i$values*

Thus, when the dependent variable y_i takes “0” and “1”, binary logit takes the name of the model. When the dependent variable is “1”, the probability is expressed by Equation 3:

$$P_i = E(y = 1|x_i) = \frac{1}{1+e^{-(\alpha+\beta x_i)}} = \frac{1}{1+e^{-Z_i}} \qquad (3)$$

In this model, P_i provides information about the argument x_i while the first individual expresses the probability of making a particular choice. Thus P_i also takes values between “0” and “1”. The equations given in Equation 4 and Equation 5 can be written here:

$$P_i = 1 \; by \;\; ln\left(\frac{P_i}{1-P_i}\right) = ln\left(\frac{1}{1-1}\right) = ln\left(\frac{1}{0}\right) = +\infty \quad (4)$$

$$P_i = 0 \; by \;\; ln\left(\frac{P_i}{1-P_i}\right) = ln\left(\frac{0}{1-0}\right) = ln\left(\frac{0}{1}\right) = -\infty \quad (5)$$

To determine the logit function, α and β parameters cannot be directly predicted by OLS and Equation 6 is used to estimate the model:

$$1 - P_i = 1 - \frac{1}{1+e^{-(\alpha+\beta x_i)}} = 1 - \frac{1}{1+e^{-Z_i}} \quad (6)$$

If equations (3) and (6) are proportional,

$$\frac{P_i}{1-P_i} = e^{Z_i} \quad (7)$$

Equation 7 is obtained. It is also the odds or odds ratio (Odds Ratio, OR). Variables close to 1 among these OR values are not the factors that have a significant effect on the change of y. For OR values greater than 1, it is interpreted that the factor is an important risk factor provided that the coefficient is significant. Values close to zero indicate that the factor is an important risk factor, provided that the coefficient is significant, but that it is a negative factor that causes the y to take low values [7]. Equation 8 can be written by taking the natural logarithm of this model according to "e" base:

$$L_i = ln\left(\frac{P_i}{1-P_i}\right) = Z_i = \alpha + \beta x_i \quad (8)$$

L_i is the difference rate logarithm and is linear with respect to both x_i and parameters. Here L_i is called the "logit model" [1]. This model is a semi-logarithmic function. Therefore, the logit model is one of the best known models among generalized linear models.
In order to estimate the parameters in the model, when the L_i function, $P_i = 1$ and $P_i = 0$ are put in their places in logit L_i, then $ln(1/0)$ and $ln(0/1)$ values are obtained which are insignificant. Estimates of the parameters in the L_i function cannot be found by OLS but these parameters can be estimated by the maximum likelihood model (ML). However, the following points should be taken into consideration in research using logit model [8]:

- All appropriate independent variables should be included in the model: Failure to include some variables in the model may cause the error term to grow and the model to be inadequate.
- All unsuitable independent variables should be excluded: Inclusion of causally inappropriate variables in the model can complicate the model.
- Observation should be done on the same individual once and there should be no repeated measurements.
- The measurement error in the independent variables must be small: measurement errors should be small, no missing (missing) data. Errors can lead to bias in estimating coefficients and inadequacy of the model.
- There should be no multicollinearity between the independent variables: The independent variables must not be interrelated.
- There should be no extreme values: As with linear regression, extreme values can significantly affect the result.

In the Logit model, the coefficients cannot be directly interpreted as the effect of a change in independent variables on the expected value of the dependent variable. For this reason, OR values or marginal effects can be calculated in applications. Furthermore, the sign of the coefficients indicates the direction of the relationship between the argument and the probability of occurrence of the event.
The logit model is tested with the "chi-square test" and the existence of each independent variable in the model is tested by "Wald test statistics". However, in cases where there is a classification and assignment process and where normal distribution assumption and continuity assumption are not prerequisite, data should be analyzed with logit model.

2.2. Probit Model

In the linear probability model, which is one of the qualitative preference models with qualitative variables that can take two values, the most obvious problem is that the predicted probability values fall outside the range of "0" and "1". One of the models used to solve this problem is the probit model. This model is a nonlinear model in terms of coefficients that allows the probabilities to remain between "0" and "1". When the dependent variable y_i is binary, P_i is expressed in Equation 9:

$$P_i = E(y = 1|x_i) = \phi(x_i\beta) \quad (9)$$

Here ϕ is the cumulative distribution function and β maximum likelihood coefficients of the standard normal distribution. The probit model assumes that the basic dependent variable is normally distributed, whereas the y dependent variable assumes that the variable is based on the logistic curve. Therefore, the tail regions of the logit cumulative distribution function of these two models are wider than those of the probit model. Although these two models give similar results, it is not possible to directly compare the predicted main mass coefficients of the two models. However, they can be compared with a coefficient proposed by Amemiya [9].

Provided that it does not fall outside the range "0" and "1", a model should be found so that the relationship between P_i and x_i is curvilinear: increases in x_i also increase P_i. The illustration of the model with the above two features is given in Fig. 2:

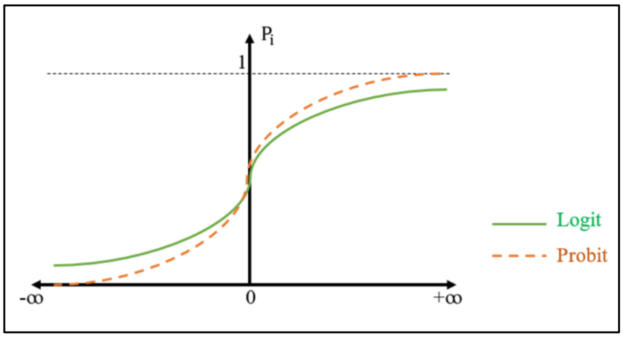


Figure 2. *Logit and probit cumulative distribution*

The probit model utilizes the cumulative normal distribution function and is called the "normit model" in the literature. Since the probit model is based on the utility theory developed by [10], the model depends on the unobservable utility index (I_i). If adapted as a model with a latent variable, the probit probability model based on the normal cumulative distribution function can be represented by Equation 10:

$$y_i^* = I_i = \alpha + \beta x_i \quad (10)$$

Where x_i is observable but y_i^* is not observable. As in the Logit model, if $y_i = 1$ then $y_i^* > 0$, but if $y_i^* < 0$ then $y_i = 0$. When assigning the result of the variable y_i, the value of τ used as the threshold value is generally taken as "0" and another number value can be used instead of zero [11]. Considering that y_i has a threshold value that cannot be observed as it is and is expressed as y_i^*, it can be said that if y_i exceeds the value $y_{i,}^*$ the event will occur and if it does not, the event will not occur (Equation 11).

$$y_i = \begin{cases} 1, & y_i^* > 0 \\ 0, & otherwise \end{cases} \quad (11)$$

The case that y_i^* is less than or equal to y_i is calculated from standardized cumulative distribution functions under the assumption of normality. If $\phi(Z)$ cumulative normal distribution function is defined as $\phi(Z) = P(Z \leq z)$ for the normal standard variable Z, then Equation 12 and Equation 13 are expressed as follows:

$$P(y_i = 1) = 1 - \phi\left(\frac{-\alpha - x_i\beta}{\sigma}\right) \quad (12)$$

$$P(y_i = 0) = \phi\left(\frac{-\alpha - x_i\beta}{\sigma}\right) \quad (13)$$

The variable Z here is a standardized normal variable with a mean of "0" and a variance of "1". Thus, the model can be represented by Equation 14:

$$F^{-1}(P_i) = F^{-1}(I_i) = \alpha + \beta x_i \quad (14)$$

In this model, F^{-1} is the inverse of the normal cumulative distribution function. It is possible to state the following assumptions for the Probit model [2].

- $y_i \in \{0,1\},\ i = 1,2,\dots,n$
- $P_i = E(Y = 1/x) = \phi(\beta x_i)$ (Unit normal cumulative distribution function)
- $y_1, y_2, \dots, y_n$ are statistically independent
- There is no exact or multicollinearity among all x_i's

Binary probit models WLSM (Weighted Least Squares Method), ML (Maximum Likelihood Method), minimum chi-square iterative can be estimated with WLSM. In addition, the coefficient of R^2 in the probit model does not give us any idea as to whether the functional form of the model is well chosen.

2.3. Tobit Model

The sample where the information about the dependent variable is found only for some observations is known as censored sample. This model is also shown among models with a limited dependent variable because the dependent variable is limited. When censorship is applied to the dependent variable, the regression model is expressed in Equation 15:

$$y_i^* = \beta x + \varepsilon_i \quad \varepsilon_i \sim \mathrm{N}(0,1) \quad (15)$$

This model is called "Tobit model". y_i^* is the latent variable and τ is the censor point. Observed and censored for values greater than τ (Equation 16):

$$y_i = \begin{cases} y_i^*, & if\ y_i^* > \tau \\ 0, & if\ y_i^* \leq \tau \end{cases} \quad (16)$$

In the traditional Tobit model in Equation 16 when $\tau = 0$, some observations above y_i^* take the value of zero. That is, it is expressed as Equation 17;

$$y_i = \begin{cases} y_i^*, & y_i^* > 0 \\ 0, & y_i^* \leq 0 \end{cases} \quad (17)$$

The observed variable is a mixture random variable with a probability mass $P(y_i = 0/x_i) = P(y_i^* < 0/x_i) = \Phi(-x_i\beta/\sigma)$ on 0 and a continuum of values above 0 with density $f(y_i/x_i) = \sigma\phi[(y_i - x_i\beta)/\sigma]$. The expected value of the observed variable is obtained by Equation 18 [12]:

$$\begin{aligned} E(y_i/x_i) &= 0.P(y_i^* \leq 0/x_i) + P(y_i^*/y_i^* \\ &\quad > 0, x_i).P(y_i^* \leq 0/x_i) \\ &= \left[x_i\beta + \sigma\frac{\Phi\left(-\frac{x_i\beta}{\sigma}\right)}{\phi\left(-\frac{x_i\beta}{\sigma}\right)}\right]\Phi\left(\frac{x_i\beta}{\sigma}\right) \qquad (18) \\ &= x_i\beta\Phi\left(\frac{x_i\beta}{\sigma}\right) + \sigma\Phi\left(\frac{x_i\beta}{\sigma}\right) \end{aligned}$$

Here, ϕ denotes standard normal distribution function and Φ denotes the cumulative distribution function. When the data is limited below or above a certain limit, the distribution applied to the sample data becomes a mix of continuous and discontinuous distributions [13].

When it is known that the error terms for Tobit models are normally distributed (or have a generally parametric form distribution function), the maximum similarity and other similarity-based processes yield consistent and asymptotically normal distributed estimators. However, if the assumed parametric form of the similarity function is incorrectly determined, the estimators are inconsistent. The Tobit model uses a normal continuous dependent variable that is censored to a certain value. $y > \tau$ is an indicator variable equal to 1, the observation is uncensored. If $y = \tau$, that is, the observation is censored, then it is equal to 0. Since $\tau = 0$ in the traditional Tobit model, likelihood function is given by Equation 19:

$$L = \prod_i^N \left[\frac{1}{\sigma}\emptyset\left(\frac{y_i - x_i\beta}{\sigma}\right)\right]^{d_i} - \left[1 - \Phi\left(\frac{x_i\beta}{\sigma}\right)\right]^{1-d_i} \quad (19)$$

The log-likelihood function of the Tobit model is expressed in Equation 20:

$$\begin{aligned} lnL = \textstyle\sum_{i=1}^N \{d_i(-ln\sigma + ln\,\emptyset\left(\frac{y_i - x_i\beta}{\sigma}\right)) + (1 - \\ d_i)\ln(1 - \Phi\left(\frac{x_i\beta}{\sigma}\right))\} \qquad (20) \end{aligned}$$

Here the possibility of log-likelihood consists of two parts. The first section corresponds to the classical regression for uncensored observations, while the second section corresponds to the possibilities of censoring an observation. The above probability function is a mixture of discrete and continuous components and the standard ML state cannot be applied. However, it can be shown that the Tobit estimator has the usual ML characteristics. Although the log-likelihood function of the Tobit model is not entirely concave, it has a single maximum.

ML prediction of censored regression models gives strong predictions when the error term is normally distributed and has equal variance. However, several semi-parametric estimation strategies have been proposed that loosen the distribution assumption related to the error term [14].

The ML estimator is inconsistent in the presence of heteroscedasticity. [13] shows how to test different variance. Apart from the maximum similarity method, the following methods give estimates for the Tobit model.

- Heckman's two-step method (Tobit II)
- Nonlinear Least Squares method (NLLS)
- Nonlinear Weighted Least Squares method (NLWLS)
- Expectation-Maksimization method (EM)

In case of heteroscedasticity, the Censored Least Absolute Deviation (CLAD) method is also recommended for the Tobit model [15].

There are 3 expected values for the censored model when $\tau = 0$.

- The expected value of the latent variable y_i^* (Equation 21):

$$E(y_i^*) = \beta x_i \quad (21)$$

- Expected value of $(y|y > 0)$ (Equation 22):

$$E(y|y > 0) = \beta x_i + \sigma\lambda(\alpha) \quad (22)$$

Where $\alpha = (\tau - x_i\beta)/\sigma_u$ and $(\alpha) = \frac{\emptyset\left(\frac{x_i\beta}{\sigma}\right)}{\Phi\left(\frac{x_i\beta}{\sigma}\right)}$ gives the inverse Mill’s ratio.

- Expected value of y (Equation 23):

$$E(y) = \beta x_i + \Phi\left(\frac{x_i\beta}{\sigma}\right)[x_i\beta + \sigma\lambda(\alpha)] \quad (23)$$

3. Marginal Effects for Models

For the classical least squares (OLS) regression model, marginal effects are coefficients and are not dependent on x. In the logit and probit model, the coefficients differ between models due to the functional form of the F function. The relationship between coefficients is as follows [9, 16, 17];

- $\beta_{LPM} \approx 0.25\beta_{logit}$ (fixed term is beyond this)
- $\beta_{LPM} \approx 0.25\beta_{logit} + 0.5$ (for fixed term)
- $\beta_{logit} \approx 4\beta_{OLS}$
- $\beta_{probit} \approx 2.5\beta_{OLS}$
- $\beta_{logit} \approx 1.6\beta_{probit}$

The approximate values given above work well in cases where the mean probability value for the event is not far from 0.5. When $F(x_i\beta) > 0$, coefficients and marginal effects are the same.

3.1. Marginal Effect for Logit Model

The marginal effect for the Logit model ($M.E_{logit}$) is expressed in Equation 24:

$$M.E_{logit} = \frac{\partial E\left(y_i^*\middle|x_i\right)}{\partial x_{ik}} = \frac{e^{x_i\beta}}{\left(1+e^{x_i\beta}\right)^2}\beta_k \qquad (24)$$

3.2. Marginal Effect for Probit Model

The marginal effect for the probit model ($M.E_{probit}$) is expressed in Equation 25:

$$M.E_{probit} = \frac{\partial E\left(y_i^*\middle|x_i\right)}{\partial x_{ik}} = \phi(x_i\beta)\beta_k \qquad (25)$$

3.3. Marginal Effect for Tobit Model

For the Tobit model, marginal effects can be calculated in 3 possible cases [18].

- Marginal effect on latent dependent variable, y_i^*:

The interpretation of the parameters depends on the researcher. If the researcher is interested in the underlying linear relationship of the entire population, the slope coefficients β can be interpreted as marginal effects with Equation 26 [12]:

$$M.E_{tobit} = \frac{\partial E\left(y_i^*\middle|x_i\right)}{\partial x_{ik}} = \beta_k \qquad (26)$$

Tobit coefficients show how a unit in an independent x_{ik} variable changes the latent dependent variable.

- For uncensored observations, the marginal effect of the expected value of the y dependent variable is expressed in Equation 27:

$$M.E_{tobit} = \frac{\partial E(y|y>0)}{\partial x_{ik}} = \beta_k\left\{1-\lambda(\alpha)\left[\frac{x_i\beta}{\sigma} + \lambda(\alpha)\right]\right\} \qquad (27)$$

These coefficients show how the change of a unit in an independent x_{ik} variable affects uncensored observations.

- For the marginal effect (censored and uncensored) of the expected value of the dependent variable; if the researcher is interested in the effect of the observed (censored) value on the expected value, the marginal effect [12, 13] is expressed in Equation 28:

$$M.E_{tobit} = \frac{\partial E(y_i|x_i)}{\partial x_{ik}} = \phi\left(\frac{x_i\beta}{\sigma}\right)\beta_k \qquad (28)$$

This marginal effect has an interesting decomposition [19] (Equation 29):

(1) effect of fully observed values on expectations,
(2) impact on the probabilities of the fully observed:

$$\frac{\partial E(y_i|x_i)}{\partial x_{ik}} = \frac{\partial E\left(y_i^*\middle|y_i^*>0,x_i\right)}{\partial x_{ik}}P(y_i^*>0) + \frac{\partial P(y_i^*>0)}{\partial x_{ik}}E(y_i^*|y_i^*>0,x_i) \qquad (29)$$

(1) $\frac{\partial E\left(y_i^*\middle|y_i^*>0,x_i\right)}{\partial x_{ik}} = \beta_k(1-\lambda^2-\alpha_i\lambda_i)$

(2) $\frac{\partial P(y_i^*>0)}{\partial x_{ik}} = \frac{\partial\phi\left(\frac{x_i\beta}{\sigma}\right)}{\partial x_{ik}} = \beta_k\sigma^{-1}(x_i\beta/\sigma)$

These marginal effects depend on the individual characteristics of the x_i values and are shown as average effects in the sample population.

4. Model Selection Criteria

AIC and BIC are information criteria that allow comparison of both logit and probit models. Therefore, these are briefly explained below. [2] recommend that the use of certainty coefficient R2 as a statitics should be avoided to explain and summarize the model in cases where there are models with a dependent variable that take two values.

4.1. Akaike Information Criterion (AIC)

Akaike Information Criterion is used to select the most appropriate one among different models (Equation 30). In model comparisons, the model that always gives the lowest AIC value is preferred.

$$AIC = -2log(L) + 2k \quad (30)$$

In cases where the number of parameters is larger than the sample size, AICc given by Equation 31 proposed by Hurvich and Tsai should be used instead of AIC [20-22].

$$AICc = AIC + 2k(k + \frac{1}{(n-k-1)} \quad (31)$$

4.2. Bayes Information Criterion (BIC)

Akaike derived the BIC model selection criterion for selected model problems in linear regression [23]. Equation of Bayesian Information Criterion is expressed in Equation 32:

$$BIC = -2log(L) + klog(n) \quad (32)$$

The BIC differs from AIC in the second part, which depends on the sample size on the right side of the equation. However, in spite of the superficial similarity between AIC and BIC, it was later revealed that they differed within the Bayesian structure [24-25]. Among the existing models as with the Akaike Information Criterion, the model with the smallest BIC value is selected as the appropriate model.

5. An Application

In this study, coefficient estimations and marginal effects were calculated with Stata 14 program for logit, probit and Tobit model and the models were compared according to AIC and BIC criteria. Family income and expenditure survey (FIES) data prepared by Philippine Statistical Institute (PSA) was used. This survey was conducted on 41544 people in 2015 and consists of 60 variables. However, 9 variables were taken into consideration for comparison of 3 models. Medical Care Expenditures were taken as the dependent variable. Independent variables;

Age	= Household Head Age
Floor	= House Floor Area
NumLT5	= Members with age less than 5 year old
Bedrooms	= Number of bedrooms
Electricity	= Electricity
Car	= Number of Cars, Jeeps, Vans
Phone	= Number of Cellular phones
Computer	= Number of Personal Computers

Descriptive statistics related to these variables are given in Table 1.

Table 1. *Descriptive Statistics (x_i)*

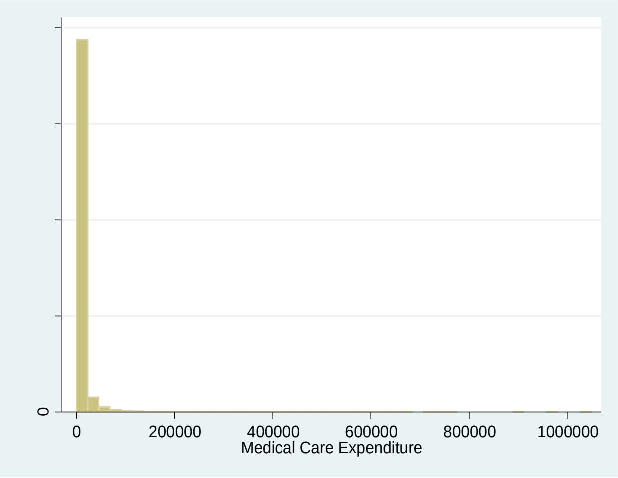
Variables	Obs	Mean	Std. Dev.	Min	Max
age	41544	51.38145	14.16608	9	99
floor	41544	.4102157	.6943901	0	5
numLT5	41544	55.60336	55.02316	5	998
bedrooms	41544	1.788.008	1.105664	0	9
electricity	41544	.8908146	.3118755	0	1
car	41544	.0812151	.3467859	0	5
phone	41544	1.905.738	1.55813	0	10
computer	41544	.3150154	.7396982	0	6

In Fig. 3, medical care expenditures of households are shown graphically. Household medical care expenditures of 1478 households out of 41544 households have zero value. In other words, these households did not spend any medical care expenditure.

5.1. Results for Logit Model

In the analysis, the dependent variable showing whether the household is making medical care expenditure is included as the dummy variable. In the creation of the dependent variable, making medical care expenditure was defined as “1” and not making as “0”. When the binary logit model results are examined in Table 2, numLT5 and car variables are insignificant and the other variables are significant ($p<0.05$).

According to the data used in the current study, model was determined as follows on the basis of the correlation between medical care expenditure and age, floor, numLT5, bedrooms, electricity, car, phone and computer,



Medical Care Expenditure	(y)
Obs	41544
Mean	7160.23
Std. Dev.	26902,002
Variance	723717726,384
Skewness	15,049
Kurtosis	357,474

Figure 3. *Histogram of Medical Care Expenditures and Some Statistics*

Table 2. *Binary Logit Regression Model*

Medical Care Expenditure	**Coef.**	**Std. Err.**	**z**	**P>t**	**[95% Conf.**	**Interval]**
age	.0111197	.0019606	5.67	0.000	.0072769	.0149624
floor	.5587745	.052456	10.65	0.000	.4559627	.6615863
numLT5	.0001076	.0006123	0.18	0.860	-.0010925	.0013078
bedrooms	.0794709	.0303316	2.62	0.009	.020022	.1389197
electricity	.3044671	.0743389	4.10	0.000	.1587655	.4501688
car	.089623	.1242187	0.72	0.471	-.1538412	.3330872
phone	.1921643	.0242938	7.91	0.000	.1445494	.2397793
computer	.1492993	.0592707	2.52	0.012	.0331308	.2654678
constant	1.801.686	.120015	15.01	0.000	1.566.461	2.036.911
N=41544 AIC= 12412.63 BIC=12490.34 Log likelihood = -6197.3163 LR χ^2 (8) = 369.55 prob> χ^2=0.0000 Pseudo R^2 = 0.0290						

$$Z = 1.801.686 + 0.0111197 * \text{age} + 0.5587745 * \text{floor} + 0.0001076 * \text{numLT5} + 0.0794709 * \text{bedrooms} + 0.3044671 * \text{electricity} + 0.089623 * \text{car} + 0.1921643 * \text{phone} + 0.1921643 * \text{computer}$$

All of the independent variables in our model have a direct relationship with the probability of occurrence (medical care expenditure). The obtained LR statistic was obtained according to 8 degrees of freedom $\chi 2$. Since the coefficients obtained in the Logit model estimates cannot be interpreted directly, marginal effects are calculated for the coefficient interpretation and the results are given in Table 3.

Table 3. *Marjinal Effect for Logit Regression Model*

Variable	**dy/dx**	**Std. Err.**	**z**	**P>z**	**[95%**	**C.I.]**	**X**
age	.0003341	.00006	5.73	0.000	.00022	.000448	513.815
floor	.0167882	.0015	11.22	0.000	.013855	.019721	.410216
numLT5	3.23e-06	.00002	0.18	0.860	-.000033	.000039	556.034
bedrooms	.0023877	.00091	2.62	0.009	.000605	.004171	178.801
electricity	.0102547	.0028	3.66	0.000	.004767	.015743	.890815
car	.0026927	.00373	0.72	0.470	-.004617	.010003	.081215
phone	.0057735	.00071	8.12	0.000	.004379	.007168	190.574
computer	.0044856	.00177	2.53	0.011	.001012	.007959	.315015

Coefficient interpretation for logit model according to marginal effects: While other variables are constant; 1 unit increase in the household age increases the probability of medical care spending 0.0003 units. Likewise, it is possible to increase the cost of medical care by 0.006 units according to the absence of a private phone. Having a private computer increases the probability of medical care spending by 0.02 units.

Table 4. *OR Binary Logit Regression Model*

MedicalCareExpenditure	Coef.	Std. Err.	z	P>t	[95% Conf.	Interval]
age	1.011182	.0019826	5.67	0.000	1.007303	1.015075
floor	1.748528	.0917207	10.65	0.000	1.577692	1.937864
numLT5	1.000108	.0006124	0.18	0.860	.9989081	1.001309
bedrooms	1.082714	.0328404	2.62	0.009	1.020224	1.149032
electricity	1.355902	.1007963	4.10	0.000	1.172063	1.568577
car	1.093.762	.1358657	0.72	0.471	.8574082	1.395269
phone	1.21187	.0294409	7.91	0.000	1.155519	1.270969
computer	1.16102	.0688145	2.52	0.012	1.033686	1.304041
constant	6.059856	.7272739	15.01	0.000	4.789667	7.666891

In logit models, besides marginal effects, OR values can be used for coefficient interpretation. Table 4 shows the OR values. All of the variables identified in the study have OR values greater than 1. Therefore, since numLT5 and car variables are insignificant it can be interpreted that other variables are an important risk factor.

5.2. Results for Probit Model

Probit model is based on benefit theory and rational choice approach. According to the rational choice approach, individuals choose the ones that will benefit the most from the options they face. The dependent variable was defined as “1” for medical care expenditure and “0” for non-medical care expenditure as in the logit model. When the results of binary probit model are examined in Table 5, it is seen that as in logit model, numLT5 and car variables are insignificant and other variables are significant (p<0.05).

Table 5. *Binary Probit Regression Model*

MedicalCareExpenditure	Coef.	Std. Err.	z	P>t	[95% Conf.	Interval]
age	.0049271	.0008648	5.70	0.000	.0032322	.006622
floor	.2372132	.0214667	11.5	0.000	.1951393	.2792871
numLT5	.0000412	.0002608	0.16	0.875	-.0004699	.0005522
bedrooms	.0341087	.0131134	2.60	0.009	.0084068	.0598105
electricity	.1506318	.0346956	4.34	0.000	.0826297	.2186339
car	.0415185	.050702	0.82	0.413	-.0578555	.1408926
phone	.0798869	.0102014	7.83	0.000	.0598925	.0998813
computer	.0586167	.0240934	2.43	0.015	.0113945	.1058389
constant	113.478	.0547335	20.73	0.000	1.027504	1.242055
N=41544 AIC= 12416.47 BIC=12494.18 Log likelihood = -6199.2359 LR χ^2 (8) = 365.71 prob> χ^2=0.0000 Pseudo R^2 = 0.0287						

Correlation between medical care expenditure and other variables according to probit model was found to be as follows;

$$Z = 113.478 + 0.0049271 * age + .2372132 * floor + 0.0000412 \ * numLT5 + 0.03410870 bedrooms + 0.1506318 * electricity + 0.0415185 * car + 0.0798869 * phone + 0.0586167 * computer$$

The obtained LR statistic was found to be significant according to $\chi 2$ value with 8 degress of freedom In the probit model, the numLT5 and car variables were found to be statistically insignificant similar to the logit model. On the other hand other variables were obtained significantly. Marginal effects were calculated following the probit model estimate. To interpret the coefficients of the probit model as in the Logit model; The mean of the independent variables was evaluated and marginal effects were used. According to this, when the age of the household increases by 1 year, the probability of spending on medical care increases by 0.005 units.

Table 6. *OR Binary Logit Regression Model*

Variable	dy/dx	Std. Err.	z	P>z	[95%	C.I.]	X
age	.0003506	.00006	5.74	0.000	.000231	.00047	513.815
floor	.0168809	.00148	11.40	0.000	.013979	.019783	.410216
numLT5	2.93e-06	.00002	0.16	0.875	-.000033	.000039	55.6034
bedrooms	.0024273	.00093	2.60	0.009	.0006	.004255	1.78801
electricity	.0119621	.00306	3.91	0.000	.005965	.017959	.890815
car	.0029546	.00361	0.82	0.413	-.004114	.010023	.081215
phone	.005685	.00072	7.94	0.000	.004282	.007088	1.90574
computer	.0041714	.00171	2.44	0.015	.000819	.007524	.315015

Having electricty was found to lead to a 15 units increase in the medical care expenditure.

5.3. Results for Tobit Model

The results for the Tobit model are given in Table 7. The lower limit of the expenditure was zero and left censored. Accordingly, 1478 observations were censored from the left and 40066 observations were not censored. When the results for Tobit model are examined, it is seen that all variables are significant.

Table 7. *Tobit Regression Model*

Medical Care Expenditure	Coef.	Std. Err.	z	P>t	[95% Conf.	Interval]
age	175.5149	9.6917	18.11	0.000	156.5189	194.5108
floor	1279.26	193.7298	6.60	0.000	899.5455	1658.975
numLT5	27.09796	2.705572	10.19	0.000	21.79498	32.40094
bedrooms	1403.485	141.3207	9.93	0.000	1126.494	1680.477
electricity	2512.529	442.4302	5.68	0.000	1645.356	3379.702
car	7558.064	426.3697	17.73	0.000	6722.371	8393.758
phone	605.0307	99.2418	6.10	0.000	410.5146	799.5467
computer	3111.545	215.5903	14.43	0.000	2688.984	3534.107
constant	-12089.91	642.4839	-18.82	0.000	-13349.19	-10830.63
/sigma	26558.79	93.95331			26374.64	26742.94
Obs. summary:	1478 left-censored observations at MedicalCar~e<=0 40066 uncensored observations 0 right-censored observations					
N=41544 AIC=932271.2 BIC=932357.5 Log likelihood = -466125.59 LR χ^2 (8) = 2847.34 prob> χ^2=0.0000 Pseudo R^2 = 0.0030						

The parameter given as sigma is the estimated standard error of the regression. The obtained 26558.79 value can be compared with the mean square error (Root MSE = 26022) in the regression model. Correlation between medical care expenditure and other variables according to Tobit model was found to be as follows:

$$y_i = -12089.91 + 175.5149 * \text{age} + 1279.26 * \text{floor} + 27.09796 * \text{numLT5} + 1403.485 * \text{bedrooms} + 2512.529 * \text{electricity} + 7558.064 * \text{car} + 605.0307 * \text{phone} + 3111.545 * \text{computer}$$

Interpretation of the coefficients for the Tobit model: Older people, people whose house has more square meters, households having more people under the age of 5, households whose houses have more rooms, households using more electricity, households having more personal phones, personal computers were found to spend more more money on medical care. If the household owner is one year older, then the expected or latent medical care spending increases by $175.5. Marginal effects for the Tobit model are given in Tables 8 and 9. Interpreting marginal effects for the censored sample: If the household owner is one year older, they spend an additional $ 74 more on medical care expenditure. Other coefficients can be interpreted similarl It is expected that the higher the number of household members the greater the expenditure. On the other hand the greater the number of individuals that make up the family the less likely the family will spend on medical expenses.

Table 8. *Marjinal Effect for Tobit Regression Model (predict(e(0,.)))*

Variables	dy/dx	Std. Err.	z	P>z	[95%	C.I.]
age	73.66675	4.073.022	18.09	0.000	65.68377	81.64972
floor	536.9284	81.32036	6.60	0.000	377.5434	696.3134
numLT5	11.3735	1.136055	10.01	0.000	9.146874	13.60013
bedrooms	589.0679	59.33781	9.93	0.000	472.7679	705.3679
electricity	1054.554	185.7133	5.68	0.000	690.562	1418.545
car	3172.255	179.1964	17.70	0.000	2821.037	3523.474
phone	253.9422	41.658	6.10	0.000	172.2941	335.5904
computer	1305.971	90.56519	14.42	0.000	1128.467	1483.476
margins, dydx(*) atmeans predict(e(0,.))						

y.

Table 9. *Marjinal Effect for Tobit Regression Model (predict(ystar(0,.)))*

Variables	dy/dx	Std. Err.	z	P>z	[95%	C.I.]
age	104.6082	5.785614	18.08	0.000	93.26861	115.9478
floor	762.4487	115.4805	6.60	0.000	536.111	988.7864
numLT5	16.15059	1.613382	10.01	0.000	12.98842	19.31276
bedrooms	836.4878	84.26893	9.93	0.000	671.3237	1001.652
electricity	1497.486	263.7239	5.68	0.000	980.5971	2.014.376
car	4504.664	254.5433	17.70	0.000	4005.768	5003.56
phone	360.6029	59.15704	6.10	0.000	244.6572	476.5485
computer	1854.505	128.631	14.42	0.000	1602.392	2106.617
margins, dydx(*) atmeans predict(ystar(0,.))						

6. Conclusion and Suggestions

In this study, logit, probit and Tobit models are compared using a questionnaire. When dummy variables that take two or more values are included in regression models as dependent variables, dependent variables indicate preference or decision. The most commonly used models among these preference models are logit and probit models. Both logit and probit model analyses are very similar and the probability estimates obtained are close to each other. However, while log-odds (likelihood ratios) are used in logit model analysis, the cumulative normal distribution of probit model is used.

The structural models of Logit, probit and Tobit are similar, but the models are different. In the Tobit model, the observed values of the dependent variable are known when $y_i^* > \tau$. In the probit and logit model, if only $y_i^* > \tau$, y value is “1”. However, if the data are below the threshold (τ), they cannot be known and the value y is assumed to be zero. More information is available on the Tobit model. Therefore, it is expected that coefficient estimations obtained from Tobit model will be more effective than those obtained from probit model.

When all the results obtained are evaluated together, it is more important that the coefficients give expected signs and the explanatory variables are statistically significant in binary models such as logit and probit than the goodness of fit measure. This value may be too low when R^2 is calculated for these models. This will not be an indication that the model is weak. The reason for this is that most of the Pseudo R^2 values calculated in qualitative preference models are based on similarity ratios, not variances explained. In the current study, Pseudo $R^2 = 0.0290$ for logit model, Pseudo $R^2 = 0.0287$ for probit model and Pseudo $R^2 = 0.0030$ for Tobit model.

Logit and probit models can be compared with the log-likelihood value instead of R^2. As the log-likelihood value approaches zero (always negative), the model works well. In the current study, Log-likelihood was found to be -6197.3163 for logit model, -6199.2359 for probit model and -466125.59 for Tobit model.

When the models were compared, the lowest AIC and BIC values (932271.2 and 932357.5) were found for the Tobit model. Therefore, it can be said that Tobit model is better than probit and logit models in estimating regression models used with censored data. In the current study, the variables found to be insignificant in logit and probit models were found to be significant in Tobit model. Even if the coefficients of the Logit model and the probit model are not the same, the information obtained from the marginal effects is quite similar. The Tobit model uses all data, including censored information, and allows for the estimation of consistent parameters.

References

[1] Gujarati, D. N., “Basic Econometrics”, 5th ed. Boston: McGraw-Hill, 2009.

[2] Aldric J. H. and Nelson, F. D., "Linear Probability, Logit and Probit Models", Sage Publications, USA, 1984.

[3] Long, J. S., “Regression Models for Categorical and Limited Dependent Variables”, Sage Pubications, USA, 1977.

[4] Tobin, J., “Estimation of Relationships for Limited Dependent Variables”, Econometrica, 46-1, 24-36 (1958).

[5] Üçdoğruk, Ş., Akın, F. and Emeç, H., “Türkiye Hane Halkı Eğlence Kültür Harcamalarında Tobit Modelin Kullanımı”, G.Ü. İ.İ.B.F. Dergisi, 3, 13-26 (2001).

[6] Freese, J. and Long, J. S., “Regression Models for Categorical Dependent Variables Using Stata”, College Station: Stata Pres, 2006.

[7] Özdamar, K., “Paket Programlar ile istatistiksel Veri Analizi-I”, Kaan Kitabevi, Eskişehir, 1999.

[8] Sümbüloğlu, K., “Lojistik Regresyon Analizi”, 2009, http://78.189.53.61/-/bs/ess/k_sumbuloglu.pdf .

[9] Amemiya, T., "Qualitative Response Models: A Survey", Journal of Economic Literature, 19-4, 481-536 (1981).

[10] McFadden, D., “Conditional Logit Analysis of Qualitative Choice Behavior”, Academic Press, New York, 1973.

[11] Demaris, A., “Regression with Social Data: Modeling Continuous and Limited Response Variables”, John Wiley & Sons, Inc. Hoboken, New Jersey, 2004.

[12] Schmidheiny, K., “Limited Dependent Variable Models”, Unversitat Pompeu Fabra, Lecture Notes in Microeconometrics, 2007.

[13] Greene, W. H., “Econometric Analysis”, New Jersey: Prentice Hall, 2002.

[14] Chay, K. Y. and Powell, J. L., “Semiparametric Censored Regression Models, Journal of Economic Perspectives, 15-4, 29-42 (2001).

[15] Powell, J. L., "Symmetrically Trimmed Least Squares Estimation for Tobit Models", Econometrica, 54-6, 1435-60 (1986).

[16] Cameron, A., “Limited Dependent Variable Models (Brief) Binary, Multinomial, Censored, Treatment Effects”,2011, http://cameron.econ.ucdavis.edu/bgpe2011/bgpev2_ldv.pdf .

[17] Katchova, A. L. and Miranda, M. J., “Two-Step Econometric Estimation of Farm Characteristics Affecting Marketing Contract Decisions”, American Journal of Agricultural Economics, 86-1, 88-102 (2004).

[18] Sigelman, L. and Langche, Z., “Analyzing Censored and Sample-Selected Data with Tobit and Heckit Models” Political Analysis, 8, 167–182 (1999).

[19] McDonald, J. F. and Mofitt, R. A., “The Uses of Tobit Analysis”, The Review of Economics and Statistics, 62-2, 318-321 (1980).

[20] Akaike, H., “Information Theory and an Extension of the Maximum Likelihood Principle”, 2nd International Symposium on Information Theory, 267-281 (1973).

[21] Sugiuna, N., “Further Analysis of the Data by Akaike’s Information Criterion and the Finite Corrections”, Communication in Statistics-Theory and Methods, 57, 13-26 (1978).

[22] Hurvich, C. M. and Tsai, C., “Regression and Time Series Model Selection in Small Samples”, Biometrika, 76, 297-307 (1989).

[23] McQuarrie, A. D. and Tsai, C. L., “Regression and Time Series Model Selection”, World Sciencetific, 1998.

[24] Raftery, A. E., “Bayesian Model Selection in Social Research (with Discussion by Andrew Gelman, Donald B. Rubin and Robert M. Hauser)”, In P.V. Marsdn (Ed.), Socialogical Methodology, 111-196 (1995).

[25] Wasserman, L., “Bayesian Model selection and Model averaging,” Jounal of Mathematical Psychology, 44, 92-107 (2000).