Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 799-807
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.5448

Comparison some censored regression models with application to renal failure data

Sunbul Rasheed Mohammed[a,*], Anwaar Dhiaa Abdulkareem[b]

[a] *College of Administration and Economics, University of Kirkuk, Iraq*
[b] *College Education for Pure Sciences, University of Kirkuk, Iraq*

(Communicated by Madjid Eshaghi Gordj)

Abstract

This article dealt with the topic of using some functions in the censored regression model, such as the Tobit model and Log-BurrXIIEE model and estimating the parameters of the two models. Based on real data for patients with renal failure, the two models were compared. Results are showed that the LBXIIEE model is the best model compared to the Tobit model according to the values of the model selection criteria, as the values of these criteria for the LBXIIEE model were less than the values of the same criteria for the Tobit model when Urea as a dependent variable in the model. We also note that the values of the criteria (AIC, BIC, H-QIC) change and decrease according to the steps used for estimation and analysis of the two models (Tobit, LBXIIEE) according to the specific dependent variable

Keywords: censored regression model,Tobit model, Log-BurrXIIEE model,criteria

1. Introduction

As it is known that each phenomenon has special data and with different types of data, the models that should be used in accordance with those data and according to the goal of the study, for the quantitative data of the independent variable and the availability of the required hypotheses, the optimal model will be the traditional regression model (simple or multiple). But if the data for the dependent variable are specific data in one part and free in another part (unspecified), then the optimal model is the censored regression model, which is the subject of our interest in this article.

*Corresponding author
Email addresses: Sunbulr79@uokirkuk.edu.iq (Sunbul Rasheed Mohammed), Anwaar71@uokirkuk.edu.iq (Anwaar Dhiaa Abdulkareem)

Received: February 2021 *Accepted:* August 2021

In the various fields of scientific research, there are many different scientific researches according to their specializations. Therefore, any researcher, before going into his research, must review what he has gone through from previous studies in support of his research from both the scientific and practical sides. Since censored regression models and censored data are among the important topics in the fields of statistical studies and research, and in light of the technological development that has taken place, researchers have begun conducting many studies and researches that delve into censored regression models, censored data, and failure times, some of these researches and studies in estimation and in 1972, researchers **Chen and Dixon** (**Chen & Dixon, 1972**)[4] dealt with the estimates of the parameters of the censored regression model sample in two different ways. In the first method, the observations below or above a certain point are subject to censoring, while the second is one or more than one observation of two sides (upper and lower), i.e., double censoring while the researcher **Siddhartha Chib (Chib S.,1992)**[5] in 1992 used the Monte Carlo method based on multivariate symmetric distributions and Laplacian estimates in certain parameters of the censored regression model in the Tobit model and estimating the model in the Bayesian method. The idea here is to increase the sample size and the two methods were compared with two examples and different combinations of sample sizes and degrees of censored, so the researchers **Bang and Tsiatis (Bang & Tsiatis , 2000)**[2] also focused in 2000 on the problem of estimating the average medical cost in a sample of individuals whose medical costs may be subject to censored (Right Censoring). A class of weighted estimators is presented, which is suitable for censoring, and its estimators have been shown to be consistent and asymptotically normal with variables that can be easily estimated. The efficiency of these estimators was also studied in order to find an effective estimator of the average medical cost and simulation was used to show that the estimators perform well in censored samples. In 2006, **Zhiliang Ying (Zhiliang & Ying, 2006)**[12] was interested in simple inference procedures and reliability using the principle of least squares for a semi-parametric failure time model that is linearly related to the covariates while leaving the error distribution undefined in the presence of data subject to Right Censoring and the proposed estimator for the vector-value regression coefficient is an iterative solution to equalize the Buckley James estimate using a consistent initial estimator as the starting value and the results using simulations that the estimator are consistent and asymptotically normal, a new estimation of the finite covariance matrix was formed and medical studies were used in this research. And the researchers **Paranaibe etal.** (**Paranaíba et al., 2011**) [11] in 2010 proposed the ML method and Bayesian analysis for estimating parameters of a new model of BurrXII system which contains some distributions known in the analysis and study of censored data such as logistic distribution and Weibull distribution. And they used simulations to show the results of their own study.

The current article aims to illustrate censored regression models as a problem faced by many data ,as well as estimation of parameters of the Tobit model and Log-BurrXII Exponentiated Exponential Model (LBXIIEE model), in the presence of data subject to the Maximum Likelihood Estimation (MLE) method, and comparison between two models in the presence of censored data and choosing the best model between them through the use of some statistical criteria in this field by applying the estimators to real data for patients with renal failure.

2. Censored Regression Models

When the data or observations are restricted in one part (specific) and free in the other part (unspecified), these data are called censored data, and therefore the use of the traditional model of regression here will lead to biased and inconsistent estimations, so it is necessary to define a suitable model for that data and this model is called a model censored regression. Censoring in statistics is a state in which the value of the measurement or observation is only partially known.

There are many censored regression models including Probit model, Logit model and Tobit model which are the most common and used. Recently, many models have been proposed through location regression models. The researcher Cordeiro presented a class of regression models based on the log-Gamma model extended from the Weibull distribution. The researcher Hashimoto also proposed a log-BurrXII regression model for a set of survival data, among others (Cordeiro, et al.,2018)[6].

There are several definitions that explain the concept of censored data, " It is the data or values that were not observed, or the impossibility of measuring their observation, or the failure to record complete information about the observation during the study period(Karim, A. A. (2018))[10].It can also be defined "It is the process of determining the number of failed units for a specific period of time for the experiment, or specifying a specific time for the experiment, and then knowing the number of failed units" (Al-Rubaie, 2019)[1].

2.1. Tobit Model

It is a model proposed by **James Tobit** in 1958. It describes the relationship between the dependent variable and the independent variables, and deals with the data of the dependent variable on the basis that it consists of two parts, and each part of it takes a specific distribution function. Observations with values equal to or close to zero take the Distribution function (c.d.f). And observations that take quantities greater than zero will take the probability density function (p.d.f), and by multiplying the (cdf) and (pdf) functions, we get the mixed function that expresses the Tobit model. (Carson & Sun, 2007). (Hadi, 2017)[3, 8].

The general form of the Tobit model is:

$$y_i{}^* = \alpha + \beta x_i + \varepsilon_i \tag{2.1}$$

$$y_i = \{\ y^* \quad if \quad y^* > \lambda\ 0 \quad if \quad y^* \leq \lambda\ \} \tag{2.2}$$

$$\varepsilon_i \ \sim \ N\ \left(0, \sigma^2\right) \qquad\qquad y^* \ \sim \ N\ (x\beta, \sigma^2)$$

whereas

λ : restriction point

y_i : dependent variable

y^*: latent variable

$_\beta$: Parameters of the model

x_i: explanatory variables (independent)

By multiplying equations (2.2) and (2.1), we get the mixed function for the following:

$$f_{(y)} = \left[\frac{1}{\sqrt{2\pi\sigma^2}}\, exp(-\frac{(y_i - x\beta)^2}{2\sigma^2})\right] \left[1 - \Phi(\frac{\lambda - x\beta}{\sigma})\right] \tag{2.3}$$

when $\lambda = 0$

$$f_{(y)} = \left[\frac{1}{\sqrt{2\pi\sigma^2}}\, exp(-\frac{(y_i - x\beta)^2}{2\sigma^2})\right] \left[1 - \Phi(\frac{-x\beta}{\sigma})\right] \tag{2.4}$$

And it can be expressed as follows

$$f_{(y)} = \left[\frac{1}{\sigma}\, \phi(-\frac{(y_i - x\beta}{\sigma})\right] \left[1 - \Phi(\frac{-x\beta}{\sigma})\right] \tag{2.5}$$

where

$\phi_{(.)}$: probability density function (pdf)
$\Phi_{(.)}$: cumulative function (cdf)
Applying the maximum likelihood method, we get the following:

$$L\left(f_{(y)}\right)=\prod_{i=}^{n}\left[\frac{1}{\sigma}\phi\frac{(y_i-x\beta)}{\sigma}\right]\left[1-\Phi\frac{(-x\beta)}{\sigma}\right] \tag{2.6}$$

Taking the natural logarithm of both sides of the above equation, we get the following:

$$\ln L \ =\sum_{i=1}^{n}\left[\left(-\ln\sigma \ +\ln\phi \ \frac{(y_i-x\beta)}{\sigma}\right)+\ln\left(1-\Phi(-\frac{x\beta}{\sigma})\right)\ \right] \tag{2.7}$$

For the difficulty of finding the estimations of the greatest possibility of the above equation we use numerical methods to obtain estimations of the Tobit model

2.2. Log-BXII Exponentiated Exponential Model

It is a new continuous distribution of families of the exponential distribution which was discussed by **Gupta** (1998) and this distribution contains two parameters which are the scaling parameter and the shape parameter similar to the Weibull distribution family and the Gamma distribution.
It was noted that the properties of this distribution are somewhat similar to the properties of the Weibull and Gamma distributions, and it can also be used as a possible alternative to them. (Gupta,R.2001)[7]
The probability density function pdf of the EE distribution can be expressed as follows:

$$f_{(x;\alpha,\lambda)}=\alpha\lambda(1-e^{-\lambda x})^{\alpha-1}e^{-\lambda x} \tag{2.8}$$

And the cumulative distribution function cdf for the distribution is as follows:

$$F_{(x;\alpha,\lambda)}=(1-e^{-\lambda x})^{\alpha} \tag{2.9}$$

whereas
λ: the scale parameter of the EE distribution
α: shape parameter of the EE distribution
Another recently proposed model is the Log-BXIIEE model reviewed by Mohamed and Haitham (Ibrahim, et al.,2020) which is an extension of the Exponentaited Exponential distribution with the BXII distribution using the location regression model. And the mathematical formula for BXIIEE is (2.9):

$$f_{(x)}=ab\alpha\lambda e^{-\lambda x}\frac{(1-e^{-\lambda x})^{a\alpha-1}}{[1-(1-e^{-\lambda x})^{\alpha}]^{a+1}}\left\{1+\left[\frac{(1-e^{-\lambda x})^{\alpha}}{1-(1-e^{-\lambda x})^{\alpha}}\right]^{a}\right\}^{-b-1} \tag{2.10}$$

$x \ \sim \ BXIIEE(a,b,\alpha,\lambda)$
In order to obtain the Log-BXIIEE regression model, we have to model the new BXIIEE distribution in a similar manner to the LBXIIW model as follows:

When x ∼ BXIIEE, and Y=log(x), the equation (2.10) is as follows:

$$
\begin{aligned}
f_{(y)} \;&=\; \frac{ab\alpha\lambda}{\sigma} expexp\ \left(-\lambda exp\left(\frac{y-\mu}{\sigma}\right)\right)\ \left(1-\left(-\lambda exp\left(\frac{y-\mu}{\sigma}\right)\right)\right)^{\alpha-1} \\
&\times\frac{\left\{\left(1-exp\left(-\lambda exp\left(\frac{y-\mu}{\sigma}\right)\right)\right)^{\alpha}\right\}^{a-1}}{\left\{1-\left(1-exp\left(-\lambda exp\left(\frac{y-\mu}{\sigma}\right)\right)\right)^{\alpha}\right\}^{a+1}} \\
&\times\left\{1+\left[\frac{\left\{(1-expexp\ \left(-\lambda expexp\ \left(\frac{y-\mu}{\sigma}\right)\ \right)\ )^{\alpha}\right\}}{1-(1-expexp\ \left(-\lambda expexp\ \left(\frac{y-\mu}{\sigma}\right)\ \right)\ )^{\alpha}}\right]^{a}\right\}^{-b-1}
\end{aligned}
\tag{2.11}
$$

whereas
μ: location parameter
σ>0: scale parameter
The above equation represents the LBXIIEE distribution and Y∼ LBXIIEE(a,b,α,λ,μ,σ)
Using the location - scale regression model that links the explanatory variable vector $\nu_{\mathrm{i}}{}^{\mathrm{T}}$= ($\nu_{\mathrm{i1}}$, ...,$\nu_{\mathrm{ip}}$) with the average response variable y_i , and the model can be expressed as follows:

$$y_i = {\nu_i}^T\beta + \sigma z_i \qquad\qquad , \ i = 1, \ldots, n \tag{2.12}$$

where y_i follows the LBXIIEE distribution and T is a random variable that follows the BXIIEE distribution

Assuming that M, N are a set of items for y_i which represents log-lifetime or log-censoring respectively, and assuming lifetimes and censoring times are independent, the LBXIIEE regression model and by applying the MLE method to estimate the parameters of the vector θ= $\left(\mathrm{a}, \mathrm{b}, \alpha, \lambda, \beta^{\mathrm{T}}\right)$ and taking the natural logarithm will be as follows (2.8):

$$
\begin{aligned}
\ln(\theta) \;&=\; r\ln\left(\frac{ab\alpha\lambda}{\sigma}\right)\ -\lambda\sum_{i\in M}u_i + (\alpha-1)\sum_{i\in M}\ln(1-\exp\left(-\lambda u_i\right)\ ) \\
&+(\alpha-1)\sum_{i\in M}\ln(1-\exp\left(-\lambda u_i\right)\ )^{a} \\
&+(\alpha+1)\sum_{i\in M}\ln(1-(1-\exp\left(-\lambda u_i\right)\ )^{\alpha}) \\
&-(b+1)\sum_{i\in M}\ln\left(1+\left[\frac{(1-\exp\left(-\lambda u_i\right)\ )^{\alpha}}{1-(1-\exp\left(-\lambda u_i\right)\ )^{\alpha}}\right]^{a}\right) \\
&+\sum_{i\in N}\ln\left(1-\left[1+\left[\frac{(1-\exp\left(-\lambda u_i\right)\ )^{\alpha}}{1-(1-\exp\left(-\lambda u_i\right)\ )^{\alpha}}\right]^{a}\right]^{-b}\right)
\end{aligned}
\tag{2.13}
$$

Whereas

$$u_i = e^{zi}\ , \qquad\quad z_i = \frac{(y_i - {\nu_i}^T\beta)}{\sigma}$$

Using NLMixed in SAS software, the estimations for the feature vector can be obtained.

3. Application to Renal failure Data

For the purpose of conducting a practical application to compare the two models of Tobit and LBX11EE, real data was taken regarding patients with renal failure in the Kirkuk General Hospital/Industrial College Unit in the city of Kirkuk - Iraq. A sample of 131 people was taken from the patient records, and the data were representative of the variables: age, urea and creatine. By taking Urea as a dependent variable for the models,. as we mentioned previously in support of the censoring models, the importance of which is one of the aims of this article. The mechanism of analyzing these models includes extracting the standard error values S.E, the T-test values and the P-values, on the basis of which the variables that significantly affect the dependent variable will be known.

4. Results and Discussion

In the table (1) below, the Tobit model was estimated, considering that Urea is the dependent variable and according to the P-value. We note that in the first step, all independent variables were estimated to find out which of them had the most impact on the variable Urea, and in the second step, age was excluded as it was the least Influencing the dependent variable, and in the third step, gender was excluded, as it is the least influential variable after age, and the variable Creatine is the most influential on the variable Urea. We also note from the table that the gender variable has an effect on the second-order variable after Creatine.

Table 1: Parameters Estimation by Tobit model for the Urea variable

Steps	Variable's in model	β	S.E	T	P-value
1	Constant	26.4128	6.8897	3.834	0.0001
	Age	-0.0150	0.08449	-0.1771	0.8594
	Gender	-2.6313	2.6290	-1.001	0.3169
	Creatine	14.8963	3.1420	4.741	0.0000
2	Constant	25.5435	4.8473	5.270	0.0000
	Gender	-2.5280	2.5650	-0.9856	0.3243
	Creatine	14.9107	3.1484	4.736	0.0000
3	Constant	21.9604	3.3307	6.593	0.0000
	Creatine	14.9481	3.1842	4.694	0.0000

Table (2) below represents a table of criteria values (AIC, BIC, H-QIC) for the Tobit model for the dependent variable Urea. We note that the criteria value, AIC , BIC , H-QIC for each step of the table (1) and the values of those criteria change according to the steps, when all criteria have the highest values in the first step and the lowest value in the third step, on the basis of which they will be compared with the values of the same criteria for other models.

Table 2: Comparative criteria values for the Urea variable by Tobit model

Steps	AIC	BIC	H-QIC
1	180	194.4	185.8
2	178	189.5	182.7
3	177	185.6	180.5

Now we present results for application using the LBXIIEE model, table (3) show that the LBXIIEE model was estimated as the variable Urea was the dependent variable, and based on the P-value, the effect of all variables was estimated and calculated in the first step, in the second step, age was excluded as it proved to be less influential. In the third step, the gender variable was excluded because it is the least influential in this step, and the Creatine variable is more influential on the dependent variable Urea for the LBXIIEE model according to the three steps shown in the table.

Table 3: Parameters Estimation by LBXIIEE model for the Urea variable

Steps	Variable's in model	β	S.E	T	P-value
1	Constant	25.7281	20.6209	1.25	0.2144
	Age	-0.0116	0.0962	-0.12	0.9044
	Gender	-1.8090	9.4501	-0.19	0.8485
	Creatine	5.0792	0.8585	5.92	0.0000
	a	19.7866	1.6719	11.83	0.0000
	b	3.4332	0.4438	7.74	0.0000
	α	29.2014	26.3378	1.11	0.2596
	β	0.1525	0.1324	1.15	0.2514
	σ	153.01	21.1483	7.24	0.0000
2	Constant	53.9976	83.0287	2.76	0.0066
	Gender	-0.6735	3.3793	-0.20	0.8423
	Creatine	5.1814	0.7824	6.62	0.0000
	a	6.6397	0.3359	19.77	0.0000
	B	2.8219	0.0260	108.63	0.0000
	α	45.6659	0.3155	144.76	0.0000
	β	0.1564	0.1588	0.98	0.3265
	σ	157.76	83.0287	1.90	0.0596
3	Constant	106.03	30.6724	3.46	0.0007
	Creatine	5.1379	0.8836	5.81	0.0000
	a	26.6919	1.4207	18.79	0.0000
	b	6.0600	0.0093	652.33	0.0000
	α	32.9634	0.0404	816.13	0.0000
	β	0.1480	0.0976	1.52	0.1320
	σ	167.06	62.8935	2.66	0.0089

The following table (4) represents the criteria table of the LBXIIEE model for the Urea dependent variable, in which we note that the values of the standards AIC, BIC, H-QIC are variable according to the steps. These criteria reach their lowest value in the last step. We also note that these values are the lowest values of the criteria when compared with the values of the same criteria for the Tobit model and the LBXIIW model.

Table 4: Comparative criteria values for the Urea variable by LBXIIEE model

1	176.9	202.1	186.8
2	165.3	188.3	174.7
3	148.7	168.9	156.9

We notice from tables (1), (3), that the two variables, Urea and Creatine, both have a clear effect on the other., the gender variable had a second degree effect on the Urea variable according to the Tobit model and the LBXIIEE model,

5. Conclusions

Through the results obtained from application some censored regression models for renal failure data, we conclude that the LBXIIEE model is the best model compared to the Tobit model according to the model selection criteria values, where the values of these criteria for the LBXIIEE model were less than the values of the same criteria for the Tobit model when urea as dependent variable in the model. Also we note that the criteria values (AIC, BIC, H-QIC) change and decrease according to the steps used to estimate and analyze the two models (Tobit, LBXIIEE) according to the dependent variable specified Urea.

References

[1] Al-Rubaie, M. A. (2019). “Some nonparametric capabilities of the first type of observational data.” Master’s Thesis - College of Administration and Economics - University of Baghdad.

[2] Bang, H., & Tsiatis, A. A. (2000), Estimating medical costs with censored data, Biometrika,87(2), 329-343.

[3] Carson, R. T., & Sun, Y. (2007). The Tobit model with a non-zero threshold. The Econometrics Journal, 10(3), 488-502.

[4] Chen, E. H., & Dixon, W. J. (1972). Estimates of parameters of a censored regression sample. Journal of the American Statistical Association, 67(399), 664-671.

[5] Chib, S. (1992). Bayes inference in the Tobit censored regression model. Journal of Econometrics, 51(1-2), 79-99.

[6] Cordeiro, G. M., Yousof, H. M., Ramires, T. G., & Orte ga, E. M. (2018). The Burr XII system of densities: properties, regression model and applications. Journal of Statistical Computation and Simulation, 88(3), 432-456.

[7] Gupta, R. D., & Kundu, D. (2001). Exponentiated exponential family: an alternative to gamma and Weibull distributions. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 43(1), 117-130.

[8] Hadi, F. H. 2017. “Studying the variables affecting the number of active sperms using the Tobit Model.” Al-Qadisiyah University - College of Administration and Economics - Statistics Department.

[9] Ibrahim, M., Ea, E. A., & Yousof, H. M. (2020). A new distribution for modeling lifetime data with different methods of estimation and censored regression modeling. Statistics, Optimization & Information Computing, 8(2), 610-630.

[10] Karim, A. A. (2018) “Comparison between the contraction method and the maximum likelihood method for estimating the survival function of the Weibull distribution in the case of type I censored data using simulation.” Master’s Thesis in Statistics - College of Administration and Economics - University of Karbala.

[11] Paranaíba, P. F., Ortega, E. M., Cordeiro, G. M., & Pescim, R. R. (2011). The beta Burr XII distribution with application to lifetime data. Computational Statistics & Data Analysis, 55(2), 1118-1136.

[12] Z., Lin, D. Y., & Ying, Z. (2006). On least-squares regression with censored data. Biometrika, 93(1), 147-161.