Sigma J Eng & Nat Sci 36 (2), 2018, 513-521

Sigma Journal of Engineering and Natural Sciences
Sigma Mühendislik ve Fen Bilimleri Dergisi

Research Article

ESTIMATION OF CENSORED REGRESSION MODEL IN THE CASE OF NON-NORMAL ERROR

İsmail YENİLMEZ*[1], Yeliz MERT KANTAR[2], Şükrü ACITAŞ[3]

[1]*Department of Statistics, Anadolu University,ESKISEHİR;* ORCID:0000-0002-3357-3898
[2]*Department of Statistics, Anadolu University,ESKISEHİR;* ORCID:0000-0001-7101-8943
[3]*Department of Statistics, Anadolu University,ESKISEHİR;* ORCID:0000-0002-4131-0086

Received: 19.12.2017 Accepted: 12.02.2018

ABSTRACT

For the censored regression model, it is well-known that while classical least squares estimation yields biased and nonconsistent estimator, maximum likelihood estimator (MLE) is consistent and efficient. Tobit estimator (Tobit model) based on MLE of normal error distribution is commonly-used estimation method for estimating censored regression in econometric literature. However, while the Tobit estimator works well for normal error distribution, its estimates may be inefficient in the case of non-normal errors. To solve this problem, different error distributions for the censored regression model have been proposed and tested in the literature. In this study, we consider the censored regression model based on the generalized logistic distribution. Generalized logistic distribution is very flexible distribution and approximates normal distribution for the special parameter cases. The considered estimator for the censored regression is evaluated by means of a simulation study designed in different combination of various error distributions and sample sizes. The results of the simulation show that the estimator of the censored regression model based on the generalized logistic distribution provides good performance for different error distributions and it is particularly good for small sample sizes. Moreover, when it is compared to classical Tobit estimator, efficiency loss of the considered estimator is very small for normal error distribution.
Keywords: Censored dependent variable, censored regression model, generalized logistic distribution, maximum likelihood estimation.

1. INTRODUCTION

In the regression model, the dependent variable, which takes only two values or do not take a negative value or takes values at certain intervals, is expressed as the limited dependent variable. These situations can be classified into three categories: i. Truncated regression models, ii. Censored regression models and iii. Dummy endogenous models [1]. If the continuous dependent variable is only observable under certain conditions, this situation is expressed as the censored variable. In this case, ordinary least squares (OLS) estimator gives biased and inconsistent results. To solve a part of this problem, the censored normal regression model or tobit model, was proposed by Tobin [2]. However, the maximum likelihood estimation (MLE) of the tobit model, which depends on the assumption of normality, yields inconsistent results when the errors are

* Corresponding Author: e-mail: ismailyenilmez@anadolu.edu.tr, tel: (222) 335 05 80 / 4688

non-normally distributed. In such cases, new estimators which are less sensitive to normality assumption should be investigated. For this purpose, various estimators have been proposed to relax the normality assumption in the literature. One of the proposed estimators is partially adaptive estimator (PAE), which is known as the semi-parametric estimator based on density estimation [3]. On the other hand, Powell's censored least absolute deviations estimator (CLAD) and symmetrically trimmed least squares estimator (STLS) are well-known non-density based estimators [4, 5]. For censored regression, the density based estimators contain the unknown underlying error as well as regression parameters [6] and they are categorized as fully adaptive and partially adaptive estimators (PAEs). While the fully adaptive estimator is usually based on a non-parametric estimate of the unknown distribution, a partially adaptive estimator takes into account a parametric approximation of the true unknown error distribution. In that case, it can be concluded that the PAEs can be used for relaxing the normality assumption. For instance, the PAE based on an error structure described by a location-scale mixture of normal distributions has been introduced in [6]. The PAEs based on some flexible distributions are used for a comparison with other estimators for the censored regression in the case of skewed and leptokurtic error distributions in [7]. Furthermore, several PAEs that cover a wide range of distributional characteristics are proposed for the censored regression model in [8]. In addition, extensions of the classical normal censored model are developed in [9, 10]. On the other hand, different estimation procedure such as modified maximum likelihood methodology is used to estimate the unknown parameters of the censored regression model [11].

In this study, classical estimation method for censored variable (Tobin's censored normal regression estimator or MLE for censored normal regression – hereafter, Tobit) and a partially adaptive estimator based on the generalized logistic distribution (hereafter, PAEGLD) are examined in the case of non-normality problem[†]. Thus, the PAEGLD is introduced for the censored regression model and also a simulation study is used to compare the performances of the Tobit and PAEGLD estimators under different error distribution assumptions. It should be noted that a partially adaptive estimator based on the generalized normal distribution (PAEGND) is introduced by authors of this study for the censored regression model [24].

The paper is organized as follows: Section 2 briefly reviews Tobit. Section 3 presents general framework of PAE and introduces PAEGLD for the censored dependent variable. Section 4 contains the simulation study carried out under the scope of the study. Finally, the obtained results are presented and discussed regarding the relevant literature in conclusion section.

2. THE CENSORED REGRESSION MODEL

Let probability density function (PDF) and cumulative distribution function (CDF) of Y^* be $f(.)$ and $F(.)$, respectively. We can record only those values of y^* greater than constant c or for $(y^* \le c)$, these values are recorded as c [1]. In this case, the probabilities of $(y^* \le c)$ is given as follows:

$$P(y_i = c) = P(y_i^* \le c) \tag{2.1}$$

For singly left censored at $x = c$, the likelihood function is given as follows:

$$L = K_1 \left[F(c) \right]^{n_c} \prod_{i=1}^{n-n_c} f(x_i) \tag{2.2}$$

[†] An earlier version of this study was presented in International Workshop on Mathematical Methods in Engineering (MME2017).

where n is sample of size, n_c and K_1 respectively denote the number of censored observation and ordering constant that do not depend on the parameters, respectively [12].

To limit the study to a certain frame, samples to be considered in this study include only singly left censored cases although different types of censoring exist (singly right or left censored, doubly censored, centrally censored and progressively censored). In addition, censored samples are generally classified as Type I and Type II. In type I, the measurement threshold is fixed and the number of censored data points varies. In type II, the number of censored data points is fixed and the implicit threshold varies [13].

The Tobit model is defined as a latent variable model. Regression model defines as $y_i^* = x_i^{'}\beta + u_i \ \left(u_i \sim N\left(0,\sigma^2\right)\right)$ and c denotes censoring point.

$$\begin{aligned} y_i &= y_i^* \qquad y_i^* > c \\ y_i &= c \qquad d.d. \end{aligned} \tag{2.3}$$

where y_i which is observed value equals y_i^* when $y_i^* \geq c$ on the other hand $y_i = c$ when $y_i^* < c$. The latent variable y_i^* satisfies the classical linear model assumption. It has a normal and homoscedastic distribution with a linear conditional mean [14].

An iteration method for obtaining the MLE of parameters is suggested in [15]. However, using advanced computer technology, the MLE for the censored models are not much more difficult than the OLS estimator.

Another important point in estimating the Tobit model is error distribution assumption. The distribution of error terms in Tobit model is assumed to be normal. Therefore, resulting Tobit estimator is biased and inconsistent under non-normality. In the related literature, fully adaptive or quasi-maximum likelihood estimators are proposed as an alternative to Tobit when non-normality occurs.

3. PARTIALLY ADAPTIVE ESTIMATION

The partially adaptive estimator depends on the flexible distribution. The PAEs can be defined in (3.1) when *f* and *F* are considered as PDF and CDF of a flexible distribution family respectively:

$$\ln L\left(\beta;\Theta\right) = \sum_{Y_t \leq c} \ln F\left(c - X_t\beta;\Theta\right) + \sum_{c<Y_t} \ln f\left(Y_t - X_t\beta;\Theta\right) \tag{3.1}$$

In this study, we use the generalized logistic distribution (GLD) for the censored regression model as an alternative to normal distribution since GLD is a flexible enough to accommodate the different characteristics of data such as skewness and kurtosis.

3.1. The Generalized Logistic Distribution (GLD)

Suppose that the random variable X has the logistic distribution (LD), then the PDF and CDF of *X* are given as follows:

$$f\left(x;\mu,\sigma\right) = \frac{\exp\left[-\left(x-\mu\right)/\sigma\right]}{\sigma\left(1+\exp\left[-\left(x-\mu\right)/\sigma\right]\right)^2} \qquad -\infty < x < \infty \tag{3.1.1}$$

$$F\left(x;\mu,\sigma\right) = \frac{1}{1+\exp\left[-\left(x-\mu\right)/\sigma\right]} \qquad -\infty < x < \infty \tag{3.1.2}$$

respectively.

There are different generalizations of LD. The skew logistic distribution based on idea of Azzalini's skew normal distribution and the proportional reversed hazard logistic distribution based on idea of proportional reserved hazard family are introduced in [16]. The beta generalized logistic distribution previously introduced in the literature is examined with details in [17]. On the other hand, the generalized logistic distribution accommodates several different families of probability distribution. In general, different forms for generalizations of the logistic distribution have been listed as four types [18]. According to this list, the type I, or GLD, is used in this study. The PDF and CDF for the general form of type I GLD are presented as follows:

$$f(x;\alpha)=\frac{\alpha\exp(-x)}{(1+\exp(-x))^{\alpha+1}} \qquad -\infty<x<\infty \tag{3.1.3}$$

$$F(x;\alpha)=\frac{1}{(1+\exp(-x))^{\alpha}} \qquad -\infty<x<\infty \tag{3.1.4}$$

respectively.

The location-scale case of GLD is given as follows:

$$f(y;\mu,\sigma,\alpha)=\frac{\alpha}{\sigma}\frac{\exp\left[-(y-\mu)/\sigma\right]}{\left(1+\exp\left[-(y-\mu)/\sigma\right]\right)^{\alpha+1}} \quad -\infty<y<\infty,\ \sigma>0,\ \alpha>0 \tag{3.1.5}$$

$$F(y;\mu,\sigma,\alpha)=\left(1+\exp\left[-(y-\mu)/\sigma\right]\right)^{-\alpha} \quad -\infty<y<\infty,\ \sigma>0,\ \alpha>0 \tag{3.1.6}$$

See also [19]. Also it should be mentioned that some of the features of the GLD are discussed by Zelterman [20] and Johnson et al [18].

The MLE for the GLD is investigated in [21]. Also point and interval estimation for the GLD under progressive type II censoring is presented [22]. In case of left censoring, the likelihood function is given in equation (3.1.7). For singly left censored at $c=0$, the likelihood function for the censored regression under the GLD is presented as follows:

$$L(y;\mu,\sigma,\alpha)=\prod_{y\le 0}F(y)\prod_{0<y}f(y)$$

$$logL(y;\mu,\sigma,\alpha)=\sum_{y\le 0}^{n_c}\ln F(y)+\sum_{0<y}^{n-n_c}\ln f(y)$$

$$logL(y;\mu,\sigma,\alpha)=\sum_{y\le 0}^{n_c}\left[(-\alpha)\ln\left(1+\exp\left[-(y-\mu)/\sigma\right]\right)\right]$$

$$+\sum_{0<y}^{n-n_c}\left[\ln(\alpha)-\ln(\sigma)-(y-\mu)/\sigma-(\alpha+1)\ln\left(1+\exp\left[-(y-\mu)/\sigma\right]\right)\right] \tag{3.1.7}$$

where n_c is the number of the censored observations at $c=0$ and n is the total sample size. PAE under the GLD for the censored regression model is denoted as PAEGLD in this study.

4. SIMULATION RESULTS

The simulation study was conducted to compare the relative bias and mean square error (MSE) of OLS_C that use only the observed data, TOBIT and PAEGLD under different error distribution. To simplify of computation, censoring point is taken as zero. In the simulation procedure, 1000 data sets are generated and the sample size is taken as 50, 100, 200 and 500.

The linear regression model is designed as:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ for } i = 1,...,n \quad (4.1)$$

without loss of generality by taking $\beta_0 = 0$ and $\beta_1 = 1$, where x is distributed as uniform distribution $(U(0,1))$ and ε_i is distributed as standard normal, student's t distribution with three degrees of freedom, scale-contaminated normal distribution $0.8N(0,1/3)+0.2N(0,3)$ and standard Laplace distribution [23]. The simulation results are reported in Tables 1-4 for each error distributions, respectively.

It is obvious from Table 1 according to MSE criterion that TOBIT performs better than the other estimators when errors are normally distributed, as expected for all sample cases. However, MSE values of PAEGLD are very close to TOBIT as can be seen in Table 1. When Table 2 is considered for the student-t distribution which is fat-tailed distribution, it can deduce that good performances of PAEGLD in terms of MSE and Bias are still seen. Similarly, Table 3 shows that the PAEGLD apparently performs the best with respect to Bias and MSE criteria under the mixture-normal error distribution. Lastly, when Table 4 is considered under the Laplace error distribution, the PAEGLD shows a better performance relative to the TOBIT and OLS_C according to MSE criterion.

Table 1. Simulation results for censored regression under Normal error distribution

$y = a + b*x$	Normal			
Slope $b = 1$	n=50		n=100	
	Bias	MSE	Bias	MSE
OLS_C	0.31452	0.22642	0.31138	0.16560
TOBIT	0.00279	0.23130	0.00313	0.12195
PAEGLD	0.02616	0.32516	0.01693	0.15521
	n=250		n=500	
	Bias	MSE	Bias	MSE
OLS_C	0.31970	0.13020	0.32395	0.11713
TOBIT	0.00731	0.04835	0.01224	0.02351
PAEGLD	0.00301	0.06016	0.00509	0.02694

Table 2. Simulation results for censored regression under Student-t error distribution

$y = a + b*x$	Student-t			
Slope $b = 1$	n=50		n=100	
	Bias	MSE	Bias	MSE
OLS_C	0.32485	0.45065	0.33114	0.33319
TOBIT	0.00106	0.70323	0.00274	0.41282
PAEGLD	0.06565	0.48676	0.01521	0.24149
	n=250		n=500	
	Bias	MSE	Bias	MSE
OLS_C	0.32845	0.17273	0.33218	0.14307
TOBIT	0.00246	0.12978	0.00121	0.06853
PAEGLD	0.00683	0.07462	0.04064	0.04595

Table 3. Simulation results for censored regression under Mixture-normal %80 error distribution

$y = a + b*x$	Mixture-normal			
Slope $b = 1$	n=50		n=100	
	Bias	MSE	Bias	MSE
OLS_C	0.16714	0.26442	0.16895	0.13961
TOBIT	0.01122	0.47333	0.00328	0.21816
PAEGLD	0.00361	0.06928	0.02834	0.03754
	n=250		n=500	
	Bias	MSE	Bias	MSE
OLS_C	0.17625	0.07117	0.17134	0.04852
TOBIT	0.00970	0.07559	0.00096	0.04404
PAEGLD	0.04174	0.02204	0.04299	0.00907

Table 4. Simulation results for censored regression under Laplace error distribution

$y = a + b*x$	Laplace			
Slope $b = 1$	n=50		n=100	
	Bias	MSE	Bias	MSE
OLS_C	0.06468	0.10450	0.08130	0.07535
TOBIT	0.00114	0.12575	0.01925	0.08437
PAEGLD	0.00262	0.09169	0.01593	0.06233
	n=250		n=500	
	Bias	MSE	Bias	MSE
OLS_C	0.06612	0.04291	0.06926	0.02466
TOBIT	0.00373	0.04982	0.00595	0.02520
PAEGLD	0.00470	0.03359	0.00950	0.01803

Finally, Figs. 1-4 demonstrate the PDF and CDF of uncensored and censored data under different error distributions. For economy, several graph examples for small and big sample sizes are shown here (*n=50 and n=500*). It can be seen from Figs. 1-4 that the censoring affects the distribution of the data.

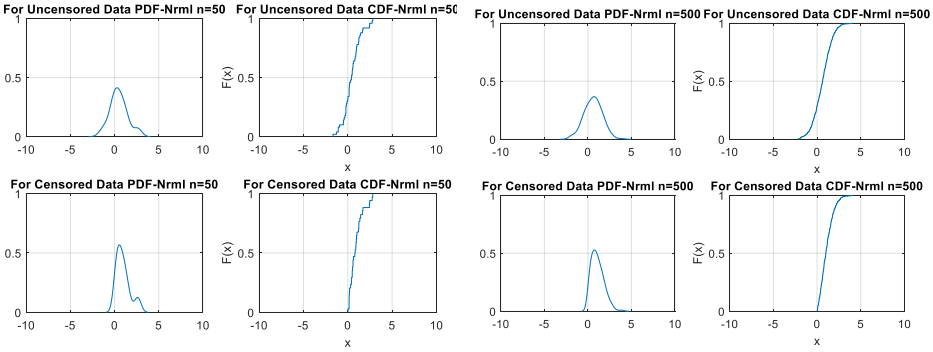

Figure 1. The PDF and CDF of uncensored and censored data under Normal error distribution

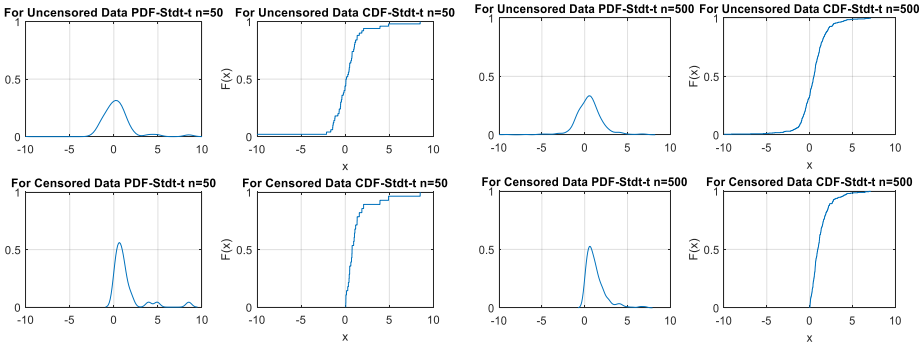

Figure 2. The PDF and CDF of uncensored and censored data under Student-t error distribution

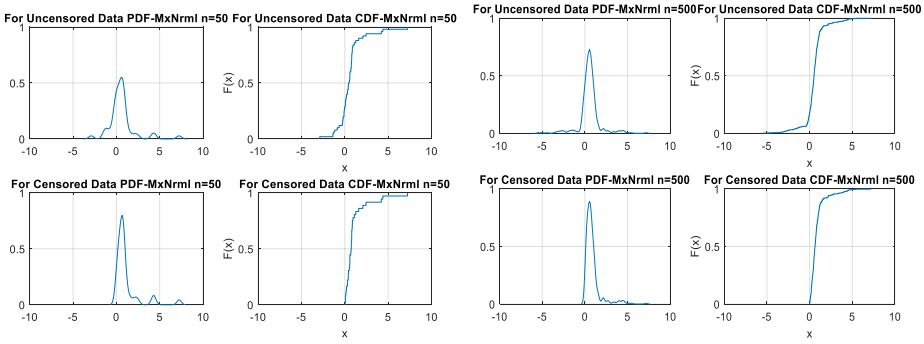

Figure 3. The PDF and CDF of uncensored and censored data under Mixture-normal %80 error distribution

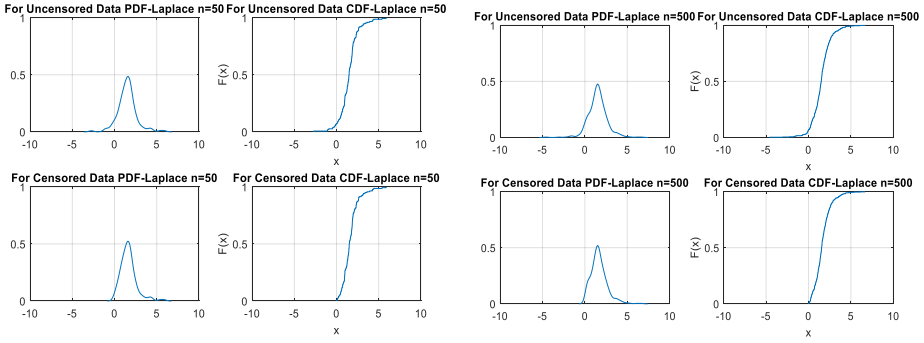

Figure 4. The PDF and CDF of uncensored and censored data under Laplace error distribution

5. CONCLUSION

Several methods for censored dependent variable have been proposed and compared in the literature. Tobit is well-known estimator for censored regression model. However, the Tobit estimation depends on normal distribution, thus it may produce inconsistent results in the case of non-normal errors. To solve this problem, partially adaptive estimators have been proposed in literature. For the censored regression model, we consider a partially adaptive estimator based on the generalized logistic distribution, which is flexible than normal distribution and able to accommodate the skewness and/or kurtosis. We compare the PAEGLD and Tobit via a simulation study for different error distributions and different sample sizes. Simulation results show that the PAEGLD presents very little loss of efficiency relative to the Tobit estimator for a normal error distribution. Furthermore, the PAEGLD is seen to be useful in small samples and more robust to underlying distributional assumptions than TOBIT. Finally, The PAEGLD is clearly the preferred estimator relative to TOBIT in the case of non-normal error distribution.

Acknowledgments

This study was supported by Anadolu University Scientific Research Projects Commission under the grant no: 1610F661 and also 1705F419.

REFERENCES

[1] Maddala, G.S. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge University Press, 1983.

[2] Tobin, J. Estimation of Relationships for Limited Dependent Variables. *Econometrica*, 26, 24-36, 1958.

[3] Pagan, A., Ullah, A. *Nonparametric Ecometrics*. Cambridge, UK: Cambridge University Press, 1999.

[4] Powell, J.L. Least Absolute Deviations Estimation of the Censored Regression Model. Econom 25:303-325, 1984.

[5] Powell, J.L. Symetrically trimmed least squares estimation for Tobit models. Econometrica 54:1435-1460, 1986.

[6] Caudill, S. B. A Partially Adaptive Estimator for The Censored Regression Model Based on A Mixture of Normal Distributions. *Stats Methods Appl*, 21:121-137, 2012.

[7] McDonald, J. B., Xu, Y. J. A Comparison of Semi-parametric and Partially Adaptive Estimators of the Censored Regression Model with Possibly Skewed and Leptokurtic Error Distributions. *Economics Letter*, 51(2), 153-159, 1996.

[8] Lewis, R. A., McDonald J. B. Partially Adaptive Estimation of the Censored Regression Model. *Economic Reviews*, 33 (7), 732-750, 2014.

[9] Arellano-Valle, R.B., Castro, L.M., González-Farías, G., Munoz-Gajardo, K.A. Student-t Censored Regression Model: Properties and Inference. *Stat Methods Appl*,21:453-473, 2012.

[10] Kantar, Y.M., Yenilmez, I., Acitas, S. Estimation based on generalized logistic distribution for the censored regression model. In *Proceeding of the International Workshop on Mathematical Methods in Engineering*. ISBN 978-975-6734-19-3, Cankaya University Press, 2017.

[11] Acitas, S., Yenilmez, I., Senoglu, B. and Kantar, Y. M. Modified Maximum Likelihood Estimation for the Censored Regression Model. The 13th IMT-GT International Conference on Mathematics, Statistics and Their Applications, Universiti Utara Malaysia December 4-7, 2017.

[12] Cohen, A.C. *Truncated and Censored Samples: Theory and Applications.* NY: Taylor & Francis Group, 1991.
[13] David, H. A. and Nagaraja, H. N.: Order Statistics, 3rd Edn.,Wiley, Hoboken, NJ, 458 pp., 2003.
[14] Wooldridge, J. *Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press, 2002.
[15] Fair, R. C. A Note on the Computation of the Tobit Estimator. *Econometrica*, 45(7):1723-7, 1977.
[16] Gupta, R. D., Kundu, D. Generalized logistic distributions. *J. Appl. Statist.*, 18(1):51–66, 2010.
[17] Nassar, M.M., Elmasry, A. A study of generalized logistic distribution. *Journal of the Egyption Mathematical Society*, 20 126-133, 2012.
[18] Johnson, N.L., Kotz, S., Balakrishnan, N. *Continuous Univariate Distributions*, vol. 2, Wiley, New York, second ed., 1995.
[19] Balakrishnan, N., Leung, M. Y. Order statistics from the Type I generalized logistic distribution. *Commun. Statist. Simulation Comput.* 17(1):25–50, 1988.
[20] Zelterman, D. Parameter Estimation in the generalized logistic distribution. *Comput. Stat. Data An.* 5, 177–184, 1987.
[21] Shao, Q. Maximum likelihood estimation for Generalised logistic distributions. *Communications in Statistics - Theory and Methods*. 31:10, 1687-1700, 2002.
[22] Asgharzadeh, A. Point and interval estimation for a generalized logistic distribution under progressive type II censoring, *Communications in Statistics - Theory and Methods*, 35:9, 1685-1702, 2006.
[23] Kantar YM, Usta I, Acitas S. A Monte Carlo simulation study on partially adaptive estimators of linear regression models. *J Appl Stat*. 38(8), 1681–99, 2011.
[24] Yenilmez I., (2017) Limited Dependent Variable Models and Estimation Methods, MS Thesis, *Graduate School of Sciences, Anadolu University*, Eskisehir, Turkey.