An Alternative Estimation Method Based on Alpha Skew Logistic Distribution for Parameters of Censored Regression Model

Ismail Yenilmez[1,*], Yeliz Mert Kantar[1]
[1]*Department of Statistics, Science Faculty, Eskisehir Technical University, Yunus Emre Campus, 26470, Eskisehir, Turkey*

***Abstract*— In the case of censored data, it is often seen that the error distribution is skewed and multimodal. Ordinary least squares (OLS) estimator, which often gives biased and inconsistent results for censored data, and Tobit estimator, which is frequently used in censored data estimation and gives inconsistent results when some assumptions are not met, are also problematic in the presence of skewed and multimodal distribution of error terms. A new estimator is proposed as an alternative to the two conventional estimators used in the case of censored data. For censored regression model, an estimation method, known as partial adaptive or quasi-maximum likelihood estimator, has been introduced based on the alpha skewed logistic distribution, which is a flexible error distribution. According to the bias and mean square error (MSE), new estimator is superior for estimating the coefficients of the censored regression model under the skewed and multimodal error distribution.**

***Keywords*— partially adaptive estimator, maximum likelihood estimator, censored data, tobit, ordinary least squares estimator.**

I. INTRODUCTION

Limited dependent variables (LDVs) are restricted in some way. For instance; binary choice variables, likert scale's ordered values, non-negative integer values, truncated values and censored values are seen as LDV. Various methods have been developed for the analysis of LDVs. For boolean-binary choice values, Logit and Probit models can be used. The ordered logit can be considered as another model used in the analysis of values with LDVs. The ordered logit model also known as ordered logistic regression or proportional odds model can be used to analysis a survey data whose responses are likert-scale ordered values. In this study, censored data, which is one of the LDVs, is concerned. For censored data, first model has been proposed in a pioneering article [1] by Nobelist Tobin. This is the first study that takes censored data into account. Tobin has named the model as the model for the limited dependent variables, Goldberger has named it as Tobit model [2]. Estimation procedure for Tobit model is the maximum likelihood method, called as Tobit estimator, and modification of this method is studied in [3]. The Tobit estimator depends on the assumption that the error terms are normally distributed. If this assumption is not met, Tobit gives inconsistent results. In addition, it is known that the ordinary least squares (OLS) method gives biased and inconsistent results for censored regression. The normal distributed error terms are an assumption but often cannot be achieved. Some new estimators that are less sensitive to the assumptions are being investigated. Alternative semi-parametric estimators are categorized as non-density and density based estimators [4]. Censored least absolute deviation (CLAD) and symmetrically trimmed least squares (STLS) are respectively proposed as non-density based alternative estimator by [5,6]. Partial adaptive estimators (PAEs) and fully adaptive estimators (FAEs) can be seen as density based estimators for censored regression [7]. FAEs is usually based on a non-parametric estimate of the unknown distribution, whereas a PAEs is based on a parametric approximation to the true unknown error distribution. So PAEs may be more advantageous than FAEs. A detailed discussion of this is presented in [7]. PAEs based on specific distributions determined for special cases are presented by [7-9]. Generalized normal, generalized logistic and maximum entropy distributions are have used in PAEs procedure for censored data [10-12]. In addition to these modifications, the use of flexible distributions involving the normal distribution allows the generalization of the Tobit model. A generalized Tobit model based on power-normal distribution is proposed [13].

In this study, a modification is proposed for the case where the error distribution is not normally distributed. Moreover, the cases where the error distribution is skewed and bimodal are discussed. PAE procedure based on the alpha skew logistic (ASLG(α)) distribution is introduced for censored regression model. In this context, Tobit and OLS estimators are mentioned in the second part. In the third section, the PAE process and the ASLG(α) distribution are summarized and the PAE based on this distribution is introduced. In the fourth chapter, simulation conducted in the context of the study are presented. Finally, the findings are shared in the conclusion section.

II. ESTIMATION FOR CENSORED DATA

The Tobit estimator and OLS are often used when the dependent variables are censored. If normality assumption is not met, Tobit gives inconsistent results. In addition, it is known that OLS yields biased and inconsistent estimates for the censored regression model. In such cases, alternative estimators are used. PAE, one of the density-based estimators, is an alternative method used in this study. Tobit

Manuscript received November 02, 2019; accepted December 27, 2019.
*Corresponding author: ismailyenilmez@eskisehir.edu.tr

and OLS are summarized in this section.

A. Tobit Estimation Procedure

The basic idea for Tobit is to consider the probability of different sampling for each observation, depending on whether the latent dependent variable falls above or below the specified censoring point. The censored regression model is defined as

$$Y_i = max(c, X_i\beta + \varepsilon_i) \quad (1)$$

where Y_i is the observed value of the dependent variable, c is the censoring and ε_i random disturbance is assumed to be ($\varepsilon_i \sim N(0, \sigma^2)$). In addition, this model is often shown with the help of latent variable. Regression model is defined as

$$Y_i^* = X_i\beta + \varepsilon_i \quad (2)$$

and $\varepsilon_i \sim N(0, \sigma^2)$

$$Y_i = Y_i^* \qquad Y_i^* > c \quad (3a)$$

$$Y_i = c \qquad Y_i^* \leq c \quad (3b)$$

The following equation can then be expressed.

$$P(Y_i = c) = P(Y_i^* \leq c) \quad (4)$$

and likelihood function is written as fallows

$$\mathrm{L} = \left(\Phi(c)\right)^{n_c} \prod_{i=1}^{n-n_c} \phi(X_i) \quad (5)$$

where $\Phi(.)$ and $\phi(.)$ are the cumulative distribution function (CDF) and probability density function (PDF) of standard normal distribution, respectively. n and n_c are sample size and number of censored observation, respectively. To emphasize the normal distribution assumption, the Tobit model is also called a censored normal regression model. The maximum likelihood estimation of this model is the Tobit estimator.

B. Ordinary Least Squares Estimates for Censored Data

OLS based on data in (3a) means that estimation procedure has been applied to only uncensored data ($OLS_{Uncensored}$). In addition, OLS can be applied to all data obtained from (3) (OLS_{Full}). It is clear that if the censor point c is 0, only positive data will be obtained from the (3a). In the literature, censorship point is often taken as 0. Therefore, the method abbreviated as $OLS_{Uncensored}$ in this study has been abbreviated as POLS by [8]. In addition, OLS_{Full} and $OLS_{Uncensored}$ yields biased and inconsistent estimates for the parameters of the censored regression model [14-15].

III. PARTIALLY ADAPTIVE ESTIMATION PROCEDURE

It is known that the Tobit estimate is strictly dependent on the normal distribution assumption, but this assumption is not always satisfied. At this stage, a flexible distribution and estimation procedure based on this distribution can be used. In that case partially adaptive estimation (PAE) or quasi-Maximum likelihood (Q-ML) is a good alternative estimation procedure: i. PAE can estimate both regression parameters (β) and distribution parameters (Θ). ii. PAE is based on flexible distribution.

General concept of PAE is based on minimizing the following likelihood function:

$$L = \prod_{i=0}^{n} f(Y_i - X_i\beta;\ \Theta) \quad (6)$$

where *f(.)* is a flexible PDF. For censored regression, PAE procedure can be defined based on following likelihood function

$$\mathrm{L} = \left(F(c - X_i\beta;\ \Theta)\right)^{n_c} \prod_{i=1}^{n-n_c} f(Y_i - X_i\beta;\ \Theta) \quad (7)$$

where *f(.)* and *F(.)* are a flexible PDF and CDF. The alpha skew logistic (ASLG(α)) distribution is used in this study.

A. The Alpha Skew Logistic (ASLG(α)) Distribution

It was emphasized that, besides the importance and prevalence of the normal distribution, the assumption of normality could not be provided for some data sets. Skewed, kurtosis, bimodal, and multimodal behaviours are reported to be common in the frequency curve. Various skewed and multimodal distributions have been developed as alternative to the normal distribution and logistic distribution in the literature. A detailed literature survey is presented by [16]. The asymmetry and bimodality behaviours can be seen in censored data. Because of the similarity of the normal distribution frequency graph and also suitability of its functional form (it can be mentioned that the alpha skew logistic distribution converges to the alpha skew normal distribution for certain parameters), the alpha skew logistic distribution ASLG(α) is used in this study. The PDF and CDF of ASLG(α) is defined as

$$f(x;\alpha) = \frac{3[(1-\alpha x)^2+1]exp(-x)}{[6+(\alpha\pi)^2](1+exp(-x))} \quad \alpha, x \in R \quad (8)$$

$$F(x;\alpha) = \frac{3[(1-\alpha x)^2+1]exp(-x)}{[6+(\alpha\pi)^2]} \left[\frac{(1-\alpha x)^2+1}{exp(-x)+1} + 2\alpha(1 - \alpha x)log[exp(x)+1] - 2\alpha^2 Li_2\left(-exp(x)\right)\right] \quad \alpha, x \in R \quad (9)$$

where $Li_n(x) = \sum_{k=1}^{\infty} \frac{x^k}{k^n}$ is poly-logarithm functions.

B. Partially Adaptive Estimator based on the Alpha Skew Logistic (ASLG(α)) Distribution

If the distribution of the observed data is skewed or bimodal, normal error distribution cannot be suitable. In such cases, it may be useful to use ASLG(α) distribution in the PAE procedure for CR model. If functions in (8) and (9) are used in (7), a PAE based on ASLG(α) ($PAE_{ASLG(\alpha)}$) is obtained for the censored data structure. For lower (left) censorship, the log-likelihood function based on the ASLG(α) distribution is defined as

$$\ell(\beta;\Theta) = (n-n_c)ln\left(\frac{3}{6+(\alpha\pi)^2}\right) + \sum_{Y_i>c} ln\left(\frac{[(1-\alpha x)^2+1]exp(-x)}{(1+exp(-x))}\right) +$$

$$(n_c)ln\left(\frac{3}{6+(\alpha\pi)^2}\right)+\sum_{Y_i\le c} ln\left(\left[\frac{(1-\alpha x)^2+1}{exp(-x)+1}+2\alpha(1-\alpha x)log[exp(x)+1]-2\alpha^2 Li_2(-exp(x))\right]\right) \quad (10)$$

where n and n_c are sample of size and number of censored observations, respectively. By minimizing (10), the obtained estimates are called as PAE estimates.

IV. SIMULATION

For comparing bias and mean square error (MSE) of conventional estimators (OLS and Tobit) and proposed $PAE_{ASLG(\alpha)}$ estimator, a simulation study have been conducted under scale-contaminated mixture-normal (MixNrml) distribution with different percentages. The sample size (n) is taken as 50, 250 and 500. $100000/n$ data sets are generated. Censoring point is taken as 0. The linear regression model is defined as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, \dots, n \quad (11)$$

where β_0 and β_1 is taken as 0 and 1, respectively. x is distibuted as uniform distribution $(x_i \sim U(0,1))$.

The analysis results of the MixNrml distribution with different percentages are provided in Table 1. In addition, the graphs of the error distributions used are presented in Figure 1. Graphs of different percentages are only presented for small sample for the sake of the brevity.

It is obvious from Table 1, $PAE_{ASLG(\alpha)}$ often performs better than the other conventional estimator when errors are from MixNrml distribution according to MSE and bias.

From Figure 2-5, it can be deduced that PDF and CDF of uncensored and censored empirical data is not normally distributed under different error distributions. Several graph examples for small sample sizes are shown here (n=50). However, other graphs obtained can be provided upon request of the authors. It can be seen from Figs. 2-5 that the censoring procedure affects the distribution of the data.

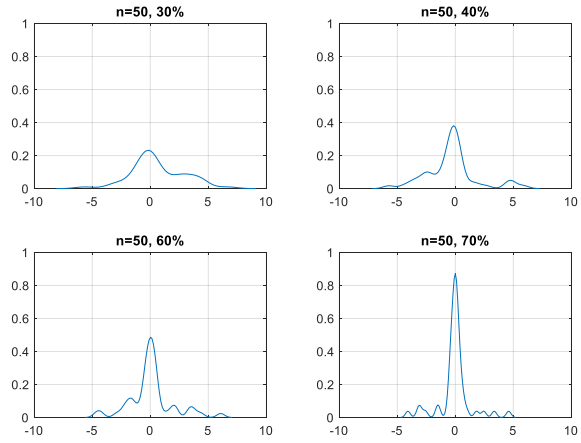

Fig. 1. Graphs of the error distributions with different percentages

TABLE I. SIMULATION RESULTS FOR CENSORED REGRESSION UNDER MIXNRML ERROR DISTRIBUTION WITH DIFFERENT PERCENTAGES.

		n=50		n=250		n=500	
Per	Est.	Bias	MSE	Bias	MSE	Bias	MSE
20%	OLS	0.4684	0.7551	0.4183	0.2787	0.3301	0.1761
	TOBIT	0.3195	2.2207	0.1927	0.2530	-0.1626	0.1480
	$PAE_{ASLG(\alpha)}$	0.0865	1.1445	0.0283	0.1179	-0.1190	0.0812
40%	OLS	0.0497	0.5195	0.1561	0.2083	0.3246	0.1840
	TOBIT	-0.2938	1.5183	-0.1738	0.2741	0.0131	0.1962
	$PAE_{ASLG(\alpha)}$	-0.1935	0.6889	-0.3201	0.3201	-0.1078	0.1194
60%	OLS	0.5177	0.7376	0.2606	0.1546	0.2583	0.1197
	TOBIT	0.3568	0.6815	0.0648	0.1804	0.0179	0.1174
	$PAE_{ASLG(\alpha)}$	0.2385	0.2748	-0.1140	0.0633	-0.0793	0.0216
80%	OLS	0.1403	0.2743	0.1076	0.1064	0.2648	0.0858
	TOBIT	0.0610	0.2991	-0.1300	0.1755	0.0951	0.0781
	$PAE_{ASLG(\alpha)}$	-0.1928	0.1383	-0.2017	0.1024	-0.0563	0.0264

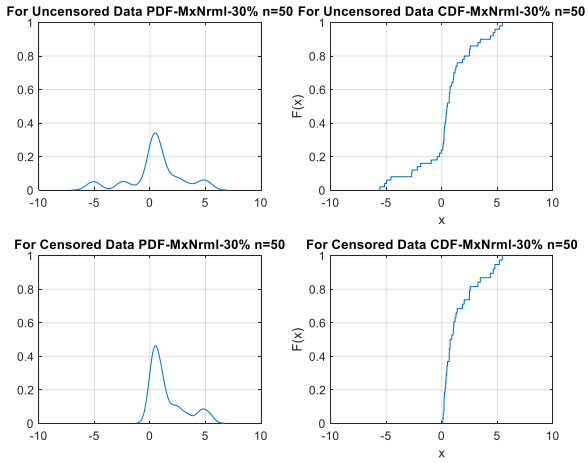

Fig. 2. The PDF and CDF of uncensored and censored data under Mixture-normal 20% error distribution

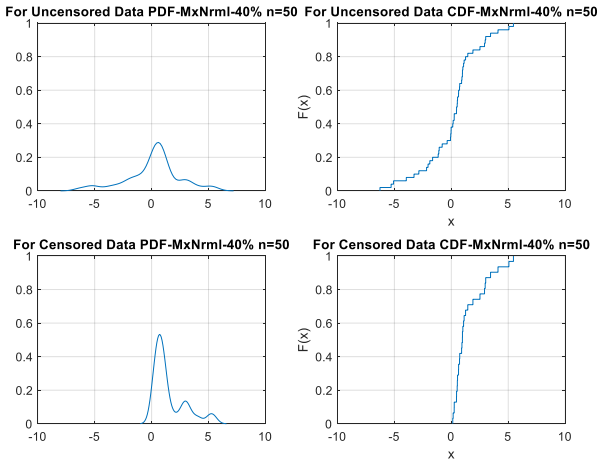

Fig. 3. The PDF and CDF of uncensored and censored data under Mixture-normal 40% error distribution

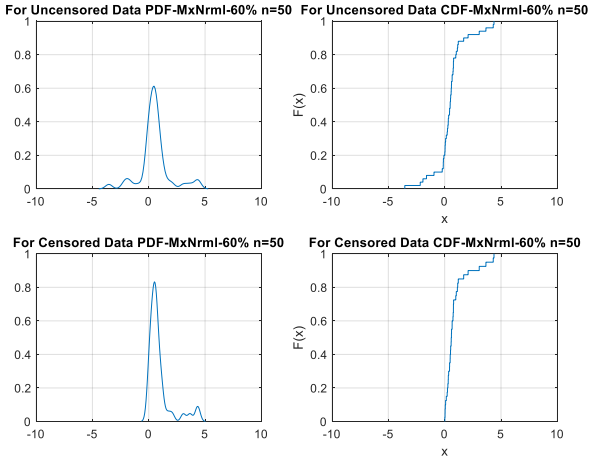

Fig. 4. The PDF and CDF of uncensored and censored data under Mixture-normal 60% error distribution

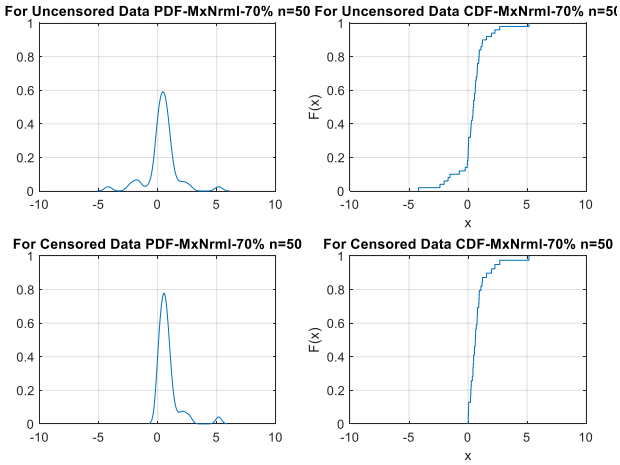

Fig. 5. The PDF and CDF of uncensored and censored data under Mixture-normal 80% error distribution

V. CONCLUSION

PAE based on the ASLG(α) distribution is proposed against conventional estimators for censored regression in case of skew and bimodal errors. OLS, which is considered as traditional methods, gives biased and inconsistent results. Tobit gives inconsistent results in case of violation of the assumption of normality. Sometimes normality assumption is not provided and skewed-bimodal frequency graphs for error terms can be encountered. In this study, especially skewed-bimodal errors are discussed for censored regression. The error term distribution of the censored regression model has been taken from the mixture normal distribution. In this way, skewness and bimodality have been included in the simulation. Similarity to the normal distribution frequency graph and the presence of thick tails make logistic distribution attractive. Furthermore, the functional form of logistic distribution is relatively soluble. According to the Bias and MSE, the superiority of the adaptive estimator based on alpha-skewed-logistic distribution over classical methods supports the idea in the motivation of the study. The proposed distribution would be an alternative for skewed and multimodal situations. In future studies, different modifications of logistic and normal distributions can be proposed for more general cases including skewed and bimodality.

ACKNOWLEDGMENT

This study was presented orally as abstract paper at the ICONDATA 2019 conference. This study was supported by Eskisehir Technical University Scientific Research Projects Commission under the grant no: 19ADP093.

REFERENCES

[1] Tobin, J. Estimation of Relationships for Limited Dependent Variables. Econometrica, 26, 24-36, 1958.
[2] Goldberger, A. S. Econometric Theory. New York: Wiley, 1964.
[3] Kennedy, P., A Guide to Econometrics (Fifth ed.). Cambridge: MIT Press. pp. 283–284. ISBN 0-262-61183-X, 2003.
[4] Pagan A, Ullah A, Nonparametric econometrics. Cambridge University Press, Cambridge, 1999.
[5] Powell J. L., Least absolute deviations estimation of the censored regression model. J Econom 25:303– 325, 1984.
[6] Powell J. L., Symmetrically trimmed least squares estimation for Tobit models. Econometrica 54:1435–1460, 1986.
[7] 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.
[8] 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.
[9] Lewis, R. A. and McDonald, J. B., "Partially Adaptive Estimation of the Censored Regression Model", Economic Reviews, 33 (7), 732-750, 2014.
[10] Yenilmez, I., Kantar, Y.M., "A Partially Adaptive Estimator for the Censored Regression Model Based on Generalized Normal Distribution", Proceedings of the 3rd IRSYSC-2017, Konya, Turkey, 2017.
[11] Yenilmez, I., Kantar, Y.M., Acitas, S., "Estimation of Censored Regression Model in the case of Non-Normal Error", Sigma J. Eng & Nat Sci 36 (2), 513-521, 2018.
[12] Usta, I., Kantar, Y.M., Yenilmez, I., "Estimation of Censored Regression Model with Maximum Entropy Distributions", Proceedings of the 4th IRSYSC-2018, İzmir, Turkey, 2018.
[13] Martínez-Flórez, G., Bolfarine, H. and Gómez, H. W., "The Alpha-power Tobit Model", Communications in Statistics Theory and Methods, 42:4, 633-643, 2013,
[14] Arabmazar, A., Schmidt, P., "An investigation of the robustness of the Tobit estimator to non-normality". Econometrica 50:1055–1069, 1982.
[15] Greene, W. H., "On the asymptotic bias of the ordinary least squares estimator of the Tobit model". Econometrica 49:505–513, 1981.
[16] Hazarika, P.J., Chakraborty, S., "Alpha-Skew-Logistic Distribution". IOSR-JM, 10(4):36-46, 2014.