A statistical analysis of the Type II Tobit model maximum Likelihood in case of non-ignoring missing data

Mona Emad El-Din Mohamed, Mervat El-Gohary
Department of Statistics- Faculty of commerce-Al-Azhar University (Girls Campus), Egypt.
Ahmed Amin El-Sheikh. Dept. of Applied Statistics and Econometrics Faculty of Graduate Studies for Statistical Research Cairo University. Egypt.

تاريخ استلام البحث: 2023/05/12

تاريخ قبول البحوث: 2023/06/22

نشر البحث في العدد الواحد والعشرين: ديسمبر / كانون اول 2023

رمز التصنيف ديوي / النسخة الالكترونية (Online): 2522-64X/511.8

رمز التصنيف ديوي / النسخة الورقية (Print): 2519-948X/511.8

A statistical analysis of the Type II Tobit model maximum Likelihood in case of non-ignoring missing data

Mona Emad El-Din Mohamed, Mervat El-Gohary
Department of Statistics- Faculty of commerce-Al-Azhar University (Girls Campus), Egypt.
Ahmed Amin El-Sheikh. Dept. of Applied Statistics and Econometrics Faculty of Graduate Studies for Statistical Research Cairo University. Egypt.

Abstract
Credit scoring is regarded as a core competence of commercial banks during the last few decades. A number of credit scoring models have been developed to evaluate credit risk of new loan applicants and existing loan clients. The main purpose of the present paper is to evaluate credit risk in banks using credit scoring models. Statistical techniques are used: maximum likelihood for one can use linear models and for, one can use Type II Tobit model, a Monte Carlo simulation study is employed, under non-ignorable missing data. The credit scoring task is performed on one bank's personal loans data-set. The results show that Tobit type-II model is more fitted than linear models.
Key words: Credit scoring, Type II Tobit, loan prediction, missing data, linear models, credit risk, maximum likelihood function.

التحليل الإحصائى للبيانات المفقودة بإستخدام دالة امكان نموذج توبيت من النوع الثانى

المستخلص:

يعتبر نظام التصنيف الائتماني من الكفاءات الأساسية للبنوك التجارية خلال العقود القليلة الماضية. تم تطوير عدد من نماذج التصنيف الائتماني لتقييم مخاطر الائتمان لمقدمي طلبات القروض الجدد وعملاء القروض الحاليين. الغرض الرئيسي من هذه الورقة هو تقييم مخاطر الائتمان في البنوك باستخدام نماذج التصنيف الائتماني. يتم استخدام التقنيات الإحصائية: أقصى احتمالية يمكن للمرء استخدام النماذج الخطية ومن أجل ، يمكن استخدام نموذج توبيت من النوع الثاني ، يتم استخدام دراسة محاكاة مونت كارلو ، في ظل بيانات مفقودة غير قابلة للتجاهل. يتم تنفيذ مهمة تسجيل الائتمان على مجموعة بيانات القروض الشخصية لأحد البنوك. أظهرت النتائج أن نموذج Tobit type-II أكثر ملاءمة من النماذج الخطية.

الكلمات المفتاحية: التصنيف الإئتمانى، نموذج توبيت من النوع الثاني، التنبؤ بالقروض، البيانات المفقودة،النماذج الخطية،مخاطر الإئتمان، دالة الإمكان الأعظم.

Introduction

Credit scoring is regarded as a core competence of commercial banks during the last few decades. A number of credit scoring models have been developed to evaluate credit risk of new loan applicants and existing loan clients. The main purpose of the present study is to evaluate credit risk in Egyptian banks using credit scoring models. Three statistical techniques are used: discriminant analysis, probit analysis and logistic regression. The credit scoring task is performed on one bank's personal loans data-set. The results so far revealed that all proposed models gave a better average correct classification rate than the one currently used. Also both type I and type II errors had been calculated in order to evaluate the misclassification costs.

Latterly, credit risks have become one of the most important financial topics of interest, especially in the banking sector. The role of credit risks has changed dramatically over the last ten decades, from passive automation to a strategic device. The process of credit risk evaluation has the interest of many researchers nowadays. Recently, bankers have come to realize that banking operations affect and affected by the natural environment and that consequently the banks might have an important role to play in helping to raise environmental standards. Although the environment presents significant risks to banks, in particular environmental credit risk, it also perhaps presents profitable opportunities (Thompson, 1998).

Credit scoring is the use of statistical models to determine the likelihood that a prospective borrower will default on a loan. Credit scoring models are widely used to evaluate business, real estate, and consumer loans (Gup & Kolari, 2005,). Credit scoring models (see, for example: Lewis, 1992; Bailey, 2001; Mays, 2001; Malhotra & Malhotra, 2003; Thomas et al., 2004; Sidique, 2006; Chuang & Lin, 2009; Sustersic et al, 2009) are some of the most successful applications of research modelling in finance and banking. Harris (2015) investigated the practice of credit scoring and introduced the use of the clustered support vector machine (CSVM) for credit scorecard development. Abbod, et al. (2016) during the last few years there has been marked attention towards hybrid and ensemble systems development, having proved their ability more accurate than single classifier models. Kozodo, et al. (2019) Credit scoring models support loan approval decisions in the financial services industry.

Judgemental techniques and/or credit scoring models can support making a decision about accepting or rejecting a client's credit. The judgemental techniques rely on the knowledge and both past and present experience of credit analysts, who evaluate the required requisites, such as the personal reputation of a client, the ability to repay credit, guarantees and client's character. Due to the rapid increase in fund-size invested through credit granted by Egyptian banks, and the need for quantifying credit risk, financial institutions including banks have started to apply credit-scoring models. Abdou, Etal (2009).

The present study is concerned with evaluating credit risk in banks using credit-scoring models. Statistical techniques are used: maximum likelihood for one can use linear models and for, one can use Type II Tobit model, a Monte Carlo simulation study is employed, under non-ignorable missing data. The credit scoring task is performed on one bank's personal loans data-set. The results show that Tobit type-II model is more fitted than linear models.

In this paper, a simulation study to examine the behaviour of the suggested methods: using linear model in case of ignoraing missing and Type II Tobit model in case of non-ignoring missing data. Results of the Monte Carlo experiments show strange behavior that has never been reported before for the Type II Tobit MLE. A real life data is also presented.

Type II Tobit Model Estimation

The models considered in this paper, classified as Type 2 Tobit models by Amemiya (1984), have the following structure:

$$Y_{1i} = X_{1i}\beta + \sigma\varepsilon_{i1} \qquad (1)$$

Where $(\varepsilon_{1i} . \varepsilon_{2i})$ is bivariate standard normal with correlation ρϵ. The first equation is a regression equation and the second a selection equation. In a typical economic application, the regression equation is a pricing or expenditure function, and the selection equation is a decision function that governs the occurrence of the transaction. Only qualitative information is available for the dependent variable in the selection equation, $Y_{2\text{i}}$. This is recorded as a binary variable, J_i. that takes the value one when Y_{2i} is positive. In addition, the dependent variable in the regression equation, Y_{1i} , is observed only when Y_{2i} is positive. The regressors, X_{1i} and X_{2i}. are observed regardless of J_i .

The log-likelihood function for this model is

$$\ln L(\delta . \beta . \sigma . \rho_i) = \sum_{i=1}^{n}\{J_i[-\ln(\sigma) + ln\emptyset(Z_i) + ln\emptyset(W_i)] + (1-J_i)ln[1-\emptyset(X_{2i}.\delta)]\} \qquad (3)$$

Where $Z_i = (Y_{1i} - X_{1i}\beta)/\sigma$, $W_i = (X_{2i}\delta + \rho_\epsilon Z_i)/\sqrt{1-\rho_\epsilon^2}$, and where ρ_ϵ is restricted to the open interval (-1,1). This likelihood function is highly nonlinear, and a solution to the score equations is obtained by numerical methods. Unfortunately, the log-likelihood function is not globally concave. Gradient methods may converge to a local maximum likelihood estimator (MLE). One can only be assure of obtaining a global

MLE, assuming one exists, if the estimation processes is start in the neighborhood of the global maximum.

The two-stage method of Heckman (1976) and Lee (1976) is typically use to obtain starting values for numerical solution of the score equations. The small sample performance of this estimator can be erratic, particularly when the same regresses used in both equations. Zuehlke and Zeman (1991) show that under these conditions, the mean square error performance of the subsample OLS estimator of β is often superior to that of the Heckman-Lee estimator. Moreover, it is uncommon for the estimate of ρ_{ϵ} to exceed one in absolute value. In an attempt to circumvent these problems, some authors have added quadratic terms to one or both equations. While the Heckman-Lee estimator is consistent, its use as starting values is not sufficient to insure convergence to a global MLE. There is a solution to this problem, however. Olsen (1982) shows that the log-likelihood function of the Type II Tobit model is globally concave conditional on ρ_{ϵ}, He suggests that a grid search over the bounded parameter ρ_{ϵ}, in conjunction with the corresponding conditional MLEs, may be used to trace the profile of the maximized value of $lnL(\delta.\beta.\sigma.\rho_{\epsilon})$ over the space of ρ_{ϵ} . The location of any local or global maxima is determined, and a simultaneous estimation procedure started in the neighborhood of the global maximum. Unfortunately, this algorithm is not available in current econometric software.

Olsen (1982) observes that with the Type II Tobit model the likelihood function is often flat with a local maximum (emphasis added) near ρ = 0. This raises a question about the practice, common in empirical work, of estimating a sample selection model as a robustness check for OLS estimates. In cases with multiple roots, tests based on the global root might lead to a different conclusion than tests based on the local root are not.

Now assume that the aim is to estimate parameters in a parametric model. Usually, this can be derived from the Maximum Likelihood (ML) method. As suggested by its name, this method obtains estimators by maximizing a likelihood function.

A model for latent variable y^*, which is only partially observed:

$$y^* = \beta_0 + \beta_1 x_i + \epsilon_i \ . \qquad \epsilon_i \sim N(0.\sigma^2) \qquad ....(4)$$

The Likelihood function, L, for e the whole sample is:

$$L(\beta_0.\beta_1.\sigma) = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} \left[\frac{1}{\sigma}\varphi\left(\frac{y_i-\beta_0-\beta_1 x_i}{\sigma}\right)\right]^{D_i}\left[1-\varphi\left(\frac{\beta_0+\beta_1 x_i}{\sigma}\right)\right]^{1-D_i} \quad ...... (5)$$

The values of β_0 , β_1 and σ that maximize the likelihood function are the Tobit estimators of the parameters. As usual the ln (L) is:

$$ln\, l = \sum_{i=1}^{n} D_i \boldsymbol{ln}\left[\frac{1}{\sigma}\varphi\left(\frac{y_i-\beta_0-\beta_1 x_i}{\sigma}\right)\right] + (1-D_i)ln\left[1-\varphi\left(\frac{\beta_0+\beta_1 x_i}{\sigma}\right)\right]$$

$$= \frac{N}{2}[ln(\sigma^2)+ln(2\pi)] + \sum_{i=1}^{n} D_i\left[-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2} + (1-D_i)ln\left[1-\varphi\left(\frac{\beta_0+\beta_1 x_i}{\sigma}\right)\right]\right] \quad \dots (6)$$

The first –order partial derivatives of l with respect to $\beta_0\ and\ \beta_1$ and equating them to zero are as follows:

$$\frac{\partial l}{\partial \beta_0} = \sum_{i=1}^{n} D_i\left[\frac{(y_i-\beta_0-\beta_1 x_i)}{\sigma^2} + \frac{(1-D_i)}{\sigma\left[1-\varphi\left(\frac{\beta_0+\beta_1 x_i}{\sigma}\right)\right]}\right] = 0 \quad \dots\dots\dots (7)$$

$$\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{n} D_i\left[\frac{x_i(y_i-\beta_0-\beta_1 x_i)}{\sigma^2} + \frac{(1-D_i)(-x_i)}{\sigma\left[1-\varphi\left(\frac{\beta_0+\beta_1 x_i}{\sigma}\right)\right]}\right] = 0 \quad \dots\dots\dots (8)$$

The normal equations (7) and (8) do not have explicit solution and they have to be solved numerically.

Fisher information matrix

The elements of the Fisher information matrix are obtained by taking the negative expectation of the second derivatives of the natural logarithm of the likelihood function with respect to $\underline{\Theta}$.

Amemiya (1985) presents the following representation for the information matrix:

$$I(\Theta) = \begin{bmatrix} \sum_{i=1}^{T} a_i x_i \grave{x}_i & \sum_{i=1}^{T} b_i x_i \\ \sum_{i=1}^{T} b_i \grave{x}_i & \sum_{i=1}^{T} c_i \end{bmatrix} \quad \dots\dots. (9)$$

Where

$$z_i = \frac{\grave{x}_i \beta}{\sigma}\ .\ a_i = \frac{-1}{\sigma^2}\left[z_i f(z_i) - \frac{f(z_i)^2}{1-F(z_i)} - F(z_i)\right]\ .\ b_i$$

$$= \frac{1}{2\sigma^3}\left[{z_i}^2 f(z_i) + f(z_i) - \frac{f(z_i)^2}{1-F(z_i)}\right]$$

$$c_i = \frac{1}{4\sigma^4}\left[{z_i}^3 f(z_i) + z_i f(z_i) - \frac{{z_i}^2 f(z_i)^2}{1-F(z_i)} - 2F(z_i)\right]$$

The elements of the Fisher information matrix are obtained by taking the negative expectation of the second derivatives of the natural logarithm of the likelihood function as follows:

$$\frac{\partial l}{\partial {\beta_0}^2} = \sum_{i=1}^{n} D_i\left[\frac{-1}{\sigma^2} - \frac{(1-D_i)}{\sigma^2\left[1-\varphi\left(\frac{\beta_0+\beta_1 x_i}{\sigma}\right)\right]^2}\right] \quad \dots\dots\dots (10)$$

$$\frac{\partial l}{\partial \beta_0 \beta_1} = \sum_{i=1}^{n} D_i\left[\frac{-x_i}{\sigma^2} - \frac{(1-D_i)x_i}{\sigma^2\left[1-\varphi\left(\frac{\beta_0+\beta_1 x_i}{\sigma}\right)\right]^2}\right] \quad \dots\dots\dots\dots (11)$$

$$\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{n} D_i\left[\frac{-{x_i}^2}{\sigma^2} - \frac{(1-D_i)(x_i)^2}{\sigma^2\left[1-\varphi\left(\frac{\beta_0+\beta_1 x_i}{\sigma}\right)\right]^2}\right] \quad \dots\dots\dots\dots (12)$$

Under particular regularity conditions, the two-sided $100(1-\alpha)\%.\ 0<\alpha<1$, asymptotic CIs (Asy-CIs) for the vector of unknown parameters Θ can be obtained.

Monte Carlo Results

The purpose of the Monte Carlo portion of this study is to analyze the performance of estimation methods, including MLE for one can use linear models and for, one can use Type II Tobit model, a Monte Carlo simulation study is employed, under non-ignorable missing data. For MLE, 1000 observation, number of replications, is generated from normal distribtion for ranodm error and coveriate of independent variable X are generating from unifrom distribution. The following assumptions are hold for Monte-Carlo simulation:

Coeffiecnts $(\beta_0.\beta_1)$ assumed to be:

$(\beta_0=-1.\beta_1=1)$ And $(\beta_0=-0.5.\beta_1=0.5)$

Random error of the proosed model is generated from normal distribtion with mean zero and standard devation (σ) 1.

Sample sizes of $n=25.\,50.\,100.\,200.500$ and 100.

Steps for simualtion:

Step 1: Genenerate independent varariable (X) and error part (U) as follows:

$$x_i \sim U(0.2) \quad .i=1.2.\dots.n$$
$$u_i \sim N(0.1) \quad .i=1.2.\dots.n$$

Step 2: Compute dependent varaible Y as follows:

$$y_i=\beta' x_i+u_i$$

Step 3: Converte dependent variable (Y) to Tobit II model variable (Y^*) as follows:

$$y_i^*=\begin{cases} y_i & \text{if } y_i\geq 0 \\ \text{ignorable} & \text{if } y_i<0 \end{cases}$$

and define indicatro variable d_i as:

$$d_i=\begin{cases} 1 & \text{if } y_i\geq 0 \\ 0 & \text{if } y_i<0 \end{cases}$$

Step 4: Find estimates of $\beta_0.\beta_1$ and σ from:

Traditional model: $\boldsymbol{Y\sim\beta_0+\beta_1X}$ as linear regression model

By solving likelhood equations of Tobit II mode (3.7) and (3.8) to obtain maximum likelihood estimates of $\beta_0.\beta_1$ and σ which denoted by: $\widehat{\beta_0}.\widehat{\beta_1}$, and $\hat{\sigma}$.

Measures of AIC and BIC also computed as:

$$AIC(\hat{\theta})=-2\text{loglikelhood}(\hat{\theta})+2\text{q}$$
$$BIC(\hat{\theta})=-2\text{loglikelhood}(\hat{\theta})+\text{q}\log n$$

where q is the number of parameters and n is the proposed sample size.
Step 5: Repeat step 1 to step 4 number of times $B = 1000$.
Step 6: Compute the following statistical measures:
Mean sqaure error (MSE)

$$MSE(\hat{\theta}) = \frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_i - \theta)^2 \quad \text{.......... (13)}$$

Relative baises (RBias)

$$\text{RBias}(\hat{\theta}) = \frac{1}{B}\sum_{i=1}^{B}\frac{|\hat{\theta}_i - \theta|}{\theta} \times 100 \quad \text{............ (14)}$$

Based on generated data and assumed two cases for β_1, all statistical measures are computed and repoterted in Table 3.1 for the initial parameter of $\beta_1 = 1$ and Table 3.2 for the initial parameter of $\beta_1 = 0.5$. Form the tabluted resluts, one can indicate that: With increasing in sample size n, MSEs and Rbiases are decreasing for all parameters in two different models AIC and BIC are decreasing in two models.
In compsion with two differnet proposed models namely; linear (traditional) models and Tobit type-II models, one can indicate that:
MSEs in linear models is samller than MSEs inTobit type-II models.
AIC and BIC in linear model is greather than in Tobit type-II models which indicate that Tobit type-II model is more fitted than linear models.

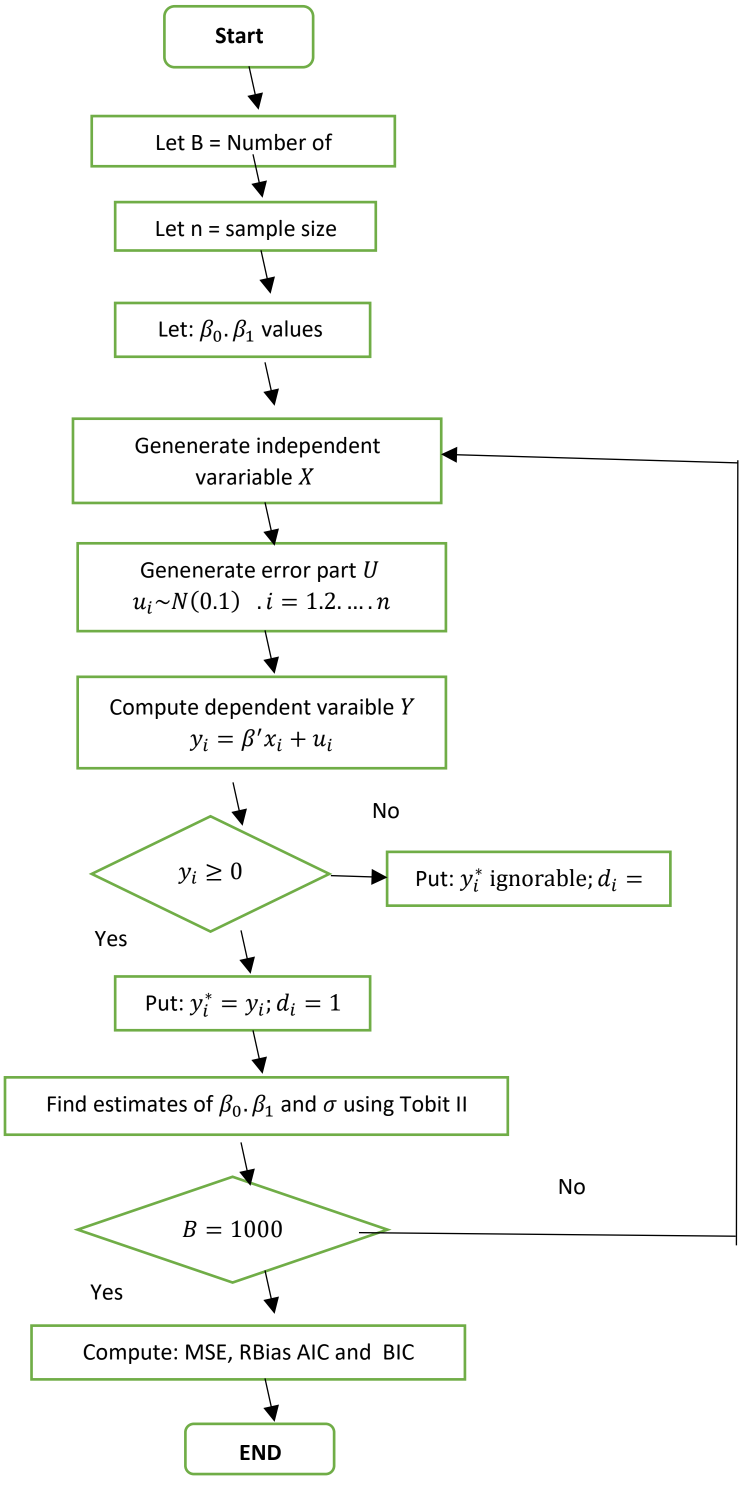
Start
Let B = Number of
Let n = sample size
Let: $\beta_0 . \beta_1$ values
Genenerate independent vararariable X
Genenerate error part U
$u_i \sim N(0.1) \quad . i = 1.2. \dots . n$
Compute dependent varaible Y
$y_i = \beta' x_i + u_i$
No
$y_i \geq 0$
Put: y_i^* ignorable; $d_i =$
Yes
Put: $y_i^* = y_i ; d_i = 1$
Find estimates of $\beta_0 . \beta_1$ and σ using Tobit II
No
$B = 1000$
Yes
Compute: MSE, RBias AIC and BIC
END

Table (1): Estimated values, MSEs, RBias (in %)and model criteria of the linear model and Tobit Type-II model for different sample size n and initial parameters: $\beta_0 = -1. \beta_1 = 1$, and $\sigma = 1$

n	Parm →	Linear Model			Tobit Type-II Model		
		$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\hat{\sigma}$	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\hat{\sigma}$
25	MSE	0.16725	0.12277	0.02303	0.36533	0.20101	0.04845
	RBias	1.57	0.96	4.68	0.63	0.02	5.32
	AIC	73.96026			53.63765		
	BIC	77.61688			57.29428		
50	MSE	0.08383	0.06734	0.01130	0.17190	0.10168	0.02678
	RBias	0.64	0.02	3.03	1.63	1.17	2.87
	AIC	144.26289			103.84047		
	BIC	149.99896			109.57654		
100	MSE	0.04247	0.03079	0.00508	0.08215	0.04698	0.01238
	RBias	0.77	0.65	1.41	1.39	0.82	1.36
	AIC	286.45275			204.75444		
	BIC	294.26826			212.56995		
200	MSE	0.01943	0.01573	0.00255	0.03826	0.02327	0.00618
	RBias	0.41	0.38	0.57	1.14	0.74	0.33
	AIC	570.77785			406.50665		
	BIC	580.67281			416.40161		
500	MSE	0.00748	0.00585	0.00109	0.01481	0.00887	0.00270
	RBias	0.16	0.01	0.29	0.39	0.18	0.39
	AIC	1421.45016			1012.75682		
	BIC	1434.09399			1025.40065		
1000	MSE	0.00407	0.00292	0.00051	0.00780	0.00431	0.00129
	RBias	0.04	0.18	0.17	0.49	0.45	0.09
	AIC	2839.95861			2022.42687		
	BIC	2854.68187			2037.15013		

Table (2): Estimated values, MSEs, RBias (in %) and model criteria of the linear model and Tobit Type-II model for different sample size n and initial parameters: $\beta_0 = -0.5$. $\beta_1 = 0.5$, and $\sigma = 1$

n	Parm→	Linear Model			Tobit Type-II Model		
		$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\hat{\sigma}$	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\hat{\sigma}$
25	MSE	0.16050	0.12106	0.02382	0.29825	0.18340	0.05621
	RBias	0.26	0.67	5.76	5.18	1.25	5.99
	AIC	73.38586			54.58556		
	BIC	77.04249			58.24218		
50	MSE	0.08745	0.06276	0.00970	0.13281	0.08102	0.02441
	RBias	0.26	0.56	2.57	3.83	1.08	2.30
	AIC	144.81639			106.98088		
	BIC	150.55246			112.71695		
100	MSE	0.03877	0.02847	0.00516	0.06300	0.03811	0.01245
	RBias	0.08	0.25	1.26	2.42	1.19	0.69
	AIC	286.74406			211.91490		
	BIC	294.55957			219.73041		
200	MSE	0.01924	0.01468	0.00260	0.03134	0.01902	0.00670
	RBias	2.49	2.18	0.78	2.70	2.20	0.89
	AIC	569.90682			419.07701		
	BIC	579.80178			428.97196		
500	MSE	0.00818	0.00614	0.00098	0.01265	0.00783	0.00240
	RBias	0.10	0.16	0.30	0.92	0.46	0.15
	AIC	1421.46692			1045.79543		
	BIC	1434.11075			1058.43926		
1000	MSE	0.00400	0.00295	0.00053	0.00630	0.00398	0.00128
	RBias	0.21	0.29	0.19	0.51	0.40	0.11
	AIC	2839.56523			2088.65979		
	BIC	2854.28850			2103.38306		

Applications

A private datasets with different characteristics was employed in the process of empirical model evaluation. The data is studied from two way, one for independent variable and second for four independent variables. This data set is related to the loan completion process for customers details provided while filling out the online application form. These details are gender, marital statues, education, number of dependents, income, loan amount, credit history and others. The data set was taken from the online website (http://www.Kaggle.com).

Case I: One indepndent varaible

Define variables of the proposed model from dataset:

Dependent variable: Co-applicant income (y_i), thus,

$$y_i^* = \begin{cases} y_i & \text{if } y_i > 0 \\ \text{ignorable} & \text{if } y_i = 0 \end{cases}$$

and define indicatro variable d_i as:

$$d_i = \begin{cases} 1 & \text{if } y_i > 0 \\ 0 & \text{if } y_i = 0 \end{cases}$$

Where. $i = 1.2. \dots .614$. Thus, we have 273 ignorable observations ($y_i = 0$) and 341 un-censored observations ($y_i^* = y_i[y_i > 0]$).

Independent variable: Applicant income (x)

In Table (3), maximum likelihood estimates of $\beta_0 . \beta_1 . \sigma$ are obtained from the given real data set for two different models, (Linear and Tobit type-II). Note that, in linear model we have an estimate of the standard error ($\hat{\sigma}$) but in Tobit type-II model we have an estimate for parameter σ. From tabulated values of real data set, we notice that the measures of fitting (AIC and BIC) in Tobit type-II model is less than those values in linear model which indicate that the proposed Tobit type-II model is better than linear models.

Table (3): Estimated values, standard errors (St.Er), and model criteria of the linear model and Tobit Type-II model for given real data set of loan prediction: Case I.

Model	Parameter	Estimate	St.Er	AIC	BIC
Linear Model	$\widehat{\beta_0}$	1923.0502	156.7689	11540.30	11553.56
	$\widehat{\beta_1}$	-0.0559	0.0192		
	$\hat{\sigma}$	2908.66	----		
Tobit Type-II Model	$\widehat{\beta_0}$	1570.7325	325.8952	7047.868	7061.128
	$\widehat{\beta_1}$	-0.3139	0.0576		
	$\hat{\sigma}$	4394.941	0.0410		

Case II: Four indepndent varaibles

Define variables of the proposed model from dataset:

Dependent variable: same as in case I.

Independent variables:

- Applicant income (x_1)
- Loan amount in thousands (x_2)
- Term of loan in months (x_3)
- Credit History (x_4)

In Table (4), maximum likelihood estimates of coefficients: $\beta_0.\beta_1.\beta_2.\beta_3.\beta_4.\sigma$ are obtained from the given real data set for two different models, (Linear and Tobit type-II). Note that, in linear model we have an estimate of the standard error ($\hat{\sigma}$) but in Tobit type-II model we have an estimate for parameter σ. From tabulated values of real data set, we notice that the measures of fitting (AIC and BIC) in Tobit type-II model is less than those values in linear model which indicate that the proposed Tobit type-II model is better than linear models.

Table (4): Estimated values, standard errors (St.Er) and model criteria of the linear model and Tobit Type-II model for given real data set of loan prediction: Case II.

Model	Parameter	Estimate	St.Er	AIC	BIC
Linear Model	$\widehat{\beta_0}$	1187.7616	640.2544	9748.179	9773.805
	$\widehat{\beta_1}$	-0.1261	0.0200		
	$\widehat{\beta_2}$	10.2707	1.5238		
	$\widehat{\beta_3}$	-1.0995	1.6259		
	$\widehat{\beta_4}$	-84.896	294.3358		
	$\hat{\sigma}$	2412.087	---		
Tobit Type-II Model	$\widehat{\beta_0}$	-0.9765	1054.4152	5893.004	5918.63
	$\widehat{\beta_1}$	-0.5111	0.0611		
	$\widehat{\beta_2}$	22.7578	2.8029		
	$\widehat{\beta_3}$	-1.8414	2.7001		
	$\widehat{\beta_4}$	113.1238	486.3130		
	$\hat{\sigma}$	3569.4791	0.0444		

Conclusions

In this paper Type II Tobit (sample selection) model studied statistically point of view depending on maximum likelyhood. A Monte Carlo simulation study is introduced to examine the behaviour of the suggested methods: using linear model in case of ignoraing missing and Type II Tobit model in case of non-ignoring missing data. Results show that, strange behavior that has never been reported before for the Type II Tobit MLE. In addition, a real data set is studied from two way, one for independent variable and second for four independent variables. The results show that the measures of fitting (AIC and BIC) in Tobit type-II model is less than those values in linear model which indicate that the proposed Tobit type-II model is better than linear models.

References

1. **Abdou, H. A. H. (2009).** Credit scoring models for Egyptian banks: Neural nets and genetic programming versus conventional techniques (Ph.D. Thesis). The University of Plymouth, UK.
2. **Abbod, M. F and Ala'raj, M. (2016).** Classifiers consensus system approach for credit scoring. Knowledge-Based Systems, 104 , 89–105 .
3. **Amemiya, T., & AMEMIYA, T. A. (1985).** Advanced econometrics. Harvard university press.
4. **Bailey, M. 2001**. Credit scoring: the principles and practicalities. Kingswood, Bristol: White Box Publishing.
5. **Chuang, C., Lin, R. 2009.** Constructing a reassigning credit scoring model. Expert Systems with Applications 36 (2/1): 1685-1694.
6. **Gup, B. E., Kolari, J. W. 2005**. Commercial Banking: The management of risk. Alabama: John Wiley & Sons, Inc.
7. **Harris, T. (2015).** Credit scoring using the clustered support vector machine. Expert Systems with Applications, 42 , 741–750
8. **Lewis, E. M. 1992.** An Introduction to Credit Scoring. California: Fair, Isaac & Co., Inc.
9. **Kozodoi, Nikita, et al.** "Shallow self-learning for reject inference in credit scoring." *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part III*. Springer International Publishing, 2020.
10. **Malhotra, R., Malhotra, D. K. 2003.** Evaluating consumer loans using Neural Networks. Omega the International Journal of Management Science 31 (2): 83-96.
11. **Mays, E. 2001.** Handbook of Credit Scoring. Chicago: Glenlake Publishing Company, Ltd.
12. Olsen, Randall J., "Distributional Tests for Selectivity Bias and a More Robust Likelihood Estimator," *International Economic Review*, 1982, *23*, 223–240.
13. **Owen, A. B. (2001).** Empirical Likelihood. Chapman & Hall/CRC, Boca Raton.
14. **Siddiqi, N. 2006.** Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. New Jersey: John Wiley & Sons, Inc.

15. **Sustersic, M., Mramor, D., Zupan J. 2009.** Consumer credit scoring models with limited data. Expert Systems with Applications 36 (3): 4736-4744.
16. **Thomas, L. C., Edelman, D. B., Crook, J. N. 2004**. Readings in Credit Scoring: recent developments, advances, and aims. New York: Oxford University Press
17. **Thompson, Paul (1998)**. "Bank lending and the environment: policies and opportunities." *International Journal of Bank Marketing* 16.6 : 243-252.