The Estimation of Regression Models with Censored Data Using Logistic and Tobit Models

Atefeh younesi [a], Elham Kamangar [b]

[a] Department of Mathematics, Master of Science, Zanjan University, Zanjan, Iran
(atefehyounesi1363@gmail.com)
[b] Department of Mathematics, Payam Noor University of Tehran

Abstract: In statistics, censoring occurs when the value of an observation is only partially known. The aim of this paper is estimation of a regression model with censored data using Logistic and Tobit models. We have compared two models based on goodness of fit and forecasting accuracy criteria. We have used the data of rate of return and volatility of Tehran Stock Exchange. Results indicate that Based on Akaike info criterion, Schwarz criterion, Hannan-Quinn criterion and Log likelihood, the model of Tobit has better goodness of fit than Logistic model. Criteria of RMSE and MAE indicate that the Tobit model has more accuracy of forecasting than Logistic Model.
[Atefeh younesi, Elham Kamangar. **The Estimation of Regression Models with Censored Data Using Logistic and Tobit Models.** *Life Sci J* 2013;10(7s):545-550] (ISSN:1097-8135). http://www.lifesciencesite.com. 85

Keywords: Regression Models, Censored Data, Logistic, Tobit

1. Introduction

In statistics and other applied sciences, censoring occurs when the value of an observation is only partially known. In other word, some observations have censored. Censored regression models were first suggested in Econometrics by Tobin (1958) and have since been a topic of active research [DeMaris, 2004; Greene, 2012; Wooldridge, 2002].

Within the past decade a variety of techniques have been proposed for handling regression problems in which the dependent variable is subject to censoring [Miller & Halpern (1982)]. Some of these techniques rely upon normal theory, but others make virtually no assumption about the underlying distribution [Miller & Halpern (1982)]. In Tsiatis (1990), a class of estimates for regression parameters in a linear model with right censored data is proposed [Tsiatis (1990)]. These estimates are derived by using linear rank tests for right censored data as estimating equations [Tsiatis (1990)]. They are shown to be consistent and asymptotically normal with covariance matrix for which estimates are proposed [Tsiatis (1990)]. Efficient estimates within this class are derived together with conditions when they are fully efficient [Tsiatis (1990)]. In Viveros & Balakrishnan (1994), a conditional method of inference is used to derive exact confidence intervals for several life characteristics such as location, scale, quantiles, and reliability when the data are Type II progressively censored [Viveros & Balakrishnan (1994)]. The method is shown to be feasible and practical, although a computer program may be required for its implementation [Viveros & Balakrishnan (1994)]. The method is applied for the purpose of illustration to the extreme-value and the one- and two-parameter exponential models. Prediction limits for the lifelength of future units are also discussed [Viveros & Balakrishnan (1994)]. An example consisting of data from an accelerated test on insulating fluid reported by Nelson is used for illustration and comparison [Viveros & Balakrishnan (1994)]. In the paper of Gentleman & Geyer (1994), Standard convex optimization techniques are applied to the analysis of interval censored data [Gentleman & Geyer (1994)]. These methods provide easily verifiable conditions for the self-consistent estimator proposed by Turnbull (1976) to be a maximum likelihood estimator and for checking whether the maximum likelihood estimate is unique [Gentleman & Geyer (1994)]. A sufficient condition is given for the almost sure convergence of the maximum likelihood estimator to the true underlying distribution function [Gentleman & Geyer (1994)]. In Kooperberg & Stone (1992), Logspline density estimation is developed for data that may be right censored, left censored, or interval censored [Kooperberg & Stone (1992)]. A fully automatic method, which involves the maximum likelihood method and may involve stepwise knot deletion and either the Akaike information criterion (AIC) or Bayesian information criterion (BIC), is used to determine the estimate [Kooperberg & Stone (1992)]. In solving the maximum likelihood equations, the Newton–Raphson method is augmented by occasional searches in the direction of steepest ascent [Kooperberg & Stone (1992)]. Also, a user interface based on S is described for obtaining estimates of the density function, distribution function, and quantile function and for generating a random sample from the fitted distribution [Kooperberg & Stone (1992)].

In Jin, Lin, & Ying (2006), the semi-parametric accelerated failure time model relates the logarithm of the failure time linearly to the covariates while leaving the error distribution unspecified [Jin, Lin, & Ying (2006)]. They described simple and reliable inference procedures based on the least-squares principle for

this model with right-censored data [Jin, Lin, & Ying (2006)]. The proposed estimator of the vector-valued regression parameter is an iterative solution to the Buckley–James estimating equation with a preliminary consistent estimator as the starting value [Jin, Lin, & Ying (2006)]. The estimator is shown to be consistent and asymptotically normal. A novel resampling procedure is developed for the estimation of the limiting covariance matrix [Jin, Lin, & Ying (2006)]. Extensions to marginal models for multivariate failure time data are considered [Jin, Lin, & Ying (2006)].

Groot & Lucas (2012) have proposed an extension of Gaussian process regression models to data in which some observations are subject to censoring. Since the model is not analytically tractable they used Expectation propagation to perform approximate inference on it [Groot & Lucas (2012)]. In Lin, He, & Portnoy (2012), a new algorithm is proposed to estimate the regression quantile process when the response variable is subject to double censoring [Lin, He, & Portnoy (2012)]. The algorithm distributed the probability mass of each censored point to its left or right appropriately, and iterated towards self-consistent solutions [Lin, He, & Portnoy (2012)]. Numerical results on simulated data and an unemployment duration study are given to demonstrate the merits of the proposed method [Lin, He, & Portnoy (2012)]. In Van Zyl, & Schall (2012), using large sample properties of the empirical distribution function and order statistics, weights to stabilize the variance in order to perform weighted least squares regression are derived [Van Zyl, & Schall (2012)]. Weighted least squares regression is then applied to the estimation of the parameters of the Weibull, and the Gumbel distribution [Van Zyl, & Schall (2012)]. The weights are independent of the parameters of the distributions considered. Monte Carlo simulation showed that the weighted least-squares estimators outperformed the usual least-squares estimators totally, especially in small samples [Van Zyl, & Schall (2012)].

The aim of this paper is estimation of a regression model with censored data using Logistic and Tobit models. Then, we have compared two models based on goodness of fit criteria.

This paper is organized by 5 sections, the next section is devoted to methods, section 3 introduces the data and the model, empirical results is shown by section 4 and final section is devoted to conclusion.

2. Methods

In some settings, the dependent variable is only partially observed. For example, in survey data, data on incomes above a specified level are often top-coded to protect confidentiality. Similarly desired consumption on durable goods may be censored at a small positive or zero value.

For estimation these regression model, we have surveyed two models consist of Logistic and Tobit models.

Consider the following latent variable regression model:

$$y_i = x_i\beta + \vartheta\varepsilon_i \qquad (1)$$

where ϑ is a scale parameter. The scale parameter ϑ is identified in censored and truncated regression models, and will be estimated along with theβ.

In the canonical censored regression model, known as the tobit (when there are normally distributed errors), the observed data y_i are given by:

$$y_i = \begin{cases} 0 & if\ y_i^{\#} \le 0 \\ y_i^{\#} & if\ y_i^{\#} > 0 \end{cases} \qquad (2)$$

In other words, all negative values of $y_i^{\#}$ are coded as 0. We say that these data are left censored at 0. Note that this situation differs from a truncated regression model where negative values of $y_i^{\#}$ are dropped from the sample. More generally, we allows for both left and right censoring at arbitrary limit points so that:

$$y_i = \begin{cases} \underline{c_i} & if\ y_i^{\#} \le \underline{c_i} \\ y_i^{\#} & if\ \underline{c_i} < y_i^{\#} \le \overline{c_\iota} \\ \overline{c_\iota} & if\ y_i^{\#} > \overline{c_\iota} \end{cases} \qquad (3)$$

where $\underline{c_i}$, $\overline{c_\iota}$ are fixed numbers representing the censoring points. If there is no left censoring, then we can set $\underline{c_i} = -\infty$. If there is no right censoring, then $\overline{c_\iota} = \infty$. The canonical tobit model is a special case with $\underline{c_i} = 0$ and $\overline{c_\iota} = \infty$.

The parameters β, ϑ are estimated by maximizing the log likelihood function:

$$l(\beta,\vartheta) = \sum_{i=1}^{N} \log f\left(\frac{y_i - x_i'\beta}{\vartheta}\right).1\left(\underline{c_i} < y_i < \overline{c_\iota}\right) + \sum_{i=1}^{N} \log F\left(\frac{\underline{c_i} - x_i'\beta}{\vartheta}\right).1\left(y_i = \underline{c_i}\right) + \sum_{i=1}^{N} \log\left(1 - F\left(\frac{\overline{c_\iota} - x_i'\beta}{\vartheta}\right)\right).1\left(y_i = \overline{c_\iota}\right)$$

where f, F are the density and cumulative distribution functions of ε_i, respectively.

3. Data and model

We have used the data of rate of return and volatility of Tehran Stock Exchange. Data are available on website of Tehran Stock Exchange. The sample period is daily data of stock price index of TSE during 2007-2010 period.

Some data of volatility are censored by authors for estimation the following model:

$r_t = \alpha + \beta v_t + \varepsilon_t$

Where r_t is rate of return of TSE, v_t is volatility of return, α and β are parameters of the model and ε_t is error term of the regression model.

4. Empirical Results

First of all, we have estimated the model with TOBIT model. Table 1 indicates the estimation result.

Table 1. Estimation Results of The model (TOBIT)				
Method: ML - Censored Normal (TOBIT) (Quadratic hill climbing)				
Date: 01/11/13 Time: 12:01				
Sample: 1/01/2007 12/31/2010				
Included observations: 1045				
Left censoring (value) at zero				
Convergence achieved after 2 iterations				
Covariance matrix computed using second derivatives				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	0.507351	0.017536	28.93253	0.0000
V	0.001689	0.030273	0.055805	0.9555
	Error Distribution			
SCALE:C(3)	0.285313	0.006241	45.71677	0.0000
Mean dependent var	0.508196	S.D. dependent var		0.285450
S.E. of regression	0.285756	Akaike info criterion		0.335283
Sum squared resid	85.08581	Schwarz criterion		0.349498
Log likelihood	-172.1853	Hannan-Quinn criter.		0.340674
Avg. log likelihood	-0.164771			
Left censored obs	0	Right censored obs		0
Uncensored obs	1045	Total obs		1045

Then, we have estimated the model with Logistic model. Table 2 indicates the estimation results of Logistic model.

Table 2. Estimation Results of The model (Logistic)		
Method: ML - Censored Logistic (Quadratic hill climbing)		
Date: 01/11/13 Time: 12:01		
Sample: 1/01/2007 12/31/2010		
Included observations: 1045		
Left censoring (value) at zero		
Convergence achieved after 3 iterations		

Covariance matrix computed using second derivatives				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	0.509057	0.018935	26.88445	0.0000
V	-0.004260	0.032753	-0.130070	0.8965
	Error Distribution			
SCALE:C(3)	0.173601	0.004327	40.11758	0.0000
Mean dependent var	0.508196	S.D. dependent var		0.285450
S.E. of regression	0.285837	Akaike info criterion		0.426622
Sum squared resid	85.13402	Schwarz criterion		0.440838
Log likelihood	-219.9102	Hannan-Quinn criter.		0.432014
Avg. log likelihood	-0.210440			
Left censored obs	0	Right censored obs		0
Uncensored obs	1045	Total obs		1045

The above results indicate that:

1. Based on Akaike info criterion, Schwarz criterion and Hannan-Quinn criterion, the model of Tobit has better than Logistic model.
2. Based on Log likelihood, the model of Tobit has better than Logistic model.

Figure 1. Forecasted Return by Tobit Model

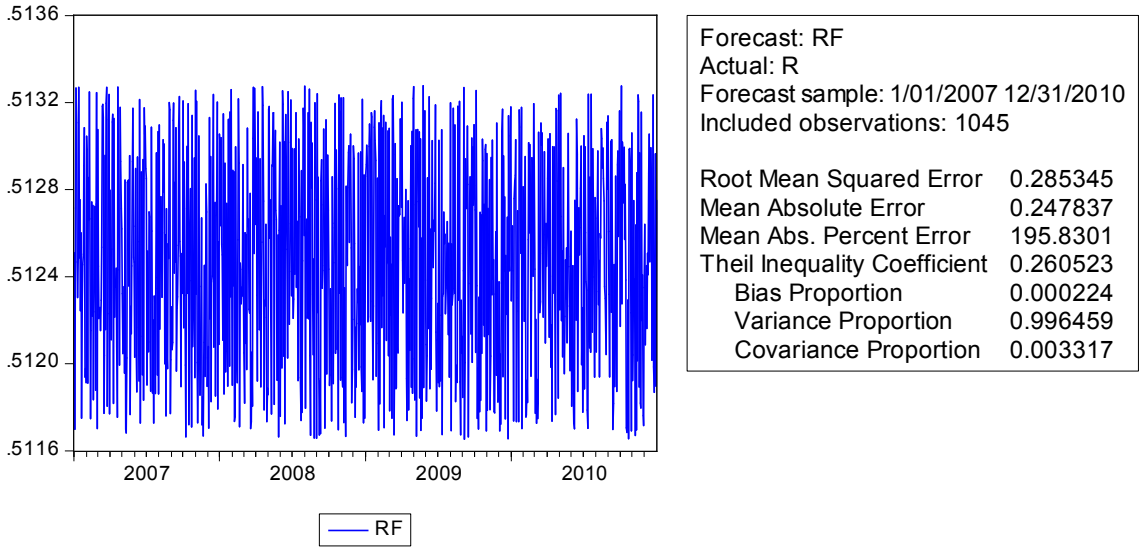
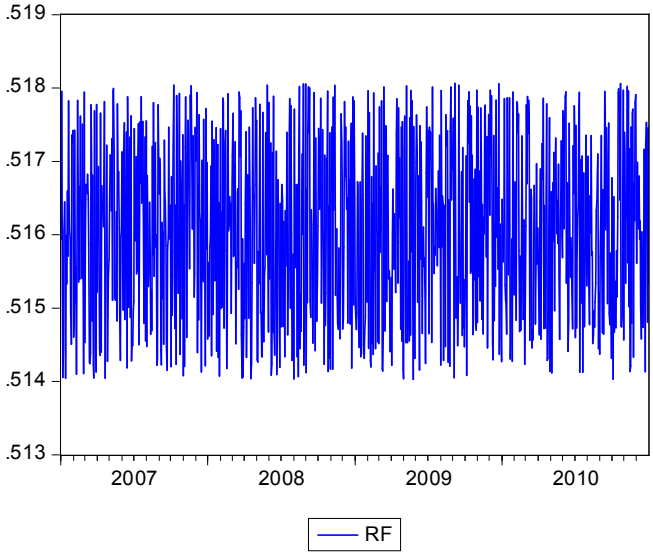


Figure 2. Forecasted Return by Logistic Model



Forecast: RF
Actual: R
Forecast sample: 1/01/2007 12/31/2010
Included observations: 1045

Root Mean Squared Error	0.285426
Mean Absolute Error	0.247849
Mean Abs. Percent Error	197.1884
Theil Inequality Coefficient	0.259749
Bias Proportion	0.000756
Variance Proportion	0.990975
Covariance Proportion	0.008269

Figure 1 and 2 indicate Forecasted Return by Tobit and Logistic Models respectively. Based on RMSE and MAE the Tobit model has more accuracy of forecasting than Logistic Model.

5. Conclusion

The method of data gathering and recording is a central issue in data analysis. A typical practical problem is censoring, which occurs when the value of a measurement or observation is only partially known [Groot, & Lucas (2012)]. Within the past decade a variety of techniques have been proposed for handling regression problems in which the dependent variable is subject to censoring [Miller & Halpern (1982)]. Some of these techniques rely upon normal theory, but others make virtually no assumption about the underlying distribution [Miller & Halpern (1982)].

The aim of this paper is estimation of a regression model with censored data using Logistic and Tobit models. Then, we have compared two models based on goodness of fit criteria. We have used the data of rate of return and volatility of Tehran Stock Exchange. Data are available on website of Tehran Stock Exchange. The sample period is daily data of stock price index of TSE during 2007-2010 period. Results indicate that Based on Akaike info criterion, Schwarz criterion and Hannan-Quinn criterion, the model of Tobit has better than Logistic model. Also, based on Log likelihood, the model of Tobit has better than Logistic model. Criteria of RMSE and MAE indicate that the Tobit model has more accuracy of forecasting than Logistic Model.

References

[1]. Buckley, J., & James, I. (1979). Linear regression with censored data. Biometrika, 66(3), 429-436.
[2]. DeMaris, A. (2004). Regression With Social Data: Modeling Continuous and Limited Response Variables. Wiley Series in Probability and Statistics. Wiley-Interscience
[3]. Efron, B. (1967). The two sample problem with censored data. Proc. 5th Berkeley Sympos. Math. Statist. Prob.", Prentice-Hall: New York.
[4]. Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. Journal of the American statistical Association, 72(359), 557-565.
[5]. Gentleman, R., & Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. Biometrika, 81(3), 618-623.
[6]. Greene. W. H. (2012). Econometric Analysis. Prentice Hall, 7th edition
[7]. Groot, P., & Lucas, P. (2012). Gaussian Process Regression with Censored Data Using Expectation Propagation. In The 6th European

Workshop on Probabilistic Graphical Models (PGM).
[8]. Jin, Z., Lin, D. Y., & Ying, Z. (2006). On least-squares regression with censored data. Biometrika, 93(1), 147-161.
[9]. Kooperberg, C., & Stone, C. J. (1992). Logspline density estimation for censored data. Journal of Computational and Graphical Statistics, 1(4), 301-328.
[10]. Lin, G., He, X., & Portnoy, S. (2012). Quantile regression with doubly censored data. Computational Statistics & Data Analysis, 56(4), 797-812.
[11]. Miller, R., & Halpern, J. (1982). Regression with censored data. Biometrika, 69(3), 521-531.
[12]. Tobin. 1958. Estimation of relationships for limited dependent variables. Econometrica, 26:24–36
[13]. Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. The Annals of Statistics, 18(1), 354-372.
[14]. Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society. Series B (Methodological), 290-295.
[15]. Van Zyl, J. M., & Schall, R. (2012). Parameter Estimation Through Weighted Least-Squares Rank Regression with Specific Reference to the Weibull and Gumbel Distributions. Communications in Statistics-Simulation and Computation, 41(9), 1654-1666.
[16]. Viveros, R., & Balakrishnan, N. (1994). Interval estimation of parameters of life from progressively censored data. Technometrics, 36(1), 84-91.
[17]. Wooldridge. J. M. (2002). Econometric Analysis of Cross Section and Panel Data. Mit Press.

3/18/2013