Scientific Research Publishing

Open Journal of Modelling and Simulation, 2021, 9, 275-280
https://www.scirp.org/journal/ojmsi
ISSN Online: 2327-4026
ISSN Print: 2327-4018

A Sequential Shrinkage Estimating Method for Tobit Regression Model

Haibo Lu, Cuiling Dong, Juling Zhou

School of Mathematical Sciences, Xinjiang Normal University, Urumqi, China
Email: andyluhaibo@foxmail.com

How to cite this paper: Lu, H.B., Dong, C.L. and Zhou, J.L. (2021) A Sequential Shrinkage Estimating Method for Tobit Regression Model. *Open Journal of Modelling and Simulation*, **9**, 275-280.
https://doi.org/10.4236/ojmsi.2021.93018

Received: June 17, 2021
Accepted: July 16, 2021
Published: July 19, 2021

Abstract

In the applications of Tobit regression models we always encounter the data sets which contain too many variables that only a few of them contribute to the model. Therefore, it will waste much more samples to estimate the "non-effective" variables in the inference. In this paper, we use a sequential procedure for constructing the fixed size confidence set for the "effective" parameters to the model by using an adaptive shrinkage estimate such that the "effective" coefficients can be efficiently identified with the minimum sample size based on Tobit regression model. Fixed design is considered for numerical simulation.

Keywords

Tobit Regression Models, Adaptive Shrinkage Estimate, Minimum Sample Size, Fixed Design

1. Introduction

Tobit regression model is called sample selection model or restricted dependent variable model, see in [1] [2]. It is a kind of models whose dependent variables satisfy certain constraints. In some cases, it is also called truncated regression model or censored regression model. Tobit regression model is widely used in Econometrics and other research fields, and plays an increasingly important role in the analysis of cross-sectional data and time series data, illustrated in [3] [4]. However, in applications data sets we encountered usually have too many explanatory variables but only a few of them contributes to the model. Methods such as LASSO and LARS, see in [5] [6], have been proposed to figure out the effective variables, however it is still intractable to know how many samples can identify the effective variables and simultaneously make the parameter estimates

achieve a pre-specified accuracy. For linear regression model, Wang and Chang propose a sequential shrinkage estimate method to identify the effective variables and attain accuracy of parameter estimate in [7]. For Tobit regression models, similar methods have not been proposed and there is still a lot of work to do. To handle the problem mentioned above, we propose a sequential procedure for constructing the fixed size confidence set for effective parameters based on an adaptive shrinkage estimate (ASE) such that the effective coefficients can be efficiently identified with the minimum sample size under fixed design.

The rest of this paper is organized as follows. In Section 2, we will give the adaptive shrinkage estimate (ASE) based on the Least Absolute Deviation Estimate (LAD) of Tobit regression models and its asymptotic properties. In Section 3, Sequential sampling strategy based on ASE and stopping rule as well as random size confident set is presented. In Section 4, an example with numerical simulation is given to illustrate the performance of the proposed method via sequential fixed size confidence estimation using synthesized data sets.

2. Sequential Adaptive Shrinkage Estimate Based on LAD

2.1. Asymptotic Properties of LAD

Suppose $a^+ = \max\{a, c\}$, where c is a known constant, we can define Tobit regression model as:

$$y_i^+ = \max\left\{x_i^{\mathrm{T}}\beta_0 + \varepsilon_i, c\right\}, i = 1, 2, \cdots, n \tag{1}$$

where y_i is dependent variable, β_0 is a p-dimensional vector of the unknown regression coefficients, x_i is a p-dimensional vector of covariates and ε_i is a random error. Without losing generality, suppose $c = 0$. Let $\varepsilon_i, i = 1, 2, \cdots, n$ be independent identically distributed and follows a standard normal distribution with mean 0 and variance σ^2, then the Likelihood function will be

$$L = \prod_0 \left(1 - \Phi\left(x_i^{\mathrm{T}}\beta/\sigma\right)\right) \prod_1 \left(\sigma^{-1}\Phi\left(x_i^{\mathrm{T}}\beta/\sigma\right)\right) \tag{2}$$

where Φ and ϕ are standard normal distribution function and density function, $\prod_0$ and $\prod_1$ are the products in $\{i : y_i \le 0\}$ and $\{i : y_i > 0\}$ separately.

Powell proposed a Least Absolute Deviation Estimate (LAD) of β_0 in [8], which is written as $\tilde{\beta}_n$, and minimize the function

$$Q_n(\beta) = \sum_{i=1}^{n} \left| y_i^+ - \max\left\{x_i^{\mathrm{T}}\beta, 0\right\} \right| \tag{3}$$

Under the assumptions (A1) Let $\sup_i \|x_i\| < \infty$ and (A2) Let the density function of the random error ε_i, satisfies $f(0) = 0$ and $med(\varepsilon_i) = 0$, then there exists some $\delta > 0$

$$\lim_{n\to\infty} \frac{\lambda}{\log n} \sum_{i=1}^{n} I\left(x_i^{\mathrm{T}}\beta > \delta\right) x_i x_i^{\mathrm{T}} = \infty \tag{4}$$

Chen and Wu proved that $\lim_{n\to\infty} \tilde{\beta}_n = \beta_0, a.s.$ and

$$\left(2f(0)M_n^{1/2}\right) \cdot \sqrt{n}\left(\tilde{\beta}_n - \beta_0\right) \xrightarrow{d} N(0, I_n) \tag{5}$$

in [9], where I_n is an identity matrix and $M_n = E\left(\frac{1}{n}\sum_i I\left(x_i^{\mathrm{T}}\beta_0 > 0\right)x_i x_i^{\mathrm{T}}\right)$.

2.2. Adaptive Shrinkage Estimate

Let $\kappa = \kappa(n)$ be a non-random function of n such that for some $0 < \delta < 1/2$ and $\gamma > 0$, $n^{1/2}\kappa \to 0$ and $n^{1/2+\gamma\delta}\kappa \to \infty$, as $n \to \infty$. Then, under the assumptions (A1) and (A2), by using Equation (4) we can see that $n^{1/2-\eta}\left(\tilde{\beta}_n - \beta_0\right) = O(1)$ almost surely as n tends to ∞ for some $\eta > 0$. Similar to Wang and Chang in [6], we can define $\hat{\beta}_n = I_n(\varepsilon)\tilde{\beta}_n$ as an adaptive shrinkage estimate (ASE) of β_0, where $I_n(\varepsilon) = diag\left\{I_{n1}(\varepsilon), I_{n2}(\varepsilon), \cdots, I_{np}(\varepsilon)\right\}$ is a $p \times p$ diagonal matrix.

So far, we get good statistical properties of the proposed ASE estimate under non-random sample size, but our goal is to determine a sample size under which the ASE attains the required accuracy. So we will introduce the sequential sampling scheme based on the ASE below. It is known that construction of the confidence set for β_0 depends on the asymptotic distribution of $\hat{\beta}_n$ and sample size under sequential analysis is a random variable. So we need to study asymptotic properties of ASE under random sample size. Fortunately, property of uniform continuity in probability, see in [10] and [11], is a sufficient condition such that the randomly stopped sequence has the same asymptotic distribution as the fixed sample size estimate. That is, $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right), n = 1, 2, \cdots$, has the property of uniform continuity in probability, which indicates the following Theorem holds.

Theorem 1. Suppose that the (A1) and (A2) are satisfied, and let $N(t)$ be a positive integer-valued random variable such that $N(t)/t$ converges to 1 in probability as $t \to \infty$. Then

$$\sqrt{N(t)}\left(\hat{\beta}_{N(t)} - \beta_0\right) \to N\left(0, I_0 \Sigma I_0^{-1}\right)$$

in distribution as $t \to \infty$.

From Theorem 1, we can construct a confidence set of β_0 and a stopping rule on sequential sampling procedure to determine final sample size. Let $\left\{(y_i, x_i) : i = 1, 2, \cdots, k\right\}$ be the first k observations and denoted by C_k. Define a stopping rule N_d as

$$N = N_d \equiv \inf\left\{k : \frac{d^2}{a_k^2} \geq \nu_k, \forall k \geq n_0\right\} \tag{5}$$

For sequential estimation procedure, one new observation is collected at a time until the stopping criterion is satisfied. When the stopping rule holds, based on N samples a confidence set of β_0 is constructed as follow,

$$R_N = \left\{Z \in R^p : \frac{S_N}{N} \leq \frac{d^2}{\nu_N}; I_{N_j}(\varepsilon) = 0 \to z_j = 0, 1 \leq j \leq p\right\} \tag{6}$$

where $S_N = \left(Z_{N_1} - \hat{\beta}_{N_1}\right)^{\mathrm{T}} \tilde{\Sigma}_{11}\left(Z_{N_1} - \hat{\beta}_{N_1}\right)$. Properties of the sequential procedure and the confidence set R_N are summarized below.

Theorem 2. Assume that the (A1) and (A2) are satisfied, and let N be the stopping time defined in Equation (5). Then: 1) $\lim_{d \to 0} d^2 N / a^2 \nu = 1$ almost surely;

2) $\lim_{d\to 0} d^2 N/a^2\nu = 1$; 3) $\lim_{d\to 0} d^2 E(N)/a^2\nu = 1$; 4) $\lim_{d\to 0} \hat{p}_0(N) = p_0$ almost surely; 5) $\lim_{d\to 0} E(\hat{p}_0(N)) = p_0$ where ν is the maximum eigen-value of matrix $I_0\Sigma^{-1}I_0$.

3. Example and Simulation

We evaluate the performance of the proposed method via sequential fixed size confidence estimation using synthesized data sets. As mentioned previously, by the definition of the stopping rule, when sampling is stopped, the final confidence ellipsoid constructed will have the prescribed precision and coverage probability. Thus, we can compare the average stopping times of procedures based on LAD and ASE. Since the proposed method ignores the non-effective variables, we expect the average stopping time to be significantly smaller than that of the procedure based on LAD with no variable identification mechanism. If the p_0 variables are known in advance, then the most efficient procedure is, of course, to use only these p_0 variables. Therefore, we also construct a sequential procedure under such a situation, and the results of the cases with known p_0 can serve as the baseline, in which the smallest sample size is achieved, asymptotically.

The synthesized data sets for the model with fixed designs are generated as follows: the regressor x_i are generated independently from a standard multivariate normal distribution with mean 0 and identity covariance matrix beforehand, and the error term e_i is independently drawn from the standard normal distribution for each $i \geq 1$. The system error is assumed to follow the standard normal distribution. The response generated by Equation (1) and the true parameter $\beta_0 = (-1.2, 2.0, 0, 0, 0, 0, 0, 0, 0, 0)$ with 8 non-effective variables. Different precisions of confidence ellipsoid $d \in \{0.3, 0.4, 0.5, 0.6\}$ are chosen with coverage probability equal to 95%, $\alpha = 0.05$ in the simulation. We choose $\gamma = 1$, $\delta = 0.55$ and $\theta = 0.70$ in analyzing simulated data. When applying the ASE method, the regularization parameter ε needs to be determined by some model selection criteria, as the AIC, BIC together with a GCV method. For convenience, we only use BIC to illustrate our method.

Table 1 state results of sequential sampling method for cox regression. In the table, we list final sample size N (stopping time), $\kappa = d^2 N/a^2\nu$ and empirical coverage probability CP of the 95% confidence set R_N. For all of the three cases: LAD, LAD_{p_0}, ASE, the value κ of is very close to 1, and the empirical coverage probability CP approaches the Normal 95% as d decreases, as stated in Theorem 2. However, the sample size N of LAD are much larger than those of the other two cases, and ASE has sample size very close to those of LAD_{p_0}. In conclusion, the proposed ASE is more efficient than LAD.

Table 2 reports powers of identity effective variables and effective variables and estimates of the regression coefficients for Tobit regression. We can see that numbers of incorrectly identified zero variables (N_{ic}^*) using ASE is almost close

Table 1. Results of sequential sampling method based on ASE, LAD with all variables and LAD_{p_0} with only p_0 non-zero variables for Tobit regression model.

		$\beta_0 = (-1.2, 2.0, 0, 0, 0, 0, 0, 0, 0, 0)$								
		LAD_{p_0}			*ASE*			*LAD*		
Design	d	N	κ^*	CP	N	κ	CP	N	κ	CP
fixed	0.6	85.44 (14.75)	1.008	0.96	92.32 (17.40)	1.044	0.94	327.8 (23.194)	1.01	0.92
	0.5	126.84 (19.75)	1.019	0.96	131.52 (19.13)	1.034	0.98	433.21 (29.586)	1.006	1
	0.4	179.13 (25.587)	1.001	0.94	190.34 (25.311)	1.021	0.93	674.26 (37.868)	1.003	0.90
	0.3	363.72 (38.211)	1.001	0.97	373.60 (37.087)	1.017	0.95	1203.05 (44.707)	1.002	0.94

$\kappa^* = d^2 N / (a^2 \nu)$; CP^+ is the empirical coverage probability of 95% confidence ellipsoid region R_N; ** Empirical standard deviations are in parentheses.

Table 2. Power of variable identification and estimation of nonzero components under sequential sampling method based on ASE and LAD with Tobit regression model.

		$\beta_1 = -1.2, \beta_2 = 2.0$							
		ASE				*LAD*			
Design	d	N_{ic}^*	N_c^*	β_1	β_2	N_{ic}^*	N_c^*	β_1	β_2
fixed	0.6	0	7.912	−1.223 (0.155)	2.16 (0.010)	-	-	−1.240 (0.13)	2.061 (0.31)
	0.5	0	7.959	−1.214 (0.129)	2.042 (0.193)	-	-	−1.224 (0.009)	2.031 (0.066)
	0.4	0	7.982	−1.208 (0.105)	2.073 (0.112)	-	-	−1.211 (0.054)	2.021 (0.077)
	0.3	0	7.933	−1.210 (0.076)	2.035 (0.102)	-	-	−1.201 (0.032)	2.004 (0.043)

N_{ic}^* and N_c^* are the average number of zero components in β correctly identified and nonzero components incorrectly estimated as zero values, respectively; $^+$ standard deviations are in parentheses.

to 0, and the number of correctly identified zero variables (N_c^*) are all very close to the true number of effective variables (2 and 8). These results suggest that $\hat{p}_0$ is a good estimator of p_0 under the sequential sampling method based on ASE. The LAD procedure does not identify the effective variables, so N_c^* and N_{ic}^* are not available. In addition, all of parameter estimates of effective variables are very close to the true values.

4. Conclusion

Based on an ASE estimate of the parameter in Tobit regression model, a sequential sampling procedure is constructed to estimate a minimum sample size to identify the effective variables and simultaneously make estimate of parameters with required accuracy. We prove that the proposed sequential procedure is asymptotically optimal in the sense of Chow and Robbins, see in [12]. Simulation studies show that the proposed method can save large sample size compared to traditional sequential sampling method. However, this paper supposes the dimension of variables is fixed, not varying as sample size. Our future work is to investigate the properties of sequential sampling method with varying number of

variables as sample size.

Support

This research was supported by Research projects of universities in Xinjiang Uygur Autonomous Region under Grant No. XJEDU2016I033 and Xinjiang Normal University postdoctoral research foundation under Grant No. XJNUBS1539.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Tobin, J. (1958) Estimation of Relationships for Limited Dependent Variables. *Econometrica*, **26**, 24-36. https://doi.org/10.2307/1907382

[2] Adams, J.D. (1980) Personal Wealth Transfers. *Quarterly Journal of Economics*, **95**, 159-179.

[3] Ashenfelter, O. and Ham, J. (1979) Education, Unemployment, and Earnings. *Journal of Political Economy*, **87**, S99-S116. https://doi.org/10.1086/260824

[4] Fair, R.C. (1978) A Theory of Extramarital Affairs. *Journal of Political Economy*, **86**, 45-61. https://doi.org/10.1086/260646

[5] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, *Series B*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[6] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least Angle Regression. *Journal of Annals of Statistics*, **32**, 407-499. https://doi.org/10.1214/009053604000000067

[7] Wang, Z.F. and Chang, Y.I. (2013) Sequential Estimate for Linear Regression Models with Uncertain Number of Effective Variables. *Metrika*, **76**, 949-978. https://doi.org/10.1007/s00184-012-0426-4

[8] Powell, J.L. (1984) Least Absolute Deviations Estimation for the Censored Regression Model. *Journal of Econometrics*, **25**, 303-325. https://doi.org/10.1016/0304-4076(84)90004-6

[9] Chen, X.R. and Wu, Y.H. (1994) Consistency of Estimates in Censored Linear Regression Models. *Communications in Statistics*, **23**, 1847-1858. https://doi.org/10.1080/03610929408831360

[10] Anscombe, F.J. (1952) Large Sample Theory of Sequential Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, **48**, 600-607. https://doi.org/10.1017/S0305004100076386

[11] Woodroofe, M. (1982) Nonlinear Renewal Theory in Sequential Analysis. Society for Industrial and Applied Mathematics, Philadelphia. https://doi.org/10.1137/1.9781611970302

[12] Chow, Y.S. and Robbins, H. (1965) On the Asymptotic Theory of Fixed-Width Sequential Confidence Intervals for the Mean. *Journal of Annals of Mathematical Statistics*, **36**, 457-462. https://doi.org/10.1214/aoms/1177700156