应用概率统计　第二十六卷
第一期　2010年2月

Chinese Journal of Applied Probability
and Statistics Vol.26 No.1 Feb. 2010

A LASSO-Type Approach to Variable Selection and Estimation for Censored Regression Model *

Wang Zhanfeng　Wu Yaohua　Zhao Lincheng

(*Department of Statistics and Finance, University of Science and Technology of China, Hefei, 230026*)

Abstract

Censored regression ("Tobit") model is one of important regression models and has been widely used in econometrics. However, studies for variable selection problem in censored regression model are rare at the present references. In this paper, for censored regression model we propose a LASSO-type approach, diverse penalty L_1 constraint method (DPLC), to select variables and estimate the corresponding coefficients. Furthermore, we obtain the asymptotic properties of nonzero elements' estimation of regression coefficient. Finally, extensive simulation studies show that DPLC method almost possesses the same performance of selecting variables and estimation as generally best subset selection method (GBSS).

Keywords: Censored regression model, least absolute deviation, variable selection, LASSO.

AMS Subject Classification: 62F05, 62G05.

§1. Introduction

Limited dependent variable (LDV) models are important regression models and have been widely used in econometrics studies. Moreover, many important advances of econometrics studies are related to LDV model. Censored regression ("Tobit") model studied in this paper is a special LDV model for which response variable has a nonnegative limitation, where only segment of response variable being not less than 0 can be measured. Details as following model,

$$Y_i^+ = (x_i'\beta_0 + e_i)^+, \qquad i = 1, 2, \cdots, n, \tag{1.1}$$

where $Y_i^+ = Y_i I(Y_i \geq 0)$ and $I(\cdot)$ denotes the indicator function of a set, $\{x_i\}$ is a sequence of vector with length p, $\{e_i\}$ is a sequence of non-observable random errors, and β_0 is the unknown p-vector of regression coefficient. The distribution function F of e_i has median

*This work was partially supported by National Natural Science Foundation of China (10471136 and 10671189) and the Knowledge Innovation Program of the Chinese Academy of Sciences (KJCX3-SYW-S02).

Received September 26, 2007. Revised February 26, 2008.

zero and positive derivative $f(0)$ at zero. This regression model with the nonnegativity constraints on the dependent variables is named as censored regression ("Tobit") model.

This model commonly presents in econometrics when the output variable is constrained from above or below. For example, we want to use data on the number of tickets sold from previous concerts to study the demand for concert tickets. The concert tickets, however, are occasionally sold out and the demand variable would be restricted by the size of the auditorium. Analogously, another example is that we want to investigate sales of a particular good, but when sales are less than preset value C, we record the sales C, so the sales variable would be censored at C.

Censored regression models have been widely studied. Powell (1984) introduced and studied the asymptotic properties of the least absolute deviations (LAD) estimate $\beta_n^{L_1}$ of β_0, which is a Borrel-measurable solution of the minimization problem

$$\sum_{i=1}^{n} |Y_i^+ - (x_i'\beta_n^{L_1})^+| = \inf_{\beta\in\mathbf{B}} \sum_{i=1}^{n} |Y_i^+ - (x_i'\beta)^+|.$$

Since $\sum_{i=1}^{n} |Y_i^+ - (x_i'\beta)^+|$ is not convex in β, the analysis of $\beta_n^{L_1}$ is quite difficult. However, by using uniform laws of large numbers, he established the strong consistency of $\beta_n^{L_1}$ when $\{x_i, i = 1, \cdots, n\}$ are independently random variables with $\mathsf{E}\|x_i\|^3$ being bounded, where $\|\cdot\|$ denotes the Euclidean norm of a vector. By extending a technique due to Huber (1967), he also established its asymptotic normal distribution under some conditions. Pollard (1990) used the maximal inequalities to improve the relevant result of Powell on asymptotic normality. He relaxed the assumptions and simplified the proof to some extent.

Chen and Wu (1993) studies the strong consistency of $\beta_n^{L_1}$ by using a different method when x_i is bound. Rao and Zhao (1993) obtained the asymptotic normality of $\beta_n^{L_1}$ under weaker conditions. Recently Zhao and Fang (2004) used randomly weighting method to derive the approximate distribution for model (1.1), and Fang, Jin and Zhao (2005) studied the strong consistency and Bahadur strong representation of $\widehat{\beta}_n$, where $\{x_i\}$ is a sequence of random variable. Linear hypothesis testings in the censored regression models have been studied by Zhao (2004) and Wang, Wu and Zhao (2009).

Model (variable) selection is an important issue of building model. For the least square regression, there are a large number of well variable selection methods, including the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), Mallows's C_p and so on. Two basic elements focused on by these variable selection criterions are goodness of model fit and model complexity. These criterions use various methods to describe the trade-offs between these two basic aspects and make the related predict error be minimized. However, for the censored regression model, there are few works on model

selection. Jin, Fang and Zhao (2005) presented some selection procedure based on the information theoretic criteria and shown these procedures to be consistent.

A novel variable selection approach, the Least Absolute Shrinkage and Selection Operator (LASSO), was proposed by Tibshirani (1996). This method simultaneously deal with variable selection and estimation by solving a single minimization problem. Fu and Knight (2000) established some asymptotic properties for LASSO-type estimators. Fan and Li (2001) proposed Smoothly Clipped Absolute Deviation (SCAD) approach and obtained its optimal properties. Efron et al. (2004) introduced the Least Angel Regression (LARS) method and discussed its connection with LASSO.

Generally, for linear regression model,

$$Z_i = T_i'\beta_0 + e_i, \qquad i = 1, 2, \cdots, n,$$

where $\{T_i\}$ is a sequence of vector with length p, $\{e_i\}$ is a sequence of non-observable random errors, and β_0 is an unknown p-vector of regression coefficient. LASSO estimator is defined as a minimizer of

$$\sum_{i=1}^{n}(Z_i - T_i'\beta)^2 + \lambda_n \sum_{j=1}^{p}|\beta_j|.$$

It can be equivalently defined as a minimizer of

$$\sum_{i=1}^{n}(Z_i - T_i'\beta)^2,$$

subject to $\sum\limits_{j=1}^{p}|\beta_j| \le s * \sum\limits_{j=1}^{p}|\beta_j^{\mathrm{LS}}|$, where β_j^{LS} is the usual least square estimator. This definition of LASSO estimator illustrates it's remarkable ability, some elements of β exactly are estimated as 0. In fact, estimated value of β goes from nonzero to 0 as the shrinkage coefficient s varies from 1 to 0.

In the present paper, we propose a parallel approach for censored regression model by LASSO method in linear regression model. Our proposed estimator of β is a minimizer of the criterion function,

$$\sum_{i=1}^{n}|Y_i^+ - (x_i'\beta)^+| + \lambda_n \sum_{j=1}^{p}|\beta_j|. \tag{1.2}$$

It can be equivalently defined as a minimizer of the objective function

$$\sum_{i=1}^{n}|Y_i^+ - (x_i'\beta)^+|,$$

subject to $\sum\limits_{j=1}^{p}|\beta_j| < s * \sum\limits_{j=1}^{p}|\beta_j^{L_1}|$, where $\beta_j^{L_1}$ is the usual LAD estimator. Since penalty parameter is constant with regards to different components of β, we name this method as constant penalty L_1 constraint method (CPLC).

Fu and Knight (2000), however, pointed out that LASSO method just correctly sets unnecessary coefficients to 0 with nonzero probability and LASSO estimator is not consistent if $\lambda_n > 0$ as sample size increases. In view of this flaw, we need to modify and extend the objective function (1.2). Note that penalty parameter λ_n plays a curial rule in offsetting between estimation of β_0 and variable selection. Large values of λ_n tend to focus on variables selection, more precisely, remove more variables as well as increasing bias of estimation. And small values tend to weaken variable selection and reduce bias of estimation. Thus we want to use diverse penalty parameters for different elements of β, a large value λ is used for regression coefficient which is close to 0 (need to be removed) and a small value is set to regression coefficient which is significantly not equal to 0. In other word, our estimator is a minimizer of the following objective function with diverse penalty parameters:

$$Z_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} |Y_i^+ - (x_i'\beta)^+| + \frac{1}{n}\sum_{j=1}^{p} \lambda_{nj}|\beta_j|, \tag{1.3}$$

where λ_{nj} is the penalty parameter. We denote this penalized L_1 estimator by $\widehat{\beta}_n$. Corresponding to CPLC, we name this method as diverse penalty L_1 constraint method (DPLC).

In this paper, main results are introduced in the next section. In section 3, extensive simulation studies are conducted to evaluate the performance of the proposed variable selection method. The proofs of main results are given in section 4.

§2. Main Results

Let $\beta_0 = (\beta_0^1, \beta_0^2)^T$, where β_0^1 is a vector of length s and β_0^2 is a vector of length $p-s$. Without loss of generality, assume that all elements of β_0^1 is nonzero and $\beta_0^2 = 0$. For simplicity, write

$$\mu_i = x_i'\beta_0, \qquad S_n = \sum_{i=1}^{n} I(\mu_i > 0) x_i x_i'.$$

We need some assumptions before presenting our main results, described as follow,

(A_1) $e_1, e_2, \cdots$ are i.i.d. random variables such that the distribution function F of e_1 has median zero and positive derivative $f(0)$ at zero.

(A_2) The parameter space B to which β_0 belongs is a bounded open convex set of R^p (with a closure $\overline{B}$) and 0 also belongs to B.

(A_3) $S_n/n \longrightarrow V^2$ as $n \to \infty$.

(A_4) For any $\gamma > 0$, there exists a finite $\alpha > 0$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\|x_i\|^2 I(\|x_i\| > \alpha) < \gamma, \qquad \text{for } n \text{ large enough.}$$

(A_5) For any $\gamma > 0$, there exists a finite $\delta > 0$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\|x_i\|^2 I(\|\mu_i\| \le \delta) < \gamma, \qquad \text{for } n \text{ large enough.}$$

Under conditions (A_1)–(A_5), we will study the asymptotic property of $\widehat{\beta}_n$. The following theorem shows that $\widehat{\beta}_n$ is a consistent estimator of β_0 when $\lambda_{nj} = o(n)$.

Theorem 2.1 Assume that in the model (1.1), the conditions (A_1)–(A_5) hold and $\lambda_{nj}/n \to \lambda_{0j} \ge 0$, $1 \le j \le p$. Then $\widehat{\beta}_n - \arg\min(Z(\beta)) \longrightarrow 0$ in probability, where

$$Z(\beta) = f(0)(\beta - \beta_0)'V^2(\beta - \beta_0) + \sum_{j=1}^{p}\lambda_{0j}|\beta_j|.$$

In particular if $\lambda_{nj} = o(n)$, $\widehat{\beta}_n$ is a consistent estimator of β_0 since β_0 is minimizer of $Z(\beta)$.

From this theorem, we can see that $Z(\beta)$ is strictly convex function in β, function $(\beta - \beta_0)'V^2(\beta - \beta_0)$ achieves its minimization at point β_0 and minimizer of $\sum_{j=1}^{p}\lambda_{0j}|\beta_j|$ is 0. If $\lambda_{0j} > 0$ for some $j \in \{1, \cdots, s\}$, then $Z(\beta)$ does not attain its minimization at points β_0 and 0. Consequently, $\widehat{\beta}_n$ is not a consistent estimator of β_0, which also indicates that CPLC is not a consistent estimation method for β_0. Furthermore, we investigate the root-n consistency of $\widehat{\beta}_n$ by the next theorem.

Theorem 2.2 Assume that in the model (1.1), the conditions (A_1)–(A_5) hold and $\lambda_{nj}/\sqrt{n} \to \lambda_{0j} \ge 0$, $1 \le j \le p$. Then $\sqrt{n}(\widehat{\beta}_n - \beta_0) \longrightarrow \arg\min(U(t))$ in distribution, where

$$U(u) = W^T u + f(0)u^T V^2 u + \sum_{j=1}^{s}\lambda_{0j}\operatorname{sgn}(\beta_{0j})u_j + \sum_{j=s+1}^{p}\lambda_{0j}|u_j|, \tag{2.1}$$

and W has a distribution with $N(0, V^2)$. In particular if $\lambda_{nj} = o(\sqrt{n})$, $\widehat{\beta}_n$ behaves like the LAD estimator of β_0.

Theorem 2.1 and Theorem 2.2 suggest that $\{\lambda_{nj}\}$ can play a different role in penalizing nonzero and zero components of β. Compared to the elements $\beta_j \ne 0$, we can give heavier penalty for $\beta_j = 0$. For example, we choose $\lambda_{nj}/\sqrt{n} \to 0$ for $1 \le j \le s$ and $\lambda_{nj}/\sqrt{n} \to M$, M is large enough for $s + 1 \le j \le p$. Function $Z_n(\beta)$ can not get its minimization until $\beta_j = 0$, $j = s + 1, \cdots, p$. So we not only get a consistent estimator of β_0, $\widehat{\beta}_n$, but also can achieve the goal of selecting the variables whose coefficients are not significantly equal to zero. These conclusions will be validated by extensive simulation results.

§3. Simulation Studies

In this section, we use extensive simulation studies to evaluate the performance of variables selection and estimation of our proposed method for censored regression model. To apply our theorems, we need to construct suitable penalty parameter $\{\lambda_{nj}\}$. Due to penalty parameters having different influence for nonzero and zero parameters, we need to use the LAD estimator $\beta_n^{L_1}$ of β_0. Denote the components of $\beta_n^{L_1}$ and their standard error by a_j and b_j respectively, $j = 1, \cdots, p$. We define λ_{nj} as,

$$\lambda_{nj} = \eta * \Big(\frac{\sqrt{n}|b_j|}{|a_j|}\Big)^{\tau}, \qquad \tau > 1,\ \eta > 0. \tag{3.1}$$

Noting that a_j/b_j converges to a normal distribution with mean 0 and variance 1 if the j-th element of β_0 equals to 0. Then $\lambda_{nj}/\sqrt{n} = \eta * n^{\tau/2-1/2}/|a_j/b_j|^{\tau}$ is very large in probability for n large enough. If the j-th element of $\beta_0 = \theta \neq 0$, then $\sqrt{n}(a_j - \theta) \to N(0, \sigma^2)$ and $\sqrt{n}b_j \to \sigma$ in probability. Therefore, $\lambda_{nj}/\sqrt{n} = \eta * n^{-1/2}((\sqrt{n}b_j)/(\varsigma_n/\sqrt{n} + \theta))^{\tau} \to 0$ in probability, where ς_n follows distribution as $N(0, \sigma^2)$.

For selection of regularization parameters (η, τ), we can use cross validation (CV) or general cross validation (GCV) to do that. Here, we respectively minimize BIC type criterion,

$$\mathrm{BIC}(\eta, \tau) = \frac{L_n(\widehat{\beta}(\eta, \tau))}{L_n(\beta_n^{L_1})} + \frac{\log n}{2} * \sharp\{i : \widehat{\beta}_i \neq 0, i = 1, \cdots, p\},$$

and GCV type criterion,

$$\mathrm{GCV}(\eta, \tau) = \frac{1}{n}\frac{L_n(\widehat{\beta}(\eta, \tau))}{(1 - \sharp\{i : \widehat{\beta}_i \neq 0, i = 1, \cdots, p\}/n)^2},$$

to selection (η, τ), where $L_n(\beta) = \sum_{i=1}^{n} |Y_i^+ - (x_i'\beta)^+|$.

Numeric examples are conducted to compare DPLC method (1.3) with CPLC method (1.2) and the generally best subset selection method (GBSS). In GBSS method, the subset which respectively minimizes BIC type criterion and GCV type criterion among all the possible subsets is chosen as the best subset. To reduce the computing complexity, here τ takes 1.5 for DPLC.

In our simulation studies, observations are generated from censored regression model (1.1). The components of x follow normal distribution with mean 1 and variance 1, and the correlation of x_i and x_j is $\rho^{|i-j|}$ with $\rho = 0.5$. we take different true parameters ($\beta_0' = (1, -1, 0, 0, 1, 0)$ and $(3, -1.5, 0, 0, 2, 0)$) and different sample sizes ($n = 50$, 100 and 150). All of the simulation procedures repeat for 1000 times. We use model error, L_n/n, to

measure the performance of each model fitting, and compare model error of each variable selected method to the full least absolute derivation estimation. The relative value is named as relative model error (RME).

Table 1 Variables selection results for true parameters $\beta_0' = (1, -1, 0, 0, 1, 0)$

n	Method	MRME	Avg. No. of 0 Coefficients	
			correct	incorrect
50	CPLC(BIC)	1.018	1.292(0.926)	0.006(0.084)
	CPLC(GCV)	1.023	1.413(0.919)	0.009(0.107)
	DPLC(BIC)	1.030	2.248(0.887)	0.017(0.144)
	DPLC(GCV)	1.035	2.350(0.828)	0.019(0.151)
	GBSS(BIC)	1.024	2.312(0.764)	0.016(0.133)
	GBSS(GCV)	1.028	2.427(0.716)	0.017(0.144)
100	CPLC(BIC)	1.012	1.501(0.905)	0(0)
	CPLC(GCV)	1.011	1.448(0.913)	0(0)
	DPLC(BIC)	1.017	2.467(0.773)	0(0)
	DPLC(GCV)	1.016	2.403(0.819)	0(0)
	GBSS(BIC)	1.015	2.465(0.688)	0(0)
	GBSS(GCV)	1.013	2.395(0.722)	0(0)
150	CPLC(BIC)	1.008	1.537(0.909)	0(0)
	CPLC(GCV)	1.006	1.414(0.901)	0(0)
	DPLC(BIC)	1.012	2.552(0.705)	0(0)
	DPLC(GCV)	1.010	2.422(0.786)	0(0)
	GBSS(BIC)	1.011	2.592(0.605)	0(0)
	GBSS(GCV)	1.010	2.477(0.686)	0(0)

*Standard deviations are in parentheses.

First, we evaluate ability of different variable selected methods in fitting model and discriminating nonzero coefficients from zero coefficients. Here the error distribution takes standard normal distribution $N(0, 1)$. Table 1 and Table 2 respectively present mean relative model error (MRME), average numbers of zero coefficients selected correctly (zero coefficients correctly estimated as zero value) and average numbers of zero coefficients selected incorrectly (nonzero coefficients incorrectly estimated as zero value) with true parameters $\beta_0' = (1, -1, 0, 0, 1, 0)$ and $(3, -1.5, 0, 0, 2, 0)$. From columns 3 in Table 1 and Table 2, the MRMEs of different variable selection procedures are very close to 1 which indicates that models constructed by various methods fit well. Columns 3 also shows that

Table 2 Variables selection results for true parameters $\beta_0' = (3, -1.5, 0, 0, 2, 0)$

n	Method	MRME	Avg. No. of 0 Coefficients	
			correct	incorrect
50	CPLC(BIC)	1.019	1.450(0.924)	0(0)
	CPLC(GCV)	1.024	1.565(0.930)	0(0)
	DPLC(BIC)	1.029	2.481(0.738)	0.001(0.032)
	DPLC(GCV)	1.033	2.577(0.679)	0.001(0.032)
	GBSS(BIC)	1.028	2.436(0.699)	0(0)
	GBSS(GCV)	1.032	2.549(0.642)	0(0)
100	CPLC(BIC)	1.012	1.631(0.894)	0(0)
	CPLC(GCV)	1.011	1.587(0.898)	0(0)
	DPLC(BIC)	1.017	2.625(0.646)	0(0)
	DPLC(GCV)	1.016	2.580(0.678)	0(0)
	GBSS(BIC)	1.016	2.637(0.572)	0(0)
	GBSS(GCV)	1.014	2.554(0.629)	0(0)
150	CPLC(BIC)	1.009	1.684(0.894)	0(0)
	CPLC(GCV)	1.007	1.567(0.911)	0(0)
	DPLC(BIC)	1.012	2.732(0.526)	0(0)
	DPLC(GCV)	1.011	2.632(0.627)	0(0)
	GBSS(BIC)	1.011	2.696(0.540)	0(0)
	GBSS(GCV)	1.009	2.568(0.635)	0(0)

*Standard deviations are in parentheses.

model fitting is better and better when sample size increases. The average numbers of zero coefficients selected incorrectly in columns 5 in Table 1 and Table 2, which approach to zero and especially equal to 0 when sample size is large (i.e. 100 and 150), suggest that nonzero coefficients can not be estimated with zero. From Columns 4 in Table 1 and 2, the average numbers of zero coefficients correctly selected by DPLC methods are almost as many as these by GBSS methods while CPLC methods choose the fewest average numbers. And the average numbers of zero coefficients correctly selected are much closer to true numbers of zero coefficients, 3, as sample size increases. Compared to GCV criterion, for small sample size (i.e. 50) BIC criterion correctly selects fewer numbers of zero coefficients while BIC correctly selects more numbers for large sample size (i.e. 100 and 150). For example, in Table 1 DPLC(BIC) selects 2.248 zero coefficients and DPLC(GCV) selects 2.350 for sample size 50, while DPLC(BIC) selects 2.467 and DPLC(GCV) selects 2.403 for sample size 100. At the same time, compared to Table 1 with $\beta_0' = (1, -1, 0, 0, 1, 0)$, Table 2 with

$\beta_0' = (3, -1.5, 0, 0, 2, 0)$ have larger average numbers of zero coefficients selected correctly, i.e. when sample size is 150, DPLC(BIC) selects 2.552 for Table 1 and 2.732 for Table 2.

Table 3 Parameters estimation results for true $\beta_0' = (1, -1, 0, 0, 1, 0)$

n	Method	$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_5$	
		Mean	SE	Mean	SE	Mean	SE
50	CPLC(BIC)	0.919	0.250	-0.884	0.306	0.931	0.266
	CPLC(GCV)	0.905	0.251	-0.864	0.306	0.921	0.265
	DPLC(BIC)	0.945	0.257	-0.924	0.302	0.962	0.243
	DPLC(GCV)	0.940	0.259	-0.916	0.302	0.957	0.245
	GBSS(BIC)	0.997	0.247	-1.005	0.309	1.014	0.264
	GBSS(GCV)	0.997	0.245	-1.005	0.308	1.014	0.257
100	CPLC(BIC)	0.919	0.164	-0.897	0.211	0.944	0.182
	CPLC(GCV)	0.924	0.164	-0.903	0.211	0.947	0.182
	DPLC(BIC)	0.970	0.166	-0.947	0.206	0.979	0.166
	DPLC(GCV)	0.973	0.165	-0.950	0.207	0.980	0.168
	GBSS(BIC)	1.011	0.164	-1.003	0.196	1.003	0.174
	GBSS(GCV)	1.010	0.164	-1.004	0.198	1.002	0.178
150	CPLC(BIC)	0.941	0.137	-0.917	0.170	0.949	0.140
	CPLC(GCV)	0.949	0.136	-0.931	0.169	0.954	0.142
	DPLC(BIC)	0.990	0.132	-0.981	0.168	0.986	0.129
	DPLC(GCV)	0.993	0.132	-0.986	0.168	0.989	0.134
	GBSS(BIC)	1.001	0.130	-0.995	0.165	1.003	0.136
	GBSS(GCV)	1.003	0.130	-0.997	0.167	1.006	0.142

Next we investigate the performance of nonzero coefficients' estimators by using 6 different variable selected procedures mentioned above. Table 3 and Table 4 respectively present mean values of nonzero coefficients' estimators (Mean) and their standard errors (SE) with true parameters $\beta_0' = (1, -1, 0, 0, 1, 0)$ and $(3, -1.5, 0, 0, 2, 0)$. From columns 3, 5 and 7 in Tables 3 and 4, DPLC methods possess more exact Mean values than CPLC, almost the same ones as GBSS. For example with sample size 100 and the nonzero values of true parameters $(3, -1.5, 2)$, Mean estimators are $(2.921, -1.405, 1.950)$ for CPLC(GCV), $(2.992, -1.482, 1.993)$ for DPLC(GCV) and $(3.001, -1.499, 2.000)$ for GBSS(GCV). Generally, standard errors SE for DPLC is a little bigger than these for GBSS, but a little less than these for CPLC. When the sample size increases, the Means are much closer to true coefficients (i.e. for DPLC(GCV) with true parameters $(3, -1.5, 2)$, $(2.979, -1.464, 1.985)$

Table 4 Parameters estimation results for true $\beta_0' = (3, -1.5, 0, 0, 2, 0)$

n	Method	$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_5$	
		Mean	SE	Mean	SE	Mean	SE
50	CPLC(BIC)	2.921	0.241	-1.392	0.286	1.931	0.244
	CPLC(GCV)	2.906	0.243	-1.371	0.285	1.922	0.242
	DPLC(BIC)	2.981	0.242	-1.469	0.267	1.985	0.218
	DPLC(GCV)	2.979	0.244	-1.464	0.270	1.985	0.214
	GBSS(BIC)	3.009	0.229	-1.504	0.257	2.011	0.233
	GBSS(GCV)	3.006	0.228	-1.503	0.256	2.011	0.226
100	CPLC(BIC)	2.916	0.171	-1.400	0.194	1.947	0.166
	CPLC(GCV)	2.921	0.172	-1.405	0.194	1.950	0.167
	DPLC(BIC)	2.991	0.158	-1.481	0.172	1.991	0.145
	DPLC(GCV)	2.992	0.158	-1.482	0.172	1.993	0.147
	GBSS(BIC)	3.000	0.154	-1.499	0.176	2.002	0.151
	GBSS(GCV)	3.001	0.155	-1.499	0.177	2.000	0.155
150	CPLC(BIC)	2.922	0.141	-1.409	0.163	1.946	0.131
	CPLC(GCV)	2.933	0.140	-1.421	0.161	1.952	0.133
	DPLC(BIC)	2.989	0.127	-1.488	0.142	1.994	0.120
	DPLC(GCV)	2.990	0.126	-1.489	0.142	1.994	0.122
	GBSS(BIC)	2.999	0.128	-1.497	0.144	2.002	0.116
	GBSS(GCV)	3.000	0.128	-1.498	0.145	2.002	0.122

for sample size 50 and $(2.990, -1.489, 1.994)$ for sample size 150) and the SEs also sharply descend (i.e. for DPLC(GCV) with true parameters $(3, -1.5, 2)$, $(0.244, 0.270, 0.214)$ for sample size 50 and $(0.126, 0.142, 0.122)$ for sample size 150).

To study robustness of different variable selected methods, error distribution is drawn from two other distributions: standard Cauchy distribution (Cauchy) and mixture distribution $(0.5 * N(0, 1) + 0.5 *$ Cauchy) which represents that observations are sampled from $N(0, 1)$ with 50% outlier from Cauchy distribution. Here we just describe results for true $\beta_0 = (3, -1.5, 0, 0, 2, 0)$ with sample size 100. Table 5 illustrates the mean relative model error (MRME), average numbers of zero coefficients selected correctly and average numbers of zero coefficients selected incorrectly. Mean values of nonzero coefficients' estimators and their standard errors are gave in Table 6. From Table 5, the MRMEs are almost equal to 1 for each variable selected methods. DPLC have the almost same average numbers of zero coefficients selected correctly as GBSS while CPLC

Tables 5 Robust property of variables selection results for true parameters $\beta_0' = (3, -1.5, 0, 0, 2, 0)$ with sample size 100

Error distr.	Method	MRME	Avg. No. of 0 Coefficients	
			correct	incorrect
Mixture	CPLC(BIC)	1.013	2.009(0.878)	0.013(0.113)
	CPLC(GCV)	1.012	1.966(0.884)	0.012(0.109)
	DPLC(BIC)	1.013	2.871(0.375)	0.012(0.126)
	DPLC(GCV)	1.012	2.837(0.420)	0.010(0.118)
	GBSS(BIC)	1.012	2.891(0.330)	0.018(0.172)
	GBSS(GCV)	1.012	2.862(0.365)	0.018(0.172)
Cauchy	CPLC(BIC)	1.013	2.208(0.786)	0.046(0.224)
	CPLC(GCV)	1.012	2.169(0.806)	0.039(0.209)
	DPLC(BIC)	1.011	2.924(0.304)	0.054(0.312)
	DPLC(GCV)	1.011	2.904(0.342)	0.050(0.303)
	GBSS(BIC)	1.011	2.940(0.242)	0.067(0.330)
	GBSS(GCV)	1.010	2.923(0.274)	0.059(0.309)

*Standard deviations are in parentheses.

Tables 6 Robust property of parameters estimation results for true $\beta_0' = (3, -1.5, 0, 0, 2, 0)$ with sample size 100

Error distr.	Method	$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_5$	
		Mean	SE	Mean	SE	Mean	SE
Mixture	CPLC(BIC)	2.827	0.234	-1.286	0.275	1.897	0.183
	CPLC(GCV)	2.834	0.232	-1.295	0.271	1.901	0.184
	DPLC(BIC)	2.977	0.200	-1.453	0.253	1.974	0.181
	DPLC(GCV)	2.981	0.193	-1.458	0.243	1.976	0.185
	GBSS(BIC)	2.989	0.220	-1.490	0.264	1.996	0.213
	GBSS(GCV)	2.988	0.220	-1.491	0.265	1.995	0.214
Cauchy	CPLC(BIC)	2.740	0.339	-1.167	0.374	1.843	0.251
	CPLC(GCV)	2.752	0.328	-1.182	0.362	1.849	0.251
	DPLC(BIC)	2.921	0.336	-1.376	0.359	1.937	0.298
	DPLC(GCV)	2.924	0.335	-1.385	0.353	1.939	0.291
	GBSS(BIC)	2.975	0.339	-1.438	0.384	1.960	0.316
	GBSS(GCV)	2.982	0.321	-1.448	0.368	1.965	0.315

have the least ones. However, average numbers selected incorrectly are a little bigger than 0 for all methods. Table 6 also shows that DPLC methods have more exact Mean values than CPLC, but a little less than GBSS. From Tables 2, 4 and Tables 5, 6, average numbers of zero coefficients selected correctly is much closer to the true number 3, and average numbers selected incorrectly and Means further deviate from true parameters when the tails of error distributions become thicker. For instance with DPLC(BIC), average numbers selected correctly and incorrectly are $(2.625, 0)$ and Means of nonzero coefficients are $(2.991, -1.481, 1.991)$ for $N(0,1)$, $(2.871, 0.012)$ and $(2.977, -1.453, 1.974)$ for Mixture, $(2.924, 0.054)$ and $(2.921, -1.376, 1.937)$ for Cauchy. We also see that the standard errors SEs augment as the tails of error distributions are heavier. For example with DPLC(BIC), SEs are $(0.158, 0.172, 0.145)$ for $N(0,1)$, $(0.200, 0.253, 0.181)$ for Mixture and $(0.336, 0.359, 0.298)$ for Cauchy.

§4. Proof of Main Theorems

Hereafter, denote by c a positive constant independent of sample size n, which may stand for various values in different places of formulae.

For simplicity, we set $\gamma = \beta - \beta_0$, $\widehat{\gamma}_n = \widehat{\beta}_n - \beta_0$,

$$L_n(\beta) = \sum_{i=1}^{n} |Y_i^+ - (x_i'\beta)^+|$$

and

$$G_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}(|(\mu_i + x_i'\gamma)^+ - Y_i^+| - |\mu_i^+ - Y_i^+|).$$

Seen in Rao and Zhao (1993), we have

$$G_n(\gamma) \longrightarrow G(\gamma) = f(0)\gamma' V^2 \gamma, \tag{4.1}$$

and for any sequence $\{\epsilon_n\}$ satisfying $\epsilon_n \to 0$, we get

$$L_n(\beta) - L_n(\beta_0) = -W_n' S_n^{1/2}\gamma + f(0)\gamma' S_n \gamma + o_p(1 + \gamma' S_n \gamma), \quad \text{uniformly for } \|\gamma\| \le \epsilon_n, \tag{4.2}$$

where $W_n = \sum_{i=1}^{n} S_n^{-1/2} x_i \operatorname{sgn}(e_i) I(\mu_i > 0)$. Hence,

$$\begin{aligned} \frac{1}{n}(L_n(\beta) - L_n(\beta_0)) &= -\frac{1}{n}\sum_{i=1}^{n} x_i'\gamma \operatorname{sgn}(e_i) I(\mu_i > 0) + f(0)\gamma' V^2 \gamma \\ &\quad + o_p\Big(\frac{1}{n} + \gamma' V^2 \gamma\Big), \qquad \text{uniformly for } \|\gamma\| \le \epsilon_n. \end{aligned} \tag{4.3}$$

Proof of Theorem 2.1 By the uniform law of large number, we know

$$\frac{1}{n}(L_n(\beta)-L_n(\beta_0))-G_n(\gamma)=o_p(1) \text{ uniformly for } \beta \text{ in any compact set } K.$$

Therefore, from definition of $Z_n(\beta)$ and (4.1),

$$Z_n(\beta)-\frac{1}{n}L_n(\beta_0)-Z(\beta)=o_p(1) \text{ uniformly for } \beta \text{ in any compact set } K. \tag{4.4}$$

From (4.4) and $\widehat{\beta}_n$ being bound, the proof of this theorem is completed. □

To prove Theorem 2.2, we need following lemma,

Lemma 4.1 For any given $u\in R^p$, we have

$$M(u)-M(\widehat{u})\geq (u-\widehat{u})'D(u-\widehat{u})/2,$$

where $M(u)=u'Du/2-a'u+\sum\limits_{j=1}^{s}\lambda_j u_j+\sum\limits_{j=s+1}^{p}\lambda_j|u_j|$, D is a positive definite matrix, $\lambda_1,\cdots,\lambda_s$ are constants, $\lambda_{s+1},\cdots,\lambda_p$ are nonnegative constants, and $\widehat{u}$ is a minimizer of $M(u)$.

Proof Proof of this lemma can be refer to Proposition 2 of Xu and Ying (2008). □

First, define

$$\begin{aligned}B_n(u) &= n^{-1/2}\sum_{i=1}^{n}x_i'u\,\mathrm{sgn}(e_i)I(\mu_i>0)+f(0)u'V^2u\\ &\quad+\sum_{j=1}^{s}\frac{\lambda_{nj}}{\sqrt{n}}\,\mathrm{sgn}(\beta_{oj})u_j+\sum_{j=s+1}^{p}\frac{\lambda_{nj}}{\sqrt{n}}|u_j|.\end{aligned}$$

Let $\widetilde{u}_n$ is a minimizer of $B_n(u)$.

Proof of Theorem 2.2 We know $\widehat{\beta}_n$ is a consistent estimator of β_0. From the definition of $Z_n(\beta)$, we have

$$\begin{aligned}&Z_n(\widehat{\beta}_n)-Z_n(\beta_0)\\ =\;&\frac{1}{n}\sum_{i=1}^{n}x_i'\widehat{\gamma}_n\,\mathrm{sgn}(e_i)I(\mu_i>0)+\frac{1}{n}f(0)\widehat{\gamma}_n'S_n\widehat{\gamma}_n\\ &+\sum_{j=1}^{s}\frac{\lambda_{nj}}{n}(|\widehat{\beta}_{nj}|-|\beta_{oj}|)+\sum_{j=s+1}^{p}\frac{\lambda_{nj}}{n}|\widehat{\beta}_{nj}|+o_p\Big(\frac{1}{n}+\widehat{\gamma}_n'V^2\widehat{\gamma}_n\Big)\\ =\;&\frac{1}{n}\sum_{i=1}^{n}x_i'\widehat{\gamma}_n\,\mathrm{sgn}(e_i)I(\mu_i>0)+f(0)\widehat{\gamma}_n'V^2\widehat{\gamma}_n\\ &+\sum_{j=1}^{s}\frac{\lambda_{nj}}{n}\,\mathrm{sgn}(\beta_{oj})(\widehat{\beta}_{nj}-\beta_{oj})+\sum_{j=s+1}^{p}\frac{\lambda_{nj}}{n}|\widehat{\beta}_{nj}|+o_p\Big(\frac{1}{n}+\widehat{\gamma}_n'V^2\widehat{\gamma}_n\Big)\\ =\;&\frac{1}{n}B_n(\widehat{u}_n)+o_p\Big(\frac{1}{n}(1+\widehat{u}_n'V^2\widehat{u}_n)\Big),\end{aligned} \tag{4.5}$$

where $\widehat{u}_n = \sqrt{n}(\widehat{\beta}_n - \beta_o)$.

Since $n^{-1/2}\sum_{i=1}^{n} x_i \operatorname{sgn}(e_i) I(\mu_i > 0) \to N(0, V^2)$, then $B_n(u) \xrightarrow{\mathcal{L}} U(u)$. We know $\widetilde{u}_n \xrightarrow{P} \arg\min(U(u))$ and $\arg\min(U(u)) = O_p(1)$ such that

$$n^{-1/2}\widetilde{u}_n \xrightarrow{P} 0. \tag{4.6}$$

From (4.5) and (4.6),

$$Z_n(\beta_0 + n^{-1/2}\widetilde{u}_n) - Z_n(\beta_0) = \frac{1}{n}B_n(\widetilde{u}_n) + o_p\Big(\frac{1}{n}(1 + \widetilde{u}_n' V^2 \widetilde{u}_n)\Big). \tag{4.7}$$

From Lemma 4.1, we know

$$B_n(\widehat{u}_n) - B_n(\widetilde{u}_n) \geq f(0)(\widehat{u}_n - \widetilde{u}_n)' V^2 (\widehat{u}_n - \widetilde{u}_n). \tag{4.8}$$

Hence, combining (4.5), (4.7) and (4.8),

$$\begin{aligned}
& Z_n(\widehat{\beta}_n) - Z_n(\beta_0 + n^{-1/2}\widetilde{u}_n) \\
= \ & \frac{1}{n}(B_n(\widehat{u}_n) - B_n(\widetilde{u}_n)) + o_p\Big(\frac{1}{n}(1 + \widetilde{u}_n' V^2 \widetilde{u}_n + \widehat{u}_n' V^2 \widehat{u}_n)\Big) \\
\geq \ & \frac{1}{n} f(0)(\widetilde{u}_n - \widehat{u}_n)' V^2 (\widetilde{u}_n - \widehat{u}_n) + o_p\Big(\frac{1}{n}(1 + \widetilde{u}_n' V^2 \widetilde{u}_n + \widehat{u}_n' V^2 \widehat{u}_n)\Big) \\
\geq \ & \frac{c}{n}(\widetilde{u}_n - \widehat{u}_n)' V^2 (\widetilde{u}_n - \widehat{u}_n) + o_p\Big(\frac{1}{n}(1 + 2\widetilde{u}_n' V^2 \widehat{u}_n)\Big).
\end{aligned}$$

And $Z_n(\widehat{\beta}_n) - Z_n(\beta_0 + n^{-1/2}\widetilde{u}_n) \leq 0$, so the next expression is validated,

$$\widetilde{u}_n \overset{d}{=} \widehat{u}_n.$$

Consequently, the theorem is proved. □

References

[1] Powell, J.L., Least absolute deviations estimates for the censored regression model, *J. Econometrics*, **25**(1984), 303–325.

[2] Huber, P.J., The behaviour of maximum likelihood estimates under nonstandard conditions, *Proc. Fifth Berkeley Symp. Math. Statist. Probab.*, Univ. of California Press, **1**(1967), 221–233.

[3] Pollard, D., Empirical process: theory and application, *NSF-CBMS Regional Conference Series in Probability and Statistics*, Vol.2, Institute of Mathematical statistics, Hayward, 1990.

[4] Chen, X.R. and Wu, Y., Consistency of L_1 estimates in censored linear regression models, *Commu. Statist. Theor. Meth.*, **23(7)**(1993), 1847–1858.

[5] Rao, C.R. and Zhao, L.C., Asymptotic normality of LAD estimators in censored regression models, *Math. Meth. Statist.*, **2**(1993), 228–239.

[6] Zhao, L.C. and Fang Y.X., Random weighting method for censored regression model, *J. Systems Science and Complexity*, **17**(2004), 262–270.

[7] Fang Y.X., Jin, M. and Zhao, L.C., Strong convergence of LAD estimates in a censored regression model, *Science in China, Ser. A, Math.*, **48**(2005), 155–168.

[8] Zhao, L.C., Linear hypothesis testing in censored regression models, *Statistica Sinica*, **14**(2004), 333–347.

[9] Wang, Z.F., Wu, Y.H. and Zhao, L.C., Approximation by randomly weighting method for linear hypothesis testing in censored regression model, *Science in China, Ser. A, Math.*, **52**(2009), 561–576.

[10] Jin, M., Fang, Y.X. and Zhao, L.C., Variable selection for censored regression models, *Chin. J. Applied Probab. Statist.*, **21**(2005), 141–149.

[11] Tibshirani, R.J., Regression shrinkage and selection via lasso, *J. Roy. Statist. Soc. Ser. B*, **58**(1996), 267–288.

[12] Knight, K. and Fu, W.J., Asymptotics for LASSO-type estimators, *Ann. Statist.*, **28**(2000), 1356–1378.

[13] Fan, J. and Li, R., Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.*, **96**(2001), 1348–1360.

[14] Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R., Least angle regression, *Ann. Statist.*, **32**(2004), 407–499.

[15] Rao, C.R. and Zhao, L.C., Approximation to the distribution of M-estimates in linear models by randomly weighted bootstrap, *Sankhyā, Ser. A*, **54**(1992), 323–331.

[16] Rao, C.R. and Zhao, L.C., Asymptotic normality of LAD estimators in censored regression models, *Math. Meth. Statist.*, **2**(1993), 228–239.

[17] Xu, J.F. and Ying, Z.L., Simultaneous estimation and variable selection in median regression using Lasso-type penalty, *Ann. Inst. Stat. Math.*, 2008, published online.

删失回归模型中一个LASSO型变量选择和估计方法

王占锋 吴耀华 赵林城

(中国科学技术大学统计与金融系, 合肥, 230026)

删失回归模型是一种很重要的模型, 它在计量经济学中有着广泛的应用. 然而, 它的变量选择问题在现今的参考文献中研究的比较少. 本文提出了一个LASSO型变量选择和估计方法, 称之为多样化惩罚L_1限制方法, 简称为DPLC. 另外, 我们给出了非0回归系数估计的大样本渐近性质. 最后, 大量的模拟研究表明了DPLC方法和一般的最优子集选择方法在变量选择和估计方面有着相同的能力.

关键词: 删失回归模型, 最小绝对偏差, 变量选择, LASSO.

学科分类号: 62F05, 62G05.