MCP-penalized Regression in High Dimensional Partially Linear Models for Right Censored Data

Xiao-Song QIU¹,a, Ying-Guang LUO²,b,*

¹,²No. 45 on Liberation Park Road, Wuhan, Hubei, China

a874415547@qq.com, blygsuiping@126.com

*Corresponding author

Keywords: Semi-parametric Regression Model, Right Censored, Minimax Concave Penalty, Selection Consistency, A Sieve Approach.

Abstract. Basing on Stute’s weighted least squares method for prognosis studies with right censored response variables, this paper first put forward a semi-parametric regression model include two covariate effects, one is for the low dimensional covariates taken a nonparametric form, the other is for the high dimensional covariates taken a parametric form. Then, the selection of parametric covariate effects was achieved by use of a minimax concave penalty (MCP) approach, it is with great consistency in the semi-parametric regression model. In succession, the nonparametric component was estimated by a sieve approach. Finally, a numerical simulation was made, the result shows that the proposed approach has satisfactory performance.

Introduction


Ma and Du [10] further generalized these results to a semi-parametric regression model whose covariate effects have two parts. The first part is for the low dimensional covariates and takes a nonparametric form, and the second part is for the high dimensional covariates and takes a parametric form. They proposed an iterated LASSO approach for variable selection with parametric component and a sieve approach for the nonparametric component.

In this paper, the minimax concave penalty [11] was adopted for variable selection, the oracle properties was enjoyed under certain regularity conditions. Basing on Stute’s weighted least squares method, the penalized weighted least square estimation was considered for building the AFT model. By use the initial estimate, the weights was constructed and a weighted MCP estimations conducted, the selection consistency and estimation consistency for MCP penalized methods for censored data was proved. The nonparametric component $\eta$ is estimated through a sieve approach.

Regularized Weighted Least Squares Regression

Consider a partially linear model for right censored data
\[ T_i = \alpha + X_i\beta + \eta(Z_i) + \varepsilon_i, \quad i = 1, \ldots, n \]  

(1)

Where \( \alpha \) is the intercept, \( \beta \in \mathbb{R}^p \) is the regression coefficient vector, \( \eta \) is an unknown multivariate smooth function, and \( \varepsilon_i \) is the error term. Since \( \eta \) is identifiable up to a constant, we adopt the constraint \( \int \eta = 0 \). As \( T_i \) is subject to right censoring, we can only observe \((Y_i, \delta_i, X_i)\), where \( Y_i = \min\{T_i, C_i\}, \) \( C_i \) is the logarithm of the censoring time, and \( \delta_i = I\{T_i \leq C_i\} \) is the censoring indicator. Assume a random sample consists of \( n \) such triplets is observed. This model combines the flexibility of nonparametric regression and parsimony of parametric regression.

Let \( \hat{K}_n \) be the Kaplan-Meier estimator of the distribution function \( K \) of \( T \). The weighted least squares loss function is defined as

\[
L(\beta, \eta) = \frac{1}{2} \sum_{i=1}^{n} \omega_m (Y_i - \alpha - X_i\beta - \eta(Z_i))^2.
\]

(2)

Now use the sieve approach to approximate \( \eta \). Suppose \( \{\phi_1, \ldots, \phi_m\} \) is a basis of \( \mathcal{N}_N \). Then any function \( \eta \in \mathcal{H}_n \) can be written as \( \eta(\bullet) = \sum_{i=1}^{m} d_i \phi_i(\bullet) + \sum_{j=1}^{n} R_j(z_j, \bullet) \equiv \Psi(\bullet)b \). Thus, the problem of estimating \( \eta \) becomes that of estimating \( b \). Thus (2) becomes,

\[
L(\beta, b) = \frac{1}{2} \sum_{i=1}^{n} \omega_m (Y_i - X_i\beta - \Psi b)^2.
\]

(3)

Where \( \Psi_j \equiv \Psi(Z_j) \).

Then estimated the baseline function by \( \hat{\alpha} = \overline{Y}_m - \overline{X}_m \hat{\beta} - \overline{\Psi}_m \hat{b} \), where \( \overline{X}_m = \sum_{i=1}^{n} \omega_m X_i / \sum_{i=1}^{n} \omega_m \), \( \overline{\Psi}_m = \sum_{i=1}^{n} \omega_m \Psi_i / \sum_{i=1}^{n} \omega_m \) and \( \overline{Y}_m = \sum_{i=1}^{n} \omega_m Y_i / \sum_{i=1}^{n} \omega_m \). Thus, substituting \( \alpha \) by \( \hat{\alpha} \) in (3) yields

\[
L(\beta, b) = \frac{1}{2} \sum_{i=1}^{n} \omega_m (Y_i^* - X_i^* \beta - \Psi_i b)^2.
\]

(4)

where \( Y_i^* = Y_i - \overline{Y}_m \), \( X_i^* = X_i - \overline{X}_m \) and \( \Psi_i = \Psi_j - \overline{\Psi}_m \). Define the penalized least squares criterion

\[
Q_{\lambda}(\beta, b) = \frac{1}{2} \sum_{i=1}^{n} \omega_m (Y_i^* - X_i^* \beta - \Psi_i b)^2 + \sum_{j=1}^{d} \rho(r_{nj} \mid \beta_j ; \lambda; \gamma).
\]

(5)

where \( \rho \) is a penalty function and \( r_{nj} = \sqrt{\sum_{i=1}^{n} \omega_m X_{ij}^2} \).

Let \( (\hat{\beta}, \hat{b}) = \text{argmin} \ Q_{\lambda}(\beta, b) \). The estimators of \( \beta \) and \( \eta \) are \( \hat{\beta} \) and \( \hat{\eta} = \Psi \hat{b} \).

For any \( \beta \), the \( \hat{b} \) that minimizes \( Q_{\lambda} \) necessarily satisfies \( \sum_{i=1}^{n} \Psi_i(Y_i^* - X_i^* \beta - \Psi_i b) = 0 \). That is, \( \hat{b} = (\sum_{i=1}^{n} \Psi_i \Psi_i^*)^{-1} \sum_{i=1}^{n} \Psi_i(Y_i^* - X_i^* \beta) \).

Let \( P = \Psi_i (\sum_{i=1}^{n} \Psi_i \Psi_i^*)^{-1} \Psi_i^* \) be the projection matrix of the column space of \( \Psi_i \). The objective function (5) can be rewritten as
\[ Q_\lambda(\beta) = \frac{1}{2} \sum_{i=1}^{n} \omega_m ((I-P)(Y_i^*-X_i^* \beta))^2 + \sum_{j=1}^{d} \rho(r_{ij} \mid \beta_j \mid \lambda; \gamma). \]  

(6)

Then, \( \hat{\beta} = \text{argmin} Q_\lambda(\beta) \). Because the profile objective function does not involve \( b \), it is useful for both theoretical investigation and computation. We will use it to established the asymptotic properties of \( \hat{\beta} \). Let \( \hat{Y}_i^* = (I-P^\top)Y_i \), \( \hat{X}_i^* = (I-P^\top)X_i^* \). Thus (6) becomes,

\[ Q_\lambda(\beta) = \frac{1}{2} \sum_{i=1}^{n} \omega_m (\hat{Y}_i^* - \hat{X}_i^* \beta)^2 + \sum_{j=1}^{d} \rho(r_{ij} \mid \beta_j \mid \lambda; \gamma), \]  

(7)

Define the oracle LSE of \( \beta \) as follows,

\[ \hat{\beta}^o \triangleq \text{argmin} \frac{1}{2} \sum_{i=1}^{n} (\hat{Y}_i^* - \hat{X}_i^* \beta)^2 : \beta_j = 0, \forall j \notin A^o \}

Asymptotic Properties

In this section, the selection of MCP estimator for censored data is a discussion.

Lemma 1. For any \( \gamma > 0 \), let \( v_\gamma = E[|e_i - E(e_i \mid \delta_i = 1, X_i)| \mid \delta_i = 1, X_i \} \), then \( v_\gamma \) is independent of \( X_i \) and only depends on \( G(t) \) and \( F(t) \), where \( G(t) \) and \( F(t) \) denote the cumulative distribution functions of \( C_i \) and \( T_i \) respectively.

With the above preparations, make the following assumptions.

(A1)

\[ v_\gamma = \int_{-\infty}^{\infty} t \int_{-\infty}^{\infty} [1-G(t)] dF(t) \int_{-\infty}^{\infty} [1-G(t)] dF(t) < \infty, \]

\[ v_{0\gamma} = \int_{-\infty}^{\infty} t \int_{-\infty}^{\infty} G(t) dF(t) dF(t) < \infty, \]

hold for \( \gamma = 2, 4 \).

(A2)

\[ \min_{\beta_j \neq 0} \frac{|\beta_j|}{\lambda} \to \infty. \]

(A3) Denote by \( M(\beta) \) the cardinality of non-zero coordinates of \( \beta \). \( M(\beta) = s \) is finite.

(A4) For any matrix \( \tilde{X} \) with rank \( \tilde{p} \), there exists constants \( 0 < \tilde{c}_1 < \tilde{c}_2 < \infty \), such that where \( \tilde{\Sigma}_A \) is as in (7) and \( c_{\text{max}}(\tilde{\Sigma}_A) \) is the largest eigenvalue of \( \tilde{\Sigma}_A \).

(A5)

\[ P^o = |A^o| \leq \tilde{p}/(\tilde{K} + 1), \]

where \( \tilde{K} \) is just the \( K \), defined in Zhang [11] with \( \tilde{c}_1, \tilde{c}_2, k \) defined in assumption (A5) and an \( \alpha \in (0, 1) \).

(A6) \( \tilde{p} - P^o \geq m \geq 1 \).

(A7) For any \( X_{B_i}^* \), \( 0 < c_i^* \leq c_{\min}(\Sigma_{B_i}^*) \). here, \( c_{\min}(\Sigma_{B_i}^*) \) is the smallest eigenvalue of \( X_{B_i}^* X_{B_i}^\top \).

(A8) For any \( X_{B_i}^* \), \( \max_{1 \leq j \leq n} X_{ij}^2 \to 0 \) as \( n \to \infty \).
Let $\hat{\beta}^*$ be the oracle LSE and $\tilde{P}_{d^*}$ be the orthogonal projection from $\mathbb{R}^n$ to the linear span of $(\tilde{X}_j, j \in A^*)_{n \times p^*}$. It is easy to see that $\tilde{P}_{d^*} = \tilde{X}_{d^*} (\tilde{X}_{d^*} \tilde{X}_{d^*})^{-1} \tilde{X}_{d^*}$ and $\tilde{Y}_j - \tilde{X}_j^* \hat{\beta}^* = (I - \tilde{P}_{d^*}) e_i$. Then
\[
\sum_{i=1}^n \omega_m \tilde{Y}_j(X_i - \tilde{X}_j^* \hat{\beta}^*) = \sum_{i=1}^n \omega_m \tilde{X}_j^* (I - \tilde{P}_{d^*}) e_i.
\]
Let $Z^*_{ij} = \tilde{X}_j^* (I - \tilde{P}_{d^*}) e_i$. Then we have
\[
\sum_{i=1}^n \omega_m \tilde{X}_j^* (\tilde{Y}_j - \tilde{X}_j^* \hat{\beta}^*) = \sum_{i=1}^n \omega_m Z^*_{ij} e_i \triangleq V^*_j.
\]
Now assume the following conditions.

(A9)
\[
\max_{1 \leq i \leq n} \{ (\omega_m) \omega_m Z^*_{ij} \} \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty.
\]

(A10)
\[
\sum_{i=1}^n \omega_m Z^*_{ij} \geq \alpha \sum_{i=1}^n (\omega_m) \omega_m Z^*_{ij}, \quad \text{where} \quad 0 < \alpha_i < 1.
\]

(A11)
\[
|EV^*_j| \geq \lambda \alpha \sqrt{\sum_{i=1}^n (\omega_m) \omega_m Z^*_{ij}^2},
\]

It is easy to see that $\hat{\beta}^* - \beta = (\tilde{X}_{d^*} W \tilde{X}_{d^*})^{-1} \tilde{X}_{d^*} W e$. Let $U^* = (\tilde{X}_{d^*} W \tilde{X}_{d^*})^{-1} \tilde{X}_{d^*}$. Then we have
\[
\hat{\beta}_j^* - \beta_j = \sum_{i=1}^n \omega_m U^*_j e_i \triangleq T^*_j.
\]

(A12)
\[
\max_{1 \leq i \leq n} \{ (\omega_m) \omega_m U^*_j \} \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty.
\]

(A13)
\[
\sum_{i=1}^n \omega_m U^*_j \geq \alpha_2 \sum_{i=1}^n (\omega_m) \omega_m U^*_j, \quad \text{where} \quad 0 < \alpha_2 < 1.
\]

(A14)
\[
|ET^*_j| \geq \lambda \alpha_2 \sqrt{\sum_{i=1}^n (\omega_m) \omega_m U^*_j^2}.
\]

(A1) requires that both the second central moment and the fourth central moment of $T_i$ are finite given $\delta_j, X_j$. It is key to the following Lemma 1 and Theorem 1. (A2) is an assumption in Fan and Peng [12]. It explicitly shows the rate at which the nonvanishing parameter can be distinguished from 0. The model is sparse under (A3). This assumption is reasonable in genomic studies, where the number of genes profiled can be large, but only a very small number of genes are associated with the response variables. (A4) is the sparse Riesz condition (SRC) that was proposed by Zhang and Huang [13]. (A5) and (A6) require that the model is sparse. (A7) and (A8) are necessary for the following Lemma 4. It is the key to the following Lemma 5 and Theorem 1. (A9) and (A10) require that not too many data are censored. Remark: If the censoring rate is too high, the values of $\omega_m$’s may be very large as $n \rightarrow \infty$. As a result, assumptions (A9) and (A10) may not hold any more. Thus, (A9) and (A10) imply that estimation consistency of MCP in the AFT model requires relatively low censoring rate. A necessary constraint on the censoring rate can be easily ignored in the AFT literature and may lead to serious problems. (A12)-(A14) are similar to the assumptions (A9)-(A11).
Lemma 2. Given $\delta, X$, suppose that (A1), (A9) and (A10) hold, we have
$$\frac{\sqrt{n} \left( V_j^* - E(V_j^* | \delta, X) \right)}{\sigma s_{nj}} \rightarrow N(0,1) \quad \text{as} \quad n \rightarrow \infty,$$
where $s_{nj} = \sqrt{\sum_{i=1}^{n} (n\omega_n) \alpha_n z_{ij}^2}$.

Lemma 3. Given $\delta, X$, suppose that (A9)-(A11) are satisfied. Fix $p \geq 3$. Let $\hat{\beta}$ be the MCP estimator defined by (5) with $\lambda = c\sigma \sqrt{\frac{\log d}{n}}$, for some $c > \sqrt{2}$. Define the event $B = \left\{ \max_{j \neq i} |V_j| \leq \lambda r_{nj} |\delta, X| \right\}$. If $n$ is sufficiently large, then $1 - P(B) \rightarrow 0$, where $A' = A(\beta)$ is the set of non-zero coefficients of $\beta$.

Lemma 4. Suppose that (A1), (A7) and (A8) hold, then for given $\delta, X$, we have
$$\left( \tilde{\varepsilon} - E(\tilde{\varepsilon} | \delta, X) \right) \left( \tilde{P}_x - \tilde{P}_x \right) \left( \tilde{\varepsilon} - E(\tilde{\varepsilon} | \delta, X) \right) \rightarrow_d \mathcal{X}_m^2 \quad \text{as} \quad n \rightarrow \infty,$$
under the condition that some data are censored, the $\left( \tilde{\varepsilon} - E(\tilde{\varepsilon} | \delta, X) \right) \left( \tilde{P}_x - \tilde{P}_x \right) \left( \tilde{\varepsilon} - E(\tilde{\varepsilon} | \delta, X) \right)$'s are not $\mathcal{X}_m^2$ any more even if the $e_i$'s are normal distribution. However, Lemma 4 implies that, under some regular conditions, the asymptotical distributions of the $\left( \tilde{\varepsilon} - E(\tilde{\varepsilon} | \delta, X) \right) \left( \tilde{P}_x - \tilde{P}_x \right) \left( \tilde{\varepsilon} - E(\tilde{\varepsilon} | \delta, X) \right)$'s are still $\mathcal{X}_m^2$. This property is the key to the following Lemma 5, Theorems 1.

Lemma 5. Suppose that (A1), (A6), (A7) and (A8) hold. Given $\delta, X$, consider the MCP estimator $\hat{\beta}$ defined by (5) with $\lambda = c\sigma \sqrt{\frac{\log p}{n}}$, for some $c > \sqrt{2}$, we have
$$P\left( \zeta(\tilde{\varepsilon}; m, A') \geq \frac{\tilde{\varepsilon}}{\sigma \sqrt{\frac{\mathcal{X}_m^2}{\alpha}}} \right) \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty.$$

Theorem 1. Suppose (A1) and (A4)-(A14) hold. Consider the MCP estimator $\hat{\beta}$ defined by (5) with $\lambda = c\sigma \sqrt{\frac{\log d}{n}}$, for some $c > 0$. Let $\lambda_1 = c_1\sigma \sqrt{\frac{\log d}{n}}$ and $\lambda_3 = c_3\sigma \sqrt{\frac{\log d}{n}}$ for two positive constants $c_1, c_3, \lambda_2 \geq \max \left\{ \lambda_1, \sqrt{\frac{\mathcal{X}_m^2}{\alpha}} \right\}$. Suppose $P(\hat{\beta}_j \neq 0 | \beta_j) \geq \gamma \lambda_2$ and $P(\lambda_1 \leq \lambda \leq \lambda_2) = 1$. Then we have
$$P(\hat{A} \neq A_0) \leq P(\text{sign}(\hat{\beta}) \neq \text{sign}(\beta) \text{ or } \hat{\beta} \neq \hat{\beta}^*) \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty.$$

Numerical Simulation

In this section, there is a simulation study done to compare the Lasso, Adaptive Lasso, SCAD and MCP. For the SCAD, let $\gamma = 2.01$ or $3.7$, and for the MCP, set $\gamma = 1.01$ or $3$, and simulated data from the AFT model such that Denote by $W(a,b)$ the Weibull distribution with shape parameter $a$ and scale parameter $b$, then generated the failure time $T$ from a $W(4, \exp(\mu(x,z)))$ distribution, where $\mu(x,z) = x^\beta + \eta(z)$. The dimensionality $p$ of covariate $X$ were 12 and 500. The censoring time $C$ had an exponential distribution whose parameter was adjusted to yield a censoring rate about $30\%$. $X_i$ were independently generated from the multivariate normal distribution with zero mean and $\text{Cov}(X_i, X_j) = 0.112 \cdot 0.5^{p \cdot i}$. The first 6 entries of the coefficient vector $\beta$ were
(1.0,0.9,0.8,-0.8,-0.9,-1.0), and the rest of the entries were all zeros. The covariate \( Z = (Z_1, Z_2, Z_3) \) had 3 dimensions, and each component was simulated from the uniform distribution on \([0,1]\). The function \( \eta(Z) = \eta_1(Z_1) + \eta_2(Z_2) + \eta_3(Z_3) \) with \( \eta_i(Z) = 0.5 \sin(2 \pi z - \pi/2) \), \( \eta_2(Z) = 2(z-0.4)^2 + 2.28e^{-z} - 1.628 \), and \( \eta_3(Z) = z - 0.5 \). Not that all \( \eta_i \)'s integrate to zero to make model identifiable.

Let \( \beta' = (\beta'_o, \beta'_n) \), where \( \beta'_o \) is the zero part and \( \beta'_n \) is the non-zero part of \( \beta \). Let \( |\beta| \) be the number of elements in \( \beta \) and \( \|\beta\|_0 \) be the number of non-zero elements in \( \beta \). Then \( \frac{\|\hat{\beta}_o\|_0}{|\beta_o|} \) is the proportion of non-zero estimates for zero parameters, and \( 1 - \frac{\|\hat{\beta}_n\|_0}{|\beta_n|} \) is the proportion of zero estimates for the non-zero parameters. Hence the smaller \( \frac{\|\hat{\beta}_o\|_0}{|\beta_o|} \) and \( 1 - \frac{\|\hat{\beta}_n\|_0}{|\beta_n|} \) are, the better the method is. Here, let \( F_o \triangleq \frac{\|\hat{\beta}_o\|_0}{|\beta_o|} \) and \( F_n \triangleq 1 - \frac{\|\hat{\beta}_n\|_0}{|\beta_n|} \).

**Table 1. Estimated coefficients by CV: \( n=200, \ p=12 \).**

<table>
<thead>
<tr>
<th>Penalty</th>
<th>Lasso Adaptive Lasso</th>
<th>SCAD(( \gamma = 2.01 ))</th>
<th>SCAD(( \gamma = 3.7 ))</th>
<th>MCP(( \gamma = 1.01 ))</th>
<th>MCP(( \gamma = 3 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\beta}_1 )</td>
<td>0.30</td>
<td>0.73</td>
<td>1.15</td>
<td>1.16</td>
<td>1.18</td>
</tr>
<tr>
<td>( \hat{\beta}_2 )</td>
<td>0.32</td>
<td>0.88</td>
<td>1.00</td>
<td>1.02</td>
<td>1.00</td>
</tr>
<tr>
<td>( \hat{\beta}_3 )</td>
<td>0.28</td>
<td>0.54</td>
<td>1.16</td>
<td>1.14</td>
<td>1.13</td>
</tr>
<tr>
<td>( \hat{\beta}_4 )</td>
<td>-0.42</td>
<td>-1.06</td>
<td>-1.60</td>
<td>-1.51</td>
<td>-1.42</td>
</tr>
<tr>
<td>( \hat{\beta}_5 )</td>
<td>-0.13</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.21</td>
<td>-0.46</td>
</tr>
<tr>
<td>( \hat{\beta}_6 )</td>
<td>-0.39</td>
<td>-1.18</td>
<td>-1.44</td>
<td>-1.34</td>
<td>-1.26</td>
</tr>
<tr>
<td>( F_o )</td>
<td>0.333</td>
<td>0.000</td>
<td>0.000</td>
<td>0.333</td>
<td>0.000</td>
</tr>
<tr>
<td>( F_n )</td>
<td>0.000</td>
<td>0.167</td>
<td>0.167</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Table 2. Estimated coefficients by BIC: \( n=200, \ P=12 \).**

<table>
<thead>
<tr>
<th>Penalty</th>
<th>Lasso SCAD(( \gamma = 2.01 ))</th>
<th>SCAD(( \gamma = 3.7 ))</th>
<th>MCP(( \gamma = 1.01 ))</th>
<th>MCP(( \gamma = 3 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\beta}_1 )</td>
<td>0.29</td>
<td>1.15</td>
<td>1.20</td>
<td>1.18</td>
</tr>
<tr>
<td>( \hat{\beta}_2 )</td>
<td>0.31</td>
<td>1.00</td>
<td>0.91</td>
<td>1.10</td>
</tr>
<tr>
<td>( \hat{\beta}_3 )</td>
<td>0.27</td>
<td>1.16</td>
<td>1.19</td>
<td>1.16</td>
</tr>
<tr>
<td>( \hat{\beta}_4 )</td>
<td>-0.41</td>
<td>-1.60</td>
<td>-1.61</td>
<td>-1.39</td>
</tr>
<tr>
<td>( \hat{\beta}_5 )</td>
<td>-0.13</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.48</td>
</tr>
<tr>
<td>( \hat{\beta}_6 )</td>
<td>-0.39</td>
<td>-1.44</td>
<td>-1.44</td>
<td>-1.29</td>
</tr>
<tr>
<td>( F_o )</td>
<td>0.167</td>
<td>0.000</td>
<td>0.000</td>
<td>0.500</td>
</tr>
<tr>
<td>( F_n )</td>
<td>0.000</td>
<td>0.167</td>
<td>0.167</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Table 3. Estimated coefficients by CV: $n=200$, $p=500$.

<table>
<thead>
<tr>
<th>Penalty</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>SCAD($\gamma=2.01$)</th>
<th>SCAD($\gamma=3.7$)</th>
<th>MCP($\gamma=1.01$)</th>
<th>MCP($\gamma=3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_1$</td>
<td>0.16</td>
<td>0.09</td>
<td>0.00</td>
<td>0.07</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$\hat{\beta}_2$</td>
<td>0.26</td>
<td>0.60</td>
<td>1.41</td>
<td>1.38</td>
<td>1.40</td>
<td>1.41</td>
</tr>
<tr>
<td>$\hat{\beta}_3$</td>
<td>0.14</td>
<td>0.11</td>
<td>0.01</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$\hat{\beta}_4$</td>
<td>-0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$\hat{\beta}_5$</td>
<td>-0.18</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$\hat{\beta}_6$</td>
<td>-0.40</td>
<td>-1.49</td>
<td>-1.84</td>
<td>-1.83</td>
<td>-1.89</td>
<td>-1.82</td>
</tr>
<tr>
<td>$F_O$</td>
<td>0.142</td>
<td>0.000</td>
<td>0.018</td>
<td>0.022</td>
<td>0.000</td>
<td>0.014</td>
</tr>
<tr>
<td>$F_N$</td>
<td>0.000</td>
<td>0.333</td>
<td>0.500</td>
<td>0.333</td>
<td>0.667</td>
<td>0.667</td>
</tr>
</tbody>
</table>

Table 4. Estimated coefficients by BIC: $n=200$, $p=500$.

<table>
<thead>
<tr>
<th>Penalty</th>
<th>Lasso</th>
<th>SCAD($\gamma=2.01$)</th>
<th>SCAD($\gamma=3.7$)</th>
<th>MCP($\gamma=1.01$)</th>
<th>MCP($\gamma=3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_1$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.13</td>
<td>0.58</td>
<td>0.97</td>
</tr>
<tr>
<td>$\hat{\beta}_2$</td>
<td>0.12</td>
<td>1.40</td>
<td>1.09</td>
<td>0.62</td>
<td>0.002</td>
</tr>
<tr>
<td>$\hat{\beta}_3$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.16</td>
<td>0.83</td>
<td>0.87</td>
</tr>
<tr>
<td>$\hat{\beta}_4$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.36</td>
<td>0.00</td>
</tr>
<tr>
<td>$\hat{\beta}_5$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.40</td>
<td>-0.76</td>
</tr>
<tr>
<td>$\hat{\beta}_6$</td>
<td>-0.30</td>
<td>-1.85</td>
<td>-1.87</td>
<td>-1.40</td>
<td>-1.18</td>
</tr>
<tr>
<td>$F_O$</td>
<td>0.000</td>
<td>0.008</td>
<td>0.014</td>
<td>0.085</td>
<td>0.071</td>
</tr>
<tr>
<td>$F_N$</td>
<td>0.667</td>
<td>0.667</td>
<td>0.333</td>
<td>0.000</td>
<td>0.167</td>
</tr>
</tbody>
</table>

The variable selections and prediction results are summarized in Tables 1-4. The estimates were compared by Tables 1-2, when $n=200$ and $p=12$. From Tables 1-2, we can know the results are similar when the tuning parameter $\lambda$ is selected cross-validation or BIC, and in every mode, MCP has the best performance in view of model selection consistency. Tables 3-4 compare the estimates coefficients when $n=200$ and $p=500$. From Tables 3-4, we can see that, MCP when the tuning parameter $\lambda$ is selected by the BIC has the best performance in view of model selection consistency. Tables 1-4 also show that LASSO tends to shrink the coefficients toward zero and MCP results with adjustment are almost unbiased.

**Conclusion**

In this paper, basing on Stute’s weighted least squares method, the estimate procedures in partly linear regression model for right censored data has been investigated. The application of the AFT model has been achieved with the semi-parameter model. The theoretical study and the simulation show that the MCP approach and the sieve approach are very useful for variable selection, which are with great consistently.

Variable selection for high dimensional censored data is popular in biology, medicine, genetics and so on. There are so many work to be done to improve the accuracy of variable selection and expand the scope of the model application in the future.
References


