Empirical Likelihood for Censored Linear Regression

GENGSHENG QIN and BING-YI JING
Hong Kong University of Science and Technology

ABSTRACT. In this paper we investigate the empirical likelihood method in a linear regression model when the observations are subject to random censoring. An empirical likelihood ratio for the slope parameter vector is defined and it is shown that its limiting distribution is a weighted sum of independent chi-square distributions. This reduces to the empirical likelihood to the linear regression model first studied by Owen (1991) if there is no censoring present. Some simulation studies are presented to compare the empirical likelihood method with the normal approximation based method proposed in Lai et al. (1995). It was found that the empirical likelihood method performs much better than the normal approximation method.

Key words: censored linear regression, empirical likelihood, normal approximation.

1. Introduction

Consider the following linear regression model

$$Y_i = X_i^\beta + e_i, \quad i = 1, \ldots, n, \quad (1)$$

where $e_i$s are independent and identically distributed (i.i.d.) random variables with mean zero and finite variance, $\beta$ is a $p \times 1$ unknown parameter vector, and $X_i$s are $p \times 1$ observable random covariates or constant covariates. In some practical problems, $Y_i$s may be censored by some random variable $C_i$s and therefore are not completely observable. So instead of observing $(X_i, Y_i)$, one observes only $(X_i, Z_i, \delta_i)$, where

$$Z_i = \min(Y_i, C_i), \quad \delta_i = I\{Y_i \leq C_i\}, \quad i = 1, \ldots, n.$$ 

Here, we assume that the censoring variables $C_i$s are i.i.d. with a common distribution $G$ satisfying $\tau_G \geq \tau_{F_i}$ for each $i$, where $F_i$ denotes the distribution function of $Y_i$, $\tau_G = \inf\{t: G(t) = 1\}$, $\tau_{F_i} = \inf\{t: F_i(t) = 1\}$, and such that $\{C_i\}$s are independent of the sequence of independent random vectors $\{(X_i', Y_i)\}$

The censored linear regression model (1) with random design has received considerable attention in the statistical literature recently. There are two main trends in this body of literature: one trend is to extend the least squares method (LSE) in the complete data case to the incomplete data case. See, for instance, Buckley & James (1979), Koul et al. (1981), Zheng (1984), Leurgans (1987), Lai & Ying (1991), Zhou (1992), and Lai et al. (1995). The other trend is to extend robust estimators to incomplete data settings. See Tsiatis (1990), Lai & Ying (1992, 1994), Ying (1993), Ritov (1990) among others.

Our interest here lies in the estimation of the slope parameter vector $\beta$. To fix ideas, we shall restrict our attention here to the estimator proposed by Koul et al. (1981) although the results shown here can be easily extended to include the more general class of estimators such as those of Zheng (1984) and Leurgans (1987). Specifically, we define transformed data

$$Y_{iG} = \frac{\delta_i Z_i}{1 - G(Z_i)}, \quad i = 1, \ldots, n,$$

where $G(\cdot)$ is the common distribution of the censoring variable $C_1, \ldots, C_n$. Therefore,
(Xᵢ, YᵢG), i = 1, ..., n, form a sequence of random vectors with E(YᵢG|Xᵢ) = Xᵢβ. Hence, if G(⋅) is known, one can define the ordinary least squares estimate by

\[ \hat{\beta}_G = \left( \sum_{i=1}^{n} X_i X_i' \right)^{-1} \sum_{i=1}^{n} X_i Y_i G. \]  

(2)

When G(⋅) is unknown, Koul et al. (1981) proposed replacing G(⋅) in (2) by the Kaplan–Meier estimator \( \hat{G}_n(t) \) defined by

\[ 1 - \hat{G}_n(t) = \prod_{i=1}^{n} \left( \frac{n - i}{n - i + 1} \right)^{I\{Z(i) \leq t, \delta(i) = 0\}}. \]

Thus the Koul et al. estimator is given by

\[ \hat{\beta}_{Gn} = \left( \sum_{i=1}^{n} X_i X_i' \right)^{-1} \sum_{i=1}^{n} X_i Y_i \hat{G}_n, \]

where

\[ Y_i \hat{G}_n = \frac{\delta_i Z_i}{1 - \hat{G}_n(Z_i)}, \quad i = 1, \ldots, n. \]

Under appropriate conditions, it has been shown that (see, e.g. Lai et al., 1995)

\[ n^{1/2}(\hat{\beta}_{Gn} - \beta) \overset{D}{\rightarrow} N(0, A^{-1}VA^{-1}), \]

where \( A = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} X_i X_i' \) and \( V \) is given in theorem 1 in section 2. Then a large sample (1 − α)-level confidence region for \( \beta \) based on the above normal approximations is given by

\[ \{ \beta : n(\hat{\beta}_{Gn} - \beta)' \hat{A}^{-1} (\hat{\beta}_{Gn} - \beta) \leq \chi^2_p(\alpha) \}, \]

(3)

where \( \hat{A} = n^{-1} \sum_{i=1}^{n} X_i X_i' \), \( \hat{V} = \hat{V}_1 - \hat{V}_2 \) with \( \hat{V}_1 \) and \( \hat{V}_2 \) defined in (11) and (12) in section 2, and \( \chi^2_p(\alpha) \) is the (1 − α)th quantile of the chi-square distribution with degree of freedom \( p \).

However, some simulation studies we have conducted seem to indicate that the coverage probabilities obtained by (3) very often fall far short of the nominal level (1 − α). For more details, see section 3 in the present paper. One possible reason why (3) yields such poor performance in coverage probabilities is that the covariance matrix estimator \( \hat{V} \) of \( V \) is very unstable, due to the unstable behaviour of \( (1 - \hat{G}_n(t))^{-1} \) towards the right tail area. We have attempted to stabilize the covariance estimate by truncation. This will have some marginal success in improving the coverage probabilities. However, the results (not listed in this paper) are still far from satisfactory. Another possibility is to use Efron’s bootstrap method such as the percentile-t method. However, since the good performance of the percentile-t method also relies on choosing a stable covariance estimate, it is not difficult to imagine that the percentile-t would also suffer from the similar problems experienced by the normal approximation based method in the censored linear regression model. Finally, we should mention that the above problems are not unique to the Koul et al. estimator; other estimators such as those of Zheng (1984) and Leurgans (1987) will face similar problems as well.

To overcome the shortcomings of these earlier methods, we shall investigate the empirical likelihood method in constructing confidence regions for the slope parameter vector \( \beta \) in the censored linear regression model (1). The empirical likelihood method, introduced by Owen (1988), amounts to computing the profile likelihood of a general multinomial distribution supported on the data. Quite generally, the empirical likelihood has a number of advantages over its competitors such as the bootstrap. For example, it avoids the explicit Studentization as this is
done internally. So the empirical likelihood will be useful in cases where unstable covariance estimates might cause problems. Another advantage is that the shape of the confidence region (in two or more dimensions) is determined automatically by the data configuration. By comparison, the bootstrap requires the user to pre-specify the shape of the region in high dimensions. For these reasons, the empirical likelihood has found many applications such as in regression models (Owen, 1991; Chen, 1993, 1994) in generalized linear models (Kolaczyk, 1994), in quantile estimation (Chen & Hall, 1993), general estimating equation (Qin & Lawless, 1994), dependent processes (Kitamura, 1997). On the other hand, the applications to the censored data are relatively few, which include Thomas & Grunkemeir (1975) and Adimari (1997) among others.

In section 2, we shall introduce the empirical likelihood method to the censored linear regression model (1). There, we shall define an empirical likelihood ratio for the unknown parameter $\hat{\beta}$ and show that its limiting distribution is a weighted sum of $p$ independent chi-square distributions with 1 degree of freedom. This is different from the usual chi-square distributions with $p$ degrees of freedom when the censoring is absent. In section 3, some simulation studies are presented to compare the empirical likelihood method with the Studentized-$t$ method given in Lai et al. (1995) and also mentioned earlier in this section. Finally, proofs of our main results are given in section 4.

2. Methodology and main results

In section 1, we have seen that $E(Y_iG_jX_i) = X_i\beta$, from which we can get

$$E(X_i(Y_i - X_i\beta)) = 0,$$

$i = 1, \ldots, n.$

Therefore, the problem of testing whether $\beta_0$ is the true parameter of $\beta$ is equivalent to testing whether $E(X_i(Y_iG_j - X_i\beta_0)) = 0$, $i = 1, \ldots, n$. This can be done by using Owen’s empirical likelihood method (1991). Let $p = (p_1, \ldots, p_n)$ be a probability vector, i.e. $\sum_{i=1}^n p_i = 1$ and $p_i \geq 0$ for all $i$. For $1 \leq i \leq n$, we define

$$W_i = X_i(Y_i - X_i\beta_0),$$

$$W_{ni} = X_i(Y_{ni}G_n - X_i\beta_0).$$

Then, the empirical likelihood, evaluated at true parameter value $\beta_0$, is defined by

$$L(\beta_0) = \sup\left\{ \prod_{i=1}^n p_i: \sum p_i = 1, \sum_{i=1}^n p_iW_i = 0 \right\}.$$

Since $W_i$s depend on unknown $G()$, we can replace them by $W_{ni}$s. Therefore, an estimated empirical likelihood, evaluated at the true value $\beta_0$ of $\beta$, is defined by

$$L(\beta_0) = \sup\left\{ \prod_{i=1}^n p_i: \sum p_i = 1, \sum_{i=1}^n p_iW_{ni} = 0 \right\}.$$

Then, by the Lagrange multiplier technique, we can easily get

$$p_i = \frac{1}{n}\left\{1 + \lambda W_{ni}\right\}^{-1}, \quad i = 1, \ldots, n,$$

where $\lambda = (\lambda_1, \ldots, \lambda_p)'$ is the solution of

$$\frac{1}{n} \sum_{i=1}^n \frac{W_{ni}}{1 + \lambda W_{ni}} = 0.$$
Note that $\prod_{i=1}^{n} p_i$, subject to $\sum_{i=1}^{n} p_i = 1$, attains its maximum $n^{-n}$ at $p_i = n^{-1}$. So we define the empirical likelihood ratio at $\beta_0$ by

$$R(\beta_0) = \prod_{i=1}^{n} (np_i) = \prod_{i=1}^{n} \{1 + \lambda \mathbf{W}_m\}^{-1},$$

and the corresponding empirical log-likelihood ratio is defined as

$$l(\beta_0) = -2 \log R = 2 \sum_{i=1}^{n} \log \{1 + \lambda \mathbf{W}_m\},$$

where $\lambda$ is the solution of (4).

The above definition of the empirical log-likelihood ratio $l(\beta_0)$, defined by (4) and (5), looks rather similar to the uncensored case with one essential difference. That is, $\mathbf{W}_m$ are not independent random variables, which makes the limiting behaviour of $l(\beta_0)$ slightly different from other cases previously studied; see theorem 1 below.

Let $B$ be any Borel subset of $\mathbb{R}^d$ and define

$$F_n(y) = n^{-1} \sum_{i=1}^{n} P(Y_i \leq y), \quad \mu_n(B, y) = n^{-1} \sum_{i=1}^{n} P(X_i \in B, Y_i \leq y),$$

$$\Gamma_0(t) = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} P(Y_i \geq t), \quad \Gamma_1(t) = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} E(X_i I\{Y_i > t\}).$$

The following conditions will be needed in this paper.

**C1.** $\Gamma_1(t) < \infty$ for every $t$.

**C2.** For all $s \leq \tau = \inf\{t: \Gamma_0(t) = 0\}$, $\Delta G(s) \cdot \Delta \Gamma_i(t) = 0$ for $i = 0, 1$, where we define $\Delta g(s) = g(s) - g(s^-)$ for any function $g(\cdot)$, and

$$\int_{-\infty}^{\tau} \left( \int_{t > y} \left( \frac{\Gamma_0(y) dG(y)}{(1 - G(y))\Gamma_0(y^-)} \right)^2 \right) \leq \infty.$$

**C3.** Letting $\tau_n = \inf\{t: F_n(t) = 1\}$, there exists $y(\epsilon) < \tau$ for every $\epsilon > 0$ such that for all large $n$,

$$\int_{x \in \mathbb{R}^d} \int_{y = y(\epsilon)}^{\tau_n} ||x|| \left| \frac{d\mu_n(x, y)}{(1 - G(y))(1 - F_n(y))} \right| < \epsilon.$$

**C4.**

$$\lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} E||X_i (Y_i \beta - X_i' \beta)||^2 I\{|X_i (Y_i \beta - X_i' \beta)|^2 \geq \epsilon n\} = 0 \text{ for every } \epsilon > 0.$$

**C5.**

$$\lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} E[(Y_i \beta - X_i' \beta)^2 X_i X_i'] \text{ is a positive definite matrix}.$$

**C6.**

$$\lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} X_i X_i' = A \text{ in probability for some positive definite matrix } A.$$
C7. 
\[ \max_i \|X_i\|^2 = o(n^{1/2}), \quad \text{and} \quad \max_i \|X_i Y_i\| = o(n^{1/2}), \quad \text{a.s.} \]

C8. 
\[ E\|X_i(Y_{iG} - X_i\beta)\|^4 < \infty, \quad \text{for each } i. \]

We note that C1–C6 are the conditions used in Lai et al. (1995) in deriving the central limit theorem for \( \hat{\beta}_{\mathcal{G}_n} \). For simplicity, denote \( x^{\otimes 2} = xx' \) for \( x \in \mathbb{R}^p \). Define

\[ V_1 = \lim_{n \to \infty} n^{-1} \sum_{i=1}^n E[(Y_{iG} - X_i\beta)^2 X_i X_i'], \]

\[ V_2 = \int_{-\infty}^{t} \left( \int_{t > y} \Gamma_1(t) \, dt \right)^2 \frac{\Gamma_0(y) dG(y)}{\Gamma_0^2(\beta - \beta_0)} \frac{\Gamma_0(y -) [1 - G(y)][1 - G(y -)]}{\Gamma_0^2(\beta - \beta_0)[1 - G(y)][1 - G(y -)]}. \]

We now state the main theorem.

**Theorem 1**

Assume the conditions C1–C8 hold. If \( \beta_0 \) is the true value of \( \beta \), then \( l(\beta_0) \) has a weighted sum of independent chi-square distributions with 1 degree of freedom as its limiting distribution, that is,

\[ l(\hat{\beta}_0) \rightsquigarrow l_1 \chi^2_{1,1} + \cdots + l_p \chi^2_{p,1}, \]

where the weights \( l_i, 1 \leq i \leq p \), are the eigenvalues of \( V_1^{-1} V \) with \( V = V_1 - V_2 \), and \( \chi^2_{1,1} \) for \( 1 \leq i \leq p \) are independent chi-square distributions with one degree of freedom.

In order to apply theorem 1, one has to first estimate the weights \( l_i, 1 \leq i \leq p \). Toward that end, define \( Y_n(s) = \sum_{i=1}^n I\{Z_i \geq s\}, \quad N_n(s) = \sum_{i=1}^n I\{Z_i \leq s, \delta_i = 0\} \) and \( \Delta N_n(s) = N_n(s) - N_n(s-). \) Then under the assumptions of theorem 1, the limiting covariance matrix \( V_1, V_2 \) (hence \( V \)) can be consistently estimated by

\[ \hat{V}_1 = n^{-1} \sum_{i=1}^n (Y_{iG_n} - \hat{\beta}'X_i)^2 X_i X_i', \]

\[ \hat{V}_2 = n^{-1} \sum_{i=1}^n \left( \sum_{j=1}^n X_i Y_{iG_n} I\{Z_i > Z_j\} \right)^2 \left( \sum_{i=1}^n I\{Z_i > Z_j\} \right) \frac{\Delta N_n(Z_j)}{Y_n(Z_j)}. \]

See also rem. (ii) of Lai et al. (1995). Then \( l_i, 1 \leq i \leq p \) can be consistently estimated by the eigenvalues \( \hat{l}_i \) of \( \hat{V}_1^{-1} \hat{V} \) with \( \hat{V} = \hat{V}_1 - \hat{V}_2 \).

Confidence regions for \( \beta \) can be constructed as follows. Let

\[ R_\alpha(\beta) = \{ \beta : l(\beta) \leq c_\alpha \}, \]

where \( c_\alpha \) is the \((1 - \alpha)\)th quantile of the weighted chi-square distribution \( \hat{l}_1 \chi^2_{1,1} + \cdots + \hat{l}_p \chi^2_{p,1} \). Then from the earlier discussion, \( R_\alpha(\beta) \) gives an approximate confidence region for \( \beta \) with asymptotically correct coverage probability \( 1 - \alpha \), i.e.

\[ P(\beta_0 \in R_\alpha(\beta)) = 1 - \alpha + o(1). \]

Before we end this section, we remark that when there is no censoring in the observations (i.e.
3. Numerical studies

In this section we shall conduct some simulation studies to compare the performances of empirical likelihood method and the normal approximation based method proposed by Lai et al. (1995).

In our first example, we consider the following 2-dimensional linear regression model:

Model A: \( Y_i = 5 + 10x_i + \epsilon_i, \quad i = 1, \ldots, n, \)

where \( x_i \)'s are drawn from the uniform distribution \( U[0, 2] \) and \( \epsilon_i \)'s are generated from the distribution \( \chi^2_{20} - 20 \). We can calculate the simulated values for \( Y_i \)'s from the above model. On the other hand, the censoring times \( C_i \)'s are generated from the exponential distribution with rate \( 0.012 \) (the censoring proportion is about 16%). The sample size \( n \) has been chosen to be 30, 60, 100 and 200, respectively. Finally, simulated observations from the above censored linear model are \( (X'_i, Z_i, \delta_i) \) for \( i = 1, \ldots, n \), where

\[
X'_i = (1, x_i), \quad Z_i = \min(Y_i, C_i), \quad \delta_i = I\{Y_i \leq C_i\}.
\]

We can repeat the above process \( M = 2000 \) times to generate \( M \) sets of data. Then the approximate coverage probabilities for the empirical likelihood and normal approximation methods based on these \( M \) simulated data sets are simply the proportions of these data sets which satisfy the inequalities (8) and (3), respectively. The nominal confidence level \( \alpha \) has been taken to be 0.90, 0.95 and 0.99, respectively. Because the denominator \( Y_n(Z_{(n)} - \Delta N_n(Z_{(n)}) \) in the definition of \( \hat{V} \) can be zero when \( i = n \), we add 0.5 to \( Y_n(Z_{(n)} - \Delta N_n(Z_{(n)}) \) if it is zero.

The results of the simulations are presented in Table 1.

In our second example, we consider a 3-dimensional regression model:

Model B: \( Y_i = 2 + 5x_{1i} + x_{2i} + \epsilon_i \)

where \( x_{1i} \)'s are drawn from the uniform distribution \( U[0, 2] \), \( x_{2i} \)'s are drawn from the standard

<table>
<thead>
<tr>
<th>Nominal levels</th>
<th>n</th>
<th>Normal approximation</th>
<th>Empirical likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90</td>
<td>30</td>
<td>0.836</td>
<td>0.848</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.838</td>
<td>0.871</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.862</td>
<td>0.891</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.880</td>
<td>0.895</td>
</tr>
<tr>
<td>0.95</td>
<td>30</td>
<td>0.892</td>
<td>0.905</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.890</td>
<td>0.932</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.923</td>
<td>0.941</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.938</td>
<td>0.945</td>
</tr>
<tr>
<td>0.99</td>
<td>30</td>
<td>0.949</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.946</td>
<td>0.979</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.971</td>
<td>0.988</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.980</td>
<td>0.994</td>
</tr>
</tbody>
</table>
logistic distribution, and $e_i$s are generated from the distribution $\chi^2_{10} - 10$. Here the censoring times $C_i$s are generated from the exponential distribution with rate 0.03 (the censoring proportion is about 18%). The other quantities such as sample size, nominal confidence level and the number $M$ are chosen to be the same as those in the first example. The results are presented in Table 2.

We make the following observations from Table 1 and Table 2.

(1) As expected, at each nominal level and for both the empirical likelihood and normal approximation methods, the coverage accuracies for $\beta$ in both models increase as the sample size $n$ increases.

(2) The empirical likelihood outperforms the normal approximation methods in both models.

Particularly, the empirical likelihood performs much better than the normal approximation method in high dimension case.

We also investigate the performances of empirical likelihood method and the normal approximation based method in fixed sample size ($n = 100$) but with different censoring proportions being controlled by choosing different censoring times (Here the censoring times are generated from the exponential distributions with different rates). The results are reported in Table 3.

Table 3 also indicates that

(1) The empirical likelihood outperforms the normal approximation methods in both models. Particularly, the empirical likelihood performs much better than the normal approximation method in high censoring case.

(2) At each nominal level and for both the empirical likelihood and normal approximation methods, the coverage accuracies for $\beta$ in both models decrease as the censoring proportions increase. However, the coverage accuracies for the empirical likelihood are acceptable even in high censoring case.

Overall, our simulation studies indicate that the empirical likelihood method is superior to the normal approximation based method in the context of the censored linear regression model, particularly in high dimensional models or high censoring proportions.

### Table 2. Comparisons of empirical likelihood and normal approximation. Coverage probabilities for $\beta$. Model B: $p = 3$

<table>
<thead>
<tr>
<th>Nominal levels</th>
<th>$n$</th>
<th>Normal approximation</th>
<th>Empirical likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90</td>
<td>30</td>
<td>0.619</td>
<td>0.768</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.665</td>
<td>0.816</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.697</td>
<td>0.864</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.739</td>
<td>0.876</td>
</tr>
<tr>
<td>0.95</td>
<td>30</td>
<td>0.708</td>
<td>0.851</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.745</td>
<td>0.884</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.801</td>
<td>0.923</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.839</td>
<td>0.932</td>
</tr>
<tr>
<td>0.99</td>
<td>30</td>
<td>0.829</td>
<td>0.942</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.873</td>
<td>0.965</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.909</td>
<td>0.978</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.939</td>
<td>0.987</td>
</tr>
</tbody>
</table>
4. Proof of Theorem 2.1

We shall introduce a few lemmas needed in proving the main theorem.

**Lemma 1** (See Lai et al., 1995.)

Under the same conditions as in theorem 1, we have

\[(i) \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_{ni} \xrightarrow{D} N(0, V), \quad (ii) \quad \hat{\beta} - \beta_0 = O_p(n^{-1/2}).\]

**Lemma 2**

Under the same conditions as in theorem 1, we have

\[(i) \quad \frac{1}{n} \sum_{i=1}^{n} W_{ni}W_{ni}' \xrightarrow{p} V_1, \quad (ii) \quad \tilde{\beta}_1 \xrightarrow{p} V_1, \quad (iii) \quad \tilde{V} \xrightarrow{p} V.\]

**Proof of Lemma 2.** Let

\[\hat{S}_n = \frac{1}{n} \sum_{i=1}^{n} W_{ni}W_{ni}', \quad S = \frac{1}{n} \sum_{i=1}^{n} W_iW_i'.\]

By **C4**, **C5** and the Lindeberg’s central limit theorem, \(S \xrightarrow{p} V_1\). In order to prove (i), we only need to prove \(\hat{S}_n = S + O_p(1)\). For any \(a \in \mathbb{R}^p\), we have the following decompositions:
\[ a'(\hat{S}_n - S)a = \frac{1}{n} \sum_{i=1}^{n} (a'(W_{ni} - W_i))^2 + \frac{2}{n} \sum_{i=1}^{n} (a'W_i)(a'(W_{ni} - W_i)) \]

\[ \leq \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} |a'(W_{ni} - W_i)| \right) \left( \frac{1}{\sqrt{n}} \max_{i} |a'(W_{ni} - W_i)| + \frac{2}{\sqrt{n}} \max_{i} |a'W_i| \right) \]

\[ \equiv J_0(J_1 + 2J_2). \quad (9) \]

From C7 and the following result due to Srinivasan & Zhou (1991)

\[ \sup_{r \leq \max_i Z_i} \left| \frac{G_n(t) - G(t)}{1 - G_n(t)} \right| = o_p(1), \quad (10) \]

we get

\[ J_1 \leq n^{-1/2}||a|| \max_i \|X, Y_{iG}\| \sup_{r \leq \max_i Z_i} \left| \frac{G_n(t) - G(t)}{1 - G_n(t)} \right| = o_p(1), \]

\[ J_2 \leq n^{-1/2}||a|| \max_i \|X, Y_{iG}\| + \|\beta_0\| \max_i \|X_i\|^2 = o_p(1). \]

Now let us look at the term \( J_0 \). From Gill (1980), we have the following martingale representation

\[ \frac{G_n(t) - G(t)}{1 - G(t)} = \int_{-\infty}^{t} D_n(t) Y_{iG}^{-1}(t) dM(t), \quad (11) \]

where

\[ M(t) = I\{Z_i \leq s, \delta_i = 0\} - \int_{-\infty}^{s} I\{C_i \geq s, Y_i > s\}(1 - G(s-))^{-1} dG(s), \]

\[ M(s) = \sum_{i=1}^{n} M_i(s), \quad D_n(s) = \frac{1 - G_n(s-)}{1 - G(s)}. \]

Also we have

\[ |Y_{iG} - Y_{iG}| = \left| Y_{iG} \left( \left( \frac{G_n(Z_i) - G(Z_i)}{1 - G(Z_i)} \right)^{-1} - 1 \right) \right| \]

\[ = \left| Y_{iG} \left( \left( G_n(Z_i) - G(Z_i) \right) \left( \frac{G_n(Z_i) - G(Z_i)}{1 - G_n(Z_i)} \right)^2 \right) \right| \]

\[ \leq \left( 1 + \sup_{r \leq \max_i Z_i} \left| \frac{G_n(t) - G(t)}{1 - G_n(t)} \right| \right) \left| \frac{G_n(Z_i) - G(Z_i)}{1 - G(Z_i)} \right| |Y_{iG}|. \quad (12) \]

It follows from (10)–(12) that, for any fixed \( u < \tau \),
\[ J_0 \leq Cn^{-1/2} \| a \| \sum_{i=1}^{n} \| X_i Y_{iG} \| \int_{-\infty}^{Z_i} D_n(t) Y_n^{-1}(t) \, dM(t) \]
\[ \leq Cn^{-1/2} \| a \| \sum_{j} \int_{-\infty}^{Z_j} \sum_{i=1}^{n} \| X_i Y_{iG} I \{ Z_i > t \} \| D_n(t) Y_n^{-1}(t) \, dM_j(t) \]
\[ + Cn^{-1/2} \| a \| \sum_{i=1}^{n} \max \| X_i Y_{iG} \| \int_{u}^{Z_i} D_n(t) Y_n^{-1}(t) \, dM(t) \]
\[ \equiv J_{o1} + J_{o2}, \]

where \( C \) is a generic constant. By a similar argument to the proof of (2.28)–(2.29) in Lai et al. (1995), we can easily show that \( J_{o1} = O_p(1) \) and \( J_{o2} = o_p(1) \). So we have shown \( a^*(\hat{\Delta}_n - S)a = o_p(1) \), hence \( \hat{\Delta}_n = S + o_p(1) \). Lemma 2(i) is proved.

To prove lemma 2(ii), we only need to show \( \hat{\Delta}_n = \hat{V}_1 + o_p(1) \) in the light of lemma 2(i). For any \( a \in \mathbb{R}^p \), we have the following decompositions:
\[ a^*(\hat{\Delta}_n - \hat{V}_1)a = \frac{1}{n} \sum_{i=1}^{n} ((a'X_i)^2(Y_{iGn} - X_i\hat{\beta})^2 - (a'W_n)^2) \]
\[ = \frac{2}{n} \sum_{i=1}^{n} (a'X_i)^2(X_i(\beta_0 - \hat{\beta}))(Y_{iGn} - Y_{iG}) + \frac{2}{n} \sum_{i=1}^{n} (a'X_i)^2(X_i(\beta_0 - \hat{\beta}))Y_{iG} \]
\[ + \frac{1}{n} \sum_{i=1}^{n} (a'X_i)^2(X_i(\hat{\beta} - \beta_0))^2 + \frac{1}{n} \sum_{i=1}^{n} (a'X_i)^2(X_i(\hat{\beta} - \beta_0))(X_i\beta_0) \]
\[ \equiv D_1 + D_2 + D_3 + D_4, \quad \text{say}. \]

From C7, (12) and lemma 1, we get
\[ |D_1| \leq 2 \max_i \| X_i Y_{iG} \| \| \beta_0 - \hat{\beta} \| \sup_{t: \max Z_i} \left| \frac{G_n(t) - G(t)}{1 - G_n(t)} \right| \cdot \frac{1}{n} \sum_{i=1}^{n} (a'X_i)^2 = o_p(1), \]
\[ |D_2| \leq 2 \max_i \| X_i Y_{iG} \| \| \beta_0 - \hat{\beta} \| \cdot \frac{1}{n} \sum_{i=1}^{n} (a'X_i)^2 = o_p(1), \]
\[ |D_3| \leq \max_i \| X_i \|^2 \| \beta_0 - \hat{\beta} \| \cdot \frac{1}{n} \sum_{i=1}^{n} (a'X_i)^2 = o_p(1), \]
\[ |D_4| \leq \max_i \| X_i \|^2 \| \beta_0 - \hat{\beta} \| \| \beta_0 \| \cdot \frac{1}{n} \sum_{i=1}^{n} (a'X_i)^2 = o_p(1). \]

Therefore \( a^*(\hat{\Delta}_n - \hat{V}_1)a = o_p(1) \). Hence \( \hat{\Delta}_n = \hat{V}_1 + o_p(1) \) and lemma 2(ii) is proved. Lemma 2(iii) follows from Lai et al. (1995). The proof of lemma 2 is thus complete.

**Lemma 3**
Let \( Z \sim N(0, I_p) \), where \( I_p \) is the \( p \times p \) identity matrix. Let \( U \) be a \( p \times p \) non-negative definite matrix with eigenvalues \( l_1, \ldots, l_p \). Then,
\[ Z'UZ \overset{d}{=} l_1 \chi^2_{1,1} + \cdots + l_p \chi^2_{p,1}, \]
where \( \chi^2_{1,1} \) are as defined in theorem 1.
Proof of lemma 3. By the assumption, there exists an orthonormal matrix $P$ such that $U = P^TDP$, where $D = \text{diag}(l_1, \ldots, l_p)$ is a diagonal matrix with diagonal elements $l_1, \ldots, l_p$. Let $\tilde{Z} = P^T\mathbf{z} = (\tilde{Z}_1, \ldots, \tilde{Z}_p)'$. Clearly, $\tilde{Z} \overset{d}{\sim} N(0, I_p)$. Therefore,

$$Z^T U Z = (P \tilde{Z})' D (P \tilde{Z}) = l_1 \tilde{Z}_1^2 + \cdots + l_p \tilde{Z}_p^2,$$

where $\tilde{Z}_1, \ldots, \tilde{Z}_p$ are i.i.d. random variables with the common limiting chi-square distribution with 1 degree of freedom. This completes the proof.

Proof of theorem 1. Applying Taylor’s expansion to (5), we have

$$l(\beta_0) = 2 \sum_{i=1}^n \log \{1 + \lambda' W_{ni}\} = 2 \sum_{i=1}^n \left( \lambda' W_{ni} - \frac{1}{2} (\lambda' W_{ni})^2 \right) + r_n, \quad (13)$$

with

$$|r_n| = C \sum_{i=1}^n (\lambda' W_{ni})^3 \text{ in probability.}$$

Write $\lambda = \rho \theta$ where $\rho \geq 0$ and $\|\theta\| = 1$. From the proof of lemma 2(i), we can get

$$\theta' S\theta = \theta' S\theta + o_p(1).$$

By C5, C7 and C8,

$$n^{-1} \sum_{i=1}^n E[\|X_i(Y_iG - X_i\beta_0)\|^4]$$

$$\leq 2n^{-1} \sum_{i=1}^n E((\max_i \|X_iY_iG\|^2 + \max_i \|X_i\|^2 \|\beta_0\|)^2) ||X_i(Y_iG - X_i\beta_0)\|^2)$$

$$\leq o(n^{-1}) \cdot \sum_{i=1}^n E(\|X_i(Y_iG - X_i\beta_0)\|^2)$$

$$= o(1).$$

Then, applying the same arguments used in Owen (1991), we can prove

$$\|\tilde{\lambda}\| = O_p(n^{-1/2}). \quad (14)$$

From C7, it follows that

$$\max_{1 \leq i \leq n} \|W_{ni}\| \leq \max_{1 \leq i} \|X_i(Y_iG - Y_iG_i)\| + \max_{1 \leq i} \|X_i(Y_iG - X_i\beta_0)\|$$

$$\leq \max_{1 \leq i} \|X_iY_iG\| \left( 1 + \sup_{t \leq \max_i Z_i} \left| \frac{G_n(t) - G(t)}{1 - G_n(t)} \right| \right) + \max_{1 \leq i} \|X_i\|^2 \beta_0$$

$$= o_p(n^{1/2}). \quad (15)$$

Hence, (14), (15) and lemma 2(i) together give

$$|r_n| \leq C \|\tilde{\lambda}\|^3 \max_{1 \leq i \leq n} \|W_{ni}\| \sum_{i=1}^n \|W_{ni}\|^2 = o_p(1). \quad (16)$$

Note that
From (18) and (19), we get

\[
\sum_{i=1}^{n} \frac{W_{ni}}{1 + \lambda W_{ni}} = \sum_{i=1}^{n} W_{mi} \left[ 1 - \lambda W_{ni} + \frac{(\lambda W_{ni})^2}{1 + \lambda W_{ni}} \right]
\]

\[
= \sum_{i=1}^{n} W_{mi} - \left( \sum_{i=1}^{n} W_{mi} W_{mi}' \right) \lambda + \sum_{i=1}^{n} W_{ni}(\lambda W_{ni})^2.
\]

Again by (4), we get that

\[
\lambda = \left( \sum_{i=1}^{n} W_{mi} W_{mi}' \right)^{-1} \sum_{i=1}^{n} W_{mi} + \left( n^{-1} \sum_{i=1}^{n} W_{mi} W_{mi}' \right)^{-1} \left( n^{-1} \sum_{i=1}^{n} W_{ni}(\lambda W_{ni})^2 \right)
\]

\[
= \left( \sum_{i=1}^{n} W_{mi} W_{mi}' \right)^{-1} \sum_{i=1}^{n} W_{mi} + O_p \left( \frac{\max_{1 \leq i \leq n} ||W_{ni}|| n^{-1} \sum_{i=1}^{n} (\lambda W_{ni})^2} {1 + \lambda W_{ni}} \right)
\]

\[
= \left( \sum_{i=1}^{n} W_{mi} W_{mi}' \right)^{-1} \sum_{i=1}^{n} W_{mi} + O_p(n^{-1/2}). \tag{17}
\]

By (14), (15) and lemma 2(i),

\[
\sum_{i=1}^{n} \frac{\lambda W_{mi}}{1 + \lambda W_{ni}} = o_p(1). \tag{19}
\]

From (18) and (19), we get

\[
\sum_{i=1}^{n} \lambda W_{mi} = \sum_{i=1}^{n} (\lambda W_{ni})^2 + o_p(1). \tag{20}
\]

By (13), (16), (17) and (20) and lemma 2(i), we get

\[
L(\beta_0) = \sum_{i=1}^{n} \lambda W_{mi} W_{mi}' \lambda + o_p(1)
\]

\[
= \left( n^{-1/2} \sum_{i=1}^{n} W_{mi} \right)^t \left( n^{-1} \sum_{i=1}^{n} W_{mi} W_{mi}' \right)^{-1} \left( n^{-1/2} \sum_{i=1}^{n} W_{mi} \right) + o_p(1)
\]

\[
= \left( n^{-1/2} \sum_{i=1}^{n} W_{mi} \right)^t V_1^{-1} \left( n^{-1/2} \sum_{i=1}^{n} W_{mi} \right) + o_p(1)
\]

\[
= \left( V^{-1/2} n^{-1/2} \sum_{i=1}^{n} W_{mi} \right)^t \left( V^{1/2} V^{-1} V^{1/2} \right) \left( V^{-1/2} n^{-1/2} \sum_{i=1}^{n} W_{mi} \right) + o_p(1).
\]

By lemma 1, we have \( V^{-1/2} n^{-1/2} \sum_{i=1}^{n} W_{mi} \rightarrow N(0, I_p) \). Also note that \( V^{1/2} V^{-1} V^{1/2} \) and \( V^{-1} V \) have the same eigenvalues. Then theorem 1 follows from lemma 3 straightforward.
References


Received March 1999, in final form January 2001

Bing-Yi Jing, Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.

E-mail: majing@ust.hk.