STATISTICAL DIAGNOSTIC FOR PARTIALLY LINEAR VARYING COEFFICIENT MODEL WITH RANDOM RIGHT CENSORSHIP BASED ON EMPIRICAL LIKELIHOOD METHOD

WANG SHULING, LIAO DAQING, LIU MAN

Abstract. In this paper, the empirical likelihood method to study the statistical diagnostic for partially linear varying coefficient model with random right censorship. First the primary model is transformed to partially linear varying coefficient model; then the parameter estimation equation based on the experience of application of likelihood methods to estimate. Experience based on the deletion model proposed experience like natural Cook distance, likelihood distance, then find outliers and strong influence point; At last, an example is given to illustrate our results.

I. INTRODUCTION

Nearly thirty years, the diagnosis and influence analysis of linear regression model has been fully developed (R.D.Cook and S.Weisberg 1982, Bo-cheng Wei, Guo-bin Lu & Jian-qing Shi 1990); the varying-coefficient model is a useful extension of classical linear model. Regarding the varying coefficient model, especially for the B-spline estimation of parameter, diagnosis and influence analysis have some results (Cai Z,Fan J,Li R (2000),Fan J,Zhang W(2008)). However, all the above results are obtained under the uncensored case. In many applications, some of the responses and/or covariates may not be observed, but are censored. For censored data,

2010 Mathematics Subject Classification. 62G08
Key words and phrases. random right censorship; Kaplan-Meier product-limit estimator; empirical likelihood; outliers; influence analysis

©2013 Authors retain the copyrights of their papers, and all open Access articles are distributed under the terms of the Creative Commons Attribution License.
the usual statistical techniques for complete data situations are not readily applicable.


So far the diagnosis of partially linear varying coefficient model with random right censorship based on empirical likelihood method has not yet seen in the literature, this paper attempts to study it.

The rest of the paper is organized as follows. The primary model is transformed into partially linear varying coefficient model in Section 2. Empirical likelihood and estimation equation are presented in Section 3. The main results are given in Section 4 and Section 5. An example is given to illustrate our results in Section 6.

II. THE TRANSFORMATION OF MODEL

Let $Y$ be the response variable and $(X^T, Z^T, T)$ be its associated covariates. The partially linear varying coefficient regression model assumes the following structure:

$$Y = X^T \beta + Z^T a(T) + \varepsilon$$  \hspace{1cm} (1)

Where $X = (X_1, \ldots, X_n)^T$ is of dimension $n \times 1$ and $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p$-dimensional vector of unknown coefficient functions. $a(T) = (a_1(T), \ldots, a_q(T))^T$ is unknown function vector. $T$ is one-dimensional vector. $\varepsilon$ is a stochastic error with $E(\varepsilon | X^T, Z^T, T) = 0, \text{Var}(\varepsilon | X^T, Z^T, T) = \sigma^2$.

Consider the model (1), where $Y$ is the survival time. Let $C$ be the censoring time associated with the survival time $Y$. Assume that $Y$ and $C$ are conditionally independent given the associate covariates $(X^T, Z^T, T)$. Denote $\Delta = \min(Y, C)$ and $\delta = I(Y \leq C)$, where $I(\cdot)$ is the index function. The observations are $\{(x_i^T, z_i^T, t_i, \Delta_i, \delta_i)\}$ which are random samples
from \((X', Z', T, \Delta, \delta)\), where \(x_k' = (x_{k1}, \cdots, x_{kp})\). Thus instead of observing \(Y_i\), we observe the pairs \((\Delta_i, \delta_i)\), where \(\Delta_i = \min(Y_i, C_i)\) and \(\delta_i = I(Y_i \leq C_i)\). Observations on \(\Delta_i\) for which \(\delta_i = 1\) are uncensored, and observations on \(\Delta_i\) for which \(\delta_i = 0\) are censored. Model (5) is called partially linear varying coefficient regression model with random right censorship right now. Let \(F_i\) is the distribution function of \(Y_i\), \(G\) is the common distribution function of \(C_i\), and \(r_i = \inf(t : F_i(t) = 1)\). Note that \(\bar{F}_i = 1 - F_i\) and \(\bar{G} = 1 - G\).

**Lemma** \(E \delta_i \Delta_i \bar{G}^{-1}(\Delta_i) = \sum_{k=1}^p \beta_k x_{ik} + \sum_{k=1}^q z_{ik} a(t_i), \quad i = 1, 2, \cdots, n\).

**Proof.** Since

\[
E \delta_i \Delta_i \bar{G}^{-1}(\Delta_i) = \int_0^{r_i} \int_{y_i}^{r_i} \frac{y}{1 - G(y)} dG(t_i) dF_i(y)
\]

and

\[
EY_i = \sum_{k=1}^p \beta_k x_{ik} + \sum_{k=1}^q z_{ik} a(t_i)
\]

thus \(E \delta_i \Delta_i \bar{G}^{-1}(\Delta_i) = \sum_{k=1}^p \beta_k x_{ik} + \sum_{k=1}^q z_{ik} a(t_i), \quad i = 1, 2, \cdots, n\).

Now we consider \((\delta_i, \Delta_i, \bar{G}^{-1}(\Delta_i), 1 \leq i \leq n)\) follow the model

\[
\frac{\delta_i \Delta_i}{G(\Delta_i)} = \sum_{k=1}^p \beta_k x_{ik} + \sum_{k=1}^q z_{ik} a(t_i) + \epsilon_i' \quad i = 1, 2, \cdots, n
\]

where \(\epsilon_i'\) is i.i.d. and \(E \epsilon_i' = 0\), \(Var(\epsilon_i') = \sigma^{-2}\). In practice, we replace \(\bar{G}\) with \(\hat{G}\) which is the Kaplan-Meier product-limite estimator of \(\bar{G}\). (Wang qihua [9]). The expression of \(\hat{G}\) is given as follows:

\[
\hat{G}(t) = \begin{cases} \prod_{j=1}^n \left( \frac{1 + N^{-}(\Delta_j)}{2 + N^{-}(\Delta_j)} \right) & , \text{ if } t \leq \Delta_{(n)} \\ 0, & \text{ if } t > \Delta_{(n)} \end{cases}
\]
where \( \Delta_{(\alpha)} = \max\{\Delta_1, \Delta_2, \ldots, \Delta_n\} \), \( \Delta_{(j)} = \sum_{i=1}^{n} I[\Delta_i \geq \Delta_j], j = 1, 2, \ldots, n \).

Let \( L_i = \frac{\delta_i \Delta_i}{G(\Delta_i)} \), model (1) is transformed to following partially linear varying coefficient regression model

\[
L_i = x_i^T \beta + z_i^T a(t_i) + \varepsilon_i, \quad i = 1, \ldots, n \quad (4)
\]

III. EMPIRICAL LIKELIHOOD AND ESTIMATION EQUATION

For partially linear varying coefficient regression model (4), Fan J Q, Huang T (2005) has proposed the empirical likelihood ratio function for \( \beta \) can be defined by

\[
L(\beta) = \sup \left\{ \prod_{i=1}^{n} p_i \left| \sum_{i=1}^{n} p_i \hat{\beta}_i(\beta) = 0, p_i \geq 0 \right\},
\]

where

\[
\hat{\beta}_i(\beta) = \left( x_i - \hat{\mu}(t_i) \right) \left[ L_i - x_i^T \beta - z_i^T \hat{a}(t_i) \right], \hat{\mu}(t) = \left[ \hat{S}(t) \right]^{-1} \hat{G}(t), \hat{S}(t) = \sum_{i=1}^{n} W_i(t) x_i x_i^T,
\]

\[
\hat{G}(t) = \sum_{i=1}^{n} W_i(t) x_i y_i^T, \hat{\mu}(t) = \sum_{i=1}^{n} W_i(t) x_i, W_i(t) = K_h(t-t_i) / \sum_{i=1}^{n} K_h(t-t_i), K_h(\cdot) = K(\cdot/h),
\]

\( K(\cdot) \) is the kernel function, \( h \) is the bandwidth. \( \omega(\cdot) \) is index function \([a, b], \quad 0 < a < b < 1 \).

By Qin and Lawless[11], Owen[6], when

\[
p_i = \frac{1}{n} \frac{1}{1 + x_i^T \hat{\beta}_i(\beta)},
\]

the empirical likelihood statistic equal to the maximum, then

\[
L(\beta) = \prod_{i=1}^{n} \frac{1}{1 + x_i^T \hat{\beta}_i(\beta)}.
\]

It is easy to obtain the empirical log-likelihood ratio statistic of \( \beta \) is

\[
l_\lambda(\beta) = - \sum_{i=1}^{n} \log(1 + x_i^T \hat{\beta}_i(\beta)),
\]

Where \( \lambda \in \mathbb{R} \) and \( \sum_{i=1}^{n} \frac{\hat{\beta}_i(\beta)}{1 + x_i^T \hat{\beta}_i(\beta)} = 0 \).

Regarding \( \lambda \) and \( \beta \) as independent variable and defining
\[ Q_{n}(\lambda, \beta) = -n^{-1} \sum_{i=1}^{n} \log(1 + \lambda^T \hat{y}_i(\beta)) \, . \]

Obviously, the maximum empirical likelihood estimates \( \hat{\beta} \) and \( \hat{\lambda} \) are the solutions of follow equations:

\[
\begin{align*}
Q_{1,\lambda}(\lambda; \beta) & = \frac{\partial Q_{n}(\lambda, \beta)}{\partial \lambda} = -n^{-1} \sum_{i=1}^{n} \hat{y}_i(\beta) \{1 + \lambda^T \hat{y}_i(\beta)\}^{-1} = 0, \\
Q_{1,\beta}(\lambda; \beta) & = \frac{\partial Q_{n}(\lambda, \beta)}{\partial \beta} = -n^{-1} \sum_{i=1}^{n} \frac{\partial \hat{y}_i(\beta)}{\partial \beta} \{1 + \lambda^T \hat{y}_i(\beta)\}^{-1} = 0. 
\end{align*}
\]

**IV. CASE-DELETION INFLUENCE MEASURES**

Consider the model (4), where the \( j \)-th case \((x_j^T, L_j)\) is deleted.

\[
L_i = x_i^T \beta + z_i^T \alpha(t_i) + \varepsilon_i, \quad i \neq j. \quad (5)
\]

This model is called case-deletion model. Let \( \hat{\beta}_{(ij)} \) is the maximum empirical likelihood estimate of \( \beta \) in model (5). In order to study the influence of the \( j \)-th case \((x_j^T, z_j^T, t_j, L_j)\), and compare the difference between \( \hat{\beta} \) and \( \hat{\beta}_{(ij)} \), the important result as follows theorem.

By Zhu, et al[7], for model (5), the maximum empirical likelihood estimator of \( \beta \) is

\[
\hat{\beta}_{(ij)} = \hat{\beta} - n^{-1} S_{22}^{-1} S_{21} \cdot S_{11}^{-1} \hat{\beta}_i(\beta) \{1 + o_p(1)\} \, , \quad (6)
\]

where \( S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} E_x(\hat{\beta}_i(\beta) \hat{\beta}_i(\beta)^T) & -E_x \left( \frac{\partial \hat{\beta}_i(\beta)}{\partial \beta} \right) \\ -E_x \left( \frac{\partial \hat{\beta}_i(\beta)}{\partial \beta} \right)^T & 0 \end{pmatrix}, S_{22} = -S_{21} S_{11}^{-1} S_{12}. \)

**(1) Empirical Cook Distance**

Zhu, et al[7] proposed empirical cook distance. Let \( M \) is a nonnegative matrix. The empirical cook distance is defined as follows

\[
ECD_{ij}(M) = (\hat{\beta} - \hat{\beta}_{(ij)})^T M (\hat{\beta} - \hat{\beta}_{(ij)}) \, , \quad (7)
\]
Where \( M = \left. \hat{\beta}^{\hat{\beta}^{\hat{\beta}}} \right|_{\beta, j} I_{\beta}(\beta) \).

(2) Empirical Likelihood Distance

Empirical likelihood distance is advanced from the view of data fitting. Considering the influence of deleting the \( j \)-th case. In order to eliminate the influence of scale, it is also need to divide the variance of estimator \( \text{Var}(\hat{\beta}) \). Because the keystone is to review the influence of deleting the \( j \)-th case. Hence, \( \hat{\sigma}^{2} \) is substituted by \( \hat{\sigma}^{2}(j) \). Then, the W-K statistic can be expressed as follows

\[
ELD_{j}(M) = 2 \left\{ I_{\beta}(\hat{\beta}) - I_{\beta}(\hat{\beta}_{(j)}) \right\}.
\]  

(8)

V. LOCAL INFLUENCE ANALYSIS OF MODEL

We consider the local influence method for a case-weight perturbation \( \omega \in \mathbb{R}^{*} \), for which the empirical log-likelihood function \( I_{\beta}(\beta | \omega) \) is defined by

\[
l_{\beta}(\beta | \omega) = \sum_{i=1}^{n} \omega_{i} l_{\beta,i}(\beta).
\]

In this case, \( \omega = \omega^{a} \), defined to be an \( n \times 1 \) vector with all elements equal to 1, represents no perturbation to the empirical likelihood, because \( l_{\beta}(\beta | \omega^{a}) = l_{\beta}(\beta) \). Thus, the empirical likelihood displacement is defined as \( l_{\omega}(\omega) = 2[I_{\beta}(\hat{\beta}) - I_{\beta}(\hat{\beta}(\omega))] \), where \( \hat{\beta}(\omega) \) is the maximum empirical likelihood estimator of \( \beta \) based on \( l_{\beta}(\beta | \omega) \). Let \( \omega(a) = \omega^{a} + ah \) with \( \omega(0) = \omega^{a} \) and \( d\omega(a)/da \bigg|_{a=0} = h \), where \( h \) is a direction in \( \mathbb{R}^{*} \). Thus, the normal curvature of the influence graph \( \omega \mapsto L_{\beta}(\omega) \) is given by

\[
C_{h}(\omega^{a}) = H \left|_{\omega^{a}} \right|^{T} H \left|_{\omega^{a}} \right| h,
\]

where

\[
H \left|_{\omega^{a}} \right| = -2 \left. \frac{\partial l_{\omega}(\omega)}{\partial \omega} \right|_{\omega^{a}} \left|_{\omega^{a}} \right|^{-2} \left. \frac{\partial l_{\omega}(\omega)}{\partial \omega} \right|_{\omega^{a}} \left|_{\omega^{a}} \right|^{-1} A \left|_{\omega^{a}} \right|,
\]

in which \( A = \partial^{2} l_{\omega}(\omega) \). This is a \( p \times n \) matrix with \( (k,i) \)-th element given by \( \partial_{\omega,i} l_{\omega}(\beta) \).

We consider two local influence measures based on the normal curvature \( C_{h}(\omega^{a}) \) as follows. Let \( \lambda_{1} \geq \cdots \geq \lambda_{p} \geq \lambda_{p+1} = \cdots = \lambda_{n} = 0 \) be the ordered
eigenvalues of the matrix $H_{LD_k(\alpha^*)}$ and let $\{v_m = (v_{m1}, \ldots, v_{mn})^T : m = 1, \ldots, n\}$ be the associated orthonormal basis, that is, $H_{LD_k(\alpha^*)} v_m = \lambda_m v_m$. Thus, the spectral decomposition of $H_{LD_k(\alpha^*)}$ is given by

$$H_{LD_k(\alpha^*)} = \sum_{m=1}^{n} \lambda_m v_m v_m^T.$$ 

The most popular local influence measures include $v_i$, which corresponds to the largest eigenvalue $\lambda_i$, as well as $C_{e_j} = \sum_{m=1}^{n} \lambda_m e_m e_m^T$, where $e_j$ is an $n \times 1$ vector with $j$-th component 1 and 0 otherwise. The $v_i$ represents the most influential perturbation to the empirical likelihood function, whereas the observation $(x_j^T, z_j^T, t_j)$ with a large $C_{e_j}$ can be regarded as influential.

As the discuss of Zhu, et al[7], for the partially linear varying coefficient regression model with random right censorship, we can deduce that

$$C_{e_j} = 2ELD_j \{1 + o_p(1)\} = 2ECD_j \{1 + o_p(1)\}$$

$$= -2n^{-1} \Delta_j S_{22}^{-1} A_j \{1 + o_p(1)\},$$

where $A_j = \partial_x I_{k,j}(x_j, \beta)|_{x = \hat{x}} = S_{21} S_{11}^{-1} \hat{\beta}(\beta) + o_p(1),$ 

$$S_{11} = \partial_x Q_{1,n} = 1 \sum_{i=1}^{n} \hat{\beta}(\beta) \hat{\beta}(\beta)^T (1 + \hat{x}(\beta)^2)^{-1},$$

$$S_{12} = \hat{\beta} Q_{1,n} = 1 \sum_{i=1}^{n} \hat{\beta}(\beta) \hat{\beta}(\beta)^T (1 + \hat{x}(\beta)^2)^{-1},$$

$$S_{21} = \hat{\beta} Q_{1,n} = 1 \sum_{i=1}^{n} \hat{\beta}(\beta) \hat{\beta}(\beta)^T (1 + \hat{x}(\beta)^2)^{-1},$$

$$S_{22} = \partial_x Q_{2,n} = 1 \sum_{i=1}^{n} \hat{\beta}(\beta) \hat{\beta}(\beta)^T (1 + \hat{x}(\beta)^2)^{-1},$$

$S_{22} = -S_{21} S_{11}^{-1} S_{12}$. 

\begin{equation}
0 = S_{22} = -S_{21} S_{11}^{-1} S_{12}.
\end{equation}
VI. AN NUMERICAL EXAMPLE

We choose \( p = 2 \) and varying coefficient functions

\[
\begin{align*}
\alpha_1(t) &= 2 \sin \left( \frac{2}{3} \pi t \right), \\
\alpha_2(t) &= \frac{1}{2} (1.5 - t)^2,
\end{align*}
\]

where \( t \sim U(0, 1) \). Let \( q = 2 \), \( \beta = [1.0, 1.5]^T \). Covariables \( X \sim N(0, \Sigma) \), \( Z \sim N(7, \Sigma) \), where

\[
\Sigma = \begin{pmatrix}
5 & \sqrt{5} \\
\sqrt{5} & 5
\end{pmatrix}, \quad \epsilon \sim N(0, 1). \text{Then according as}
\]

\[
Y = X^T \beta + Z^T \alpha(T) + \epsilon
\]

build the censored data \( C \),

\[
C_i = Y_i \times \hat{\lambda}_i, \quad i = 1, 2, \ldots, n
\]

where \( \{\hat{\lambda}_i\}_{i=1}^n \sim U(0, c) \) and \( n = 1000 \). If \( c = 5 \), the censored proportion is about 20\%. Supposed that bandwidth \( h_n = 5(n \log n)^{1/2} \) (Wang & Jing[13]) and \( K(t) = \frac{15}{16} (1 - t^2)^3 (|t| < 1) \).

In order to checkout the validity of this method, we change the response variable value of the first, 115th, 550th, 725th, 999th data.

For every \( i \), it is easy to obtain \( \eta_i(\beta) \). For the parameters \( \beta \) and \( \hat{\lambda} \), using the samples, we evaluated their maximum empirical likelihood estimators are \( \hat{\beta} = -6.6171 \), \( \hat{\lambda} = 0.0072 \).

Consequently, it is easy to calculate the value of \( S_{11}, S_{12}, S_{21}, S_{22} \) and \( C_{e_i} \). The result of \( C_{e_i} \) is as follows figure.
It can be seen from the result of $c_i$ that the first, 115th, 550th, 725th, 999th data are strong influence point. Indeed, our results are illustrated.

REFERENCES


DEPARTMENT OF FUNDAMENTAL COURSE, AIR FORCE LOGISTICS COLLEGE, XUZHOU, CHINA