International Journal of Analysis and Applications
ISSN 2291-8639
Volume 3, Number 1 (2013), 25-34
http://www.etamaths.com

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 patially 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 varing-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

the usual statistical techniques for complete data situations are not readily applicable.

The empirical likelihood method origins from Thomas & Grunkemeier (1975). Owen (2001) first proposed the definition of empirical likelihood and expounded the system info of empirical likelihood. Zhu and Ibrahim (2008) utilized this method for statistical diagnostic. Liugen Xue and Lixing Zhu (2010) summarized the application of this method.

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:

$$Y = X^T\beta + Z^T\alpha(T) + \varepsilon \qquad (1)$$

Where $X = (X_1, \cdots, X_n)^T$ is of dimension $n \times 1$ and $\beta = (\beta_1, \cdots, \beta_p)^T$ is a p-dimensional vector of unknown coefficient functions. $\alpha(T) = (\alpha_1(T), \cdots, \alpha_q(T))^T$ is unknown function vector. T is one-dimension vector. ε is a stochastic error with $E(\varepsilon \mid X^T, Z^T, T) = 0, Var(\varepsilon \mid X^T, Z^T, T) = \sigma^2$.

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 $\delta = I(Y \le C)$, where $I(\cdot)$ is the index function. The observations are $\{(x_k^T, z_k^T, t_k, \Delta_k, \delta_k)\}$ which are random samples

from $(X^T, Z^T, T, \Delta, \delta)$,where $x_k^T = (x_{k1}, \cdots, x_{kp})^T$ .Thus instead of observing Y_k ,we observe the pairs (Δ_k, δ_k) ,where $\Delta_k = \min(Y_k, C_k)$ and $\delta_k = I(Y_k \le C_k)$,Observations on Δ_k for which $\delta_k = 1$ are uncensored, and observations on Δ_k for which $\delta_k = 0$ 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 $\tau_{F_i} = \inf\{t : F(t) = 1\}$ .Note that $\bar{F}_i = 1 - F_i$ and $\bar{G} = 1 - G$.

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

Proof. Since

$$E\delta_i \Delta_i \bar{G}^{-1}(\Delta_i) = \int_0^{\tau_{F_i}} \int_y^{\tau_G} \frac{y}{1 - G(y)} dG(t) dF_i(y)$$

$$= EY_i$$

and

$$EY_i = \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{k=1}^{q} z_{ik}\alpha(t_i)$$

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

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

$$\frac{\delta_i \Delta_i}{\bar{G}(\Delta_i)} = \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{k=1}^{q} z_{ik}\alpha(t_i) + \varepsilon_i^* , \quad i = 1,2,\cdots,n \qquad (2)$$

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

$$\hat{\bar{G}}(t) = \begin{cases} \prod_{j=1}^{n} \left(\frac{1 + N^+(\Delta_j)}{2 + N^+(\Delta_j)} \right)^{I[\delta_j = 0, \Delta_j \le t]}, & if \quad t \le \Delta_{(n)} \\ 0, & if \quad t > \Delta_{(n)} \end{cases} \qquad (3)$$

where $\Delta_{(n)} = \max\{\Delta_1, \Delta_2, \cdots, \Delta_n\}$, $N^+(\Delta_j) = \sum_{i=1}^{n} I[\Delta_i \geq \Delta_j]$, $j = 1, 2, \cdots, n$.

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

$$L_i = x_i^T \beta + z_i^T \alpha(t_i) + \varepsilon_i^*, \quad i = 1, \cdots, n \qquad (4)$$

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 β can be defined by

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

where

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

$\hat{G}(t) = \sum_{i=1}^{n} W_i(t) x_i y_i^{*T}$, $\hat{\mu}(t) = \sum_{i=1}^{n} W_i(t) x_i$, $W_i(t) = K_h(t - t_i) \Big/ \sum_{i=1}^{n} K_h(t - t_i)$, $K_h(\bullet) = K(\bullet/h)$,

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

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

$$p_i = \frac{1}{n} \frac{1}{1 + \lambda^T \hat{\eta}_i(\beta)},$$

the empirical likelihood statistic equal to the maximum, then

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{1 + \lambda^T \hat{\eta}_i(\beta)}.$$

It is easy to obtain the empirical log-likelihood ratio statistic of β is

$$l_E(\beta) = -\sum_{i=1}^{n} \log(1 + \lambda^T \hat{\eta}_i(\beta)),$$

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

Regarding λ and β as independent variable and defining

$$Q_n(\lambda,\beta) = -n^{-1}\sum_{i=1}^{n}\log(1+\lambda^T\hat{\eta}_i(\beta)) \quad .$$

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

$$\begin{cases} Q_{1,n}(\lambda,\beta) = \dfrac{\partial Q_n(\lambda,\beta)}{\partial\lambda} = -n^{-1}\sum_{i=1}^{n}\hat{\eta}_i(\beta)\left\{1+\lambda^T\hat{\eta}_i(\beta)\right\}^{-1} = 0 \\ Q_{2,n}(\lambda,\beta) = \dfrac{\partial Q_n(\lambda,\beta)}{\partial\beta} = -n^{-1}\sum_{i=1}^{n}\dfrac{\partial\hat{\eta}_i(\beta)}{\partial\beta}\lambda\left\{1+\lambda^T\hat{\eta}_i(\beta)\right\}^{-1} = 0 \end{cases} .$$

IV. CASE-DELETION INFLUENCE MEASURES

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

$$L_i = x_i^T\beta + z_i^T\alpha(t_i) + \varepsilon_i^* , \qquad i \neq j . \qquad (5)$$

This model is called case-deletion model. Let $\hat{\beta}_{(j)}$ is the maximum empirical likelihood estimate of β 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 $\hat{\beta}_{(j)}$.The important result as follows theorem.

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

$$\hat{\beta}_{(j)} = \hat{\beta} - n^{-1}S_{22.1}^{-1}S_{21}S_{11}^{-1}\hat{\eta}_i(\beta)\{1+o_p(1)\} , \qquad (6)$$

where $S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} E_F(\hat{\eta}_i(\beta)\hat{\eta}_i(\beta)^T) & -E_F\left(\dfrac{\partial\hat{\eta}_i(\beta)}{\partial\beta}\right)^T \\ -E_F\left(\dfrac{\partial\hat{\eta}_i(\beta)}{\partial\beta}\right) & 0 \end{pmatrix}$, $S_{22.1} = -S_{21}S_{11}^{-1}S_{12}$.

(1)Empirical Cook Distance

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

$$ECD_j(M) = (\hat{\beta} - \hat{\beta}_{(j)})^T M(\hat{\beta} - \hat{\beta}_{(j)}) , \qquad (7)$$

Where $M = \left.\frac{\partial^2 l_E(\beta)}{\partial \beta^2}\right|_{\beta=\hat{\beta}}$.

(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 $Var(\hat{y}_j^*)$.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

$$ELD_j(M) = 2\left\{l_E(\hat{\beta}) - l_E(\hat{\beta}_{(j)})\right\} . \quad (8)$$

V. LOCAL INFLUENCE ANALYSIS OF MODEL

We consider the local influence method for a case-weight perturbation $\omega \in R^n$,for which the empirical log-likelihood function $l_E(\beta|\omega)$ is defined by $l_E(\beta|\omega) = \sum_{i=1}^{n} \omega_i l_{E,i}(\beta)$.In this case, $\omega = \omega^0$,defined to be an $n \times 1$ vector with all elements equal to 1,represents no perturbation to the empirical likelihood, because $l_E(\beta|\omega^0) = l_E(\beta)$.Thus, the empirical likelihood displacement is defined as $l_{DE}(\omega) = 2[l_E(\hat{\beta}) - l_E\{\hat{\beta}(\omega)\}]$,where $\hat{\beta}(\omega)$ is the maximum empirical likelihood estimator of β based on $l_E(\beta|\omega)$.Let $\omega(a) = \omega^0 + ah$ with $\omega(0) = \omega^0$ and $d\omega(a)/da|_{a=0} = h$,where h is a direction in R^n.Thus, the normal curvature of the influence graph $(\omega^T, LD_E(\omega))^T$ is given by

$C_h(\omega^0) = h^T H_{LD_E(\omega^0)} h$, where

$H_{LD_E(\omega^0)} = -2\left.\frac{\partial^2 LD_E\{\hat{\beta}(\omega)\}}{\partial\omega\partial\omega^T}\right|_{\omega^0} = 2\Delta^T\{-\partial_\beta^2 l_E(\beta)\}^{-1}\Delta\Big|_{\omega^0,\hat{\beta}}$ in which $\Delta = \partial^2_{\beta\omega} LD_E(\beta,\omega)$ is a $p \times n$ matrix with (k,i)-th element given by $\partial_{\beta_k} l_{E,i}(\beta)$.

We consider two local influence measures based on the normal curvature $C_h(\omega^0)$ 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 $H_{LD_E(\omega^0)}$ and let $\{v_m = (v_{m1}, \cdots v_{mn})^T : m = 1, \cdots n\}$ be the associated orthonormal basis, that is, $H_{LD_E(\omega^0)} v_m = \lambda_m v_m$ .Thus, the spectral decomposition of $H_{LD_E(\omega^0)}$ is given by

$$H_{LD_E(\omega^0)} = \sum_{m=1}^{n} \lambda_m v_m v_m^T \quad ..$$

The most popular local influence measures include v_1 ,which corresponds the largest eigenvalue λ_1 ,as well as $C_{e_j} = \sum_{m=1}^{p} \lambda_m v_{mj}^2$,where e_j is an $n \times 1$ vector with j -th component 1 and 0 otherwise. The v_1 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

$$\begin{aligned} C_{e_j} &= 2ELD_j\{1 + o_p(1)\} = 2ECD_j\{1 + o_p(1)\} \\ &= -2n^{-1}\Delta_j^T S_{22.1}^{-1}\Delta_j\{1 + o_p(1)\} \quad , \end{aligned} \tag{9}$$

Where $\Delta_j = \partial_\beta l_{E,j}(x_j, \beta)\Big|_{\beta=\hat{\beta}} = \frac{S_{21}S_{11}^{-1}\hat{\eta}_i(\beta)}{1 + \hat{\lambda}^T\hat{\eta}_i(\beta)} + o_p(1)$,

$$S_{11} = \partial_\lambda Q_{1,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\eta}_i(\beta)\hat{\eta}_i^T(\beta)}{(1 + \lambda^T\hat{\eta}_i(\beta))^2}\Bigg|_{\beta=\hat{\beta},\lambda=\hat{\lambda}},$$

$$S_{12} = \partial_\beta Q_{1,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\eta}_i(\beta)\lambda^T\partial_\beta(\hat{\eta}_i(\beta)) - \partial_\beta(\hat{\eta}_i(\beta))(1 + \lambda^T\hat{\eta}_i(\beta))}{(1 + \lambda^T\hat{\eta}_i(\beta))^2}\Bigg|_{\beta=\hat{\beta},\lambda=\hat{\lambda}},$$

$$S_{21} = \partial_\lambda Q_{2,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\eta}_i(\beta)\lambda^T\partial_\beta(\hat{\eta}_i(\beta)) - \partial_\beta(\hat{\eta}_i(\beta))(1 + \lambda^T\hat{\eta}_i(\beta))}{(1 + \lambda^T\hat{\eta}_i(\beta))^2}\Bigg|_{\beta=\hat{\beta},\lambda=\hat{\lambda}},$$

$$S_{22} = \partial_\beta Q_{2,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial_\beta^T(\hat{\eta}_i(\beta))\lambda\lambda^T\partial_\beta(\hat{\eta}_i(\beta))}{(1 + \lambda^T\hat{\eta}_i(\beta))^2}\Bigg|_{\beta=\hat{\beta},\lambda=\hat{\lambda}}, \quad S_{22.1} = -S_{21}S_{11}^{-1}S_{12} .$$

VI. AN NUMERICAL EXAMPLE

We choose $p = 2$ and varying coefficient functions

$$\alpha_1(t) = 2\sin\left(\frac{2}{3}\pi t\right), \quad \alpha_2(t) = \frac{1}{2}(1.5 - t)^2,$$

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

$\Sigma = \begin{pmatrix} 5 & \sqrt{5} \\ \sqrt{5} & 5 \end{pmatrix}$, $\varepsilon \sim N(0,1)$ .Then according as

$$Y = X^T\beta + Z^T\alpha(T) + \varepsilon$$

build the censored data C ,

$C_i = Y_i \times \lambda_i$, $i = 1, 2, \cdots, n$,

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

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 β and λ ,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.

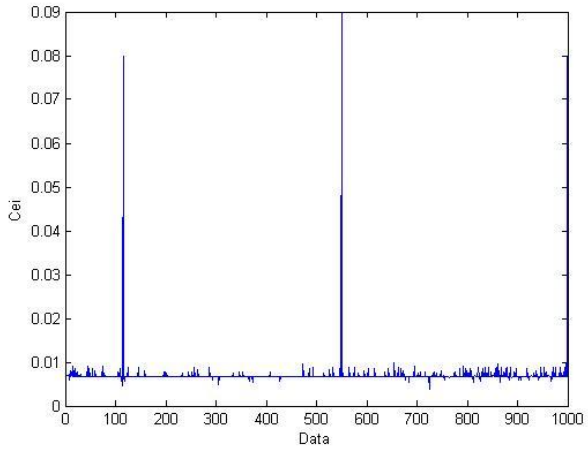

Fig 1 The influence c_{e_i}

It can be seen from the result of c_{e_i} that the first、115th、550th、725th、999th data are strong influence point. Indeed, our results are illustrated.

REFERENCES

[1] R,D.Cook and S.Weisberg.,Residuals and Influence in Regression. New York:Chapman and Hall,1982

[2] Wei Bocheng, Lu Guobin，Shi Jianqing., Statistical Diagnostics. Nanjing：Publishing House of Southeast University,1990

[3] Cai Z,Fan J,Li R.Efficient Estimation and Inferences for Varying-coefficient Models[J].Journal of American Statistical Association,2000,95(451):888-902.

[4] Fan J,Zhang W.Statistical Methods with Varying Coefficient Models[J].Statistics and Its Interface,2008,1(1):179-195.

[5] Thomas,D.R.and Grunkemeier,G.L.,Confidence Interval Estimation of Survival Interval Estimation of Survival Probabilities for Censored Data.Journal of the American Ststistical Association.1975:865-871.

[6] Owen A.,Empirical Likelihood.New York:Chapman and Hall,2001

[7] Zhu H T,Ibrahim J G,Tang N S,et al,Diagnostic Measures for Empirical Likelihood of Generalized Estimating Equations,Biometrika,2008,99:489-507

[8] Liugen Xue,Lixing Zhu.,Empirical Likelihood in Nonparametric and Semiparametric Models.Beijing.Science Press.2010.

[9] Wang Qihua.,Analysis of Survival Data. Beijing：Science Press,2006

[10] Fan J Q,Huang T.Profile likelihood inferences on semiparametric varying-coefficient partially linear models[J].Bernoulli,2005,1:1031-1057.

[11] Qin J, Lawless J., Empirical Likelihood and General Estimating Equations. Ann.Statist.,1994,22:300-325

[12] Eubank R L., Diagnostics for Smoothing Spline.J.R.Statist.Soc. (Series B) , 1985(47):322-341

[13] Wang,Q.&Jing,B.Empirical Likelihood for Partial Linear Model with Fixed Designs.Statist.Probab.Lett.1999,no.41:425-433.

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