Applied Mathematics, 2013, 4, 854-858
http://dx.doi.org/10.4236/am.2013.45117 Published Online May 2013 (http://www.scirp.org/journal/am)

Local Influence Analysis of Varying-Coefficient Model with Random Right Censorship

Shuling Wang, Man Liu, Daqing Liao, Ting Wang
Department of Fundamental Course, Air Force Logistics College, Xuzhou, China
Email: wangshuling2007@yahoo.com.cn

Received October 19, 2012; revised April 3, 2013; accepted April 10, 2013

ABSTRACT

For this model, this paper studies the method and application of the diagnostic mostly. Firstly, the primary model is transformed to varying-coefficient model by using a general transformation method. Secondly, a simple estimation form of the coefficient functions is obtained by employing the B spline. Then, local influence is discussed and concise influence matrix is obtained. At last, an example is given to illustrate our results.

Keywords: Random Right Censorship; B Splines; Local Influence

1. Introduction

Local influence analysis is proposed from the viewpoint of differential geometry [1]. Nearly thirty years, the diagnosis and influence analysis of linear regression model have been fully developed (Ref. [2,3]). The varing-coefficient model is a useful extension of classical linear model. It has been widely applied in statistical modelling, for example, see Ref. [1,4-6]. 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, the usual statistical techniques for complete data situations are not readily applicable. When the response is censored, the relationship between the response and the covariate has been widely studied in the literature [7-10].

So far the local influence analysis of varying-coefficient model with random right censorship has not yet seen in the literature, this paper attempts to study it. The paper is organized as follows: The introduction of local influence is given in Section 2; The model and the estimators are introduced in Section 3; The statistical diagnostics are given in Section 4; The example to illustrate our results is given in Section 5.

2. Local Influence

Ref. [2,3] have discussed the method of local influence analysis. Let α be an unknown k-dimensional parameter, whose domain is an open subset of Euclidean space R^k. $l(\alpha)$ is a object function (for example, likelihood function, punishment log-likelihood function). ω is a *n*-vector which denotes disturbed factor, for example weighted or tiny shift. Let $M(\omega)$ be the disturbed model, whose object function is $l(\alpha\,|\,\omega)$. $\hat{\alpha}_\omega$ is the estimate which is from $M(\omega)$. Given ω_0 makes $l(\alpha\,|\,\omega_0)=l(\alpha)$ and $\hat{\alpha}=\hat{\alpha}_{\omega_0}$, where $l(\alpha\,|\,\omega)$ has continuous second-order partial derivatives, $l(\hat{\alpha}_\omega)$ is the function of ω. In geometry, $l(\hat{\alpha}_\omega)$ denotes *n*-dimentional surface

$$\eta(\omega)=\left(\omega^{\mathrm{T}},l(\hat{\alpha}_\omega)\right)^{\mathrm{T}} \tag{1}$$

This image is called influence image, which varies with ω. The variation rate in ω_0 of influence image reflects that the sensitivity of model, where ω_0 corresponds to the primary model. This method is called local influence. COOK advanced that utilize influence curvature to measure the change of influence image near ω_0.

Ref. [2,3] pointed out that the influence curvature of $\eta(\omega)$ is given by

$$C_d=2\left|d^{\mathrm{T}}D^{\mathrm{T}}\ddot{l}Dd\right|=2\left|d^{\mathrm{T}}\Delta^{\mathrm{T}}\ddot{l}^{-1}\Delta d\right| \tag{2}$$

where $\ddot{l}$ is second derivatives of $l(\alpha)$ with respect to α, and

$$\boldsymbol{D}=\frac{\partial\hat{\alpha}_\omega}{\partial\omega},\ \Delta=\frac{\partial^2 l(\alpha\,|\,\omega)}{\partial\alpha\partial\omega^{\mathrm{T}}} \tag{3}$$

$\boldsymbol{D}$ and Δ are $k\times n$ matrix, where $\alpha=\hat{\alpha}$, $\omega=\omega_0$.

The influence matrix is given by

$$\ddot{\boldsymbol{F}}=D^{\mathrm{T}}\ddot{l}D=\Delta^{\mathrm{T}}\ddot{l}^{-1}\Delta \tag{4}$$

Formula (2) shows that the maximal influence curvature $C_{\max}=2\lambda_1$, where λ_1 is the eigenvalue of $\ddot{F}$ whose absolute value is maximal, and $d_{\max}$ is the corresponding eigenvector which is called the direction of maximal influence curvature. Ref. [5] pointed out that the diagonal value of influence matrix also is the important diagnostic statistics.

3. The Model and Estimators

Let *Y* be the response variable and $\left(T,X^{\mathrm{T}}\right)$ be its associated covariates. The varying-coefficient regression model assumes the following structure:

$$Y=\beta^{\mathrm{T}}(T)X+\varepsilon \tag{5}$$

where $X=\left(X_1,\cdots,X_n\right)^{\mathrm{T}}$ is of dimension $n\times 1$ and $\beta(\bullet)=\left(\beta_1(\bullet),\cdots,\beta_p(\bullet)\right)^{\mathrm{T}}$ is a p-dimensional vector of unknown coefficient functions. ε is a stochastic error with

$$E(\varepsilon\,|\,T,X)=0, Var(\varepsilon\,|\,T,X)=\sigma^2(T,X).$$

Consider the model (5), 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 $\left(T,X^{\mathrm{T}}\right)$. Denote $\Delta=\min(Y,C)$ and $\delta=I(Y\le C)$, where $I(\cdot)$ is the index function. The observations are $\left\{\left(t_k,x_k^T,\Delta_k,\delta_k\right):k=1,2,\cdots,n\right\}$ which are random samples from $\left(T,X^{\mathrm{T}},\Delta,\delta\right)$, where $x_k^{\mathrm{T}}=\left(x_{k1},\cdots,x_{kp}\right)^{\mathrm{T}}$. Thus instead of observing Y_k, we observe the pairs $\left(\Delta_k,\delta_k\right)$, where $\Delta_k=\min\left(Y_k,C_k\right)$ and $\delta_k=I\left(Y_k\le C_k\right)$. 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 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 $E\delta_i\Delta_i\bar{G}^{-1}(\Delta_i)=\sum_{k=1}^{p}\beta_k(t_i)x_{ik}$, $i=1,2,\cdots,n$.

Proof. Since

$$\begin{aligned}&E\delta_i\Delta_i\bar{G}^{-1}(\Delta_i)\\&=\int_0^{\tau_{F_i}}\int_y^{\tau_G}\frac{y}{1-G(y)}\mathrm{d}G(t)\mathrm{d}F_i(y)=EY_i\end{aligned}$$

and

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

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

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

$$\frac{\delta_i\Delta_i}{\bar{G}(\Delta_i)}=\sum_{k=1}^{p}\beta_k(t_i)x_{ik}+\varepsilon_i^*,\ i=1,2,\cdots,n \tag{6}$$

where ε_i^* is i.i.d. and $E\varepsilon_i^*=0$, $Var\left(\varepsilon_i^*\right)=\sigma^{*2}$. In practice, we replace $\bar{G}$ with $\hat{\bar{G}}$ which is the Kaplan-Meier product-limited estimator of $\bar{G}$ (Ref. [11]). The expression of $\hat{\bar{G}}$ is given as follows:

$$\hat{\bar{G}}(t)=\begin{cases}\prod_{j=1}^{n}\left(\dfrac{1+N^+(\Delta_j)}{2+N^+(\Delta_j)}\right)^{I\left[\delta_j=0,\Delta_j\le t\right]}, & \text{if } t\le\Delta_{(n)}\\ 0, & \text{if } t>\Delta_{(n)}\end{cases} \tag{7}$$

where $\Delta_{(n)}=\max\left\{\Delta_1,\Delta_2,\cdots,\Delta_n\right\}$,

$$N^+(\Delta_j)=\sum_{i=1}^{n}I\left[\Delta_i\ge\Delta_j\right],\ j=1,2,\cdots,n.$$

Let $y_i^*=\dfrac{\delta_i\Delta_i}{\bar{G}(\Delta_i)}$, model (5) is transformed to following varying-coefficient regression model

$$y_i^*=\sum_{k=1}^{p}\beta_k(t_i)x_{ik}+\varepsilon_i^*,\ i=1,\cdots,n \tag{8}$$

Now we want to estimate the unknown coefficient function vector based on the transformed data. In varying-coefficient model, there are a lot of estimates for $\beta(t)$. Here we use the B-spline estimate $\hat{\beta}(t)$.

Let $z=(z_1,\cdots,z_k)$, $a<z_1<\cdots<z_k<b$ are the knots in $[a,b]$, $\pi(t)=\left(\pi_1(t),\cdots,\pi_N(t)\right)^{\mathrm{T}}$ and $N=m+k+1$ are the basis functions of *m*-th B-spline, $S(m,z)=\left\{\pi^{\mathrm{T}}(t)\alpha;\alpha\in R^N\right\}$ is the space of *m*-th B-spline function. We use the lemma 1.2 of Ref. [3], every smooth coefficient function $\beta_l(t)$ can be approximated by B-spline function $s_l(t)\in S(m,z)$. The B-spline estimator of the coefficient function $\beta_l(t),l=1,\cdots,p$ in model (8) is the solution of following formula

$$\begin{aligned}&\sum_{i=1}^{n}\left(y_i^*-\left(x_{i1}\hat{\beta}_1(t_i)+\cdots+x_{ip}\hat{\beta}_p(t_i)\right)\right)^2\\&=\min_{s_l(t)\in S(m,z)}\sum_{i=1}^{n}\left(y_i^*-\left(x_{i1}s_1(t_i)+\cdots+x_{ip}s_p(t_i)\right)\right)^2\\&=\min_{\alpha_j\in R^N,j=1,\cdots,p}\sum_{i=1}^{n}\left[y_i^*-\left(x_{i1}\pi^T(t_i)\alpha_1+\cdots+x_{ip}\pi^T(t_i)\alpha_P\right)\right]^2\end{aligned} \tag{9}$$

In order to depict conveniently, supposed that

$$Y^*=\left(y_1^*,\cdots,y_n^*\right)^{\mathrm{T}},\ \varepsilon^*=\left(\varepsilon_1^*,\cdots,\varepsilon_n^*\right)^{\mathrm{T}},$$

$$A=\begin{pmatrix}x_{11}\pi^{\mathrm{T}}(t_1) & \cdots & x_{1p}\pi^{\mathrm{T}}(t_1)\\ \vdots & \ddots & \vdots\\ x_{n1}\pi^{\mathrm{T}}(t_n) & \cdots & x_{np}\pi^{\mathrm{T}}(t_n)\end{pmatrix}=\begin{pmatrix}A_1^{\mathrm{T}}\\ \vdots\\ A_n^{\mathrm{T}}\end{pmatrix},$$

$$X=\mathrm{diag}\left(x_1^{\mathrm{T}},\cdots,x_n^{\mathrm{T}}\right),\quad x_i^{\mathrm{T}}=\left(x_{i1},\cdots,x_{ip}\right),$$

$$\beta(t)=\left(\beta_1(t),\cdots,\beta_p(t)\right)^{\mathrm{T}},\quad \alpha=\left(\alpha_1^{\mathrm{T}},\cdots,\alpha_P^{\mathrm{T}}\right)^{\mathrm{T}},$$

$$\bar{\beta}(t)=\left(\beta^{\mathrm{T}}(t_1),\cdots,\beta^{\mathrm{T}}(t_n)\right)^{\mathrm{T}},$$

$$\beta(t_l)=\left(\beta_1(t_l),\cdots,\beta_p(t_l)\right)^{\mathrm{T}},$$

then $Y^*=X\bar{\beta}(t)+\varepsilon^*$, and Formula (9) can be transformed to following minimize problem

$$S(\beta)=\min_{\alpha}\left(Y^*-A\alpha\right)^{\mathrm{T}}\left(Y^*-A\alpha\right) \tag{10}$$

Utilize the least-square method, the estimator of α is

$$\hat{\alpha}=\left(A^{\mathrm{T}}A\right)^{-1}A^{\mathrm{T}}Y^*$$

The estimator of the *l*-th coefficient function $\beta_l(t)$, $l=1,2,\cdots,p$ is

$$\hat{\beta}_l(t)=\pi^{\mathrm{T}}(t)\hat{\alpha}_l$$

Then, the estimator of the coefficient function $\beta(t)$ is

$$\hat{\beta}(t)=\boldsymbol{I_p}\otimes\pi^{\mathrm{T}}(t)\cdot\hat{\alpha}=\boldsymbol{I_p}\otimes\pi^{\mathrm{T}}(t)\cdot\left(A^{\mathrm{T}}A\right)^{-1}A^{\mathrm{T}}Y^* \tag{11}$$

where $\boldsymbol{I_p}$ is an $p\times p$ unit matrix, and $A\otimes B=\left(a_{ij}B\right)$ is Kronecker product of matrix.

4. The Local Influence of the Model

4.1. Weighted Perturbation Model

Suppose that $\omega=\left(\omega_1,\omega_2,\cdots,\omega_n\right)^{\mathrm{T}}$, then the weighted perturbation model can be shown that

$$S(\alpha|\omega)=\sum_{i=1}^{n}\omega_i\left[y_i^*-\left(x_{i1}\pi^{\mathrm{T}}(t_i)\alpha_1+\cdots+x_{ip}\pi^{\mathrm{T}}(t_i)\alpha_P\right)\right]^2 \tag{12}$$

Substituting this result into (3) yields

$$\Delta=\left.\frac{\partial^2 S(\alpha|\omega)}{\partial\alpha\partial\omega^{\mathrm{T}}}\right|_{\hat{\alpha},\omega_0}=-2A^{\mathrm{T}}D\left(\varepsilon^*\right) \tag{13}$$

where $D\left(\varepsilon^*\right)=\mathrm{diag}\left(\hat{\varepsilon}_1^*,\hat{\varepsilon}_2^*,\cdots,\hat{\varepsilon}_n^*\right)$ and $\omega_0=(1,1,\cdots,1)$, the second derivatives of $S(\alpha|\omega)$ with respect to α is given by

$$\ddot{S}=E\ddot{S}=2A^{\mathrm{T}}A \tag{14}$$

Substituting (13) and (14) into (4), we obtain the corresponding influence matrix

$$\boldsymbol{F}_{\omega}(\alpha)=2D^{\mathrm{T}}\left(\varepsilon^*\right)A\left(A^{\mathrm{T}}A\right)^{-1}A^{\mathrm{T}}D\left(\varepsilon^*\right) \tag{15}$$

Here d_w denotes the direction of maximal influence curvature.

4.2. Response Variable Perturbation Model

Suppose that $Y_{\omega}=Y+\omega$, then the response variable perturbation model can be shown that

$$S(\alpha|\omega)=\sum_{i=1}^{n}\left[y_i^*+\omega_i-\left(x_{i1}\pi^{\mathrm{T}}(t_i)\alpha_1+\cdots+x_{ip}\pi^{\mathrm{T}}(t_i)\alpha_P\right)\right]^2 \tag{16}$$

Substituting this result into (3) yields

$$\Delta=\left.\frac{\partial^2 S(\alpha|\omega)}{\partial\alpha\partial\omega^{\mathrm{T}}}\right|_{\hat{\alpha},\omega_0}=-2A^{\mathrm{T}} \tag{17}$$

the second derivatives of $S(\alpha|\omega)$ with respect to α is given by

$$\ddot{S}=E\ddot{S}=2A^{\mathrm{T}}A \tag{18}$$

Substituting (17) and (18) into (4), we obtain the corresponding influence matrix

$$\boldsymbol{F}_r(\alpha)=2A\left(A^{\mathrm{T}}A\right)^{-1}A^{\mathrm{T}} \tag{19}$$

Here d_r denotes the direction of maximal influence curvature.

5. An Illustrative Example

(Vicious Tumour Data) Now we consider an example as the illustration for the above results. Considering a clinical research trial data (see Ref. [4]), there are 205 cancer patients who have been treated in Odense university hospital and tracked until the end of 1977. The survival time of some individuals due to death or end of the trial for other reasons were censored. Ref. [11] utilized a linear semi-parametric model to fit this test data. We utilized varying-coefficient model to fit the data of 57 patients. Where x_j denoted the thickness of tumour, t_i denoted the sex (1 is male, 0 is female). Considering that there was

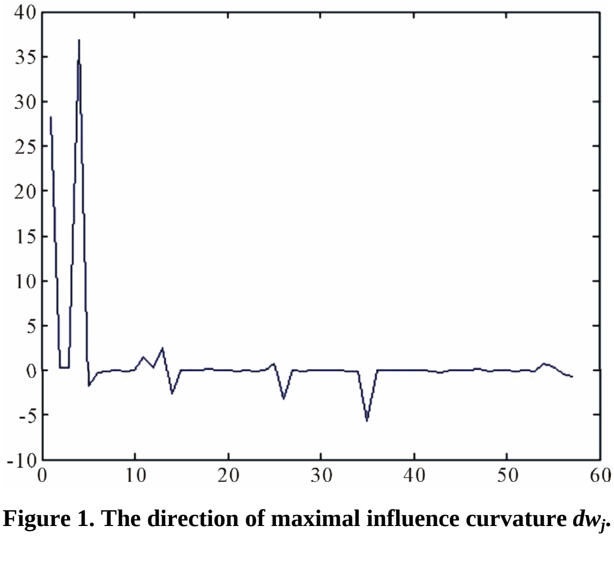

Figure 1. The direction of maximal influence curvature *dw_j*.

Table 1. The value of static.

No.	Fw_j	dw_j	Fr_j	dr_j
1	28.2231	0.5329	0.4035	0.3979
2	0.2392	0.0490	0.0648	0.1595
3	0.3537	0.0596	0.0737	0.1701
4	36.8723	0.6094	0.4592	0.4247
5	−1.7874	0.1713	−0.0528	−0.1703
6	−0.3430	0.0754	−0.0176	−0.0977
7	−0.0784	0.0359	−0.0056	−0.0554
8	0.0212	0.0146	0.0415	0.1276
9	−0.1190	0.0443	−0.0074	−0.0640
10	0.0075	−0.0087	0.0162	0.0798
11	1.5449	0.1246	0.1380	0.2330
12	0.3135	0.0561	0.0831	0.1807
13	2.5216	0.1591	0.1652	0.2545
14	−2.5727	0.2055	−0.0614	−0.1827
15	0.0120	−0.0110	0.0152	0.0772
16	0.0253	0.0159	0.0485	0.1380
17	0.0147	0.0121	0.0451	0.1329
18	0.1052	0.0325	0.0648	0.1595
19	−0.0003	0.0023	−0.0000	−0.0042
20	−0.0290	0.0219	−0.0021	−0.0341
21	−0.0641	0.0324	−0.0040	−0.0470
22	−0.0033	0.0074	−0.0003	−0.0128
23	−0.1388	0.0475	−0.0074	−0.0639
24	0.0204	−0.0143	0.0073	0.0534
25	0.7643	0.0877	0.1148	0.2123
26	−3.2453	0.2309	−0.0687	−0.1933
27	0.0137	−0.0117	0.0184	0.0850
28	−0.0823	0.0367	−0.0048	−0.0511
29	0.0000	−0.0004	0.0347	0.1165
30	0.0030	0.0055	0.0415	0.1276
31	−0.0071	0.0109	−0.0006	−0.0177
32	0.0213	−0.0146	0.0139	0.0738
33	−0.0850	0.0375	−0.0048	−0.0511
34	−0.0688	0.0336	−0.0040	−0.0468
35	−5.7540	0.3085	−0.0974	−0.2301
36	0.0061	−0.0078	0.0287	0.1062
37	0.0229	0.0152	0.0563	0.1488
38	0.0238	−0.0155	0.0184	0.0850
39	−0.0127	0.0144	−0.0008	−0.0214
40	−0.0229	0.0195	−0.0014	−0.0278
41	0.0144	−0.0120	0.0018	0.0267
42	−0.2025	0.0579	−0.0085	−0.0681
43	−0.2996	0.0702	−0.0109	−0.0764
44	0.0423	−0.0206	0.0073	0.0534
45	−0.0052	0.0093	−0.0003	−0.0128
46	−0.0322	0.0231	−0.0012	−0.0257
47	0.1642	−0.0406	0.0141	0.0745
48	−0.1150	0.0434	−0.0030	−0.0405
49	0.0432	−0.0208	0.0018	0.0267
50	−0.0366	0.0248	−0.0010	−0.0232
51	−0.0746	0.0353	−0.0016	−0.0298
52	0.0622	−0.0250	0.0018	0.0267
53	−0.0960	0.0396	−0.0016	−0.0298
54	0.7052	−0.0841	0.0347	0.1165
55	0.4619	−0.0682	0.0104	0.0639
56	−0.4003	0.0814	−0.0040	−0.0467
57	−0.6435	0.1039	−0.0033	−0.0426

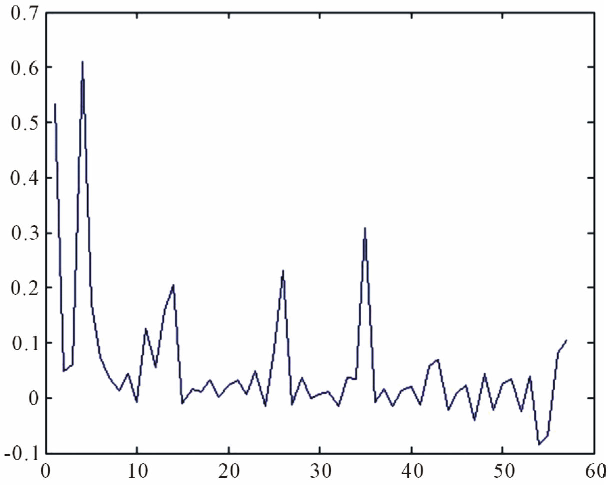

Figure 2. The diagonal value of influence matrix Fw_j.

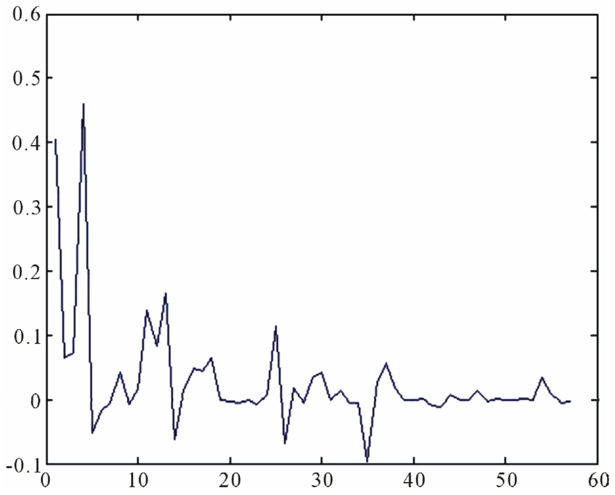

Figure 3. The diagonal value of influence matrix *Fr*$_j$.

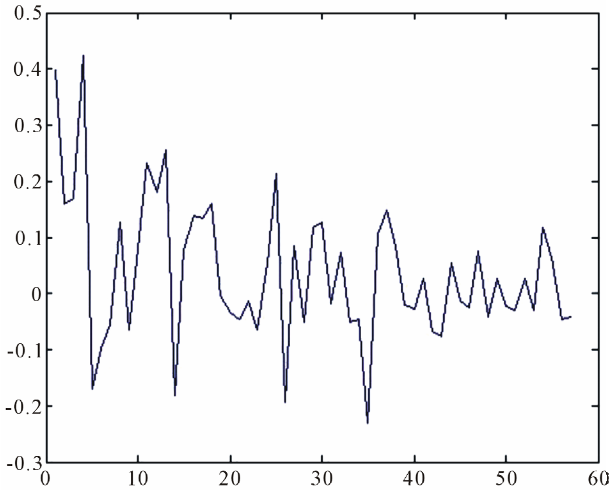

Figure 4. The direction of maximal influence curvature *dr*$_j$.

relation between the thickness of tumor and the sex, so we supposed that there was a relation between the coefficient β and t_j. Hence, we utilized the varying-coefficient model $y=\beta(t)x+\varepsilon$ to analyze these data. The results are as **Table 1** and **Figures 1-4**.

Figures 1 and **2** show that the first and the fourth data are the outlier, **Figures 3** and **4** show that the first and the fourth data are the outliers. Indeed, the diagnostic effect of the diagonal value is identical with the direction of maximal influence curvature and this result is similar to Li Yali [12].

REFERENCES

[1] R. D. Cook, "Assessment of Local Influence (with Discussion)," *Journal of the Royal Statistical Society*: *Series B*, Vol. 48, No. 2, 1986, pp. 133-169.

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

[3] B. C. Wei, G. B. Lu and J. Q. Shi, "Statistical Diagnostics," Publishing House of Southeast University, Nanjing, 1990.

[4] P. K. Andersen, O. Borgan, R. D. Gill and N. Keiding, "Statistical Models Based on Counting Processes," Springer-Verlag, New York, 1993. doi:10.1007/978-1-4612-4348-9

[5] L. A. Escobar and W. Q. Meeker, "Assessing Influence in Regression Analysis with Censored Data," *Biometrics*, Vol. 48, No. 2, 1992, pp. 507-528. doi:10.2307/2532306

[6] R. L. Eubank, "Diagnostics for Smoothing Spline," *Journal of the Royal Statistical Society*: *Series B*, Vol. 47, No. 1, 1985, pp. 322-341.

[7] R. L. Eubank, "The Hat Matrix for Smoothing Spline," *Statistics & Probability Letters*, Vol. 2, No. 1, 1984, pp. 9-16. doi:10.1016/0167-7152(84)90029-4

[8] R. L. Eubank and R. F. Gunst, "Diagnostic for Penalized Least-Squares Estimators," *Statistics & Probability Letters*, Vol. 4, No. 5, 1986, pp. 265-272. doi:10.1016/0167-7152(86)90101-X

[9] P. J. Green and B. W. Silverman, "Nonparametric Regression and Generalized Linear Models," Chapman and Hall, London, 1994.

[10] C. Kim, "Cook's Distance in Spline Smoothing," *Statistics & Probability Letters*, Vol. 31, No. 2, 1996, pp. 139-144. doi:10.1016/S0167-7152(96)00025-9

[11] Q. H. Wang, "Analysis of Survival Data," Science Press, Beijing, 2006.

[12] Y. L. Li, "Statistical Diagnostics of Partial Linear Model with Random Right Censorship," Nanjing University of Science and Technology, Nanjing, 2009.