第 35 卷 第 4 期 应 用 数 学 学 报 Vol. 35 No. 4
2012 年 7 月 ACTA MATHEMATICAE APPLICATAE SINICA July, 2012

纵向数据半参数 Beta 回归模型的影响分析 *

赵为华

(华东师范大学金融与统计学院，上海 200241)

(南通大学理学院，南通 226007)

(E-mail: zhaowhstat@163.com)

李泽安

(南通大学计算机学院，南通 226007)

徐相建

(南通大学理学院，南通 226007)

摘 要 本将随机效应当作是缺失数据，基于 Q 函数和 EM 算法并利用 P- 样条拟合非参数部分，得到了纵向数据半参数 Beta 回归模型估计方法. 基于数据删除模型，我们得到了模型参数部分的广义 Cook 距离以及非参数部分的广义 DFIT. 此外，本文还研究了在四种不同扰动情形下模型的局部影响分析，得到了相应的影响矩阵. 最后，我们通过两个数值实例验证了所得诊断统计量的有效性.

关键词 Beta 回归；纵向数据；半参数；影响分析； P- 样条； EM 算法

MR(2000) 主题分类 62G05

中图分类 O212.2

1 引言

纵向数据 (longitudinal data) 主要是指对同一对象或同一个个体 (试验单位) 在不同时刻上而得到的由截面和时间序列融合在一起的数据，其特点是将截面数据和时间序列数据结合在一起. 近年来，对纵向数据各种模型的研究引起了国内外统计学者的广泛

本文 2009 年 1 月 4 日收到. 2011 年 12 月 8 日收到修改稿.

* 国家自然科学基金 (11171112) 和南通大学自然科学基金 (10Z008) 资助项目.

关注，成为当今统计学的热点课题之一，并且它在生物医学、计量经济学和社会科学领域中都有着广泛的应用.

在回归分析中，若响应变量为比例的连续型数据，即其取值范围为 (0,1) 区间时，利用经典的线性回归模型或数据变换后的回归模型 (如 logistic 回归模型) 进行预测研究时常常都不尽人如意，原因在于其拟合值经常超出上下界，即预测值往往在 (0,1) 区间之外，或者模型中的参数不易根据原有的响应变量进行合理的解释. 此外，由于比例数据之间还往往表现为异方差，利用常规的线性回归模型进行统计分析，会带来很多问题，且分析、预测的效果比较差. 利用 Beta 回归模型对响应变量为比例的连续型数据进行建模分析，能克服以上问题，且拟合效果非常好，因而受到人们的青睐.

纵向数据半参数 Beta 回归模型是半参数混合效应模型和 Beta 回归模型的结合，它适合于响应变量是连续型比例数据的纵向数据分析. 对于此类模型的参数和非参数的估计关键在于条件期望的计算，本文根据 Zhu et al[1] 将随机效应看作缺失数据，进而引入算法，并在 E 步中使用 MCMC 方法来计算条件期望，再利用 P- 样条对非参数部分进行逼近，从而得到参数和非参数的估计.

Cook 于 1977 年和 1986 年先后提出了回归模型的全局影响分析方法 [2] 和局部影响分析法 [3], 成为统计诊断中最重要的两个方法. 全局影响分析法刻画了删除个别数据点对回归分析的影响；局部影响分析法是在某种扰动模式下，用基于似然函数替换的影响图的正则曲率以及相应的最大扰动方向以识别数据中的强影响点. 经过近三十年的发展，这两种统计诊断方法已推广至各种模型. Wing 等 [4] 对半参数混合效应模型的统计诊断进行了研究； Zhu[1] 和 Zhu, Lee[5] 分别研究了缺失数据模型和广义线性混合模型的影响分析；张浩，朱仲义 [6] 对半参数广义线性混合效应模型进行了影响分析；Ferrari 等 [7] 研究了 Beta 回归模型参数的极大似然估计，并给出了 Cook 距离和广义杠杆值这两个诊断统计量； Espinheira 等 [8] 以及李爱萍等 [9] 分别研究了 Beta 回归模型的影响诊断问题. 本文基于惩罚似然函数，将利用 Q 函数和 EM 算法，并用 P- 样条逼近非参数函数对纵向数据半参数 Beta 回归模型进行分析，基于惩罚完全对数似然函数的条件期望进行统计诊断和影响分析，并得到相应的诊断统计量，最后通过一个实例分析证实了我们提出方法的可行性和有效性.

2 纵向数据半参数 Beta 回归模型及其估计方法

2.1 纵向数据半参数 Beta 回归模型

假设第 i 个受试单元第 j 次的观察值 y_{ij} 关于随机效应 u_i 的条件密度为:

$$\begin{cases} p_{y_{ij}|u_i}(y_{ij}|u_i,\beta,f,\phi) \\ \quad = \dfrac{\Gamma(\phi)}{\Gamma(\mu_{ij}\phi)\Gamma((1-\mu_{ij})\phi)} y_{ij}{}^{\mu_{ij}\phi-1}(1-y_{ij})^{(1-\mu_{ij})\phi-1}, \qquad y_{ij}\in(0,1), \\ u_i \sim N\big(0,\Sigma(\gamma)\big), \\ \eta_{ij} = g(\mu_{ij}) = x_{ij}^T\beta + f(t_{ij}) + z_{ij}^T u_i, \end{cases} \tag{1}$$

其中 $\Gamma(\cdot)$ 为 Gamma 函数； $g(\cdot)$ 是已知的单调连续函数， $g^{-1}(\cdot)$ 是 $g(\cdot)$ 的反函数，在应用

中通常取 g 为 logistic 形式, 即 $g(x)=\log\{x/(1-x)\}$; y_{ij} 在给定 u_i 下的条件期望和条件方差为 $E(y_{ij}|u_i)=\mu_{ij}$, $\mathrm{Var}\ (y_{ij}|u_i)=\mu_{ij}(1-\mu_{ij})/(1+\phi)$, $i=1,2,\cdots,m$, $j=1,2,\cdots,n_i$ 且 $\sum_{i=1}^{m} n_i=n$, ϕ 是个散度参数; u_i 为 q 个随机效应子, 假设 u_i 独立同分布且服从 $\Sigma(\gamma),\gamma$ 是 $p_2\times 1$ 维未知方差成分参数, 不失一般性, 可设 $\Sigma(\gamma)=\sigma_u^2 I_q$, 此时 $\gamma=\sigma_u^2$; $f(\cdot)$ 是未知的单变量函数; $x_{ij}=(x_{ij1},\cdots,x_{ijp_1})^T\in R^{p_1}$ 和 $z_{ij}=(z_{ij1},\cdots,z_{ijq})^T\in R^q$, y_{ij} 和 t_{ij} 是已知非随机设计点列. 此外, $y_i=(y_{i1},\cdots,y_{in_i})^T$, $x_i=(x_{i1},\cdots,x_{in_i})$, t_i, u_i, z_i 类似定义. 我们称模型 (1) 为纵向数据半参数 Beta 回归模型.

纵向数据半参数 Beta 回归模型有着广泛应用, 但目前对于此模型的统计诊断问题还未见报道, 本文将具体研究此模型的影响分析问题. 为此, 先介绍模型中非参数的 P-样条估计.

2.2 非参数函数的 P- 样条估计

对于未知单变量函数 $f(\cdot)$, 本文采用 P- 样条估计. 由 [10] 知, 假设

$$f(t_{ij})=\delta_0+\delta_1 t_{ij}+\cdots+\delta_l t_{ij}^l+\sum_{r=1}^{K}\delta_{l+r}(t_{ij}-\kappa_r)_+^l, \tag{2}$$

其中 $\{\kappa_r\}_{r=1}^K$ 为 K 个样条节点, $l\geq 1$ 且为整数. Yu, Ruppert[10] 详细研究了节点的选取方法. 对于光滑函数, 通常情况下, 我们取预测变量的等分位点为节点, 一般取节点个数 $5\sim 10$ 个. 设样条系数为 $\delta=(\delta_0,\delta_1,\cdots,\delta_{l+K})^T$, 样条基为

$$B(t_{ij})=\big(1,t_{ij},\cdots,t_{ij}^l,(t_{ij}-\kappa_1)_+^l,\cdots,(t_{ij}-\kappa_K)_+^l\big)^T, \tag{3}$$

则函数 $f(\cdot)$ 的样条函数估计为 $\widehat{f}(t_{ij})=B^T(t_{ij})\delta$.

令 $B(t_i)=(B(t_{i1}),\cdots,B(t_{in_i}))$, 我们可将上述向量结合写成矩阵的形式, $X=(x_1,\cdots,x_m)^T$, $t=(t_1,\cdots,t_m)^T$, $y=(y_1^T,\cdots,y_m^T)^T$, $Z=\mathrm{diag}\,(z_1,\cdots,z_m)$, $u=(u_1^T,\cdots,u_m^T)^T$, $B(t)=(B(t_1),\cdots,B(t_m))^T$ 和 $f(t)=B(t)\delta$, 则可将模型 (1) 中的第 3 式写成如下矩阵形式:

$$\eta=g(\mu)=X\beta+B(t)\delta+Zu, \tag{4}$$

令 Y_o 表示观测数据集, 则纵向数据半参数 Beta 回归模型关于 φ 的惩罚对数似然函数为:

$$PL_o(\beta,\delta,\phi,\gamma|Y_o)=L_o(\beta,\delta,\phi,\gamma|Y_o)-\frac{1}{2}\lambda\int(\ddot{f}(t))^2\,\mathrm{d}t,\qquad \lambda>0, \tag{5}$$

$$L_o(\beta,\delta,\phi,\gamma|Y_o)=\sum_{i=1}^{m}\log\Big\{\prod_{j=1}^{n_i}p_{y_{ij}|u_i}(y_{ij}|u_i,\beta,\delta,\phi)\cdot\gamma^{-q/2}\exp\Big(-\frac{1}{2\gamma}u_i^T u_i\Big)\,\mathrm{d}u_i\Big\}, \tag{6}$$

其中 $\lambda>0$ 是光滑参数.

由 [10] 知, $\int(\ddot{f}(t))^2\,\mathrm{d}t$ 可以表示为 $\delta^T G\delta$, 其中 G 是与节点 t 有关的矩阵, 这里取

G 为对角矩阵, 且只取最后 K 个对角元素的值为 1, 其他为 0, 因此式 (5) 又可表示为:

$$PL_o(\beta,\delta,\phi,\gamma|Y_o)=L_o(\beta,\delta,\phi,\gamma|Y_o)-\frac{1}{2}\lambda\delta^TG\delta, \tag{7}$$

有关光滑参数可通过 GCV 方法选取，在具体计算时，用格子点方法获得最优的 λ .

2.3 模型的估计

由于式 (6) 计算相当复杂, 一般情况下难以得到显示表达式. 类似于 [1] 和 [6], 我们可利用 Q 函数和 EM 算法对纵向数据半参数 Beta 回归模型中的参数和非参数进行估计，而方差参数 γ 的估计可采用类似于 [11] 中 REML(residual maximum likelihood) 估计方法. 具体的做法是将随机效应 u_i 看作缺失数据 Y_m, 以 $Y_c=\{Y_o,Y_m\}$ 表示完全数据， $\varphi=(\beta^T,\delta^T,\phi)^T$ 为 $p\times 1$ $(p=p_1+p_2+l+K)$ 维有兴趣参数，则完全数据的惩罚对数似然函数为:

$$\begin{aligned}PL_c(\varphi|Y_c)=&\sum_{i=1}^{m}\Big\{\sum_{j=1}^{n_i}\big[\log\Gamma(\phi)-\log\Gamma(\phi\mu_{ij})-\log\Gamma(\phi(1-\mu_{ij}))\\&+(\phi\mu_{ij}-1)\log y_{ij}+(\phi(1-\mu_{ij})-1)\log(1-y_{ij})\big]-\frac{q}{2}\log\gamma-\frac{1}{2\gamma}u_i^Tu_i\Big\}\\&-\frac{1}{2}\lambda\delta^TG\delta,\end{aligned} \tag{8}$$

根据 [1], 我们可利用 EM 算法 [12] 求解式 (8), 标准的 EM 算法包含 E 步和 M 步, 给定初值 $\varphi^{(r)}$:

$$\begin{aligned}&E-\text{step}:\ Q_\varphi(\varphi|\varphi^{(r)})=E\big\{PL_c(\varphi|Y_c)|Y_o,\varphi^{(r)}\big\},\\&M-\text{step}:\ \varphi^{(r+1)}=\max Q_\varphi(\varphi|\varphi^{(r)}),\end{aligned} \tag{9}$$

其中 (r) 为迭代次数.

在一般的正则条件下, 迭代系列 $\{\varphi^{(r)}\}$ 能收敛到的极大似然估计 $\hat{\varphi}$[13]. 由上可知, 每次迭代计算 φ 的极大似然估计方程为:

$$\dot{Q}_\varphi(\varphi|\varphi^{(r)})=E\Big[\frac{\partial PL_c(\varphi|Y_c)}{\partial\varphi}\Big|Y_o,\varphi^{(r)}\Big]=0, \tag{10}$$

显然估计方程 (10) 中涉及到后验期望的计算，为了避免复杂的积分计算，我们采用 Metropolis 算法 [14] 得到条件分布 $u|y$ 的随机样本，该方法不需要 u_i 的具体分布，然后再利用 Monte Carlo 方法近似计算条件期望.

令 u_i 是前一次从条件分布 $u_i|y_i$ 中的抽样，用候选分布 (这里取为 $N(0,\gamma^{(r)}I_q)$) 重新生成 u_i 中第 k 个元素 u_{ik}^* . 假定 $u_i^*=(u_{i1},\cdots,u_{ik-1},u_{ik}^*,u_{ik+1},\cdots,u_{iq})$, 从 (0,1) 中随机抽取一个数 rand, 若 $A_k(u_i,u_i^*)<$ rand, 则接受 u_i^*, 否则我们保留 u_i, 这里 $A_k(u_i,u_i^*)$

为接受 u_i^* 的概率，经化简后为：

$$A_k(u_i,u_i^*)=\min\left(1,\frac{\prod_{j=1}^{n_i}p_{y_{ij}|u_i}(y_{ij}|u_i^*,\beta,\delta,\phi)}{\prod_{j=1}^{n_i}p_{y_{ij}|u_i}(y_{ij}|u_i,\beta,\delta,\phi)}\right),\tag{11}$$

基于估计方程 (10), 为得到 φ 的惩罚极大似然估计，我们可采用类似于 Beta 回归模型中参数估计的 New-Raphson 迭代算法. 为此，先计算 $PL_c(\varphi|Y_c)$ 关于 φ 的前二阶导数：

$$\frac{\partial PL_c(\varphi|Y_c)}{\partial\beta}=\phi X^TT(y^*-\mu^*),\qquad \frac{\partial PL_c(\varphi|Y_c)}{\partial\delta}=\phi B(t)^TT(y^*-\mu^*)-\lambda G\delta,$$
$$\frac{\partial PL_c(\varphi|Y_c)}{\partial\beta}=a,\qquad -\frac{\partial^2PL_c(\varphi|Y_c)}{\partial\beta\partial\beta^T}=\phi X^TWX,\qquad -\frac{\partial^2PL_c(\varphi|Y_c)}{\partial\beta\partial\delta^T}=\phi X^TWB(t),$$
$$-\frac{\partial^2PL_c(\varphi|Y_c)}{\partial\beta\partial\phi}=\phi X^TTc,\qquad -\frac{\partial^2PL_c(\varphi|Y_c)}{\partial\delta\partial\delta^T}=\phi B(t)^TWB(t)+\lambda G,$$
$$-\frac{\partial^2PL_c(\varphi|Y_c)}{\partial\beta\partial\phi}=B(t)^TTc,\qquad -\frac{\partial^2PL_c(\varphi|Y_c)}{\partial^2\phi}=\mathrm{tr}\,(D),$$

其中

$$y_{ij}^*=\log\{y_{ij}/(1-y_{ij}\},\qquad y_i^*=(y_{i1}^*,\cdots,y_{in_i}^*)^T,$$
$$\mu_{ij}^*=\psi(\phi\mu_{ij})-\psi(\phi(1-\mu_{ij})),\qquad \mu_i^*=(\mu_{i1}^*,\cdots,\mu_{in_i}^*)^T,$$
$$\psi(z)=d\Gamma(z)/dz,\qquad T=\mathrm{diag}\,(T_1,\cdots,T_m),\qquad T_i=(1/\dot g(\mu_{i1}),\cdots,1/\dot g(\mu_{in_i})),$$
$$W=\mathrm{diag}\,(w_1,\cdots,w_m),\qquad w_i=(w_{i1},\cdots,w_{im}),$$
$$w_{ij}=\phi\{\dot\psi(\mu_{ij}\phi)+\dot\psi((1-\mu_{ij})\phi)\}/\dot g^2(\mu_{ij}),$$
$$c_i=(c_{i1},\cdots,c_{in_i}),\quad c_{ij}=\phi\{\dot\psi(\mu_{ij}\phi)\mu_{ij}-\dot\psi((1-\mu_{ij})\phi)(1-\mu_{ij})\},\quad c=(c_1,\cdots,c_m)^T,$$
$$d_i=(d_{i1},\cdots,d_{in_i}),\qquad d_{ij}=\phi\{\dot\psi(\mu_{ij}\phi)\mu_{ij}^2-\dot\psi((1-\mu_{ij})\phi)(1-\mu_{ij})^2-(\dot\psi)\},$$
$$D=\mathrm{diag}\,\{d_1,\cdots,d_m\},$$
$$a=\sum_{i=1}^m\sum_{j=1}^{n_i}a_{ij},\qquad a_{ij}=\mu_{ij}(y_{ij}^*-\mu_{ij}^*)+\log(1-y_{ij})-\psi((1-\mu_{ij})\phi)+\psi(\phi).$$

由上可得模型中兴趣参数 φ 的 New-Raphson 迭代公式：

$$\begin{pmatrix}\widehat\beta^{(r+1)}\\ \widehat\delta^{(r+1)}\\ \widehat\phi^{(r+1)}\end{pmatrix}=\begin{pmatrix}\widehat\beta^{(r)}\\ \widehat\delta^{(r)}\\ \widehat\phi^{(r)}\end{pmatrix}+\left(E\begin{pmatrix}\phi X^TWX & \phi X^TWB(t) & X^TTc\\ \phi B(t)^TWX & \phi B(t)^TWB(t)+\lambda G & B(t)^TTc\\ c^TTX & c^TTB(t) & \mathrm{tr}\,(D)\end{pmatrix}\Big|Y_o\right)^{-1}$$
$$\times\left(E\begin{pmatrix}\phi X^TT(y^*-\mu^*)\\ \phi B(t)^TT(y^*-\mu^*)\\ a\end{pmatrix}\Big|Y_o\right)\Big|_{\widehat\varphi^{(r)}}.$$

3　基于数据删除的影响分析

3.1 个体删除模型的影响分析

在本节中，我们侧重于讨论基于数据点和数据组删除的影响分析，主要研究特定的某几个，特别是某一个数据点对于统计分析的影响. 数据删除模型是常用的影响分析方法之一. 由于本文所讨论的是基于纵向数据下的半参数 Beta 回归模型，因此我们对模型讨论个体删除和组删除两种情形，并分别研究了参数和非参数部分的影响分析. 在下文中，凡是带有下标 "$[i]$" 或 "$[ij]$" 的量表示将原来数据中第 i 组数据或第 i 组中第 j 个数据予以删除.

在影响分析的讨论中，根据 [1], 我们定义的 Q 函数为以下形式：

$$Q(\varphi|\widehat{\varphi}) = E\big\{PL_c(\varphi|Y_c)|Y_o, \widehat{\varphi}\big\}, \tag{12}$$

其中 $\widehat{\varphi}$ 为 φ 的惩罚极大似然估计.

设 $PL_c(\varphi|Y_{c[ij]})$ 是删除了模型中第 i 组第 j 个观测值后所得到的完全数据的惩罚对数似然函数，相应的 Q 函数为 $Q_{[ij]}(\varphi|\widehat{\varphi}) = E\{PL_c(\varphi|Y_{c[ij]})|Y_o, \widehat{\varphi}\}$, 且假定 $\widehat{\varphi}$ 和 $\widehat{\varphi}_{[ij]}$ 分别是 $Q(\varphi|\widehat{\varphi})$ 和 $Q_{[ij]}(\varphi|\widehat{\varphi})$ 达到最大值时的取值. 如果 $\widehat{\varphi}$ 和 $\widehat{\varphi}_{[ij]}$ 相差很大，则我们认为第 i 组第 j 个观测值为强影响点. 实际计算中，如果对于每一个要进行迭代计算，则计算量非常大，因此我们根据 [14], 采用 $\widehat{\varphi}_{[ij]}$ 的一步近似 $\widehat{\varphi}^1_{[ij]}$ 来减少计算量，

$$\widehat{\varphi}^1_{[ij]} = \widehat{\varphi} + \{-\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})\}^{-1}\dot{Q}_{[ij]}(\widehat{\varphi}|\widehat{\varphi}), \tag{13}$$

其中

$$\dot{Q}_{[ij]}(\widehat{\varphi}|\widehat{\varphi}) = \partial\dot{Q}_{[ij]}(\varphi|\widehat{\varphi})/\partial\varphi\big|_{\varphi=\widehat{\varphi}}, \qquad \ddot{Q}(\widehat{\varphi}|\widehat{\varphi}) = \partial^2\dot{Q}(\varphi|\widehat{\varphi})/\partial\varphi\partial\varphi^T\big|_{\varphi=\widehat{\varphi}}.$$

由上一节，我们进一步可得：

$$\ddot{Q}\big\{(\widehat{\varphi}|\widehat{\varphi})\big\} = -E\left(\left(\begin{array}{ccc} \phi X^TWX & \phi X^TWB(t) & X^TTc \\ \phi B(t)^TWX & \phi B(t)^TWB(t)+\lambda G & B(t)^TTc \\ c^TTX & c^TTB(t) & \mathrm{tr}\,(D) \end{array}\right)\Big|Y_o\right)\Big|_{\widehat{\varphi}}, \tag{14}$$

$$\ddot{Q}_{[ij]}\big\{(\widehat{\varphi}|\widehat{\varphi})\big\} = -E\left(\left(\begin{array}{c} \phi x_{ij}T_{ij}(y^*_{ij}-\mu^*_{ij}) \\ \phi B(t_{ij})T_{ij}(y^-_{ij}*\mu^*_{ij}) \\ a_{ij} \end{array}\right)\Big|Y_o\right)\Big|_{\widehat{\varphi}}. \tag{15}$$

由于 $\widehat{\varphi}_{[ij]} - \widehat{\varphi}$ 是个向量，不能定量地表达影响的大小，因此，我们仿照线性模型下的 Cook 距离，基于 Q 函数构造如下的广义 Cook 距离：

$$GD_{ij}(\varphi) = (\widehat{\varphi}_{[ij]} - \widehat{\varphi})^T\big\{-\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})\big\}(\widehat{\varphi}_{[ij]} - \widehat{\varphi}), \tag{16}$$

由式 (13), 可得其一步近似公式：

$$GD^1_{ij}(\varphi) = \dot{Q}_{[ij]}(\widehat{\varphi}|\widehat{\varphi})^T\big\{-\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})\big\}^{-1}\dot{Q}_{[ij]}(\widehat{\varphi}|\widehat{\varphi}) \tag{17}$$

为了进一步分析造成强影响点的原因，我们通常分别研究参数和非参数对模型的影响分析，下面先给出参数部分的广义 Cook 距离：

$$GD_{ij}(\varphi_1) = (\widehat{\varphi}_{1[ij]} - \widehat{\varphi}_1)^T\big\{-\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})\big\}(\widehat{\varphi}_{1[ij]} - \widehat{\varphi}_1), \tag{18}$$

其中 $\varphi_1^T=(\beta^T,\phi)$, $\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})=\partial^2\dot{Q}(\varphi|\widehat{\varphi})/\partial\varphi_1\partial\varphi_1^T|_{\varphi=\widehat{\varphi}}$.

对于非参数部分 f, 同样可以构造非参数的广义 Cook 距离，但是由于样条拟合的性质，我们更加倾向于采用删除一个观测点的方法来对非参数进行影响分析. 根据 [15], 以及非参数的渐近性质，我们得到以下公式来做非参数部分的影响分析：

$$\mathrm{DFIT}_{ij}=|d_c(\widehat{f}_{[ij]}-\widehat{f})|/s_{ij}, \tag{19}$$

其中 d_c 为 n 维且第 $c=\sum_{k=1}^{i-1}n_k+j$ 个元素为 1 其余元素为 0 的向量， s_{ij} 为矩阵 $B(t)\{-\ddot{Q}_\delta(\widehat{\varphi}|\widehat{\varphi})\}^{-1}B^T(t)$ 中第 c 个对角元素， $\ddot{Q}_\delta(\widehat{\varphi}|\widehat{\varphi})=\partial^2\dot{Q}(\varphi|\widehat{\varphi})/\partial\delta\partial\delta^T$. 结合 (13), 可以分别得到参数部分广义 Cook 距离和非参数部分 DFIT 的一步近似公式为：

$$GD_{ij}^1(\varphi_1)=\dot{Q}_{[ij]}(\widehat{\varphi}|\widehat{\varphi})^T\{-\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})\}^{-1}A_1^TA_1\{-\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})\}A_1^TA_1\{-\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})\}\dot{Q}_{[ij]}(\widehat{\varphi}|\widehat{\varphi}), \tag{20}$$

$$\mathrm{DFIT}_{ij}^1=|\{-\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})\}\dot{Q}_{[ij]}(\widehat{\varphi}|\widehat{\varphi})|/s_{ij}, \tag{21}$$

其中

$$A_1=\begin{pmatrix}I_{p_1}&0&0\\0&0_{(l+K)}&0\\0&0&1\end{pmatrix},\qquad A_2=(0_{(l+K)\times p_1}\ I_{(l+K)\times(l+K)}\ 0_{(l+K)\times 1}).$$

3.2 组删除模型的影响分析

上面我们基于删除个体作了诊断分析，但在纵向数据中，由于来自同一组的观测值, 它们的协变量是相同的, 因此很有必要研究删除一组的影响诊断. 当观察数目 n 变大时，单个数据点对整体的影响相对变小，而一组数据对整体的影响相对明显. Banerjee, Frees[15] 和 Wing 等 [4] 指出前者对模型的敏感性分析方面比较适合，而后者对模型的异常点分析方面更有用. 对于组删除模型，我们同样可以推导出一步近似公式：

$$\widehat{\varphi}_{[i]}^1=\widehat{\varphi}+\big\{-\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})\big\}^{-1}\dot{Q}_{[i]}(\widehat{\varphi}|\widehat{\varphi}), \tag{22}$$

其中

$$\dot{Q}_{[i]}(\widehat{\varphi}|\widehat{\varphi})=\partial\dot{Q}_{[i]}(\varphi|\widehat{\varphi})/\partial\varphi|_{\widehat{\varphi}}=\sum_{j=1}^{n_i}\partial\dot{Q}_{[ij]}(\varphi|\widehat{\varphi})/\partial\varphi|_{\widehat{\varphi}}=-E\left(\begin{pmatrix}\phi x_iT_i(y_i^*-\mu_i^*)\\ \phi B(t_i)T_i(y_i^*-\mu_i^*)\\ \sum_{j=1}^{n_i}a_{ij}\end{pmatrix}\Big|Y_o\right).$$

同上，我们有以下参数部分广义距离的一步近似估计：

$$GD_i^1(\varphi_1)=\dot{Q}_{[i]}^T(\widehat{\varphi}|\widehat{\varphi})\{-\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})\}^{-1}A_1^TA_1\{-\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})\}A_1^TA_1\{-\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})\}\dot{Q}_{[i]}(\widehat{\varphi}|\widehat{\varphi}). \tag{23}$$

由于删除第 i 组中有 n_i 个数据，我们需要对这 n_i 个点分别计算 DFIT, 从而得到曲线 f 的影响度量，这个相当于计算 $B(t_i)\widehat{\delta}$ 与 $B(t_i)\widehat{\delta}_{[i]}$ 之间的 Cook 距离：

$$GD_i(f)=(B(t_i)\widehat{\delta}-B(t_i)\widehat{\delta}_{[i]})^T(B(t_i)\{-\ddot{Q}_\delta(\widehat{\varphi}|\widehat{\varphi})\}^{-1}B^T(t_i))^{-1}(B(t_i)\widehat{\delta}-B(t_i)\widehat{\delta}_{[i]}).$$

相应的一步近似为：

$$GD_i^1(f) = \dot{Q}_{[i]}^T(\widehat{\varphi}|\widehat{\varphi})\{-\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})\}^{-1}A_2^T B^T(t_i)(B(t_i)\{-\ddot{Q}_\delta(\widehat{\varphi}|\widehat{\varphi})\}^{-1}B^T(t_i))^{-1} \cdot B(t_i)A_2\{-\ddot{Q}(\widehat{\varphi}|\widehat{\varphi})\}\dot{Q}_{[i]}(\widehat{\varphi}|\widehat{\varphi}). \tag{24}$$

4 局部影响分析

4.1 局部影响分析简介

局部影响分析是一种广义的影响分析方法，在以往的线性模型中，我们先对模型 M 进行微小的扰动得到扰动模型 $M(\omega)$, 然后通过研究影响曲率和影响图，最后可以得到使得似然距离产生最大局部变化的方向，这个方向是影响分析中最值得人们关心的统计量. 此外，影响图比 Cook 距离提供了更多的关于扰动对于模型影响的信息. 在本节中，我们结合完全惩罚对数似然函数和上节中定义的 Q 函数，研究半参数 Beta 回归模型的局部影响分析，并推导出相关的诊断统计量.

设 $\omega = (\omega_1, \cdots, \omega_k)$ 是一个定义在 $\Omega \subset R^k$ 上的 k 维向量, 表示对模型的扰动因素. 令 $PL_c(\varphi, \omega|Y_c)$ 为扰动模型的完全惩罚对数似然函数. 假定存在 ω_0 使得 $PL_c(\varphi, \omega_0|Y_c) = PL_c(\varphi|Y_c)$ 和 $PL_o(\varphi, \omega_0|Y_o) = PL_o(\varphi|Y_o)$ 对所有 φ 都成立. 设 $\widehat{\varphi}$ 和 $\widehat{\varphi}(\omega)$ 分别使得 Q 函数 $Q(\varphi|\widehat{\varphi}) = E\{PL_c(\varphi, \omega_0|Y_c)|Y_o, \widehat{\varphi}\}$ 和 $Q(\varphi, \omega|\widehat{\varphi}(\omega)) = E\{PL_c(\varphi, \omega_0|Y_c)|Y_o, \widehat{\varphi}(\omega)\}$ 达到最大值.

根据 [6], 我们构造半参数 Beta 回归模型下的 Q 距离函数：

$$f_Q(\omega) = 2\big\{Q(\widehat{\varphi}|\widehat{\varphi}) - Q(\widehat{\varphi}(\omega)|\widehat{\varphi})\big\}, \tag{25}$$

其中 $Q(\widehat{\varphi}|\widehat{\varphi}) = Q(\widehat{\varphi}, \omega_0|\widehat{\varphi})$. 同样，我们考虑影响图 $\eta(\omega) = (\omega^T, f_Q(\omega))^T$, 则其影响曲率为：

$$C_{f_Q,h} = -2h^T\ddot{Q}_{\omega_0}h = -2h^T\Delta_{\omega_0}^T\{\ddot{Q}_\varphi(\widehat{\varphi})\}^{-1}\Delta_{\omega_0}h, \tag{26}$$

其中

$$\ddot{Q}_{\omega_0} = \frac{\partial^2 Q(\widehat{\varphi}(\omega)|\widehat{\varphi})}{\partial\omega\partial\omega^T}\Big|_{\omega_0} = \Delta_{\omega_0}^T\{\ddot{Q}_\varphi(\widehat{\varphi})\}^{-1}\Delta_{\omega_0}, \qquad \Delta_{\omega_0} = \frac{\partial^2 Q(\varphi, \omega|\widehat{\varphi})}{\partial\phi\partial\omega}.$$

令 $h_{\max}$ 为 $\ddot{Q}_{\omega_0}$ 的最大特征值所对应的最大特征向量，则 $h_{\max}$ 即为 $C_{f_Q,h}$ 的最大影响曲率. 以 $(h_{\max})_i$, $i = 1, \cdots, k$ 为指标值的散点图能够很好的反映模型的扰动情况.

在下面，我们将分别研究四种扰动情形：组内加权扰动、组间加权扰动、响应变量扰动和散度参数扰动，并求出相应的影响矩阵.

4.2 组内加权扰动

在不考虑数据结构的情况下，我们倾向于在所有观测数据中寻找强影响点，常用的方法就是给每个数据加权. 令 $\omega = (\omega_{11}, \cdots, \omega_{1n_1}, \omega_{21}, \cdots, \omega mn_m)^T$ 为 n 维扰动向量，当

$\omega_0 = 1_{n\times 1}$ 时，模型为无扰动模型，则组内加权扰动模型的似然函数可表示为：

$$\begin{aligned} PL_c(\varphi|Y_c) = &\sum_{i=1}^{m}\Big\{\sum_{j=1}^{n_i}\omega_{ij}\big[\log\Gamma(\phi)-\log\Gamma(\phi\mu_{ij})-\log\Gamma(\phi(1-\mu_{ij}))\\ &+(\phi\mu_{ij}-1)\log y_{ij}+(\phi(1-\mu_{ij})-1)\log(1-y_{ij})\big]\\ &-\frac{q}{2}\log\gamma-\frac{1}{2\gamma}u_i^Tu_i\Big\}-\frac{1}{2}\lambda\delta^TG\delta, \end{aligned} \tag{27}$$

通过对上式求导我们有：

$$\Delta_{\omega_{ij}} = E\Big(\frac{\partial^2 PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\beta^T},\frac{\partial^2 PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\delta^T},\frac{\partial^2 PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\varphi}\Big)^T,$$

其中

$$\begin{aligned} &\frac{\partial^2 PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\beta^T} = \phi x_{ij}T_{ij}(y_{ij}^*-\mu_{ij}^*),\\ &\frac{\partial^2 PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\delta^T} = \phi x_{ij}T_{ij}(y_{ij}^*-\mu_{ij}^*),\frac{\partial^2 PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\varphi} = a_{ij}. \end{aligned}$$

由此可得 $\Delta_\omega = (\Delta_{\omega_{11}},\cdots,\Delta_{\omega_{mn_m}})$.

4.3 组间加权扰动

组间加权模型通常用来寻找强影响数据组或异常数据组，令 $\omega = (\omega_1,\cdots,\omega_m)^T$ 为 m 维扰动向量，当 $\omega_0 = (1,\cdots,1)^T$ 时，模型为无扰动模型，则组内加权扰动模型的似然函数可表示为：

$$\begin{aligned} PL_c(\varphi|Y_c) = &\sum_{i=1}^{m}\omega_i\Big\{\sum_{j=1}^{n_i}\big[\log\Gamma(\phi)-\log\Gamma(\phi\mu_{ij})-\log\Gamma(\phi(1-\mu_{ij}))\\ &+(\phi\mu_{ij}-1)\log y_{ij}+(\phi(1-\mu_{ij})-1)\log(1-y_{ij})\big]-\frac{q}{2}\log\gamma-\frac{1}{2\gamma}u_i^Tu_i\Big\}\\ &-\frac{1}{2}\lambda\delta^TG\delta, \end{aligned} \tag{28}$$

通过对上式求导我们有：

$$\Delta_{\omega_i} = E\Big(\frac{\partial^2 PL_c(\varphi,\omega|Y_c)}{\partial\omega_i\partial\beta^T},\frac{\partial^2 PL_c(\varphi,\omega|Y_c)}{\partial\omega_i\partial\delta^T},\frac{\partial^2 PL_c(\varphi,\omega|Y_c)}{\partial\omega_i\partial\varphi}\Big)^T,$$

其中

$$\begin{aligned} &\frac{\partial^2 PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\beta} = \phi x_{ij}T_{ij}(y_{ij}^*-\mu_{ij}^*),\\ &\frac{\partial^2 PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\delta} = \phi x_{ij}T_{ij}(y_{ij}^*-\mu_{ij}^*),\frac{\partial^2 PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\varphi} = \sum_{j=1}^{n_i}a_{ij}. \end{aligned}$$

由此可得 $\Delta_\omega = (\Delta_{\omega_1},\cdots,\Delta_{\omega_m})$.

4.4 响应变量扰动

令 $\omega=(\omega_{11},\cdots,\omega_{1n_1},\omega_{21},\cdots,\omega_{mn_m})^T$ 为 n 维扰动向量，则响应变量扰动模型的似然函数可表示为：

$$\begin{aligned}PL_c(\varphi|Y_c)=&\sum_{i=1}^{m}\Big\{\sum_{j=1}^{n_i}\big[\log\Gamma(\phi)-\log\Gamma(\phi\mu_{ij})-\log\Gamma(\phi(1-\mu_{ij}))\\&+(\phi\mu_{ij}-1)\log(y_{ij}+\omega_{ij})+(\phi(1-\mu_{ij})-1)\log(1-(y_{ij}+\omega_{ij}))\big]\\&-\frac{q}{2}\log\gamma-\frac{1}{2\gamma}u_i^Tu_i\Big\}-\frac{1}{2}\lambda\delta^TG\delta,\end{aligned}\tag{29}$$

当 $\omega_0=0_{n\times1}$ 时，模型为无扰动模型．通过对上式求导我们有：

$$\Delta_{\omega_{ij}}=E\Big(\frac{\partial^2PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\beta^T},\frac{\partial^2PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\delta^T},\frac{\partial^2PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\varphi}\Big)^T,$$

其中

$$\begin{aligned}&\frac{\partial^2PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\beta^T}=\phi x_{ij}T_{ij}s_{ij},\qquad\frac{\partial^2PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\delta^T}=\phi x_{ij}T_{ij}s_{ij},\\&\frac{\partial^2PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\varphi}=-(y_{ij}-\mu_{ij})s_{ij}.\end{aligned}$$

由此可得 $\Delta_\omega=(\Delta_{\omega_{11}},\cdots,\Delta_{\omega_{mn_m}})$.

4.5 散度参数 ϕ 扰动

前面我们均假定散度参数 ϕ 为常数，而实际情况散度参数可能会随着观测点的不同而变化，即散度参数 ϕ 发生扰动．令 $\omega=(\omega_{11},\cdots,\omega_{1n_1},\omega_{21},\cdots,\omega_{mn_m})^T$ 为 n 维扰动向量，设 $\phi_{ij}(\omega)=\omega_{ij}\cdot\phi$, 当 $\omega_0=0_{n\times1}$ 时，模型为无扰动模型．同上我们可得：

$$\Delta_{\omega_{ij}}=E\Big(\frac{\partial^2PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\beta^T},\frac{\partial^2PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\delta^T},\frac{\partial^2PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\varphi}\Big)^T,$$

其中

$$\begin{aligned}&\frac{\partial^2PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\beta^T}=\phi x_{ij}T_{ij}(y_{ij}^*-\mu_{ij}^*-c_{ij}),\qquad\frac{\partial^2PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\delta^T}=\phi x_{ij}T_{ij}(y_{ij}^*-\mu_{ij}^*-c_{ij}),\\&\frac{\partial^2PL_c(\varphi,\omega|Y_c)}{\partial\omega_{ij}\partial\varphi}=a_{ij}-\phi d_{ij}.\end{aligned}$$

由此可得 $\Delta_\omega=(\Delta_{\omega_{11}},\cdots,\Delta_{\omega_{mn_m}})$.

5 实例分析

为了评价和说明本文所提出的诊断统计量的有效性，本节通过两个数值实例来进行分析研究.

5.1 仿真模拟数据

设定 $\beta=-0.5,\ \phi=3,\ \gamma=0.25$, 由计算机随机模拟产生 80 个数据：

$$\eta_{ij}=x_{ij}\beta+f(t_{ij})+u_i,\qquad i=1,\cdots,20,\ \ j=1,\cdots,4,$$

其中 $x_{ij}\sim U(-1,1),\ f(t_{ij})=4t_{ij}(1-t_{ij}),\ t_{ij}\sim U(0,1),\ u_i\sim N(0,\gamma)$.

由 η_{ij} 相应生成 $\mu_{ij}=g^{-1}(\eta_{ij})=\exp(\eta_{ij})/(1+\exp(\eta_{ij}))$. 由 μ_{ij} 和 ϕ, 生成 80 个响应变量数据 $y_{ij}\sim\beta(\mu_{ij}\phi,(1-\mu_{ij})\phi)$, 此处 $m=20,\ n_1=n_2=\cdots=n_m=4$.

下面我们首先利用本文提出的估计方法对该仿真模拟纵向数据进行分析，得惩罚极大似然估计为：$\widehat{\beta}=-0.4207(0.1872),\ \widehat{\phi}=3.2062(0.4351),\ \widehat{\gamma}=0.2108(0.0925)$(括号中的数为标准差), 光滑参数 $\lambda=0.0029$. 非参数函数的样条估计情况见图 1(d), 估计曲线与真实曲线非常接近.

为检验我们所提出的影响分析方法的有效性，我们将响应变量的第 4 号和第 27 号点都加上 0.2, 即事先设定第 4 号和 27 号数据点是异常点. 对修改后的数据进行惩罚极大似然估计得：$\widehat{\beta}=-0.3882(0.1928),\ \widehat{\phi}=2.6819(0.3827)$, 其数据删除模型诊断统计量的散点图见图 1(a),(b),(c) .

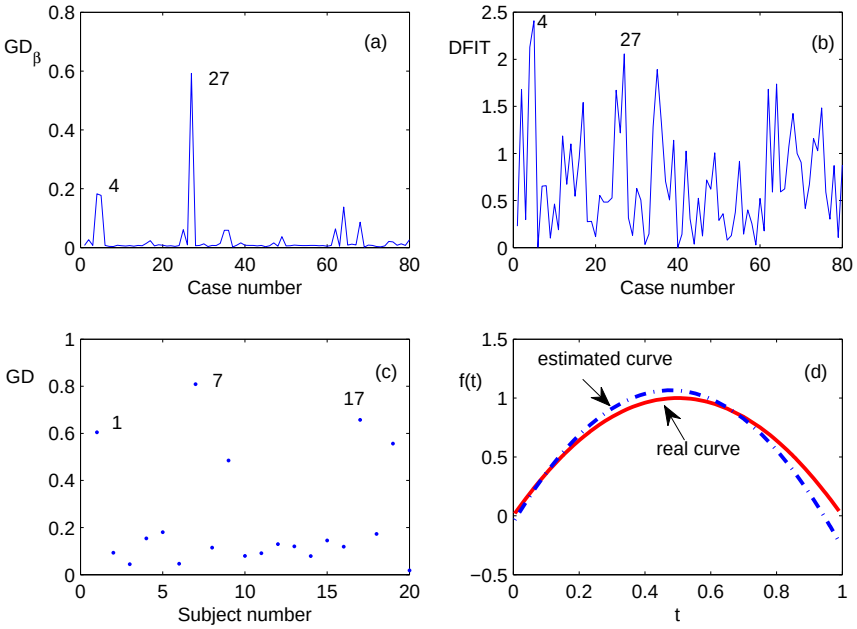

图 1　仿真数据的广义 Cook 距离

图 1(a) 和 (b) 分别是个体删除模型下的参数部分的广义 Cook 距离和非参数部分的 DFIT. 图 1(c) 是组删除模型下的广义 Cook 距离.

从图 1(a) 和 (b) 中看出，第 4 号和 27 数据点确实是强影响点. 从图 1(c) 看出，强

影响点第 27 数据点所在的第 7 组和第 4 数据点所在的第 1 组是强影响组，同时发现第 17 组数据组是强影响组，但第 17 组的每一个数据点并不是影响显著的.

图 2(a), (b), (c), (d) 分别是模型在组内扰动、响应变量扰动、组间扰动和散度参数扰动情形下的散点图.

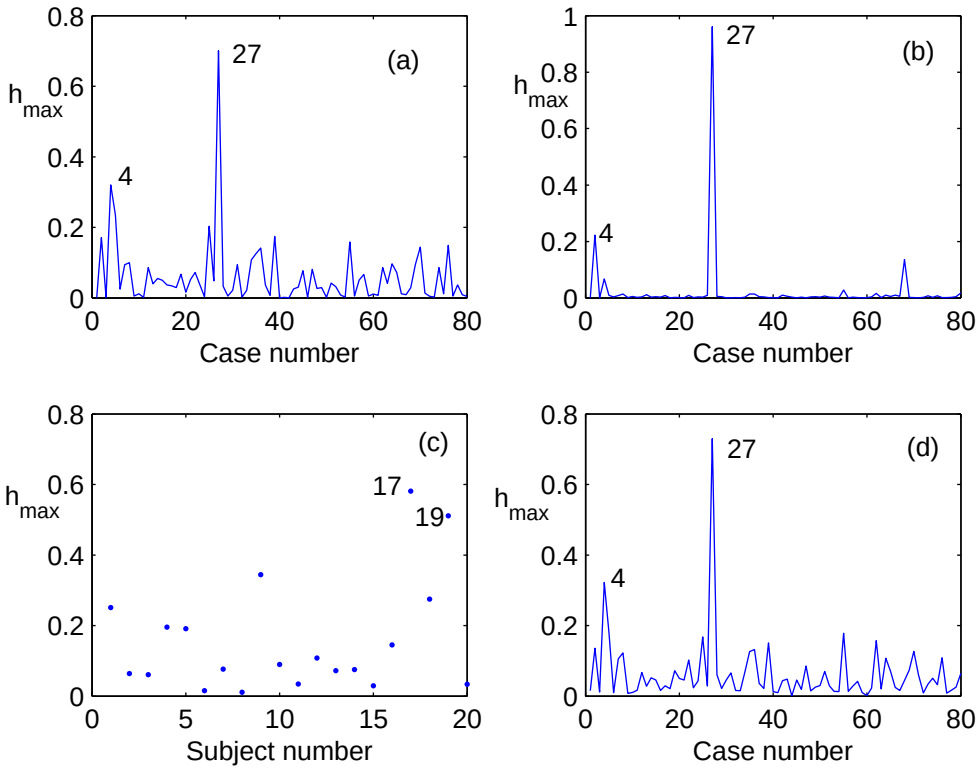

图 2 仿真数据的局部影响分析

从图中看出，第 27 号点，第 4 号对三种扰动情形都比较敏感，但由图 2(c) 看出这两个点所在的组并不是强影响组，反而第 17 组和第 19 组是影响最强的两个组. 图 2 所反映的信息与图 1 反映的情况有很多类似之处，但也有不同的地方.

从以上分析可知，对纵向数据模型进行影响分析时，一定要同时考虑数据点和数据组对模型的影响情况，否则某些数据点 (数据组) 的影响信息可能被屏蔽.

5.2 家庭食物消费数据 (Household food expenditure data)

家庭食物消费数据来自 Griffiths 等 [16]. 此数据是从美国的某个大城市中随机抽取 38 个家庭而得到的. 此数据中有两个自变量 x_1 和 x_2 分别表示某个家庭中的人数和该家庭的收入水平，响应变量是表示某个家庭食物消费支出占整个家庭收入的比例. 对此数据感兴趣的是食物消费支出比例与家庭人口和收入水平的关系. 由于此数据中响应

变量是比例数据，取值在 (0,1) 范围内，并且此数据经检验存在明显的异方差，因此利用常规的回归模型进行分析，效果很差． Ferrari 等 [7] 曾利用 Beta 回归模型对此数据进行分析，得到拟合的均方误差为 0.0053．这里我们利用半参数 Beta 回归模型对此数据进行再次分析，以说明本文所得到的诊断统计量的有效性.

$$y_i = \beta(\mu_i\phi, (1-\mu_i)\phi), \qquad i = 1, \cdots, 38,$$

这里取 $K = 5$ 个节点， $l = 1$.

$$\eta_i = \log\frac{\mu_i}{(1-\mu_i)} = x_{i1}\beta + \delta_0 + \delta_1 x_{i2} + \sum_{r=1}^{K}\delta_{r+1}(x_{i2} - \kappa_r)_+^l$$

经编程计算的参数的惩罚极大似然估计 $\widehat{\beta} = 0.1049(0.0325)$, $\widehat{\phi} = 49.2187(11.2013)$,(括号中的数为标准差), 光滑参数 $\lambda = 0.2848$, 均方误差为 0.0040. 非参数的估计情况见图 3.

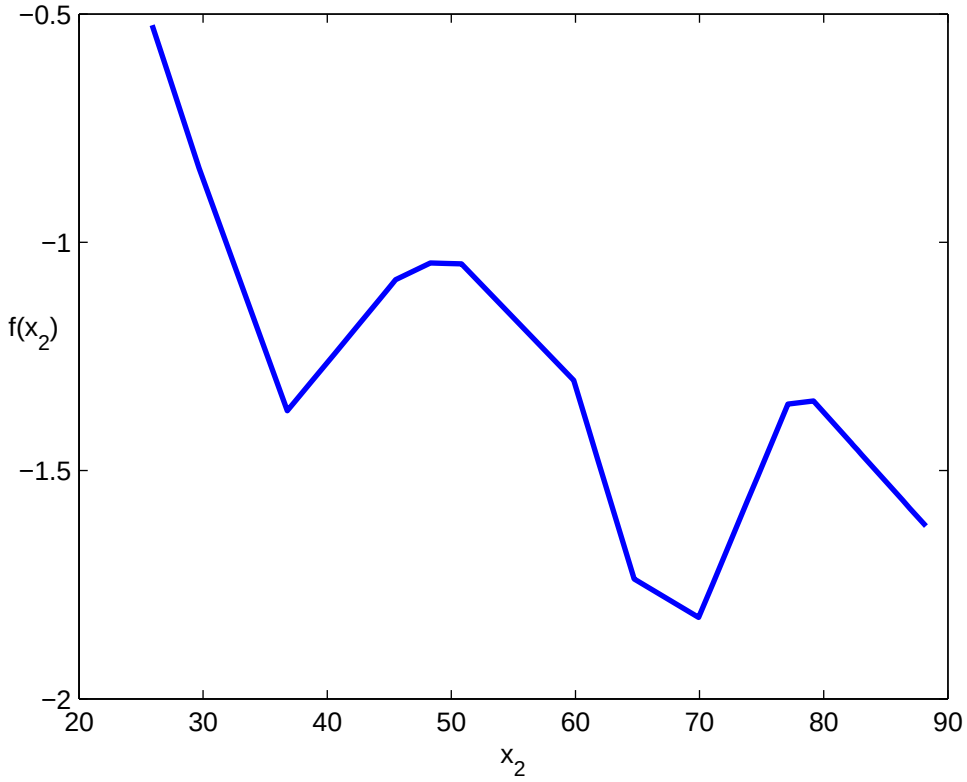

图 3　家庭消费数据的非参数估计

从图 3 中可以明显地看出食物消费结构总的变化规律：一个家庭人员固定的情形下，家庭收入越少，家庭收入中用来购买食物的支出所占的比例就越大，随着家庭收入的增加，家庭收入中用来购买食物的支出则会下降．同时，这样的下降趋势并不是完全

单调的，跟一个家庭随着收入水平的提高，受家庭消费食物结构的变化、城市化程度、食品加工、饮食业等因素影响，所占比重上升一段时期后，呈递减趋势．图 3 的估计曲线完全反映了经济学上恩格尔定律与最终食物消费变化的关系．如果用纯粹的参数的 Beta 回归模型 [7] 进行拟合，得不到食物消费结构总的变化规律，并且拟合的均方误差比本文大．

下面我们考虑在三种扰动模型：加权扰动、响应变量扰动和散度参数扰动情形下的最大影响曲率方向，见图 4(b)–(d), 图 4(a) 是估计的残差图．

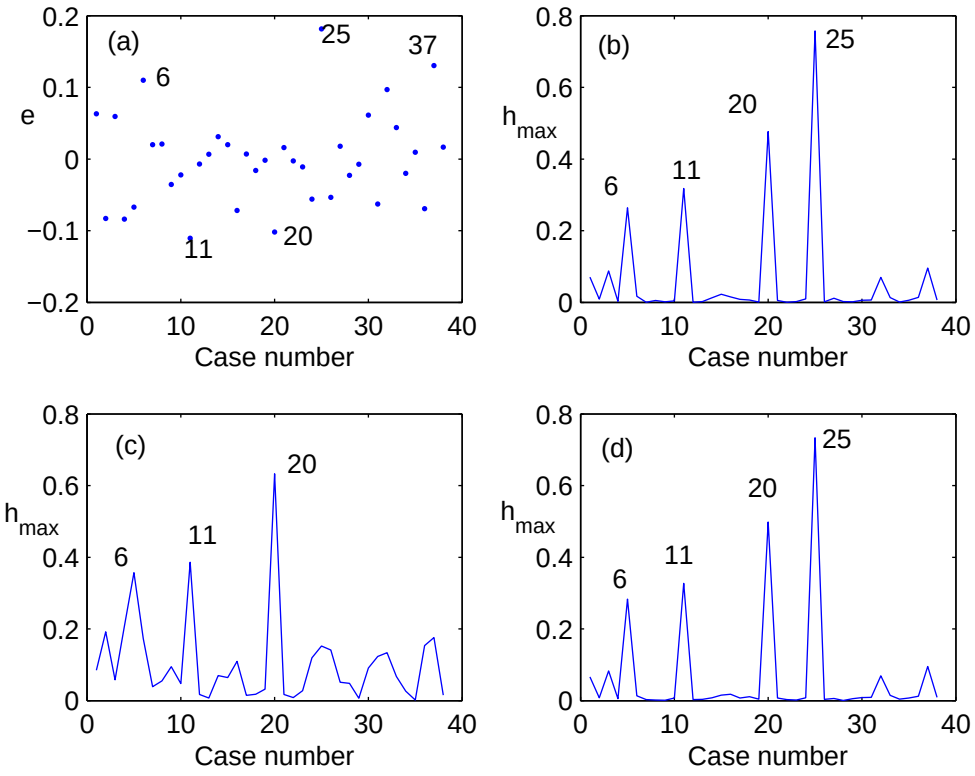

图 4 家庭消费数据的残差和影响分析图

从图 4 中看出，总体而言第 25 号，第 20 号，第 11 号和第 6 号点都是值得关注的点，但在不同扰动模式下，情况也有所差异．从图 4(c) 发现，第 25 号数据点对于响应变量扰动情况不敏感，并从图 4(a) 中显示，第 25 号点的残差最大．第 25 号点很可能既是强影响点又是异常点，另外 37 号点也值得注意，它的残差较大，但在三种扰动模式下都不是强影响点，第 37 点可能是异常点．因此，我们在进行局部影响分析时，要尽量考虑多种扰动模式并配合残差图，从而可以更多地挖掘出数据点隐含的信息，为我们作进一步统计分析提供有力的支撑．

参 考 文 献

[1] Zhu H, Lee S. Local Influence for Incomplete-data Models. *J. Journal of the Royal Statistical Society, Series B*, 2001, 63: 111–126

[2] Cook R. Detection of Influential Observation in Linear Regression. *J. Technometrics*, 1977, 19: 15–18

[3] Cook R. Assessment of Local Influence. *J. Journal of the Royal Statistical Society* (Series B), 1986, 48: 133–169

[4] Wing F, Zhu Z, Wei B, He X. Influence Diagnostics and Outlier Tests for Semi-parametric Mixed Models. *J. Journal of the Royal Statistical Society* (Series B), 2002, 64: 565–579

[5] Zhu H, Lee S. Local Influence for Generalized Linear Mixed Models. *J. the Canadian Journal of Statistics*, 2003, 31: 1–17

[6] 张浩，朱仲义. 半参数广义线性混合效应模型的影响分析. 应用数学学报, 2007, 30: 743–759
(Zhang H, Zhu Z. Influence Analysis of Generalized Partially Linear Mixed Models. *Acta Mathematicae Applicatae Sinica*, 2007, 30: 743–759)

[7] Ferrari S, Cribari-Neto F. Beta Regression for Modeling Rates and Proportions. *J. Journal of Applied Statistics*, 2004, 31: 799–815

[8] Espinheira P, Ferrari S, Cribari-Neto F. Influence Diagnostics in Beta Regression. *J. Computational Statistics and Data Analysis*, 2008, 52: 4417–4431

[9] 李爱萍，解锋昌，刘应安. Beta 回归模型的影响诊断. 高校应用数学学报, 2007, 22: 293–300
(Li A, Xie F, Liu Y. Influence Diagnostics in Beta Regression Model. *J. Applied Mathematics a Journal of Chinese Universities*, 2007, 22: 293–300

[10] Yu Y, Ruppert D. Penalized Spline Estimation for Partially Linear Single-index Models. *J. Journal of the American Statistical Association*, 2002, 97: 1042–1054

[11] McGilchrist C. Estimation in Generalized Mixed Models. *J. Journal of the Royal Statistical Society* (Series B), 1994, 56: 61–69

[12] Dempster A, Laird N, Rubin D. Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *J. Journal of the Royal Statistical Society* (Series B), 1977, 39: 1–38

[13] Wu, C. On the Convergence Properties of the EM Algorithm. *J. Annal of Statistics*, 1983, 11: 95–103

[14] Tanner M. Tools for Statistical Inference: Observed Data and Data Augmentation. Berlin: Springer-Verlag, 1993

[15] Banerjee M, Frees E. Influence Diagnostics for Linear Longitudinal Models. *J. Journal of the American Statistical Association*, 1997, 92: 999–1005

[16] Griffiths W, Hill R. Judge G. Learning and Practicing Econometrics. New York: Wiley, 1993

Influence Analysis for Semi-parametric Beta Regression Model with Longitudinal Data

ZHAO Weihua

(*School of Finance and Statistics, East China Normal University, Shanghai* 200241)

(*School of Science, NanTong University, JiangSu NanTong* 226007)

(E-mail: zhaowhstat@163.com)

LI Zean

(*School of Computer, NanTong University, JiangSu NanTong* 226007)

XU Xiangjian

(*School of Science, NanTong University, JiangSu NanTong* 226007)

Abstract This paper present several case-deletion as well as local influence measures for assessing the influence of an observation for Semi-parametric Beta Regression Model with Longitudinal Data. The essential idea is to treat the latent random effects in the model as missing data and get the estimate algorithm by adding penalized spline to estimate the non-parameters. We generate generalized Cook distance and generalized DFIT for the parametric and nonparametric part respectively based case-deletion model by Q-function. Four different perturbation schemes are discussed. Two numeric examples are presented to illustrate the results.

Key words Beta regression; longitudinal data; semi-parametric; influence analysis; P-spline; EM algorithm.

MR(2000) Subject Classification 62G05

Chinese Library Classification O212.2