应用概率统计　第二十三卷
第二期　2007 年 5 月

Chinese Journal of Applied Probability
and Statistics Vol. 23 No. 2 May 2007

部分线性混合效应模型中方差分量的稳健估计 *

秦国友　　朱仲义

(华东师范大学统计系，上海，200062)　(复旦大学统计系，上海市，200433)

摘　要

部分线性混合效应模型中方差分量是我们感兴趣的参数，文献中已经给出许多估计方法．但是其中很多方法都可以归结为广义估计方程方法 (GEE), 如：最大似然估计 (MLE), 约束最大似然估计 (REMLE) 等，而 GEE 方法对异常点很敏感．本文提出一组关于部分线性混合效应模型 (PLMM) 中均值和方差分量的稳健估计方程，对均值和方差分量同时进行稳健估计；并进行了随机模拟考察所提出稳健估计的有效性，最后通过两个实例，说明了所提方法的可行性．

关 键 词：回归样条，方差分量，混合模型，部分线性模型，稳健性．

学 科 分 类 号：O212.1.

§1. 引　　言

纵向数据具有个体内观察相关，个体间观察独立的特点，部分线性混合效应模型 (PLMM) 则是分析纵向数据常用的一种统计模型．而广义估计方程 (GEE) 是一种分析纵向数据情形下 PLMM 的有力统计推断工具．但是采用 GEE 方法对数据进行分析，受到数据中可能存在的异常点影响很大．很多研究者对线性和部分线性模型中的均值分量构造了稳健估计方程，以此达到减小异常点对统计推断的不利影响，最近的有关文献包括 He, Fung 和 Zhu (2005), Qin 和 Zhu (2005) 以及 Sinha (2004) 和 Wang, Lin 和 Zhu (2005).

混合效应模型中的方差分量是我们感兴趣的参数，因此对方差分量估计的研究具有理论和实际意义．文献中已经给出许多方法，但是其中很多方法都可以归结为广义估计方程方法 (GEE), 如：最大似然估计 (MLE), 约束最大似然估计 (REMLE) 等，而 GEE 方法受异常点的影响很大，因此有必要研究方差分量的稳健估计．

根据 Huggins (1993) 提出的针对多元正态模型的稳健估计方法，本文提出了一组稳健估计方程用以估计 PLMM 中回归参数、非参数函数和方差分量．本文第二部分给出模型和稳健的估计方法．第三部分通过随机模拟检验稳健估计的有效性．最后，采用文中提出的稳健方法分析了两组实际数据．通过模型实例分析，说明了我们所提方法的稳健性和可操作性，进一步推广了前人的结果．

§2. 模型及其稳健估计方法

2.1 部分线性混合效应模型

* 国家自然科学基金资助 (10371042).
本文 2006 年 1 月 12 日收到，2006 年 4 月 10 日收到修改稿.

本文讨论纵向数据，假定所研究的纵向数据有 m 个个体，第 i 个个体有 n_i 次观察，共有 $n=\sum_{i=1}^{m} n_i$ 次观察， y_{ij} $(i=1,\cdots,m,\ j=1,\cdots,n_i)$ 是第 i 个个体在时刻 t_{ij} 观察到的响应值，满足：

$$y_{ij}=x_{ij}^T\beta_0+f_0(t_{ij})+\sum_{s=1}^{k-1}U_{ijs}^T\xi_{is}+e_{ij}, \tag{2.1}$$

其中 β_0 是未知 p 维回归系数向量， x_{ij} 为协变量， f_0 是未知的光滑函数， ξ_{is}, $i=1,\cdots,m$ 为 $q_s\times 1$ 随机效应向量，相互独立，服从正态分布 $N(0,\gamma_s A_s)$, 假定 A_s 为已知相关矩阵．$e_i=(e_{i1},\cdots,e_{in_i})$ 是测量误差向量，相互独立，服从正态分布 $N(0,\gamma_k I_{ni})$. 于是 $Y_i=(y_{i1},\cdots,y_{in_i})^T$ 相互独立服从正态分布 $N(\mu_i,V_i)$, 其中 $V_i=\sum_{s=1}^{k-1}U_{is}A_sU_{is}^T\gamma_s+I_{ni}\gamma_k$, $U_{is}=(U_{i1s},\cdots,U_{in_is})^T$, $\gamma_0=(\gamma_1,\cdots,\gamma_k)$ 为待估的方差分量．不失一般性，假定 t_{ij} 取值于 $[0,1]$ 区间．

例 黄体酮数据 (Zhang et al, 1998)

该数据中有 34 个个体 492 个观察值，响应变量为对数变换后的黄体酮水平，除时间效应 t 外，两个协变量分别是年龄 (AGE) 和身体质量指标 (BMI). 考虑到对数变换后的黄体酮水平与时间效应 t 可能存在非线性关系，以及个体间可能存在的随机性差异，因此可以建立如下的部分线性混合效应模型：

$$y_{ij}=\beta_1\mathrm{AGE}_i+\beta_2\mathrm{BMI}_i+f(t_{ij})+u_i+e_{ij},$$

来分析该数据，其中 u_i 相互独立服从正态分布 $N(0,\gamma_1)$, e_{ij} 相互独立服从正态分布 $N(0,\gamma_2)$. $f(t_{ij})$ 用来刻画响应变量与时间效应 t 可能存在非线性关系，而 u_i 用来刻画个体间可能存在的随机性差异．Fung 等 (2002) 分析了该数据，指出数据中存在异常点和强影响点，因此考虑对此数据的稳健推断是有必要和现实意义的．

2.2 稳健估计

为估计非参数部分，类似于 He, Fung 和 Zhu (2005), 我们采用回归样条来逼近 f_0, 记 $0=s_0<s_1<\cdots<s_{k_n+1}=1$ 为样条节点， $B_1(t),\cdots,B_N(t)$ 表示 $N=k_n+l$ 个标准化阶数为 $l+1$ 的 B- 样条基函数．我们用 $\pi^T(t)\alpha_0$ 逼近 $f_0(t)$, 其中 $\pi(t)=(B_1(t),\cdots,B_N(t))^T$ 为 $N\times 1$ 维向量， $\alpha\in R^N$ 为样条系数向量．这样我们可以线性化 (2.1) 式，得到：

$$y_{ij}\approx(x_{ij}^T,\pi^T(t_{ij}))\theta_0+\sum_{s=1}^{k-1}U_{ijs}^T\xi_{is}+e_{ij}, \tag{2.2}$$

其中 $\theta_0^T=(\beta_0^T,\alpha_0^T)$ 为待估的联合回归参数向量．为了方便，我们引入几个记号，记 $\mu_i=(\mu_{i1},\cdots,\mu_{in_i})^T$, 其中 $\mu_{ij}=D_{ij}^T\theta$, $D_{ij}=(x_{ij}^T,\pi^T(t_{ij}))^T$, 并且 $D_i=(D_{in_1},\cdots,D_{in_i})^T$, X_i 和 π_i 表示类似的矩阵．

类似于 Huggins (1993), 我们提出如下的稳健估计方程分别估计参数 θ_0 和方差分量 γ_0,

$$\sum_{i}^{m}D_i^TV_i^{-1/2}(\gamma)h_i(Z_i(\theta))=0, \tag{2.3}$$

和

$$\sum_{i=1}^{m}\frac{c_1}{2}\mathrm{tr}\Big(V_i^{-1}(\gamma)\frac{\partial V_i(\gamma)}{\partial\gamma_s}\Big)-\frac{1}{2}h_i^T(Z_i(\theta))V_i^{-1/2}(\gamma)\frac{\partial V_i(\gamma)}{\partial\gamma_s}V_i^{-1/2}(\gamma)h_i(Z_i(\theta))=0,$$
$$s=1,\cdots,k, \tag{2.4}$$

其中 $\gamma = (\gamma_1, \cdots, \gamma_k)$, $Z_i(\theta) = V_i^{-1/2}(\gamma)(Y_i - \mu_i)$, $\mu_i = D_i\theta$, $h_i(Z_i) = \psi(Z_i)$, 本文中 ψ 取为 Huber 函数: $\psi(x) = \min(c, \max(-c, x))$, $c \in [1, 2]$, 文中 $c = 1.5$. 注意到 Huber 函数为有界函数, 可以用来减少异常点对于估计的影响, 从而获得稳健估计. 另外, (2.3) 和 (2.4) 中 Huber 函数中的常数 c 可以不同. 比如: 当方差分量受异常点影响比较大, 而均值分量受到的影响不大时, 可以将 (2.3) 中的常数 c 取得大一些, 以此提高均值分量稳健估计的效率. 本文为了方便, 只讨论了 (2.3) 和 (2.4) 中 Huber 函数中的常数 c 相同的情况. $Z_{0,ij}$ 表示 Z_i 的第 j 个分量 Z_{ij} 在 $\mu_i = \mu_{0,i}$ 和 $\gamma = \gamma_0$ 处计值. 由此模型可知, $Z_{0,ij}$ 为独立同分布的标准正态随机变量, 这样 $c_1 = \mathsf{E}(\psi(Z_{0,ij}))^2$ 是一个常数, 用来保证估计方程的 Fisher 相合性.

对于参数 θ_0 和 γ_0 按照下面两个步骤进行估计:

(1) 假定 $\widehat{\theta}_{WI}$ 是估计方程 (2.3) 中协方差阵 V_i 取为 $I_i\sigma^2$ 时得到的关于 θ_0 的估计, 其中 I_i 表示单位阵, σ^2 的估计可以通过如下的基于中位数绝对偏差的稳健估计来获得,

$$\widehat{\sigma}^2 = (1.4826 \times \mathrm{median}(|U - \mathrm{median}(U)|))^2,$$

其中, $U = Y - \widehat{\mu}, Y = (Y_1^T, \cdots, Y_m^T)^T$, $\widehat{\mu} = (\widehat{\mu}_1^T, \cdots, \widehat{\mu}_m^T)^T$. 把 (2.4) 中的 θ 用 $\widehat{\theta}_{WI}$ 代替, 并通过解 (2.4) 得到 γ_0 的稳健估计 $\widehat{\gamma}$.

(2) 类似于 (1), 把 (2.3) 中的 γ 用 $\widehat{\gamma}$ 代替, 并通过解 (2.3) 得到 θ 的估计 $\widehat{\theta}$.

由上述两步得到的 $\widehat{\theta}$ 和 $\widehat{\gamma}$ 就是我们提出的稳健估计量. 注意这里 (1) 、 (2) 两步之间不需要迭代.

我们采用 Matlab 6.5 中求解非线性方程的函数 fsolve 来解估计方程 (2.3) 和 (2.4), 并且我们以由对应于稳健估计方程 (2.3) 和 (2.4) 的非稳健估计方程, 也就是由经典的最大似然估计方法得到的估计方程, 得到的估计作为求解 (2.3) 和 (2.4) 的初值. 这里的非稳健估计方程与稳健的类似, 只是将 (2.3) 和 (2.4) 中的 $h(x)$ 取为 $h(x) = x$. 在我们的模拟中, 发现采用这样的初值可以得到收敛的稳健估计 $\widehat{\theta}$ 和 $\widehat{\gamma}$, 未发生发散的情况.

对于所得的稳健估计 $\widehat{\theta}$ 和 $\widehat{\gamma}$, 我们采用如下的 "三明治" (Sandwich) 估计分别估计 $\widehat{\theta}$ 和 $\widehat{\gamma}$ 的协方差阵

$$\widehat{V}_\beta = n\widehat{K}_{\beta,n}^{-1}\widehat{S}_{\beta,n}\widehat{K}_{\beta,n}^{-1}, \tag{2.5}$$

和

$$\widehat{V}_\gamma = n\widehat{K}_{\gamma,n}^{-1}\widehat{S}_{\gamma,n}\widehat{K}_{\gamma,n}^{-1}. \tag{2.6}$$

其中

$$\widehat{K}_{\beta,n} = \sum_{i=1}^m X_i^{*T}\Omega_i X_i^*, \qquad \widehat{S}_{\beta,n} = \sum_{i=1}^m X_i^{*T} h_i h_i^T X_i^*, \tag{2.7}$$

$$\widehat{K}_{\gamma,n} = \frac{\partial}{\partial\gamma}\Big[\sum_{i=1}^m (D_{i,\gamma_1}, \cdots, D_{i,\gamma_k})^T\Big], \tag{2.8}$$

$$\widehat{S}_{\gamma,n} = \sum_{i=1}^m (D_{i,\gamma_1}, \cdots, D_{i,\gamma_k})^T (D_{i,\gamma_1}, \cdots, D_{i,\gamma_k}), \tag{2.9}$$

其中 $X^* = (I - P)X$, $P = M(M^T\Omega M)^{-1}M^T\Omega$, $M = (\pi(t_1), \cdots, \pi(t_n))^T$, $\Omega = \mathrm{diag}(\Omega_i)$,

$$\Omega_i = -\frac{\partial h_i(\mu_i)}{\partial \mu_i^T}, \qquad D_{i,\gamma_s} = \frac{K}{2}\mathrm{tr}\Big(V_i^{-1}\frac{\partial V_i(\gamma)}{\partial\gamma_s}\Big) - \frac{1}{2}h_i^T(Z_i)V_i^{-1/2}\frac{\partial V_i}{\partial\gamma_s}V_i^{-1/2}h_i(Z_i),$$

(2.7)–(2.9) 均在 $\widehat{\theta}$ 和 $\widehat{\gamma}$ 处计值.

§3. 模拟研究

在这一部分我们通过随机模拟来研究文中所提出的关于 PLMM 中方差分量稳健估计的稳健性和有效性. 我们已知 PLMM 中参数 β 的稳健估计在数据存在异常点时能够有效的减少估计的偏差, 如 He, Fung 和 Zhu (2005), Qin 和 Zhu (2005) 和 Sinha (2004), 因此在这一部分, 我们对于 β 均采用稳健估计方程 (2.3) 来进行估计, 但是对于方差分量 γ_0 则分别采用稳健估计方程 (2.4) 和相应的非稳健估计方程得到稳健估计 $\widehat{\gamma}_R$ 和非稳健估计 $\widehat{\gamma}$ 进行估计. 这里非稳健估计方程指与 (2.4) 类似, 但是 $\psi(x)$ 取为 $\psi(x)=x$.

模拟中给出了在数据被污染和未被污染情况下 $\widehat{\gamma}_R$ 以及 $\widehat{\gamma}$ 的偏差、标准差、均方差 (MSE) 和 $\widehat{f}$ 的积分均方和 (IMSE), 并做了比较.

我们采用如下正态部分线性混合效应模型:

$$y_{ij}=\beta x_{1,ij}+5\sin(2t_{ij})+u_i+x_{2,ij}v_i+e_{ij},\qquad i=1,\cdots,100,\ j=1,\cdots,4,$$

其中随机效应 u_i, v_i 分别独立同分布于正态分布 $N(0,\gamma_1)$ 和 $N(0,\gamma_2)$, e_{ij} 相互独立服从正态分布 $N(0,\gamma_3)$, 模拟中 $\beta=5$, $\gamma_1=2$, $\gamma_2=1$, $\gamma_3=1$, 协变量 $x_{1,ij}$ 独立的从均匀分布 $U(-1,1)$ 中抽取, $x_{2,ij}$ 以等概率取值 0 和 1, 协变量 t_{ij} 独立的从均匀分布 $U(0,1)$ 中抽取. 总共从此模型中抽取 500 个样本.

为了研究稳健性, 我们通过以下三种方式来污染数据:

1. 通过对随机效应 u_i 的产生进行扰动来污染数据, 具体做法是从分布 $(1-\delta)N(0,\gamma_1)+\delta N(0,4)$ 中抽取 u_i, 其中 δ 分别取为 $0.05, 0.10$.

2. 通过对测量误差 e_{ij} 的产生进行扰动来污染数据, 具体做法是从分布 $(1-\delta)N(0,\gamma_3)+\delta N(0,2)$ 中抽取 e_{ij}, 其中 δ 分别取为 $0.05, 0.10$.

3. 通过对随机效应 u_i 和测量误差 e_{ij} 的产生同时进行扰动来污染数据, 具体做法是随机效应 u_i 和测量误差 e_{ij} 分别从分布 $(1-\delta)N(0,\gamma_1)+\delta N(0,4)$ 和 $(1-\delta)N(0,\gamma_3)+\delta N(0,2)$ 中抽取 u_i 和 e_{ij}, 其中 δ 分别取为 $0.05, 0.10$.

表 1, 表 2 给出了数据被以上三种方式污染后得到的关于 $\widehat{\gamma}_R$ 和 $\widehat{\gamma}$ 的统计量的计算结果. 其中表 1, 表 2 分别给出了污染比例 δ 为 0.05 和 0.10 时得到的结果. 根据 He, Fung 和 Zhu (2005), 这里样条内节点取为可区分的样本点 t_{ij} 的样本分位数, 内节点的个数取为 3, 即 $400^{1/5}$ 的整数部分.

从表 1 中, 我们可以发现在数据无污染时, 虽然采用稳健方法会导致一定的效率损失, 但是 $\widehat{\gamma}_R$ 和 $\widehat{\gamma}$ 具有接近的 MSE. 当数据被第一种方式污染后, 非稳健方法得到的 $\widehat{\gamma}_1$ 的偏差明显增大, $\widehat{\gamma}_2$ 和 $\widehat{\gamma}_3$ 受到的影响不大; 但是由稳健方法得到的 $\widehat{\gamma}_{R1}$ 具有比 $\widehat{\gamma}_1$ 显著减小的偏差和标准差, 从而有更小的 MSE. 另外, 其他两个分量的估计受到的数据污染的影响不大, 这主要由于数据的污染来自于随机效应 u_i.

从表 1 中, 我们还可以发现当数据被第二种方式污染后, 非稳健方法得到的 $\widehat{\gamma}_3$ 的偏差明显增大, $\widehat{\gamma}_1$ 和 $\widehat{\gamma}_2$ 受到的影响不大; 但是由稳健方法得到的 $\widehat{\gamma}_{R3}$ 具有比 $\widehat{\gamma}_3$ 更小的偏差和标准差. 另外, 其他两个分量的估计受到的数据污染的影响也不大, 这主要由于数据的污染来自于随机效应 e_{ij}, 与数据被第一种方式污染得到的结论类似. 而当数据被第三种方式污染后, 非稳

健方法得到的 $\hat{\gamma}_1$ 和 $\hat{\gamma}_3$ 的偏差同时增大，$\hat{\gamma}_2$ 受到的影响不大；但是由稳健方法得到的 $\hat{\gamma}_{R1}$ 和 $\hat{\gamma}_{R3}$ 具有比 $\hat{\gamma}_1$ 和 $\hat{\gamma}_3$ 更小的偏差和标准差．这主要也是由于数据的污染同时来自随机效应和测量误差．

从表 2 中可以得到与表 1 类似的结论．

通过 “三明治” 估计 (2.10) 计算了 $\hat{\gamma}_R$ 的标准差，计算结果一并列在表 1 和表 2 中．平均的 “三明治” 标准差 (AESE($\hat{\gamma}$)) 和 Monte Carlo 标准差的差别不大，说明采用的 “三明治” 标准差是可行的．

表 1　关于 $\hat{\gamma}$ 的 500 次模拟结果 ($\delta = 0.05$)

	$\hat{\gamma}_1$			$\hat{\gamma}_2$			$\hat{\gamma}_3$		
	BIAS1	MSE1	AESE1	BIAS2	MSE2	AESE2	BIAS3	MSE3	AESE3
NP NR	-0.059(0.342)	0.121	0.324(0.060)	-0.047(0.321)	0.105	0.309(0.056)	0.028(0.096)	0.010	0.101(0.014)
R	-0.079(0.365)	0.139	0.336(0.066)	-0.044(0.348)	0.123	0.308(0.064)	0.026(0.108)	0.012	0.114(0.018)
P1 NR	0.609(0.687)	**0.843**	0.576(0.278)	-0.020(0.333)	0.111	0.324(0.069)	0.035(0.101)	0.012	0.103(0.015)
R	0.359(0.544)	**0.425**	0.597(0.360)	-0.019(0.345)	0.120	0.320(0.074)	0.031(0.114)	0.014	0.116(0.019)
P2 NR	-0.125(0.352)	0.140	0.323(0.064)	-0.030(0.343)	0.119	0.333(0.065)	0.177(0.130)	**0.048**	0.128(0.026)
R	-0.142(0.373)	0.160	0.335(0.072)	-0.046(0.358)	0.130	0.338(0.073)	0.124(0.123)	**0.031**	0.141(0.028)
P3 NR	0.618(0.705)	**0.878**	0.597(0.278)	-0.028(0.347)	0.121	0.346(0.068)	0.180(0.122)	**0.047**	0.130(0.025)
R	0.334(0.544)	**0.408**	0.626(0.337)	-0.034(0.363)	0.133	0.347(0.082)	0.125(0.116)	**0.029**	0.144(0.030)

NP = 无污染；　P1 = 被方式 1 污染；　P2= 被方式 2 污染；　P3 = 被方式 3 污染；　NR = 采用非稳健方法估计 γ; R = 采用稳健方法估计 γ; AESE = 平均的 “三明治” 标准差；括号中数字表示相应估计的标准差．

表 2　关于 $\hat{\gamma}$ 的 500 次模拟结果 ($\delta = 0.10$)

	$\hat{\gamma}_1$			$\hat{\gamma}_2$			$\hat{\gamma}_3$		
	BIAS1	MSE1	AESE1	BIAS2	MSE2	AESE2	BIAS3	MSE3	AESE3
P1 NR	1.248(0.877)	**2.326**	0.773(0.348)	-0.028(0.325)	0.107	0.327(0.062)	0.033(0.100)	0.011	0.104(0.015)
R	0.839(0.694)	**1.185**	0.846(0.519)	-0.014(0.355)	0.126	0.321(0.070)	0.030(0.116)	0.014	0.117(0.020)
P2 NR	-0.083(0.332)	0.117	0.332(0.059)	-0.025(0.358)	0.129	0.356(0.072)	0.323(0.142)	**0.124**	0.149(0.029)
R	-0.115(0.349)	0.135	0.348(0.067)	-0.030(0.353)	0.126	0.363(0.084)	0.230(0.133)	**0.070**	0.168(0.037)
P3 NR	1.325(0.958)	**2.674**	0.803(0.379)	-0.036(0.377)	0.143	0.381(0.080)	0.322(0.143)	**0.124**	0.155(0.032)
R	0.875(0.741)	**1.316**	0.893(0.552)	-0.019(0.387)	0.150	0.380(0.098)	0.224(0.136)	**0.069**	0.171(0.037)

§4.　实例分析

1．黄体酮数据

我们采用文中的稳健方法分析前面例子中的黄体酮数据 (Zhang et al, 1998)．我们对数据中各变量做了与 Zhang et al (1998) 相同的变换．Fung et al (2002) 分析了该组数据，指出数据中存在异常点和强影响点．

我们采用下面的部分线性混合效应模型来分析该数据：

$$y_{ij} = \beta_1 \mathrm{AGE}_i + \beta_2 \mathrm{BMI}_i + f(t_{ij}) + u_i + e_{ij},$$

其中 u_i 相互独立服从正态分布 $N(0, \gamma_1)$, e_{ij} 相互独立服从正态分布 $N(0, \gamma_2)$．我们采用含有 2 个内节点的 4 阶回归样条逼近 f_0 (见图 1), 反映了对数变换后的黄体酮水平与时间之间非线性

趋势，与 Zhang et al (1998) 得到的非参数函数的曲线类似．表 3 给出了估计结果．由表 3 可以看出， AGE和 BMI的效应均不显著，而 γ_1, γ_2 的效应显著，与 Zhang et al (1998) 的结论一致．

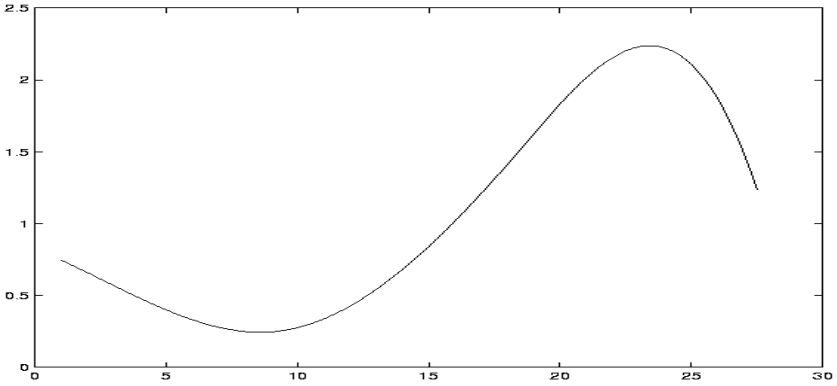

图 1　黄体酮数据中以时间为变量的非参数函数的估计

由表 3 我们还可以看到虽然稳健方法和非稳健方法得到的结论是 γ_1 和 γ_2 的效应都不显著，但是稳健估计 $\widehat{\gamma}_R$ 和非稳健估计 $\widehat{\gamma}$ 在数值上还是有较大差异，说明稳健方法在一定程度上限制了数据中异常点的影响，有助于我们在分析数据时作出更加合理和正确的结论．另外，我们注意到 $\widehat{\gamma}_R$ 和 $\widehat{\gamma}$ 的两个分量之间在数值上都有较大的差异．由我们模拟研究中的结论，似乎该模型中的随机效应和测量误差部分都受到了较大的扰动，这也有助于我们对模型和数据有更深入的认识．

表 3　PDG 数据中回归系数的估计

	Our method		Zhang et al
	$\widehat{\beta}(\widehat{\gamma}_R)$	$\widehat{\beta}(\widehat{\gamma})$	
Age	1.432(2.237)	1.484(2.278)	0.925(1.924)
BMI	-2.267(2.672)	-2.142(2.711)	-2.913(2.376)
γ_1	0.275(0.071)	0.226(0.053)	0.262(0.072)
γ_2	0.267(0.035)	0.331(0.045)	0.137(0.016)

$\widehat{\beta}(\widehat{\gamma}_R)$ 和 $\widehat{\beta}(\widehat{\gamma})$ 表示将 $\widehat{\gamma}_R$ 和 $\widehat{\gamma}$ 分别带入估计方程 (2.3) 所得的 β 估计．

2. CD4 细胞数数据

我们再采用文中的稳健方法和部分线性混合效应模型分析著名的 CD4 细胞数数据，该数据记录了 369 位男性 HIV 感染者随着时间变化 (以血清转化为时刻零点) 的 CD4+ 细胞数目的变化过程，共得到 2376 个观察数据．同时还考虑了一些协变量：年龄 (Age) 、每天抽烟的数量 (Smoking) 、是否吸毒 (Drug) 、性伙伴数目 (Sex partner) 、以及精神状态得分 (Depression). 我们以 CD4+ 细胞数的平方根作为响应变量，对时间建立非参数联系，其他五个协变量建立一般线性关系，由此建立了类似于上面分析黄体酮数据的部分线性混合效应模型．

我们采用含有 3 个内节点的 4 阶回归样条逼近 f_0 (见图 2), 反映了 CD4 细胞数与时间之间非线性趋势．表 4 列出了模型中参数的估计结果．Smoking, Drug 和 Depression 的效应是显著的，其他两个效应在显著性水平为 0.05 时则是不显著的，并且 $\gamma_0 = (\gamma_1, \gamma_2)^T$ 效应是高度显著

的．这与 Wang et al (2005) 的结论基本上是一致的．而在 Zeger 和 Diggle (1994) 的分析中只有 Sex partner 和 Depression 两个效应是显著的．

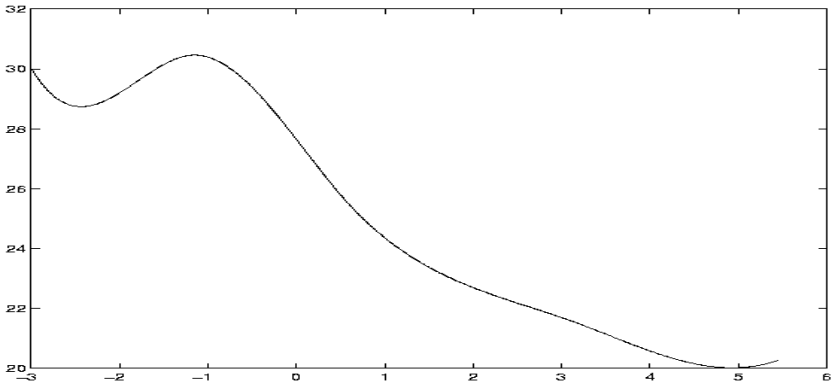

图 2　CD4 细胞数数据中以时间为变量的非参数函数的估计

由表 4, 我们发现 $\hat{\gamma}_R$ 和 $\hat{\gamma}$ 的第一个分量在数值上是接近的，而第二个分量的差别较大，由模拟的结论，我们似乎可以作出这样的判断，即：该模型中随机效应受到的影响不大，而测量误差部分可能受到扰动．

表 4　CD4 数据中回归系数的估计

	Our method		Zeger & Diggle	Wang et al
	$\widehat{\beta}(\hat{\gamma}_R)$	$\widehat{\beta}(\hat{\gamma})$		
Age	0.006(0.031)	0.005(0.031)	0.037(0.180)	0.008(0.032)
Smoking	0.574(0.123)	0.543(0.120)	0.270(0.150)	0.579(0.139)
Drug	0.529(0.226)	0.527(0.226)	0.370(0.310)	0.584(0.335)
Sex partner	0.059(0.039)	0.065(0.038)	0.100(0.038)	0.078(0.039)
Depression	-0.042(0.015)	-0.042(0.015)	-0.058(0.015)	-0.046(0.014)
γ_1	16.033(1.516)	16.519(1.589)	–	–
γ_2	16.393(0.861)	18.997(1.033)	–	–

§5.　结　　论

本文采用构造稳健估计方程的方法获得了部分线性混合效应模型中方差分量的稳健估计. 通过模拟研究, 我们发现在数据未受污染时, 稳健方法相比于非稳健方法会有一定的效率损失, 但是在数据被污染后，稳健方法的确能够有效的减少估计的偏差，同时具有更小的标准差，因而比非稳健方法要更加有效．模拟中，我们还发现当混合效应模型中的某个方差分量对应的随机效应受到污染时，主要会对该方差分量的估计有影响，对其它分量几乎没有影响，而稳健方法则能够有效的减少受污染方差分量估计的偏差和标准差．最后对两个实际数据的分析，说明了本文提出方法在实际操作中的可行性．另外，本文主要研究的是部分线性混合效应模型中方差分量的稳健估计．

参 考 文 献

[1] Fung, W.K., Zhu, Z.Y., Wei, B.C. and He, X., Influential diagnostics and outlier tests for semiparametric mixed models, *J. R. Statist. Soc. B*, **64**(2002), 565–579.

[2] He, X., Fung, W.K. and Zhu, Z.Y., Robust estimation in generalized partial linear models for clustered data, *J. Am. Statist. Assoc.*, **100**(2005), 1176–1184.

[3] Huggins, R.M., A robust approach to the analysis of repeated measures, *Biometrics*, **49**(1993), 715–720.

[4] Liang, K.Y. and Zeger, S.L., Longitudinal data analysis using generalized linear models, *Biometrika*, **73**(1986), 13–22.

[5] Qin, G.Y. and Zhu, Z.Y., Robust estimation in partial linear mixed model for longitudinal data, 2005. (submitted)

[6] Sinha, S.K., Robust analysis of generalized linear mixed models, *J. Am. Statist. Assoc.*, **99**(2004), 451–460.

[7] Wang, N., Carroll, R., and Lin, X.H., Efficient semiparametric marginal estimation for longitudinal/clustered data, *J. Am. Statist. Assoc.*, **100**(2005), 147–157.

[8] Wang, Y.G., Lin, X. and Zhu, M., Robust estimation functions and bias correction for longitudinal data analysis, *Biometrics*, **61**(2005), 684–691.

[9] Zeger, S.L. and Diggle, P.J., Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters, *Biometrics*, **50**(1994), 689–99.

[10] Zhang, D.W., Lin, X.H., Raz, J. and Sowers, M., Semiparametric stochastic mixed models for longitudinal data, *J. Am. Statist. Assoc.*, **93**(1998), 710–719.

Robust Estimation of the Variance Components in Semiparametric Linear Mixed Model

Qin Guoyou

(*Department of Statistics, East China Normal University, Shanghai, 200062*)

Zhu Zhongyi

(*Department of Statistics, Fudan University, Shanghai, 200433*)

For a partial linear mixed model, we usually focus on the estimation of the variance components, and a lot of methods can be applied. However, many of these methods, such as maximum likelihood method and restricted maximum likelihood method, can be included in the framework of generalized estimating equation (GEE). As well known, the GEE method is sensitive to outliers. So, an alterative set of robust GEEs for both mean components and correlation parameters are proposed for the partial linear mixed model for longitudinal data in this paper. Some simulations are conducted to evaluate the performance of the proposed estimators. In the end, the method is illustrated with analysis of two real data sets.