

基于M估计的具有一致相关线性混合效应模型中 相关系数的齐性检验 *

孙 慧 慧

(盐城师范学院数学科学学院, 盐城, 224002)

摘 要

本文对纵向数据的线性混合效应模型, 用Fisher Scoring方法得到了参数的M估计(稳健估计), 研究了M估计下一致相关系数的齐性检验问题, 并对检验统计量的功效进行了模拟, 最后通过葡萄糖数据的实例说明了本文方法的有效性.

关键词: M估计, 纵向数据, 一致相关性, Score检验.

学科分类号: O212.2.

§1. 引 言

纵向数据广泛存在于社会生活的各个领域, 关于纵向数据的分析也是近年来统计学的热点课题之一. Diggle等(2002)首次详细地研究了基于线性、广义线性模型的纵向数据的统计分析方法; Verbeke和Molenberghs (2000)详细讨论了线性混合效应模型的纵向数据分析. 常用刻画纵向数据协方差结构的因素有三种, 即序列相关、随机效应和随机误差. Diggle等(2002), Pinheiro和Bates (2000)等用随机效应和随机误差刻画了线性纵向数据模型, 并在随机效应和随机误差的方差齐性假设下对模型进行了统计诊断; 林金官等(2004a, 2004b, 2009)研究了非线性纵向数据模型中随机效应的存在性和相关性检验问题, 具有一阶自相关误差的非线性纵向数据模型的方差齐性及相关系数齐性检验问题, 以及广义非线性纵向数据模型中偏离名义离差的检验问题.

一致相关是另一种重要的相关形式, Diggle等(2002), Wolfinger (1996)讨论了此种相关形式; 范俊花等(2009)研究了具有一致相关协方差结构的纵向数据模型的方差和相关系数的齐性检验, 但研究的纵向模型中不含有随机效应. 本文将研究含有随机效应的线性混合效应模型的一致相关性问题.

众所周知, 极大似然估计较易受极端值的影响, 稳健性较差. 若直接采用极大似然估计对含有异常点的数据进行估计, 势必会产生不良影响甚至错误的结果. 为了消除异常值对估计的影响, 可以采用M估计的方法. M估计最早是由Huber (1981)引入回归问题, 是目前

*国家自然科学基金(11171065, 11202180)和江苏省自然科学基金(2011058)资助.

本文2012年9月28日收到, 2013年8月19日收到修改稿.

应用最广泛的稳健估计方法. 本文将Huber函数引入纵向数据的线性混合效应模型, 基于模型对数似然的稳健形式, 研究了稳健形式下一致相关性检验问题. 结构如下: 第二节介绍了模型, 用Fisher得分迭代法对参数进行了M估计, 得到了稳健极大似然估计(RMLE); 第三节基于M估计研究了一致相关性的检验问题, 给出了Score统计量; 最后用葡萄糖数据的实例说明了本文的方法.

§2. 线性混合效应模型

本文研究的线性混合效应模型如下:

$$\mathbf{y}_k = \mathbf{X}_k \boldsymbol{\beta} + \mathbf{C}_k \boldsymbol{\tau}_k + \mathbf{e}_k, \quad k = 1, 2, \dots, N, \quad (2.1)$$

其中 $\mathbf{y}_k = (y_{k1}, \dots, y_{kn_k})^T$ 是第 k 个个体的 n_k 次观测结果组成的向量, \mathbf{X}_k 是 q 维未知固定效应向量 $\boldsymbol{\beta}$ 的 $n_k \times q$ 阶设计矩阵, 且 $\mathbf{X}_k = (x_{k1}, \dots, x_{kn_k})^T$. \mathbf{C}_k 是随机效应 $\boldsymbol{\tau}_k$ 的 $n_k \times r$ 阶设计矩阵, 且假设 $\boldsymbol{\tau}_k \sim N(0, \sigma^2 \boldsymbol{\Gamma})$. $\mathbf{e}_k = (e_{k1}, \dots, e_{kn_k})^T$ 是 $n_k \times 1$ 维不可观测的随机误差向量. 一致相关性是指同一受试单元中任何两次测量之间具有相同的相关系数 ϕ , 即 $\text{corr}(e_{ki}, e_{kj}) = \phi_k$ ($i \neq j$), 则 $\text{Var}(\mathbf{e}_k) = \sigma^2 [I_{n_k} + \phi_k (J_{n_k} - I_{n_k})]$, 其中 I_{n_k} 为 $n_k \times n_k$ 的单位阵, J_{n_k} 为 $n_k \times n_k$ 阶元素全为1的矩阵. 若记 $\mathbf{V}_k = I_{n_k} + \phi_k (J_{n_k} - I_{n_k})$ 为一致相关结构, 则 $\mathbf{e}_k \sim N(0, \sigma^2 \mathbf{V}_k)$. 纵向数据由于实际情况的复杂性, 组内以及组间方差均可能变异(Sun等, 2011), 同样各组的一致相关系数 ϕ 也可能产生变异, 根据Núñez-Antón和Zimmerman (2000)的参数化方法, 可化为 $\phi_k = \phi \omega(\mathbf{v}_k, \boldsymbol{\gamma})$, 且存在 $\boldsymbol{\gamma}_0$ 使得 $\omega(\mathbf{v}_k, \boldsymbol{\gamma}_0) = c \neq 0$, 此时 $\mathbf{V}_k = I_{n_k} + \phi \omega(\mathbf{v}_k, \boldsymbol{\gamma})(J_{n_k} - I_{n_k})$. 此外假设 $\boldsymbol{\tau}_k$ 与 \mathbf{e}_k 相互独立, 从而

$$\text{Cov}(\mathbf{y}_k) = \sigma^2 \boldsymbol{\Sigma}_k = \sigma^2 \mathbf{C}_k \boldsymbol{\Gamma} \mathbf{C}_k^T + \sigma^2 \mathbf{V}_k,$$

此处及以后, T 表示矩阵或向量的转置. 我们用 $\boldsymbol{\alpha}$ 表示 $\boldsymbol{\Sigma}_k$ 中的未知参数向量, 则模型的对数似然函数为

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha} | \mathbf{y}) = \text{constant} - \frac{1}{2} M \log \sigma^2 - \frac{1}{2} \sum_{k=1}^N \log |\boldsymbol{\Sigma}_k| - \sum_{k=1}^N \frac{1}{2} \boldsymbol{\varepsilon}_k^T \boldsymbol{\varepsilon}_k, \quad (2.2)$$

其中 $\boldsymbol{\varepsilon}_k = \sigma^{-1} \boldsymbol{\Sigma}_k^{-1/2} (\mathbf{y}_k - \mathbf{X}_k \boldsymbol{\beta})$. 我们看到(2.2)式的最后一项是一个平方和的形式, 随偏差增长很快. 如果采用增长较慢的函数代替这个二次函数, 就可以提高其稳健性, 从而限制异常观察值的影响. 基于这个观点的稳健估计算法主要是M估计, 是目前应用最广泛的稳健估计方法. 本文我们选用Huber函数来代替二次函数.

$$\rho(\varepsilon) = \begin{cases} \frac{1}{2} \varepsilon^2, & \text{当 } |\varepsilon| \leq c \text{ 时;} \\ c|\varepsilon| - \frac{1}{2} c^2, & \text{当 } |\varepsilon| > c \text{ 时,} \end{cases}$$

其中 c 是固定常数, 通常 $c \in [0.7, 2]$, 本文我们取 $c = 1.345$. 对 $\rho(\varepsilon)$ 求导得

$$\psi(\varepsilon) = \partial\rho(\varepsilon)/\partial\varepsilon = \begin{cases} \varepsilon, & \text{当 } |\varepsilon| \leq c \text{ 时;} \\ c \operatorname{sign}(\varepsilon), & \text{当 } |\varepsilon| > c \text{ 时.} \end{cases}$$

因此(2.2)的稳健形式为

$$\eta(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2) = \text{constant} - \frac{1}{2}\kappa_1 M \log \sigma^2 - \frac{1}{2}\kappa_1 \sum_{k=1}^N \log |\Sigma_k| - \sum_{k=1}^N \sum_{j=1}^{n_k} \rho(\varepsilon_{jk}), \quad (2.3)$$

其中 $\kappa_1 = E(\varepsilon\psi(\varepsilon)) = P(|\varepsilon| \leq c)$ 为相合修正因子.

现在我们基于(2.3)式, 用Fisher得分迭代法对参数进行稳健估计. 关于 $\boldsymbol{\beta}$ 有

$$\begin{aligned} \frac{\partial\eta}{\partial\boldsymbol{\beta}} &= \sigma^{-1} \sum_{k=1}^N \mathbf{X}_k^T \Sigma_k^{-1/2} \psi[\sigma^{-1} \Sigma_k^{-1/2} (\mathbf{y}_k - \mathbf{X}_k \boldsymbol{\beta})], \\ \frac{\partial^2\eta}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= -\sigma^{-2} \sum_{k=1}^N \mathbf{X}_k^T \Sigma_k^{-1/2} \Lambda \Sigma_k^{-1/2} \mathbf{X}_k, \\ H_{\boldsymbol{\beta}\boldsymbol{\beta}^T} &= E\left(-\frac{\partial^2\eta}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right) = \nu\sigma^{-2} \sum_{k=1}^N \mathbf{X}_k^T \Sigma_k^{-1} \mathbf{X}_k, \end{aligned}$$

其中 Λ 为对角阵, 对角元素 $\Lambda_{ii} = \partial\psi(\varepsilon)/\partial\varepsilon$, 当 $|\varepsilon| \leq c$ 时, 其值为1, 当 $|\varepsilon| > c$ 时, 其值为0. $\nu = E(\Lambda_{ii}) = P(|\varepsilon| \leq c) = \int_{-c}^c (2\pi)^{-1/2} e^{-1/2\varepsilon^2} d\varepsilon = \kappa_1$. 由此得到 $\boldsymbol{\beta}$ 的稳健估计的迭代公式为

$$\hat{\boldsymbol{\beta}}^{(h+1)} = \hat{\boldsymbol{\beta}}^{(h)} + (\hat{H}_{\boldsymbol{\beta}\boldsymbol{\beta}^T}^{(h)})^{-1} \frac{\partial\eta^{(h)}}{\partial\boldsymbol{\beta}}. \quad (2.4)$$

当迭代序列收敛时, 得到 $\hat{\boldsymbol{\beta}}$ 即为 $\boldsymbol{\beta}$ 的稳健极大似然估计(RMLE). 注意到, 当 $c = \infty$ 时, $\boldsymbol{\beta}$ 的稳健估计 $\hat{\boldsymbol{\beta}}$ 就是传统的极大似然估计(MLE). 用类似的方法, 可以得到方差分量的稳健极大似然估计迭代公式.

§3. 一致相关系数的存在性检验

本节在组内方差和组间方差齐性的假设下, 首先研究一致相关系数的存在性检验. 此时 $\mathbf{V}_k = I_{n_k} + \phi(J_{n_k} - I_{n_k})$, $\operatorname{Cov}(\mathbf{y}_k) = \sigma^2 \Sigma_k = \sigma^2(\mathbf{C}_k \Gamma \mathbf{C}_k^T + \mathbf{V}_k)$. 令 $\boldsymbol{\theta} = (\phi, \boldsymbol{\beta}^T, \sigma^2, \boldsymbol{\delta}^T)^T$, 其中 $\boldsymbol{\delta} = (\delta_1, \dots, \delta_{r'})^T = (d_{11}, d_{12}, \dots, d_{1r}, d_{22}, \dots, d_{rr})^T$ 为 $r' = r(r+1)/2$ 维向量, d_{ij} 为 Γ 的第 (i, j) 个元素. 因此, 一致相关系数的存在性检验化为如下假设检验问题:

$$H_0 : \phi = 0; \quad H_1 : \phi \neq 0. \quad (3.1)$$

在正则条件下, 根据Cox和Hinkley (1974), 对于假设检验问题(3.1), 基于(2.3)式的检验统计量为

$$SC_1 = \left\{ \left(\frac{\partial\eta(\boldsymbol{\theta})}{\partial\phi} \right)^2 (N \mathbf{I}^{\phi\phi} (\mathbf{J}_N^{\phi\phi})^{-1} \mathbf{I}^{\phi\phi}) \right\}_{\hat{\boldsymbol{\theta}}}, \quad (3.2)$$

其中 $\hat{\boldsymbol{\theta}}$ 为 H_0 成立时 $\boldsymbol{\theta}$ 的稳健极大似然估计, 即 $\hat{\boldsymbol{\theta}} = (0, \hat{\boldsymbol{\beta}}^T, \hat{\sigma}^2, \hat{\boldsymbol{\delta}}^T)^T$. $\mathbf{I}^{\phi\phi}$ 表示 $\mathbf{I}(\boldsymbol{\theta})$ 的逆阵中与 $I_{\phi\phi}$ 对应的子块, 则 $(\mathbf{I}^{\phi\phi})^{-1} = I_{\phi\phi} - I_1 I_2^{-1} I_1^T$, $I_1 = (I_{\phi\sigma^2} \quad I_{\phi\boldsymbol{\delta}})$, $I_2 = \begin{pmatrix} I_{\sigma^2\sigma^2} & I_{\sigma^2\boldsymbol{\delta}} \\ I_{\boldsymbol{\delta}\sigma^2} & I_{\boldsymbol{\delta}\boldsymbol{\delta}} \end{pmatrix}$.

$\mathbf{J}_N^{\phi\phi}$ 的含义类似. 首先, 当 H_0 成立时, Score函数为

$$\begin{aligned} \frac{\partial\eta(\boldsymbol{\theta})}{\partial\phi} &= \frac{1}{2} \sum_{k=1}^N \psi[\hat{\sigma}^{-1} \hat{\Sigma}_k^{-1/2} (\mathbf{y}_k - \mathbf{X}_k \hat{\boldsymbol{\beta}})]^T \hat{\Sigma}_k^{-1} \frac{\partial \Sigma_k}{\partial \phi} \Big|_{\hat{\boldsymbol{\theta}}} [\hat{\sigma}^{-1} \hat{\Sigma}_k^{-1/2} (\mathbf{y}_k - \mathbf{X}_k \hat{\boldsymbol{\beta}})] \\ &\quad - \frac{1}{2} \kappa_1 \sum_{k=1}^N \text{tr} \left(\hat{\Sigma}_k^{-1} \frac{\partial \Sigma_k}{\partial \phi} \Big|_{\hat{\boldsymbol{\theta}}} \right) \\ &= \frac{1}{2} \sum_{k=1}^N \psi[\hat{\sigma}^{-1} \hat{\Sigma}_k^{-1/2} (\mathbf{y}_k - \mathbf{X}_k \hat{\boldsymbol{\beta}})]^T \hat{\Sigma}_k^{-1} (J_{n_k} - I_{n_k}) [\hat{\sigma}^{-1} \hat{\Sigma}_k^{-1/2} (\mathbf{y}_k - \mathbf{X}_k \hat{\boldsymbol{\beta}})] \\ &\quad - \frac{1}{2} \kappa_1 \sum_{k=1}^N \text{tr}(\hat{\Sigma}_k^{-1} (J_{n_k} - I_{n_k})), \end{aligned} \quad (3.3)$$

其中 $\hat{\Sigma}_k = \mathbf{C}_k \hat{\Gamma} \mathbf{C}_k + I_{n_k}$.

H_0 成立时, 关于 $\boldsymbol{\theta}$ 的Fisher信息阵为

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} I_{\phi\phi} & 0 & I_{\phi\sigma^2} & I_{\phi\boldsymbol{\delta}} \\ 0 & I_{\beta\beta} & 0 & 0 \\ I_{\sigma^2\phi} & 0 & I_{\sigma^2\sigma^2} & I_{\sigma^2\boldsymbol{\delta}} \\ I_{\boldsymbol{\delta}\phi} & 0 & I_{\boldsymbol{\delta}\sigma^2} & I_{\boldsymbol{\delta}\boldsymbol{\delta}} \end{bmatrix}. \quad (3.4)$$

其非零子块为

$$\begin{aligned} I_{\phi\phi} &= \frac{1}{2} \kappa_1 \sum_{k=1}^N \text{tr}(\hat{\Sigma}_k^{-1} (J_{n_k} - I_{n_k}) \hat{\Sigma}_k^{-1} (J_{n_k} - I_{n_k})), \\ I_{\phi\sigma^2} &= \frac{1}{4} \kappa_1 (1 + \kappa_2) \hat{\sigma}^{-2} \sum_{k=1}^N \text{tr}(\hat{\Sigma}_k^{-1} (J_{n_k} - I_{n_k})) = I_{\sigma^2\phi}^T, \\ I_{\phi\boldsymbol{\delta}} &= \left(\frac{\kappa_1}{2} \sum_{k=1}^N \text{tr}(\hat{\Sigma}_k^{-1} (J_{n_k} - I_{n_k}) \hat{\Sigma}_k^{-1} \mathbf{C}_k \frac{\partial \Gamma}{\partial d_{ab}} \mathbf{C}_k^T) \right)_{1 \times r'} = I_{\boldsymbol{\delta}\phi}^T, \\ I_{\beta\beta} &= \nu \hat{\sigma}^{-2} \sum_{k=1}^N \mathbf{X}_k^T \hat{\Sigma}_k^{-1} \mathbf{X}_k, \quad I_{\sigma^2\sigma^2} = \frac{1}{2} M \kappa_1 \hat{\sigma}^{-4}, \\ I_{\sigma^2\boldsymbol{\delta}} &= \left(\frac{1}{4} \kappa_1 (1 + \kappa_2) \hat{\sigma}^{-2} \sum_{k=1}^N \text{tr}(\hat{\Sigma}_k^{-1} \mathbf{C}_k \frac{\partial \Gamma}{\partial d_{ab}} \mathbf{C}_k^T) \right)_{1 \times r'} = I_{\boldsymbol{\delta}\sigma^2}^T, \\ I_{\boldsymbol{\delta}\boldsymbol{\delta}} &= \left(\frac{1}{2} \sum_{k=1}^N \text{tr}(\mathbf{C}_k \frac{\partial \Gamma}{\partial d_{ab}} \mathbf{C}_k^T \hat{\Sigma}_k^{-2} \mathbf{C}_k \frac{\partial \Gamma}{\partial d_{cd}} \mathbf{C}_k^T) \right)_{r' \times r'}, \end{aligned}$$

其中 $\kappa_2 = \int_{-c}^c (2\pi)^{-1/2} \varepsilon^2 e^{-1/2\varepsilon^2} d\varepsilon$, $\partial\Gamma/\partial d_{ab}$ 的元素在 (a, b) 和 (b, a) 处为1, 其余为0, $\partial\Gamma/\partial d_{cd}$ 类似. 由参数稳健极大似然估计(RMLE)的渐近正态性(孙慧慧, 2011), 经计算得

$$\mathbf{J}_N(\boldsymbol{\theta}) = N^{-1} \mathbf{I}(\boldsymbol{\theta}). \quad (3.5)$$

综合以上结果,可以得到一致相关存在性检验的Score检验统计量 SC_1 . Score检验统计量是广义似然比检验统计量的一种特殊形式,与似然比检验相比,不需要计算备择假设下的参数估计;另外,在给定的正则条件下Score检验统计量与似然比检验统计量的渐近分布相同,检验功效相当,因而应用比较广泛.

§4. 一致相关系数的齐性检验

由于实际问题的复杂性,相关系数可能变异,此时为各受试单元一致相关. 根据Núñez-Antón和Zimmerman (2000)的参数化方法,将各受试组的一致相关系数参数化为 $\phi_k = \phi\omega(\mathbf{v}_k, \gamma)$, 其中 \mathbf{v}_k 是协变量,且假设存在 γ_0 使得 $\omega(\mathbf{v}_k, \gamma_0) = h \neq 0$ 对所有的 k 都成立, h 是与 k 无关的常数. 则一致相关系数的齐性检验化为如下假设检验问题:

$$H_0 : \gamma = \gamma_0; \quad H_1 : \gamma \neq \gamma_0. \quad (4.1)$$

令 $\boldsymbol{\theta} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}^T, \phi, \sigma^2, \boldsymbol{\delta}^T)^T$, 其中 $\boldsymbol{\gamma}$ 为 q 维兴趣参数向量. 且 $\Sigma_k = I_{n_k} + \phi\omega(\mathbf{v}_k, \gamma)(J_{n_k} - I_{n_k}) + \mathbf{C}_k\boldsymbol{\Gamma}\mathbf{C}_k^T$.

在正则条件下,根据Cox和Hinkley (1974),对于假设检验问题(4.1),基于(2.3)式的检验统计量为

$$SC_2 = \left\{ \left(\frac{\partial \eta(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}} \right)^2 (N \mathbf{I}^{\boldsymbol{\gamma}\boldsymbol{\gamma}} (\mathbf{J}_N^{\boldsymbol{\gamma}\boldsymbol{\gamma}})^{-1} \mathbf{I}^{\boldsymbol{\gamma}\boldsymbol{\gamma}}) \right\}_{\hat{\boldsymbol{\theta}}}, \quad (4.2)$$

其中 $\hat{\boldsymbol{\theta}}$ 为 H_0 成立时 $\boldsymbol{\theta}$ 的稳健极大似然估计. 即 $\hat{\boldsymbol{\theta}} = (\boldsymbol{\gamma}_0^T, \hat{\boldsymbol{\beta}}^T, \hat{\phi}, \hat{\sigma}^2, \hat{\boldsymbol{\delta}}^T)^T$. 首先,当 H_0 成立时,Score函数为

$$\begin{aligned} \frac{\partial \eta(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}} &= \frac{1}{2} \sum_{k=1}^N \psi[\hat{\sigma}^{-1} \hat{\Sigma}_k^{-1/2} (\mathbf{y}_k - \mathbf{X}_k \hat{\boldsymbol{\beta}})]^T \hat{\Sigma}_k^{-1} \frac{\partial \Sigma_k}{\partial \boldsymbol{\gamma}} \Big|_{\hat{\boldsymbol{\theta}}} [\hat{\sigma}^{-1} \hat{\Sigma}_k^{-1/2} (\mathbf{y}_k - \mathbf{X}_k \hat{\boldsymbol{\beta}})] \\ &\quad - \frac{1}{2} \kappa_1 \sum_{k=1}^N \text{tr} \left(\hat{\Sigma}_k^{-1} \frac{\partial \Sigma_k}{\partial \boldsymbol{\gamma}} \Big|_{\hat{\boldsymbol{\theta}}} \right) \\ &= \frac{\hat{\phi}}{2} \left(\sum_{k=1}^N \psi[\hat{\sigma}^{-1} \hat{\Sigma}_k^{-1/2} (\mathbf{y}_k - \mathbf{X}_k \hat{\boldsymbol{\beta}})]^T \hat{\Sigma}_k^{-1} \dot{\omega}_{ka} (J_{n_k} - I_{n_k}) [\hat{\sigma}^{-1} \hat{\Sigma}_k^{-1/2} (\mathbf{y}_k - \mathbf{X}_k \hat{\boldsymbol{\beta}})] \right. \\ &\quad \left. - \kappa_1 \sum_{k=1}^N \text{tr} (\hat{\Sigma}_k^{-1} \dot{\omega}_{ka} (J_{n_k} - I_{n_k})) \right)_{q \times 1}, \end{aligned} \quad (4.3)$$

其中 $\dot{\omega}_{ka} = \partial \omega_k / \partial \gamma_a$, $\hat{\Sigma}_k = \mathbf{C}_k \hat{\boldsymbol{\Gamma}} \mathbf{C}_k + I_{n_k} + c\phi(J_{n_k} - I_{n_k})$.

H_0 成立时,关于 $\boldsymbol{\theta}$ 的Fisher信息阵为

$$I(\boldsymbol{\theta}) = \begin{bmatrix} I_{\boldsymbol{\gamma}\boldsymbol{\gamma}} & 0 & I_{\boldsymbol{\gamma}\phi} & I_{\boldsymbol{\gamma}\sigma^2} & I_{\boldsymbol{\gamma}\boldsymbol{\delta}} \\ 0 & I_{\boldsymbol{\beta}\boldsymbol{\beta}} & 0 & 0 & 0 \\ I_{\phi\boldsymbol{\gamma}} & 0 & I_{\phi\phi} & I_{\phi\sigma^2} & I_{\phi\boldsymbol{\delta}} \\ I_{\sigma^2\boldsymbol{\gamma}} & 0 & I_{\sigma^2\phi} & I_{\sigma^2\sigma^2} & I_{\sigma^2\boldsymbol{\delta}} \\ I_{\boldsymbol{\delta}\boldsymbol{\gamma}} & 0 & I_{\boldsymbol{\delta}\phi} & I_{\boldsymbol{\delta}\sigma^2} & I_{\boldsymbol{\delta}\boldsymbol{\delta}} \end{bmatrix}. \quad (4.4)$$

其非零子块为

$$\begin{aligned}
 I_{\gamma\gamma} &= \left(\frac{1}{2} \kappa_1 \hat{\phi}^2 \sum_{k=1}^N \text{tr}(\hat{\Sigma}_k^{-1} \dot{\omega}_{ka}(J_{n_k} - I_{n_k}) \hat{\Sigma}_k^{-1} \dot{\omega}_{kb}(J_{n_k} - I_{n_k})) \right)_{q \times q}, \\
 I_{\gamma\sigma^2} &= \left(\frac{1}{4} \hat{\phi} \kappa_1 (1 + \kappa_2) \hat{\sigma}^{-2} \sum_{k=1}^N \text{tr}(\hat{\Sigma}_k^{-1} \dot{\omega}_{ka}(J_{n_k} - I_{n_k})) \right)_{q \times 1}, \\
 I_{\gamma\delta} &= \left(\frac{1}{2} \kappa_1 \hat{\phi} \sum_{k=1}^N \text{tr}(\hat{\Sigma}_k^{-1} \mathbf{C}_k \frac{\partial \Gamma}{\partial d_{ab}} \mathbf{C}_k^T \Sigma_k^{-1} \dot{\omega}_{ka}(J_{n_k} - I_{n_k})) \right)_{q \times r'}, \\
 I_{\gamma\phi} &= \left(\frac{1}{2} \kappa_1 \hat{\phi} \sum_{k=1}^N \text{tr}(\hat{\Sigma}_k^{-1} \omega(\mathbf{v}_k, \gamma)(J_{n_k} - I_{n_k}) \hat{\Sigma}_k^{-1} \dot{\omega}_{ka}(J_{n_k} - I_{n_k})) \right)_{q \times 1}, \\
 I_{\phi\phi} &= \frac{1}{2} \kappa_1 \sum_{k=1}^N \text{tr}(\hat{\Sigma}_k^{-1} \omega(\mathbf{v}_k, \gamma)(J_{n_k} - I_{n_k}) \hat{\Sigma}_k^{-1} \omega(\mathbf{v}_k, \gamma)(J_{n_k} - I_{n_k})), \\
 I_{\phi\sigma^2} &= \frac{1}{4} \kappa_1 (1 + \kappa_2) \hat{\sigma}^{-2} \sum_{k=1}^N \text{tr}(\hat{\Sigma}_k^{-1} \omega(\mathbf{v}_k, \gamma)(J_{n_k} - I_{n_k})), \\
 I_{\phi\delta} &= \left(\frac{\kappa_1}{2} \sum_{k=1}^N \text{tr}(\hat{\Sigma}_k^{-1} \omega(\mathbf{v}_k, \gamma)(J_{n_k} - I_{n_k}) \hat{\Sigma}_k^{-1} \mathbf{C}_k \frac{\partial \Gamma}{\partial d_{ab}} \mathbf{C}_k^T) \right)_{1 \times r'}, \\
 I_{\beta\beta} &= \nu \hat{\sigma}^{-2} \sum_{k=1}^N \mathbf{X}_k^T \hat{\Sigma}_k^{-1} \mathbf{X}_k, \quad I_{\sigma^2\sigma^2} = \frac{1}{2} M \kappa_1 \hat{\sigma}^{-4}, \\
 I_{\sigma^2\delta} &= \left(\frac{1}{4} \kappa_1 (1 + \kappa_2) \hat{\sigma}^{-2} \sum_{k=1}^N \text{tr}(\hat{\Sigma}_k^{-1} \mathbf{C}_k \frac{\partial \Gamma}{\partial d_{ab}} \mathbf{C}_k^T) \right)_{1 \times r'}, \\
 I_{\delta\delta} &= \left(\frac{1}{2} \sum_{k=1}^N \text{tr}(\mathbf{C}_k \frac{\partial \Gamma}{\partial d_{ab}} \mathbf{C}_k^T \hat{\Sigma}_k^{-2} \mathbf{C}_k \frac{\partial \Gamma}{\partial d_{cd}} \mathbf{C}_k^T) \right)_{r' \times r'}.
 \end{aligned}$$

由参数稳健极大似然估计(RMLE)的渐近正态性(孙慧慧, 2011), 经计算得

$$J_N(\boldsymbol{\theta}) = N^{-1} I(\boldsymbol{\theta}). \quad (4.5)$$

综合以上结果, 可以得到一致相关系数齐性的Score检验统计量 SC_2 . 在正则条件的假设下, Score检验统计量 SC_1 和 SC_2 渐近服从 χ^2 分布(Sun等, 2011).

§5. 数值模拟

本节先通过随机模拟的数据扰动研究本文中M估计的稳健性, 再通过Monte Carlo模拟研究一致相关存在性和齐性的Score检验统计量的功效表现.

考虑如下具有一致相关协方差结构的纵向数据模型:

$$y_{ij} = \beta_1 + \beta_2 x_{ij} + \tau_i + e_{ij}, \quad i = 1, 2, \dots, N; j = 1, 2, \dots, m,$$

其中 $e_i \sim N(0, \sigma^2 V_i)$, $V_i = I_m + \phi_i(J_m - I_m)$, $\tau_i \sim N(0, \sigma^2 \Gamma)$. 假定一致相关系数 ϕ_i 具有如下形式:

$$\phi_i = \phi \frac{\exp(v_i \gamma)}{1 + \exp(v_i \gamma)}, \quad i = 1, 2, \dots, N.$$

协变量 x_{ij} 产生于 $[0, 40]$ 上的离散均匀随机数, 参数的真值设置为 $\beta_1 = 1, \beta_2 = 1.5, \sigma^2 = 0.05, \Gamma = 0.1$.

下面我们仅对随机效应 τ_i 的产生进行扰动来说明方差的M估计的稳健性, 具体做法是从分布 $(1 - \lambda)N(0, \sigma^2\Gamma) + \lambda N(0, 0.01)$ 中抽取 τ_i , 其中污染比例 λ 取0.1. 协变量 v_i 是 $[0, 14]$ 上的均匀随机数, $\phi = 0.02, \gamma = 0.2, N = 30, m = 20$, 抽取500个样本. 模拟结果见表1, 其中NP表示数据没有受到污染, P表示受到随机效应扰动污染, NR表示非稳健估计, R表示稳健估计. 由表1可以看到在数据无污染时, 虽然采用稳健方法会导致一定的效率损失, 但是方差的稳健估计和非稳健估计的MSE较接近. 当数据被污染后, 非稳健方法得到的随机效应方差 $\sigma^2\Gamma$ 估计的偏差增大, 而随机误差的方差 σ^2V_i 估计变化不大; 稳健方法得到的随机效应方差 $\sigma^2\Gamma$ 估计的MSE更小. 这主要由于我们是对随机效应 τ_i 进行扰动产生数据的污染.

表1 随机效应扰动对方差估计的影响模拟结果

| | σ^2V_i | | $\sigma^2\Gamma$ | |
|------|---------------|-------|------------------|-------|
| | Bias | MSE | Bias | MSE |
| NPNR | 0.078 | 0.142 | -0.056 | 0.163 |
| NPR | 0.089 | 0.171 | -0.063 | 0.182 |
| PNR | 0.069 | 0.153 | 0.703 | 0.925 |
| PR | 0.092 | 0.186 | 0.486 | 0.508 |

对应于不同的检验, 我们选择若干 N 和 m , 对检验功效进行模拟, 每种模拟均重复1000次, 在每次模拟中计算相应的统计量的值, 并与0.05显著性水平下的临界值相比较, 若该值大于临界值, 则拒绝原假设, 最后计算出1000次模拟中拒绝原假设的次数比例, 即为功效的模拟值.

表2 稳健形式下一致相关系数存在性检验的功效模拟

| $(N, m) \setminus \phi$ | -0.08 | -0.06 | -0.04 | -0.02 | 0 | 0.02 | 0.04 | 0.06 | 0.08 |
|-------------------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| (20,30) | 0.993 | 0.904 | 0.543 | 0.182 | 0.042 | 0.173 | 0.436 | 0.724 | 0.816 |
| | (1.000) | (0.982) | (0.744) | (0.346) | (0.046) | (0.216) | (0.604) | (0.819) | (0.897) |
| | (1.000) | (1.000) | (0.924) | (0.486) | (0.047) | (0.257) | (0.628) | (0.890) | (0.912) |
| (30,30) | 1.000 | 0.947 | 0.722 | 0.313 | 0.041 | 0.282 | 0.612 | 0.824 | 0.906 |
| | (1.000) | (0.983) | (0.817) | (0.399) | (0.045) | (0.307) | (0.724) | (0.882) | (0.928) |
| | (1.000) | (1.000) | (0.869) | (0.416) | (0.046) | (0.392) | (0.825) | (0.929) | (1.000) |
| (50,30) | 1.000 | 1.000 | 0.913 | 0.407 | 0.043 | 0.362 | 0.738 | 0.912 | 0.994 |
| | (1.000) | (1.000) | (0.946) | (0.438) | (0.046) | (0.495) | (0.842) | (0.983) | (1.000) |
| | (1.000) | (1.000) | (0.968) | (0.487) | (0.048) | (0.479) | (0.836) | (0.972) | (1.000) |

(1) 对一致相关存在性检验, 分别取 $\phi = 0, \pm 0.02, \pm 0.04, \pm 0.06, \pm 0.08$. 先对给定的 σ^2 和 ϕ , 从 $N(0, \sigma^2)$ 中产生具有相关性的误差序列 e_{ij} ; 再对给定的 σ^2 和 Γ , 从 $N(0, \sigma^2\Gamma)$ 中产生 τ_i , 通过真值和产生的 x_{ij}, τ_i, e_{ij} 得到 y_{ij} . 对不同的 (N, m) 和不同的 ϕ 的模拟重复1000次, 得到一致相关的存在性检验的模拟功效. 由表2知, 当 $\phi = 0$ 时, 功效接近0.05; 当 $|\phi|$ 或样本容量增加时, 功效增加, 且趋于1. 因此, 对较大的样本容量而言, SC_1 效果较好.

(2) 对一致相关系数齐性检验, 分别取 $\gamma = 0, \pm 0.1, \pm 0.2, \pm 0.3, \pm 0.4$, 对给定的 γ 和 σ^2 产生具有相关性的误差序列 e_{ij} ; 相关系数协变量 v_i 是 $[0, 14]$ 上的均匀随机数. 结合给定的 γ 产生一致相关系数序列 ϕ_i ; 通过真值和产生的 x_{ij}, τ_i, e_{ij} 得到具有一致相关误差的序列 y_{ij} . 对不同的 (N, m) 和不同的 γ 的模拟重复1000次, 得到一致相关齐性检验的模拟功效. 由表3知, 当 $\gamma = 0$ 时, 功效接近0.05, 当 $|\gamma|$ 或样本容量增加时, 功效也增加, 且趋于1, 因此统计量 SC_2 对大样本容量的检验效果较好.

表3 稳健形式下一致相关系数齐性检验的功效模拟

| $(N, m) \setminus \gamma$ | -0.4 | -0.3 | -0.2 | -0.1 | 0 | 0.1 | 0.2 | 0.3 | 0.4 |
|---------------------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| (10,10) | 0.883 | 0.704 | 0.403 | 0.172 | 0.041 | 0.153 | 0.282 | 0.614 | 0.823 |
| | (0.917) | (0.825) | (0.519) | (0.212) | (0.045) | (0.224) | (0.513) | (0.841) | (0.915) |
| | (1.000) | (0.970) | (0.792) | (0.346) | (0.046) | (0.397) | (0.708) | (0.903) | (0.988) |
| (15,10) | 0.932 | 0.879 | 0.673 | 0.384 | 0.042 | 0.314 | 0.510 | 0.827 | 0.925 |
| | (0.982) | (0.903) | (0.799) | (0.328) | (0.044) | (0.356) | (0.706) | (0.912) | (0.983) |
| | (1.000) | (0.993) | (0.826) | (0.509) | (0.045) | (0.416) | (0.830) | (0.957) | (1.000) |
| (20,15) | 1.000 | 1.000 | 0.924 | 0.583 | 0.044 | 0.483 | 0.718 | 0.983 | 1.000 |
| | (1.000) | (1.000) | (0.959) | (0.631) | (0.047) | (0.526) | (0.890) | (0.992) | (1.000) |
| | (1.000) | (1.000) | (0.979) | (0.623) | (0.048) | (0.649) | (0.916) | (1.000) | (1.000) |

以上表格中括号内为相应Huber函数中的常数分别取 $c = 2$ 和 $c = \infty$ (即非稳健形式)时的模拟值, 总体来说, 比稳健形式下的值偏大, 这也是合理的.

§6. 实例分析

葡萄糖数据由美国科罗拉多州医疗中心大学小儿科临床研究病房提供, 该数据通过对13个控制病人和20个肥胖病人测试其标准葡萄糖忍耐力. 实验过程为: 让33个病人服用葡萄糖, 分别在0, 0.5, 1, 1.5, 2, 3, 4, 5小时后测其血样. 实验目的是为了研究比较控制组的病人和肥胖组的病人是否有显著区别. 很显然, 这是一个典型的与时间有关的纵向数据. Chi和Reinsel (1989), Pan和Fang (2002)分别对该组数据进行了研究, 林金官和韦博成 (2004b)在对该数据有、无组内自相关的假设下, 进行了异方差和自相关性的分析. 本文对

该数据采用以下的线性纵向数据模型来进行分析:

$$\mathbf{y}_i = \mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{1}\tau_i + \mathbf{e}_i, \quad i = 1, 2, \dots, 33,$$

其中 $\mathbf{y}_i = (y_{ij})$, $j = 1, 2, \dots, 8$; y_{ij} 为第 i 个病人第 j 次的测量结果, 由Pan和Fang (2002)采用 \mathbf{X}_i 的设计阵如下:

$$\mathbf{X}_i = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0.5 & 1 & 1.5 & 2 & 3 & 4 & 5 \\ 0 & 0.25 & 1 & 2.25 & 4 & 9 & 16 & 25 \\ 0 & 0.125 & 1 & 3.375 & 8 & 27 & 64 & 125 \end{bmatrix}.$$

下面在方差齐性的假设下, 对这组数据进行一致相关存在性检验和一致相关齐性检验. 对于异方差的情形, 在文献(林金官等, 2004b)中已有研究. 令 $\omega_k = \exp(v_k\gamma)/[1 + \exp(v_k\gamma)]$, 其中 v_k 为各控制病人的属性量. 则当 $\gamma = 0$ 时, $\omega_k = 1/2$, 是与 k 无关的常量. 因此, 一致相关系数齐性的检验为 $H_0: \gamma = 0$ 是否成立. 首先计算在方差齐性和 $\phi = 0$ 时参数的稳健极大似然估计:

$$\hat{\boldsymbol{\beta}} = (4.3629, -1.4213, 0.4900, -0.0434)^T, \quad \hat{\sigma}^2 = 0.3530, \quad \hat{\Gamma} = 1.6892.$$

根据本文得出的一致相关存在性检验统计量 SC_1 得到结果如下表所示.

表4 葡萄糖数据一致相关存在性检验结果

| | SC1 | SC1' |
|----------|---------|---------|
| Score统计量 | 16.6204 | 31.4769 |
| p-值 | 0.00 | 0.00 |

其中SC1表示用检验统计量 SC_1 计算得到, $SC1'$ 表示普通极大似然下相应的检验统计量, 由表4结果可以看出, 葡萄糖数据存在一致相关结构, 且稳健情形下的Score检验统计量的值明显比普通情形下相应的值小. 另外, 这里取Huber函数中的常数 $c = 1.345$, 若取作 $c = 0.8$, 则一致相关存在性检验的Score检验统计量值为26.6817 (0.00), 括号内为相应的p值. 由此我们可以看出, 由于Huber函数的限制作用, 使得模型对数据的敏感性降低, 比较稳健, 且常数 c 越小, 限制越强, 敏感性越低.

其次研究方差齐性时的一致相关系数的齐性检验, H_0 成立时, 即 $\gamma = 0$ 时, 各参数的稳健极大似然估计为

$$\hat{\boldsymbol{\beta}} = (4.3834, -1.4032, 0.4888, -0.0441)^T, \\ \hat{\sigma}^2 = 0.3602, \quad \hat{\Gamma} = 1.6998, \quad \hat{\phi} = 0.3963.$$

由检验统计量 SC_2 计算结果如表5:

表5 葡萄糖数据一致相关系数齐性检验结果

| 常数 c | Score统计量 | p-值 |
|----------|----------|--------|
| 0.8 | 1.0816 | 0.2983 |
| 1.345 | 1.2047 | 0.2724 |
| ∞ | 1.3814 | 0.2399 |

由表5中结果可以看出,葡萄糖数据在方差齐性时一致相关系数具有显著齐性.且由于Huber函数的限制作用,使得一致相关齐性比非稳健情形下显著,且常数 c 愈小,一致相关齐性愈显著.

§7. 结 语

本文研究的是纵向数据的线性混合模型,我们在此模型稳健形式的对数似然函数下,用Fisher得分迭代法得到了模型参数的M估计(稳健估计).接着研究了模型在方差齐性假设下的一致相关存在性和齐性的检验问题,并得到了检验的Score统计量.实例分析说明了本文方法的有效性.此外,文中选用代替二次函数的稳健函数是Huber函数,也可以选用其他有界函数,如Hampel函数, Tukey bisquare函数等进行M估计.

参 考 文 献

- [1] Diggle, P.J., Liang, K.Y. and Zeger, S.T., *Analysis of Longitudinal Data*, New York: Oxford University Press, 2002.
- [2] Verbeke, G. and Molenberghs, G., *Linear Mixed Models for Longitudinal Data*, New York: Springer, 2000.
- [3] Pinheiro, J.C. and Bates, D.M., *Mixed-Effects Models in S and S-PLUS*, New York: Springer-Verlag, 2000.
- [4] 林金官, 韦博成, 非线性纵向数据模型中自相关性和随机效应的存在性检验, *应用数学*, **17(1)**(2004a), 42-48.
- [5] 林金官, 韦博成, 非线性纵向数据模型中方差和自相关系数的齐性检验, *应用数学学报*, **27(3)**(2004b), 466-480.
- [6] 林金官, 李勇, 韦博成, 基于连续可分的广义非线性纵向数据模型偏离名义离差的score检验及其功效, *应用数学学报*, **32(1)**(2009), 60-74.
- [7] Wolfinger, R.D., Heterogeneous variance-covariance structures for repeated measures, *Journal of Agricultural, Biological and Environmental Statistics*, **1(2)**(1996), 205-230.
- [8] 范俊花, 林金官, 韦博成, 具有一致相关的纵向数据模型中方差和相关系数的齐性检验, *应用概率统计*, **25(1)**(2009), 12-26.
- [9] Huber, P.J., *Robust Statistics*, New York: Wiley, 1981.

- [10] Sun, H.H. and Lin, J.G., Testing for heteroscedasticity in mixed effect linear models based on M-estimation, *应用数学*, **24(4)**(2011), 798–805.
- [11] Núñez-Antón, V. and Zimmerman, D.L., Modeling nonstationary longitudinal data, *Biometrics*, **56(3)**(2000), 699–705.
- [12] Cox, D.R. and Hinkley, D.V., *Theoretical Statistics*, London: Chapman and Hall, 1974.
- [13] 孙慧慧, 纵向数据线性混合模型中M估计的渐近正态性, *周口师范学院学报*, **28(5)**(2011), 21–23.
- [14] Chi, E.M. and Reinsel, G.C., Models for longitudinal data with random effects and AR(1) errors, *Journal of the American Statistical Association*, **84(406)**(1989), 452–459.
- [15] Pan, J.X. and Fang, K.T., *Growth Curve Models and Statistical Diagnostics*, New York: Springer-Verlag, 2002.

Testing for Correlation Coefficients in Uniform Correlation Longitudinal Mixed Effect Linear Models Based on M-estimation

SUN HUIHUI

(*School of Mathematics and Science, Yancheng Teachers University, Yancheng, 224002*)

In this paper, the Fisher scoring method is applied to get M-estimator (robust estimator) in the mixed effects linear model for longitudinal data. The score tests for correlation coefficients in the model with uniform correlation covariance structure based on M-estimator are also studied. Then the properties of test statistics are investigated through Monte Carlo simulations. At last, the methods and properties are illustrated by the grape sugar data example.

Keywords: M-estimation, longitudinal data, uniform correlation, score test.

AMS Subject Classification: 62J20.