

纵向数据下广义经验似然方法的有效稳健估计

徐孝琳, 樊亚莉, 苏依官

(上海理工大学 理学院 上海 200093)

摘要: 基于广义经验似然估计方法, 提出了一种有效且稳健的估计, 实现对纵向数据在线性模型下均值和协方差矩阵的联合估计。利用 Cholsky 分解将模型重参数化, 利用拉格朗日乘子法求出估计值, 再还原出均值和协方差矩阵的估计。在模拟研究中将所提方法同文献中其他稳健估计进行比较, 结果显示所提方法效率更高。最后将所提方法用于分析 CD4 细胞数据, 交叉验证结果显示所提方法更加可靠。

关键词: 纵向数据; 稳健性; 有效性; 经验似然; 交替优化

中图分类号: O 212.1 **文献标志码:** A

Effective Robust Estimation Based on the Generalized Empirical Likelihood Method for Longitudinal Data

XU Xiaolin, FAN Yali, SU Yiguan

(College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Based on the generalized empirical likelihood estimation method, a robust and effective estimation was proposed to realize the joint estimation of the mean and covariance matrix of longitudinal data in linear models. Using the Cholsky decomposition, the model was re-parameterized, and using the Lagrangian multiplier method, estimates were obtained, then, an estimate of the mean and covariance matrix was restored. Comparing the proposed method with other robust estimates in the literature in a simulation study, the results show that the proposed method is more efficient. Finally, the proposed method was used to analyze CD4 cell data. Cross-validation results show that the proposed method is more reliable.

Keywords: longitudinal data; robustness; effectiveness; empirical likelihood; alternating optimization

纵向数据是指对同一个个体或受试单位在不同时间进行重复观测得到的数据。观测不同个体所得数据是独立的, 但对同一个个体观测所得的数据往往具有相关性。因此, 纵向数据具有组间

独立、组内相关的特点, 是数理统计研究中的复杂数据类型之一, 在生物、医学及经济学等领域都有广泛的应用。在实际中经常需要对纵向数据进行统计分析和建模。

收稿日期: 2018-04-20

基金项目: 国家自然科学基金资助项目 (11401383)

第一作者: 徐孝琳(1993-), 女, 硕士研究生。研究方向: 概率论与数理统计。E-mail: 992104062@qq.com

通信作者: 樊亚莉(1978-), 女, 讲师。研究方向: 概率论与数理统计。E-mail: yalifan@usst.edu.cn

对于纵向数据的研究,学者们已经提出了各种参数模型和统计方法。参数模型易于解释和统计分析,能很好地反映协变量与响应变量的关系,但存在模型误判风险,一旦模型假设错误,直接导致得到错误的结论;非参数方法不需要对模型进行假定,具有非常高的灵活性,广义经验似然估计是非参数方法中常用的一种。Thomas 等^[1]最早阐述了经验似然方法的思想;Owen^[2]进一步系统地研究了经验似然方法,他指出经验似然是一种构造未知参数的置信区间和置信域的非参数推断方法,其本质是在约束条件下求非参数似然比的极大值,而总体参数由约束条件带入似然比中。经验似然方法自提出后即引起了很多统计学家的兴趣,他们将这一方法应用到统计的诸多领域;1991 年 Owen^[3]将经验似然方法应用于独立同分布样本总体均值的统计推断中,构造了非参数经验似然比统计量,并进一步对线性模型和广义线性模型进行了深入的研究;Bondell 等^[4]提出了一种基于最小化 C-R 距离的两阶段广义经验似然方法,该方法对残差大的样本赋予小的权重,从而达到对异常值稳健的目的。

经验似然方法与一些经典的统计方法相比,其优点在于它不需要对模型作任何参数形式的假设,也不需要估计方差,只要矩条件正确,则估计一定是渐近有效的。

本文研究的内容是纵向数据的均值和协方差的联合估计,对此前人已作了许多研究,He 等^[5]提出了一种基于半参数广义线性模型的稳健估计方程,用来估计纵向数据的均值和协方差,通过样条回归和得分函数从而达到对异常值稳健的目的;Qin 等^[6]基于广义估计方程,提出了逆概率加权法来估计纵向数据的均值和协方差,估计效果较稳健;Qin 等^[7]建立了稳健估计方程,利用 Cholesky 分解,实现了纵向数据在缺失机制下,均值和协方差的联合稳健估计,并且对数据异常值具有一定的抵抗性。Bondell 等^[4]通过广义经验似然方法中约束条件的限制,在保证稳健性的同时又提高了估计的有效性。但 Bondell 等^[4]的方法只适用于横截面数据,没有考虑纵向数据。本文在 Bondell 等^[4]的启发下,结合 Cholesky 分解将纵向数据的线性模型重参数化,利用广义经验似然估计方法,实现了纵向数据的均值和协方差的稳健联合估计,在保证估计的稳健性的同时,还提高了估计的有效性。

1 模型与估计方法

1.1 初始模型

假设纵向数据有 n 个对象,对于每一个对象有 m 次观察值,考虑线性模型

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_0 + \varepsilon_{ij}, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, m \quad (1)$$

式中: $\boldsymbol{\beta}_0$ 是 p 维未知的回归系数向量; $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{im})^T (i = 1, 2, \dots, n)$ 独立同分布,且服从于均值为 $\mathbf{0}_{m \times 1}$, 协方差 $\boldsymbol{\Sigma}_0$ 未知的分布。

记 $\mathbf{Y}_i = (y_{i1}, \dots, y_{im})^T$, $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{im})^T$, 在本文中主要研究模型(1)中 $\boldsymbol{\beta}_0$ 和 $\boldsymbol{\Sigma}_0$ 的有效稳健估计。

1.2 模型重参数化

首先对 $\boldsymbol{\Sigma}_0$ 进行 Cholesky 分解,实现 $\boldsymbol{\beta}_0$ 和 $\boldsymbol{\Sigma}_0$ 的联合估计。即存在对角线元素都是 1 的下三角矩阵 $\mathbf{C}$, 使得 $\text{cov}(\mathbf{C}\boldsymbol{\varepsilon}_i) = \mathbf{C}\boldsymbol{\Sigma}_0\mathbf{C}^T = \mathbf{B}$, 其中, $\mathbf{B}$ 是对角矩阵, 记 $\mathbf{B} = \text{diag}(b_1^2, b_2^2, \dots, b_m^2)$ 。令

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -c_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -c_{m1} & -c_{m2} & \cdots & 1 \end{pmatrix}$$

记 $\mathbf{e}_i = (e_{i1}, \dots, e_{im})^T = \mathbf{C}\boldsymbol{\varepsilon}_i$, $\text{var}(e_{ij}) = b_j^2, j = 1, 2, \dots, m$ 。则有 $\varepsilon_{i1} = e_{i1}$; $\varepsilon_{ij} = c_{j1}\varepsilon_{i1} + c_{j2}\varepsilon_{i2} + \cdots + c_{j,j-1}\varepsilon_{i,j-1} + e_{ij}$, $i = 1, 2, \dots, n, j = 1, 2, \dots, m$ 。

根据上述关系,模型(1)可以改写为

$$\begin{cases} y_{i1} = \mathbf{x}_{i1}^T \boldsymbol{\beta}_0 + e_{i1} \\ y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_0 + c_{j1}e_{i1} + \cdots + c_{j,j-1}e_{i,j-1} + e_{ij}, \\ i = 1, 2, \dots, n, \quad j = 2, 3, \dots, m \end{cases} \quad (2)$$

此处由于 ε_{ij} 未知,故先对模型 $y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_0 + \varepsilon_{ij}, i = 1, 2, \dots, n, j = 1, 2, \dots, m$ 中的数据用最小二乘估计出初始 $\hat{\boldsymbol{\beta}}_0$, 则 $\hat{\varepsilon}_{ij} = y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_0$, 令 $\varepsilon_{ij} = \hat{\varepsilon}_{ij}$, 代入模型(2)中。

当得到 $c_{ij} (i = 1, 2, \dots, m, j = 1, 2, \dots, (m-1))$, $b_j (j = 1, 2, \dots, m)$ 和 $\boldsymbol{\beta}_0$ 的相合估计值后,则可根据 Cholesky 分解得 $\hat{\boldsymbol{\Sigma}}_0 = \hat{\mathbf{C}}^{-1} \hat{\mathbf{B}} (\hat{\mathbf{C}}^T)^{-1}$, 从而实现均值和协方差矩阵的联合估计。

基于 Cholesky 分解,为了将模型(1)重参数化后改写成一般的模型,需作如下处理:

令 $\mathbf{s}_0 = (c_{21}, c_{31}, c_{32}, \dots, c_{m,m-1})^T$, $\hat{\mathbf{Z}}_{ij} = (\boldsymbol{\theta}_{(j-2)(j-1)/2}^T, \hat{\varepsilon}_{i1}, \dots, \hat{\varepsilon}_{i,j-1}, \boldsymbol{\theta}_{(m-1)m/2-(j-1)j/2}^T)^T$, $\hat{\mathbf{Z}}_i^{(-1)} = (\hat{\mathbf{Z}}_{i2}, \dots, \hat{\mathbf{Z}}_{im})^T$, $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \mathbf{s}_0^T)^T$, $\mathbf{X}_i^{(-1)} = (\mathbf{x}_{i2}, \dots, \mathbf{x}_{im})^T$, $\mathbf{D}_1 = (\mathbf{x}_{i1}^T, \boldsymbol{\theta}_{m(m-1)/2}^T)^T$, 则 $\hat{\mathbf{D}}_i^{(-1)} = (\mathbf{X}_i^{(-1)T}, \hat{\mathbf{Z}}_i^{(-1)T})^T$, $\hat{\mathbf{D}}_i = (\mathbf{D}_1, \hat{\mathbf{D}}_i^{(-1)})^T$ 。

在此变换后,原始的纵向数据的模型(1)被改

写成 $Y_i = \hat{D}_i \theta + e_i$, $i = 1, 2, \dots, n$, 其中, Y_i 和 e_i 都是 m 维列向量, 进一步记

$$\begin{aligned} Y &= (Y_1, Y_2, \dots, Y_n)^T = (y_1, y_2, \dots, y_{nm})^T \\ \hat{D} &= (\hat{D}_1, \hat{D}_2, \dots, \hat{D}_n)^T = (d_1^T, d_2^T, \dots, d_{nm}^T)^T \\ e &= (e_1, e_2, \dots, e_n)^T = (\omega_1, \omega_2, \dots, \omega_{nm})^T \end{aligned}$$

则模型最终化为

$$y_k = d_k^T \theta + \omega_k, \quad k = 1, 2, \dots, N \quad (3)$$

式中: $N = nm$; y_k 是一个数; $\omega_k (k = 1, 2, \dots, N)$ 相互独立; d_k 为 $[p + (m-1)m/2]$ 维列向量; θ 为 $[p + (m-1)m/2]$ 维列向量。

1.3 广义经验似然估计

定义广义经验似然目标函数

$$L(\theta, w) = \min \left\{ \sum_{k=1}^N w_k \log(Nw_k) \right\} \quad (4)$$

约束条件

$$\sum_{k=1}^N w_k = 1, \quad \sum_{k=1}^N w_k (y_k - d_k^T \theta) d_k = 0 \quad (5)$$

$$\sum_{k=1}^N w_k (y_k - d_k^T \theta)^2 = \tilde{\sigma}_T^2 \quad (6)$$

在上述条件中, w_k 代表权重, $N = nm$, $w = (w_1, w_2, \dots, w_N)$ 。本文提出的估计量是一个广义经验似然估计量, 它是经验似然方法的延伸, 是基于最小化 C-R 距离的统计量, C-R 距离最小化也就意味着其经验分布差异最小化。当权重无限接近于等权重时, 才能使得估计的有效性达到最好, 具体可参考文献[8-9]。

本文所提出的估计是两阶段的估计, 第一阶段, 在约束条件式(5)和式(6)下, 目标函数式(4)达到最小值时, 可得对应 θ 和 w_k 的估计值, 即

$$\tilde{\theta} = \arg \min L(\theta, w), \quad \hat{w}_k = \arg \min L(\theta, w)$$

第二阶段, θ 的最终估计值是基于模型(3)和 $\hat{w}_k$ 的加权估计量, 记 $W = \text{diag}(\hat{w}_1, \hat{w}_2, \dots, \hat{w}_N)$, $D = (d_1, d_2, \dots, d_N)$, $Y = (y_1, y_2, \dots, y_N)^T$ 。则 θ 的最终估计值为

$$\hat{\theta} = (D^T W D)^{-1} D^T W Y \quad (7)$$

在上述条件中, w_k 代表权重, 在仅考虑条件式(5)时, 目标函数将在 $w_k = 1/N$, $k = 1, 2, \dots, N$ 达到全局最小。约束条件式(5)保证了该经验似然方法的有效性, 加入条件式(6)则进一步保证本文提出方法的稳健性, 其中, $\tilde{\sigma}_T^2$ 是一个需要预估的残差尺度参数, 此处需要要求 $\tilde{\sigma}_T^2 \leq \hat{\sigma}_{OLS}^2$, $\hat{\sigma}_{OLS}^2 =$

$(1/N) \sum_{k=1}^N (y_k - d_k^T \hat{\theta}_{OLS})^2$ 。 $\tilde{\sigma}_T^2$ 的引入, 使得本文提出的方法不仅有效, 而且稳健。现进一步说明估计量式(7)稳健的原因。

记 $r_k^2(\hat{\theta}) = y_k - d_k^T \hat{\theta}$, 因为, $\sum_{k=1}^N \hat{w}_k \{r_k^2(\hat{\theta}) - \tilde{\sigma}_T^2\} = 0$, 且 $\sum_{k=1}^N \hat{w}_k = 1$ 。在假设 $\tilde{\sigma}_T^2 \leq \hat{\sigma}_{OLS}^2$ 的前提下, 可得

$$\begin{aligned} \sum_{k=1}^N \hat{w}_k r_k^2(\hat{\theta}) &= \sum_{k=1}^N \hat{w}_k \tilde{\sigma}_T^2 = \tilde{\sigma}_T^2 \leq \hat{\sigma}_{OLS}^2 = \\ (1/N) \sum_{k=1}^N r_k^2(\hat{\theta}_{OLS}) &\leq (1/N) \sum_{k=1}^N r_k^2(\hat{\theta}) \end{aligned} \quad (8)$$

因为, $(1/N) \sum_{k=1}^N w_k = 1/N$, 所以, 式(8)可变为

$$\sum_{k=1}^N \hat{w}_k r_k^2(\hat{\theta}) \leq (1/N) \sum_{k=1}^N r_k^2(\hat{\theta}) \Rightarrow \sum_{k=1}^N \{\hat{w}_k - (1/N)\} r_k^2(\hat{\theta}) \leq 0$$

因此, $r_k^2(\hat{\theta})$ 和 $\hat{w}_k$ 是负相关的。当残差 $r_k^2(\hat{\theta})$ 比较大时, $\hat{w}_k$ 比较小; 当残差 $r_k^2(\hat{\theta})$ 比较小时, $\hat{w}_k$ 比较大。这也是本文方法稳健的根本原因。

总结本文估计方法的步骤如下:

a. 根据每一组对象的第一个观测值, 利用最小二乘法, 得到 β_0 的初始估计值, 记为 $\hat{\beta}_b$, 从而有 $\hat{\varepsilon}_{ij} = y_{ij} - x_{ij}^T \hat{\beta}_b$, 并且令 $\varepsilon_{ij} = \hat{\varepsilon}_{ij}$ 。

b. 对于模型(3), $\theta = (\beta_0^T, s_0^T)^T$, 采用 Bondell 等^[4]提出的两阶段方法, 得到 $\hat{\theta}$, 提取 $\hat{\theta}$ 的前 p 个分量即为 $\hat{\beta}_0$, 再根据 $\hat{\omega}_k = y_k - d_k^T \hat{\theta}$, 利用绝对中位数离差 MAD 方法得到 b_j^2 的稳健估计。

c. 当求得 $\hat{s}_0$ 和 $\hat{b}_j^2$ 后, 可根据 Cholesky 分解还原, 就能得到纵向数据协方差矩阵 $\hat{\Sigma}_0$ 。

2 交替优化算法

为了解决带有约束条件的目标优化函数式(4), 采用拉格朗日算法。

定义 $\Lambda = (\lambda_1, \lambda_2^T, \lambda_3)^T$

$$L_0(w, \theta, \Lambda) = \sum_{k=1}^N w_k \log(Nw_k) -$$

$$\lambda_1 \left(\sum_{k=1}^N w_k - 1 \right) - \lambda_2^T \sum_{k=1}^N w_k (y_k - d_k^T \theta) d_k -$$

$$\lambda_3 \sum_{k=1}^N w_k \left\{ (y_k - d_k^T \theta)^2 - \tilde{\sigma}_T^2 \right\}$$

令 $\partial L_0 / \partial w_k = 0$, $k = 1, 2, \dots, N$, $\partial L_0 / \partial \lambda_1 = 0$, 解得

$$w_k = w_k^* / \sum_{k=1}^N w_k^*$$

$$\begin{cases} J_1(\hat{\lambda}_{3Q}, \hat{\theta}_Q) = 0 \\ J_2(\hat{\lambda}_{3Q}, \hat{\theta}_Q) = 0 \end{cases} \quad (14)$$

其中,

$$w_k^* = \exp[\lambda_2^T d_k r_k(\theta) + \lambda_3 \{r_k(\theta) - \tilde{\sigma}_T^2\}]$$

令 $\partial L_0 / \partial \lambda_2 = 0, \partial L_0 / \partial \lambda_3 = 0, \partial L_0 / \partial \theta = 0$, 可得

$$\begin{cases} \sum_{k=1}^N w_k^* r_k(\theta) d_k = 0 \\ \sum_{k=1}^N w_k^* \{r_k^2(\theta) - \tilde{\sigma}_T^2\} = 0 \\ \sum_{k=1}^N w_k^* \{d_k d_k^T \lambda_2 + 2r_k(\theta) d_k\} = 0 \end{cases} \quad (9)$$

由式(9)可以解得 $\lambda_2 = \theta$, 所以, 最终得到关于 $\lambda_3, \hat{\theta}$ 和 $\hat{w}_k^*$ 的方程组

$$\begin{cases} \sum_{k=1}^N \hat{w}_k^* \{r_k^2(\hat{\theta}) - \tilde{\sigma}_T^2\} = 0 \\ \sum_{k=1}^N \hat{w}_k^* r_k(\hat{\theta}) d_k = 0 \\ \hat{w}_k^* - \exp[\lambda_3 \{r_k^2(\hat{\theta}) - \tilde{\sigma}_T^2\}] = 0, \\ k = 1, 2, \dots, N \end{cases} \quad (10)$$

对于方程组(10), 可将其转变成交替极大值和极小值问题。记

$$J(\lambda_3, \theta) = (1/N) \sum_k \exp[\lambda_3 \{r_k^2(\theta) - \tilde{\sigma}_T^2\}] =$$

$$\{\exp(\lambda_3 \tilde{\sigma}_T^2) / N\} \sum_k \exp\{\lambda_3 r_k^2(\theta)\}$$

令 $J_1(\lambda_3, \theta) = \partial J / \partial \lambda_3, J_{11}(\lambda_3, \theta) = \partial^2 J / \partial \lambda_3^2, J_2(\lambda_3, \theta) = \partial J / \partial \theta$ 。

容易验证

$$\begin{cases} J_1(\lambda_3, \theta) = 0 \\ J_2(\lambda_3, \theta) = 0 \end{cases} \quad (11)$$

等式(11)与方程(10)等价。

容易求得 $J_{11}(\lambda_3, \theta) > 0, J_1(0, \theta) \leq 0$, 所以, 当固定 $\theta, J(\lambda_3, \theta)$ 是关于 λ_3 的凸函数, 所以, $J(\lambda_3, \theta)$ 一定能取到极小值。

现固定 θ , 求关于 λ_3 的函数 $J(\lambda_3, \theta)$ 取极小值时, λ_3 的取值 $\hat{\lambda}_3(\theta)$, 其对应点的一阶导数必为 0, 即

$$J_1(\hat{\lambda}_3(\theta), \theta) = 0 \quad (12)$$

将 $\lambda_3 = \hat{\lambda}_3(\theta)$ 代入 $J(\lambda_3, \theta)$, 得到仅与 θ 有关的函数 $Q(\theta) = J(\hat{\lambda}_3(\theta), \theta)$, 当 $Q(\theta)$ 取极大值时, θ 对应的取值记为 $\hat{\theta}_Q$, 则其对应点的一阶导数必为 0, 即

$$J_1(\hat{\lambda}_3(\hat{\theta}_Q), \hat{\theta}_Q) * (\partial \hat{\lambda}_3(\theta) / \partial \theta) + J_2(\hat{\lambda}_3(\hat{\theta}_Q), \hat{\theta}_Q) = 0 \quad (13)$$

根据定义 $J_1(\hat{\lambda}_3(\hat{\theta}_Q), \hat{\theta}_Q) = 0$, 式(13)可以表示为

计算步骤总结如下:

- a. 取定 $\hat{\sigma}_T^2 = \min(\hat{\sigma}_{OLS}^2, \hat{\sigma}_{LTS}^2)$, $\hat{\sigma}_{LTS}^2$ 是样本在截尾最小二乘估计下的残差尺度的估计。
- b. 选取截尾最小二乘估计 $\hat{\theta}^{(0)}$ 作为 θ 的初始估计值, 另外, 给定 ε 为一个很小的数。
- c. 取 $\theta = \hat{\theta}^{(0)}$, 求关于 λ_3 的函数 $J(\lambda_3, \hat{\theta}^{(0)})$ 的极小值, 即 $J_1(\lambda_3, \hat{\theta}^{(0)}) = 0$ 时对应的 λ_3 的取值, 记为 $\hat{\lambda}_3^{(0)} = \arg \min J(\lambda_3, \hat{\theta}^{(0)})$ 。
- d. 再固定 $\lambda_3 = \hat{\lambda}_3^{(0)}$, 求关于 θ 的函数 $J(\hat{\lambda}_3^{(0)}, \theta)$ 中的极大值, 即 $J_2(\hat{\lambda}_3^{(0)}, \theta) = 0$ 时对应 θ 的估计值 $\hat{\theta}^{(1)}$, 此时 $\hat{\theta}^{(1)}$ 即为更新后的参数估计值。
- e. 重复步骤 c—d, 第 i 次更新后的值记为 $\hat{\theta}^{(i)}$, 直至 $|\hat{\theta}^{(i+1)} - \hat{\theta}^{(i)}| < \varepsilon$, 或者, 循环次数 $i > 1\,000$ 截止, 此时 $\hat{\theta}^{(*)}$ 记为 $\theta^{(i)}$ 的最终迭代值。
- f. 计算样本在 $\hat{\theta}^{(*)}$ 下的残差 $r_k(\hat{\theta}^{(*)})$, 则权重 $w_k = \exp\{\hat{\lambda}_3^{(0)} r_k(\hat{\theta}^{(*)})\} / \sum_{k=1}^N \exp\{\hat{\lambda}_3^{(0)} r_k(\hat{\theta}^{(*)})\}$, 最终 θ 的估计值由式(7)可得, 提取 $\hat{\theta}$ 的前 p 个分量为均值 $\hat{\beta}_0$, 最后再利用 Cholesky 分解, 通过对 $\hat{\theta}$ 的反向分解和还原, 得到纵向数据协方差矩阵 $\hat{\Sigma}_0$ 。

3 模拟研究

根据模型 $Y_{ij} = \beta_1 + x_{ij1}\beta_2 + x_{ij2}\beta_3 + \varepsilon_{ij}, i = 1, 2, \dots, n, j = 1, 2, \dots, m$ 产生数据, 设 $n = 100, m = 3, x_{ij1} \sim N(0, 1), x_{ij2} \sim N(0, 1)$, 残差项 ε_{ij} 服从均值为 0, 协方差结构分别为 I 单位结构、 $\rho = 0.5$ 的可交换结构、 $\rho = 0.5$ 的 AR(1) 和无结构的 m 维正态分布。其中, 无结构可定义为

$$\begin{bmatrix} 1 & 0.7 & 0.4 \\ 0.7 & 1 & 0.2 \\ 0.4 & 0.2 & 1 \end{bmatrix}$$

$\beta_0 = (\beta_1, \beta_2, \beta_3)^T = (1, 1, 1)^T$, 模拟重复次数为 100 次。 θ 的初始估计值选取数据在最小二乘估计下的参数估计值 $\hat{\theta}_{OLS}$, 记 Xu 为本文提出的方法, He 是 He 等^[5]提出的稳健广义估计方程的方法, Qin 是 Qin 等^[7]提出的稳健方程的估计方法, 由于 Qin 等^[7]研究的纵向数据带有缺失, 而本文只考虑了污染, 故在模拟时将缺失示性函数全部改为 1, 代表数据不缺失。此外, 数据可能存在污染等异常情况, 本文 β_0 是最终估计式(7)中 $\hat{\theta}$ 的前 p 个分量, 协方差 Σ 通过返还函数对 $\hat{\theta}$ 反向分解和还原可

得, 在返还函数中, R 表示采用了绝对中位数离差 MAD 方法, NR 表示一般处理方法, 以此来验证本文提出的方法是否对异常值具备抵抗性。另外, 两种方法中 R 和 NR 均代表各自的稳健方法

和非稳健方法, 表 1 比较了 3 种方法得到的 $\hat{\beta}_0$ 的偏差 (BIAS)、标准差 (SE) 和均方误差 (MSE)。

表 1 数据不污染时参数估计对比

Tab.1 Parameter estimation and comparison when data are not polluted

不污染			β_1			β_2			β_3		
			BIAS	SE	MSE	BIAS	SE	MSE	BIAS	SE	MSE
I 单位	Xu	R	-0.006 0	0.049 5	0.002 5	-0.003 6	0.065 3	0.004 3	-0.002 2	0.063 9	0.004 0
	Qin	NR	-0.010 0	0.058 2	0.003 5	-0.004 8	0.059 5	0.003 6	-0.003 7	0.063 4	0.004 0
	Qin	R	-0.011 1	0.061 2	0.003 9	-0.002 9	0.061 0	0.003 7	-0.002 8	0.064 8	0.004 2
	He	NR	-0.009 3	0.058 0	0.003 5	-0.006 8	0.069 7	0.004 9	-0.010 8	0.067 0	0.004 6
	He	R	-0.008 9	0.059 4	0.003 6	-0.004 9	0.071 2	0.005 1	-0.008 2	0.067 3	0.004 6
可交换	Xu	R	-0.000 7	0.069 7	0.004 8	-0.006 8	0.049 7	0.002 5	-0.000 3	0.050 9	0.002 6
	Qin	NR	-0.009 5	0.079 5	0.006 4	-0.004 8	0.053 0	0.002 8	-0.002 9	0.057 0	0.003 3
	Qin	R	-0.009 8	0.083 5	0.007 1	-0.002 4	0.055 2	0.003 1	-0.002 1	0.057 9	0.003 4
	He	NR	-0.008 5	0.079 4	0.006 4	-0.002 5	0.048 3	0.002 3	-0.007 9	0.048 2	0.002 4
	He	R	-0.008 5	0.081 3	0.006 7	-0.001 6	0.050 4	0.002 5	-0.006 9	0.048 6	0.002 4
AR (1)	Xu	R	-0.002 5	0.069 8	0.004 8	-0.007 6	0.058 0	0.003 4	0.007 2	0.055 6	0.003 1
	Qin	NR	-0.009 3	0.075 4	0.005 8	-0.004 6	0.053 7	0.002 9	-0.003 3	0.058 4	0.003 4
	Qin	R	-0.009 8	0.079 0	0.006 3	-0.002 7	0.055 5	0.003 1	-0.002 2	0.059 6	0.003 6
	He	NR	-0.008 9	0.077 1	0.006 0	-0.000 8	0.051 2	0.002 6	-0.008 6	0.049 7	0.002 5
	He	R	-0.008 8	0.079 0	0.006 3	-0.000 1	0.054 5	0.003 0	-0.007 5	0.050 0	0.002 6
无结构	Xu	R	-0.000 7	0.064 5	0.004 2	-0.005 3	0.047 2	0.002 3	-0.001 5	0.048 5	0.002 4
	Qin	NR	-0.008 7	0.076 6	0.005 9	-0.004 2	0.050 8	0.002 6	-0.003 6	0.054 4	0.003 0
	Qin	R	-0.009 2	0.079 9	0.006 5	-0.002 2	0.052 7	0.002 8	-0.002 6	0.055 4	0.003 1
	He	NR	-0.008 8	0.077 2	0.006 0	0.000 1	0.050 5	0.002 6	-0.008 0	0.049 9	0.002 6
	He	R	-0.008 8	0.079 5	0.006 4	0.000 8	0.053 8	0.002 9	-0.007 0	0.050 4	0.002 6

为了进一步说明本文提出方法的稳健性, 对数据进行 3% 污染, 污染方式为随机挑选样本 3% 的 x_i 将其替换成 $x_i - 2$, 并将 x_i 所对应的 y_i 替换成 $y_i + 2$, 所得 β_0 估计结果如表 2 所示。

通过比较可以看出: 在数据不污染和 3% 污染的情况下, 本文提出的稳健方法均要优于其他 2 种方法, 参数的偏差和均方误差都比其他 2 种方法要小, 显示估计更准确; 当数据不污染时, 本文提出的方法与其他方法比较, 主要是参数的偏差比较小, 从而使得本文方法的有效性比其他 2 种方法高; 当数据被污染时, 本文提出的方法的有效性更为显著。相同条件下, 本文提出方法的

偏差远远小于其他 2 种方法, 以参数 β_3 为例, 固定协方差为 I 单位结构, 在数据不污染时, 相同稳健的方法, 本文 β_3 偏差分别比 Qin 和 He 小 0.000 2, 0.006; 在数据 3% 污染时, 相同稳健的方法, 本文 β_3 偏差分别比 Qin 和 He 小 0.001, 0.011 1。

此外, 本文还对数据进行了 5% 的污染, 结果显示, 数据污染程度越大, 本文的方法越有效, 为了节省篇幅, 相关表格在本文没有给出。

为了进一步说明对纵向数据协方差矩阵的估计好坏程度, 本文定义指标 QL 和 EL 来衡量 $\hat{\Sigma}$ 与 Σ 的接近程度, QL 和 EL 越小, 表明协方差矩阵的估计值越接近真实值。

表 2 数据 3% 污染时参数估计对比

Tab.2 Parameter estimation and comparison when data are 3% polluted

3% 污染			β_1			β_2			β_3		
			BIAS	SE	MSE	BIAS	SE	MSE	BIAS	SE	MSE
I 单位	Xu	R	0.048 3	0.053 3	0.005 2	-0.110 1	0.073 5	0.017 5	-0.108 2	0.072 7	0.017 0
	Qin	NR	0.122 1	0.063 9	0.019 0	-0.279 1	0.075 3	0.083 6	-0.274 0	0.087 3	0.082 7
	Qin	R	0.041 7	0.060 6	0.005 4	-0.118 7	0.068 4	0.018 8	-0.114 2	0.070 6	0.018 0
	He	NR	0.132 1	0.063 2	0.021 4	0.302 8	0.084 6	0.098 8	-0.303 4	0.094 6	0.101 0
	He	R	0.056 5	0.058 3	0.006 6	-0.145 9	0.081 6	0.027 9	-0.148 7	0.083 9	0.029 1
可交换	Xu	R	0.045 6	0.075 1	0.007 7	-0.093 4	0.061 2	0.012 5	-0.089 3	0.061 7	0.011 8
	Qin	NR	0.123 2	0.081 6	0.021 8	-0.275 8	0.071 3	0.081 2	-0.273 5	0.085 7	0.082 2
	Qin	R	0.018 4	0.082 6	0.007 2	-0.109 0	0.059 9	0.015 5	-0.107 6	0.064 9	0.015 8
	He	NR	0.133 1	0.081 6	0.024 4	-0.299 5	0.070 4	0.094 6	-0.301 4	0.081 2	0.097 5
	He	R	0.056 9	0.079 1	0.009 5	-0.142 4	0.063 5	0.024 3	-0.147 1	0.067 8	0.026 2
AR (1)	Xu	R	0.044 2	0.073 6	0.007 4	-0.096 9	0.067 3	0.013 9	-0.083 3	0.069 5	0.011 8
	Qin	NR	0.123 1	0.078 9	0.021 4	-0.275 4	0.071 4	0.080 9	-0.274 2	0.085 8	0.082 5
	Qin	R	0.023 1	0.079 5	0.006 8	-0.109 4	0.061 6	0.015 8	-0.109 7	0.064 9	0.016 2
	He	NR	0.132 6	0.079 7	0.023 9	-0.298 6	0.070 6	0.094 1	-0.302 8	0.081 2	0.098 3
	He	R	0.056 2	0.077 2	0.009 1	-0.141 4	0.064 3	0.024 1	-0.147 8	0.067 4	0.026 4
无结构	Xu	R	0.047 3	0.070 6	0.007 2	-0.086 5	0.060 7	0.011 2	-0.084 1	0.061 0	0.010 8
	Qin	NR	0.122 9	0.080 0	0.021 5	-0.273 1	0.070 8	0.079 6	-0.273 5	0.085 0	0.082 0
	Qin	R	0.024 3	0.080 9	0.007 1	-0.103 0	0.060 1	0.014 2	-0.105 1	0.062 1	0.014 9
	He	NR	0.132 9	0.079 7	0.024 0	-0.297 8	0.069 8	0.093 5	-0.302 3	0.081 0	0.097 9
	He	R	0.056 5	0.076 8	0.009 1	-0.140 7	0.062 7	0.023 7	-0.147 5	0.067 8	0.026 4

$$QL = \Delta_Q(\Sigma, \hat{\Sigma}) = \left(trace(\Sigma^{-1} \hat{\Sigma} - I) \right)^2$$

$$EL = \Delta_E(\hat{\Sigma}, \Sigma) = trace(\Sigma^{-1} \hat{\Sigma}) - \log|\Sigma^{-1} \hat{\Sigma}|$$

式中： Δ 为协方差真实值与估计值的差值； $trace$ 为矩阵的迹。

在数据不污染和 3% 污染的情况下，将本文方法和 Qin 等^[7]提出的稳健方程进行比较，对应的 QL 和 EL 结果如表 3 所示，其中， $EL_i, QL_i (i = 1, 2, 3, 4)$ 代表误差项服从第 i 种协方差结构下的 QL 和 EL 指标，对应模型中设定的 4 种协方差结构。

由表 3 可以看出：在数据不污染时，本文提出的非稳健方法所估计出的协方差的 EL 指标，一直优于 Qin 提出的方法；在数据样本受到 3% 污染时，本文提出的稳健方法所估计的协方差的 EL 和 QL 指标，都几乎一直优于 Qin 提出的方法，表明本文方法估计出的协方差结构与真实值最为接近，说明本文提出的方法具有显著稳健性和有效性，不仅能保证参数 β_0 和 Σ 估计的有效性，而且在样本受到污染时仍能保证其准确性；此处还对

数据进行了 5% 的污染，得到结果与上面类似，并且污染程度越大，本文方法的优越性越显著。

为了进一步比较经验似然方法的优越性，选取误差项服从多元 t 分布的纵向数据进行模拟研究，研究方法同上，所得结果如表 4 所示。

由表 4 可知，在误差服从 t 分布时，经验似然方法的估计结果仍然非常可观。3 种方法的稳健结果显示，在 4 种协方差结构下，本文方法的估计方差要远远小于其他 2 种方法，说明本文提出的经验似然估计方法比其他 2 种方法更有效，参数估计更趋于一致。此外，对比误差项服从多元 t 分布下本文方法和 Qin 等^[7]的 4 种协方差估计情况如表 5 所示。

由表 5 可知，在误差服从多元 t 分布时，经验似然方法的有效性更为显著。根据 QL 指标和 EL 指标显示，本文估计的纵向数据协方差矩阵与真实协方差矩阵相似度更高，与表 3 对比可看出，当模拟数据的误差服从多元 t 分布时，其协方差矩阵

表 3 *EL* 和 *QL* 对比表

Tab.3 *EL* and *QL* comparison table

污染方式		指标	Qin		Xu	
			R	NR	R	NR
不污染		EL_1	0.124 0	0.062 9	0.254 8	0.058 8
		EL_2	0.124 4	0.063 8	0.159 0	0.058 4
		EL_3	0.125 5	0.064 0	0.165 6	0.058 4
		EL_4	0.126 4	0.064 1	0.151 2	0.058 6
		QL_1	0.142 2	0.058 0	0.362 6	0.115 3
		QL_2	0.156 1	0.060 2	0.235 4	0.114 9
		QL_3	0.156 3	0.060 0	0.244 1	0.114 8
		QL_4	0.166 4	0.061 1	0.226 6	0.116 0
3% 污染		EL_1	0.128 8	0.844 6	0.199 2	0.956 0
		EL_2	0.314 5	1.658 1	0.182 2	1.861 7
		EL_3	0.316 7	1.588 1	0.203 8	1.835 3
		EL_4	0.448 9	2.335 9	0.289 1	2.550 1
		QL_1	0.292 6	6.940 1	0.293 8	3.609 5
		QL_2	0.836 9	15.527 9	0.265 3	8.852 1
		QL_3	0.798 9	14.443 6	0.319 5	9.136 3
		QL_4	1.269 9	23.030 6	0.576 0	15.336 0

表 4 *t* 分布误差下的参数估计对比表

Tab.4 Comparison of parameter estimates under *t* distribution error

残差项服从 <i>t</i> 分布			β_1			β_2			β_3		
			BIAS	SE	MSE	BIAS	SE	MSE	BIAS	SE	MSE
<i>I</i> 单位	Xu	R	−0.006 9	0.063 2	0.004 0	−0.013 7	0.076 9	0.006 1	−0.003 3	0.075 3	0.005 7
	Qin	NR	−0.009 6	0.104 6	0.011 0	−0.008 9	0.113 5	0.013 0	0.001 8	0.092 6	0.008 6
	Qin	R	0.002 0	0.075 0	0.005 6	−0.012 5	0.087 5	0.007 8	0.005 2	0.071 9	0.005 2
	He	NR	−0.010 8	0.105 9	0.011 3	−0.008 7	0.116 7	0.013 7	−0.002 6	0.114 4	0.013 1
	He	R	−0.005 2	0.081 9	0.006 7	−0.008 0	0.092 5	0.008 6	0.003 0	0.084 6	0.007 2
可交换	Xu	R	−0.007 6	0.100 3	0.010 1	−0.008 0	0.071 4	0.005 2	0.003 4	0.065 0	0.004 2
	Qin	NR	−0.017 8	0.141 6	0.020 4	−0.007 8	0.097 6	0.009 6	0.001 4	0.080 8	0.006 5
	Qin	R	−0.004 4	0.103 4	0.010 7	−0.010 4	0.077 5	0.006 1	0.003 5	0.063 7	0.004 1
	He	NR	−0.017 9	0.141 4	0.020 3	−0.013 9	0.085 7	0.007 5	0.000 7	0.092 8	0.008 6
	He	R	−0.009 8	0.111 2	0.012 5	−0.011 6	0.068 6	0.004 8	0.001 6	0.062 9	0.004 0
AR(1)	Xu	R	−0.005 7	0.095 6	0.009 2	−0.008 7	0.072 4	0.005 3	0.002 1	0.067 9	0.004 6
	Qin	NR	−0.016 1	0.133 0	0.018 0	−0.009 7	0.101 1	0.010 3	0.000 9	0.082 9	0.006 9
	Qin	R	−0.004 1	0.100 1	0.010 0	−0.012 5	0.078 4	0.006 3	0.001 9	0.073 1	0.005 4
	He	NR	−0.017 3	0.134 8	0.018 5	−0.013 5	0.093 2	0.008 9	−0.001 6	0.098 6	0.009 7
	He	R	−0.007 9	0.107 8	0.011 7	−0.013 3	0.073 0	0.005 5	0.000 8	0.072 0	0.005 2
无结构	Xu	R	−0.006 8	0.093 6	0.008 8	−0.008 1	0.071 5	0.005 2	0.006 2	0.065 8	0.004 4
	Qin	NR	−0.018 7	0.134 9	0.018 6	−0.008 0	0.093 7	0.008 8	0.000 7	0.078 3	0.006 1
	Qin	R	−0.005 8	0.099 6	0.010 0	−0.010 5	0.071 5	0.005 2	0.003 4	0.060 5	0.003 7
	He	NR	−0.019 8	0.136 4	0.019 0	−0.013 2	0.092 5	0.008 7	−0.001 3	0.096 4	0.009 3
	He	R	−0.009 2	0.108 9	0.012 0	−0.014 2	0.072 9	0.005 5	0.000 4	0.071 7	0.005 1

表 5 t 分布误差下 QL 和 EL

Tab.5 EL and QL under t distribution error

指标	Qin		Xu	
	R	NR	R	NR
EL_1	0.323 1	3.011 3	0.166 7	3.870 5
EL_2	0.332 9	2.909 6	0.125 3	3.801 7
EL_3	0.313 2	2.974 8	0.131 6	3.828 1
EL_4	0.332 2	2.959 3	0.141 5	3.722 8
QL_1	1.527 6	60.465 8	0.248 4	81.651
QL_2	1.580 8	58.989 2	0.207 0	88.798
QL_3	1.437 4	59.987 7	0.213 2	86.761
QL_4	1.598 0	59.913 4	0.255 4	91.438

的估计效果更优于误差服从多元正态分布时的情形，进一步说明了本文方法对误差具有抵抗性。

4 实证分析

为了进一步验证本文方法的实用性，分析了 CD4 细胞计数纵向数据。关于这个数据集的完整描述见 Diggle 的主页: <http://www.lancs.ac.uk/diggle/>. 这里分析所用的数据是完整数据的截取部分, 包含 340 个样本的 1 020 次观测值。响应变量 y 是 CD4 计数的算数平方根, 协变量包括血清转换的时间 x_2 , 相对于一个起点的年龄 x_3 , 由吸烟的包数刻画的吸烟状况 x_4 , 娱乐性药物使用是/否 x_5 , 性伴侣的个数 x_6 , 以及流行病中心给出抑郁状态和抑郁程度 x_7 。有许多学者研究过这个数据集。Zeger 等^[10]和 Wang 等^[11]分别对这个数据拟合了半参数模型和部分线性半参数模型, 其中, 协变量 x_2 作为非线性形式进入模型。

本文分析的主要目的是寻找 CD4 数量和协变量之间的关系, 考虑到可能存在非线性关系, 引入了除 x_5 以外的平方项, 加上截距项 x_1 , 记

$$\boldsymbol{X} = (x_1, x_2, x_2^2, x_3, x_3^2, x_4, x_4^2, x_5, x_6, x_6^2, x_7, x_7^2)^T \quad (15)$$

建立线性模型

$$y_{ij} = \boldsymbol{X}_{ij}^T \boldsymbol{\beta}, i = 1, 2, \cdots, 340, j = 1, 2, 3 \quad (16)$$

由于实际数据无法知道真实参数, 所以, 现利用交互验证的方法来比较各种方法的优劣, 交互验证过程中的均方误差

$$MSE_{CV} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_{(-i)}| \quad (17)$$

这里的 $n = 339$, $\hat{y}_{(-i)}$ 是去掉第 i 个个体外, 由

其余 339 个个体拟合所得的预测值, 数据预处理后, 代入算法中, 本文方法与 Qin 和 He 方法的比较结果如表 6 所示。

表 6 说明, 本文方法在实际应用中十分有效, CV 值要明显小于另外 2 种方法的 CV 值, 值得推广。

表 6 CV 值比较

Tab.6 Comparison of CV value

方法	Xu	Qin	He
CV	0.44	1.353 7	1.359 5

参考文献:

[1] THOMAS D R, GRUNKEMEIER G L. Confidence interval estimation of survival probabilities for censored data[J]. *Journal of the American Statistical Association*, 1975, 70(352): 865–871.

[2] OWEN A. Empirical likelihood ratio confidence regions[J]. *The Annals of Statistics*, 1990, 18(1): 90–120.

[3] OWEN A. Empirical likelihood for linear models[J]. *The Annals of Statistics*, 1991, 19(4): 1725–1747.

[4] BONDELL H D, STEFANSKI L A. Efficient robust regression via two-stage generalized empirical likelihood[J]. *Journal of the American Statistical Association*, 2013, 108(502): 644–655.

[5] HE X M, FUNG W K, ZHU Z Y. Robust estimation in generalized partial linear models for clustered data[J]. *Journal of the American Statistical Association*, 2005, 100(472): 1176–1184.

[6] QIN G Y, ZHU Z Y, FUNG W K. Robust estimation of generalized partially linear model for longitudinal data with dropouts[J]. *Annals of the Institute of Statistical Mathematics*, 2016, 68(5): 977–1000.

[7] QIN G Y, ZHU Z Y. Robust estimation of mean and covariance for longitudinal data with dropouts[J]. *Journal of Applied Statistics*, 2015, 42(6): 1240–1254.

[8] CRESSIE N, READ T R C. Multinomial goodness-of-fit tests[J]. *Journal of the Royal Statistical Society. Series B*, 1984, 46(3): 440–464.

[9] CORCORAN S A. Bartlett adjustment of empirical discrepancy statistics[J]. *Biometrika*, 1998, 85(4): 967–972.

[10] ZEGER S L, DIGGLE P J. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters[J]. *Biometrics*, 1994, 50(3): 689–699.

[11] WANG Y G, LIN X, ZHU M. Robust estimating functions and bias correction for longitudinal data analysis[J]. *Biometrics*, 2005, 61(3): 684–691.