纵向数据下广义线性模型的稳健二次推断函数估计

关晓妮 黄 彬*
(北京化工大学 理学院, 北京 100029)

摘 要: 针对纵向数据下的广义线性模型,为了有效控制离群点对估计的影响以及进一步提高估计的效率,利用二次推断函数(QIF)改进加权的指数得分函数,得到了模型参数有效且稳健的二次推断函数估计(ERQIF),并证明了在一定条件下所得估计的相合性和渐近正态性。数值计算结果进一步表明,当离群点存在或工作相关矩阵被错误指定时,所得估计有稳健的模拟结果。

关键词: 纵向数据; 加权指数得分函数; 二次推断函数法(QIF); 稳健估计
中图分类号: O212 **DOI**: 10.13543/j.bhxbzr.2018.02.017

引 言

纵向数据广泛应用在生物医药、流行病学和经济学等领域中,它通过对同一个体在不同的时间点重复测量得到,具有组内相关、组间独立的特性。在实际的统计建模中,如果忽略这种特性,将会降低估计的效率,所以如何处理组内相关性是纵向数据研究不可避免的问题。Liang 等[1]提出的广义估计方程(GEE)方法可以有效地处理组内相关问题,该方法利用含少量讨厌参数的工作相关矩阵建立估计方程,即使工作矩阵错误也可以得到回归系数的相合估计。但是 GEE 方法首先需要得到讨厌参数的相合估计,同时讨厌参数的相合估计可能不存在。为了提高估计的效率,Qu 等[2]提出了二次推断函数法(QIF),利用某些基矩阵的线性组合来逼近工作相关矩阵的逆矩阵和避免估计讨厌参数,既使所得估计具有相合性,又提高了参数估计的效率。已有许多学者将 GEE 和 QIF 方法应用于不同的模型[3-5],但是这些应用方法大多类似于加权最小二乘估计,不具有稳健性。

在纵向数据中,由于重复测量,一个个体的突变往往会产生一系列的离群点,因此在纵向数据的研究中稳健估计是一个十分重要的问题[6-8]。为了获得更好的稳健性和提高估计的效率, Wang 等[9]提出了基于指数得分函数的稳健回归估计方法,通过引入一个调节参数来提高估计的稳健性和有效性。Lv 等[10]将加权指数得分函数和 GEE 方法相结合,得到了纵向数据下广义线性模型参数的有效且稳健的估计(ERGEE)。但是估计讨厌参数必然会降低估计的效率。本文将指数得分函数和 QIF 方法相结合,利用加权的指数得分函数有效控制协变量和因变量中离群点的影响,同时基于 QIF 方法建立稳健二次推断函数,得到了纵向数据下广义线性模型参数的 ERQIF 估计,并进一步提高了估计的效率。最后通过模拟计算进一步验证了所得估计的有限样本性质。

收稿日期: 2017-08-01
基金项目: 国家自然科学基金(11471321)
第一作者: 女,1990 年生,硕士生
* 通讯联系人
E-mail: abinhuang@gmail.com

1 参数估计方法

1.1 ERQIF 方法

假定纵向数据中包含 n 个个体,对每个个体重复观测 m_i 次, 总的观测次数 $M=\sum_{i=1}^{n} m_i$。设 $\boldsymbol{Y}_i=(y_{i1},\cdots,y_{im_i})^{\mathrm{T}}$ 为第 i 个个体的响应变量($i=1,\cdots,n$),$\boldsymbol{X}_i=(\boldsymbol{x}_{i1},\boldsymbol{x}_{i2},\cdots,\boldsymbol{x}_{im_i})^{\mathrm{T}}$ 为其相应的协变量,其中 $\boldsymbol{x}_{ij}=(x_{ij}^{(1)},x_{ij}^{(2)},\cdots,x_{ij}^{(p)})^{\mathrm{T}}$。本文重点讨论纵向数据下的广义线性模型(GLM)。假定 y_{ij} 的边际均值和边际方差函数分别为

$$\mu_{ij}=E(y_{ij}|\boldsymbol{x}_{ij})=g(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_{ij})$$
$$Var(y_{ij}|\boldsymbol{x}_{ij})=\phi v(\mu_{ij}) \quad (i=1,\cdots,n;j=1,\cdots,m_i)$$

其中 $\boldsymbol{\beta}$ 为未知的 $p\times 1$ 回归系数,$g(\cdot)$ 和 $v(\cdot)$ 分别是已知的逆连接函数和方差函数,ϕ 为尺度参数。不失一般性,假设不同个体间的观测是相互独立的,

且同一个个体内部的观测具有一定的相关关系。

Lv 等[10]基于指数得分函数提出 ERGEE 方法，通过求解如式(1)所示的估计方程得到了参数 $\boldsymbol{\beta}$ 的基于 GEE 有效且稳健的估计 $\hat{\boldsymbol{\beta}}_{\text{ERGEE}}$

$$\boldsymbol{U}_{\text{n}}^{\text{ERGEE}}(\boldsymbol{\beta},\alpha)=\sum_{i=1}^{n}\boldsymbol{D}_i^{\text{T}}\boldsymbol{V}_i^{-1}h_i^{\gamma}(\mu_i(\boldsymbol{\beta}))=0 \quad (1)$$

其中 $\boldsymbol{D}_i=\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ 是 $n_i\times p$ 矩阵；$\boldsymbol{V}_i=\boldsymbol{R}_i(\alpha)\boldsymbol{A}_i^{1/2}$，$\boldsymbol{R}_i(\alpha)$ 是依赖于参数 α 的 $m_i\times m_i$ 工作相关矩阵，且 $\boldsymbol{A}_i=\varphi\text{diag}(v(\mu_{i1}),\cdots,v(\mu_{im_i}))$，$\boldsymbol{\mu}_i=(\mu_{i1},\cdots,\mu_{im_i})^{\text{T}}$；$h_i^{\gamma}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))=\boldsymbol{W}_i[\psi_{\gamma}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))-C_i(\boldsymbol{\mu}_i(\boldsymbol{\beta}))]$，$\psi_{\gamma}(\boldsymbol{\mu}_i)=\psi_{\gamma}(\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i-\boldsymbol{\mu}_i))$，$C_i(\boldsymbol{\mu}_i)=E[\psi_{\gamma}(\boldsymbol{\mu}_i)]$，$\psi_{\gamma}(t)=-\frac{2t}{\gamma}\exp(-t^2/\gamma)(\gamma>0)$ 为调节参数。

权重矩阵 $\boldsymbol{W}_i=\text{diag}(w_{i1},\cdots,w_{im_i})$ 用于限制协变量中离群点的影响，可取 Mallows－type 函数 w_{ij} 为

$$w_{ij}=w(\boldsymbol{x}_{ij})=\min\left\{1,\left(\frac{b_0}{(\boldsymbol{x}_{ij}-\boldsymbol{m}_x)^{\text{T}}\boldsymbol{S}_x^{-1}(\boldsymbol{x}_{ij}-\boldsymbol{m}_x)}\right)^{\rho/2}\right\}$$

其中 $\rho\geqslant 1$，b_0 是自由度为 p 的卡方分布的 0.95 分位点，$\boldsymbol{m}_x$ 和 $\boldsymbol{S}_x$ 分别表示关于 $\boldsymbol{x}_{ij}$ 的位置参数和尺度参数的稳健估计，可以利用 minimum covariance determinant(MCD)进行估计。

为了简单起见，假定重复观测次数是相同的，即 $m_i=m<\infty$，并令 $\boldsymbol{R}(\alpha)$ 为公共的工作相关矩阵。当重复观测次数不同时，可采用 Xue 等[11]的方法加以处理。在实际计算中，为了得到 $\hat{\boldsymbol{\beta}}_{\text{ERGEE}}$，ERGEE 方法首先要对参数 α 和 ϕ 进行估计，同时 Lv 等[10]也指出若能得到参数 α 和 ϕ 的 $\sqrt{n}$-相合估计，则 $\hat{\boldsymbol{\beta}}_{\text{ERGEE}}$ 是 $\boldsymbol{\beta}$ 的 $\sqrt{n}$-相合估计。但即使在一些简单的情形下，讨厌参数 α 的相合估计也可能不存在；而且对参数 α 进行估计还会降低参数 $\boldsymbol{\beta}$ 估计的效率。为了弥补这一缺陷同时提高 ERGEE 估计的效率，根据 Qu 等[2]提出的 QIF 方法，本文利用一组基矩阵的线性组合来逼近 $\boldsymbol{R}^{-1}(\alpha)$，即

$$\boldsymbol{R}^{-1}\approx a_1\boldsymbol{M}_1+a_2\boldsymbol{M}_2+\cdots+a_k\boldsymbol{M}_k \quad (2)$$

其中 $\boldsymbol{M}_1$ 是单位矩阵，$\boldsymbol{M}_2,\cdots,\boldsymbol{M}_k$ 是某些已知的对称的基矩阵[2]，$a_1,a_2,\cdots,a_k$ 是未知常数。

将式(2)代入式(1)构造扩展得分向量为

$$\overline{\boldsymbol{U}}_n(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{U}_i(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\boldsymbol{D}_i^{\text{T}}\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_1h_i^{\gamma}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))\\ \vdots\\ \boldsymbol{D}_i^{\text{T}}\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_kh_i^{\gamma}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))\end{pmatrix}$$

由于 $\overline{\boldsymbol{U}}_n(\boldsymbol{\beta})$ 的维数为 kp，显然高于 $\boldsymbol{\beta}$ 的维数 p，因此求解方程 $\overline{\boldsymbol{U}}_n(\boldsymbol{\beta})=0$ 会出现过度识别问题。基于 GMM 估计的思想，定义有效且稳健的 ERQIF 估计方程如式(3)

$$Q_n(\boldsymbol{\beta})=n\overline{\boldsymbol{U}}_n^{\text{T}}(\boldsymbol{\beta})C_n^{-1}(\boldsymbol{\beta})\overline{\boldsymbol{U}}_n(\boldsymbol{\beta}) \quad (3)$$

其中，$C_n(\boldsymbol{\beta})=\frac{1}{n}\sum\boldsymbol{U}_i(\boldsymbol{\beta})\boldsymbol{U}_i^{\text{T}}(\boldsymbol{\beta})$。

通过极小化目标函数 $Q_n(\cdot)$ 得到 $\boldsymbol{\beta}$ 的估计

$\hat{\boldsymbol{\beta}}_{\text{ERQIF}}=\arg\min_{\boldsymbol{\beta}}Q_n(\boldsymbol{\beta})$

利用如式(4)的 Newton－Raphson 迭代算法得到估计 $\hat{\boldsymbol{\beta}}_{\text{ERQIF}}$[2]

$$\hat{\boldsymbol{\beta}}^{(k+1)}=\hat{\boldsymbol{\beta}}^{(k)}-[\nabla^2Q_n(\hat{\boldsymbol{\beta}}^{(k)})]^{-1}\nabla Q_n(\hat{\boldsymbol{\beta}}^{(k)}) \quad (4)$$

其中，$\nabla Q_n(\boldsymbol{\beta})$ 和 $\nabla^2Q_n(\boldsymbol{\beta})$ 分别为 $Q_n(\boldsymbol{\beta})$ 对 $\boldsymbol{\beta}$ 的一阶和二阶导数，且满足

$$\begin{cases}\nabla Q_n(\boldsymbol{\beta})=2n\nabla\overline{\boldsymbol{U}}_n^{\text{T}}(\boldsymbol{\beta})C_n^{-1}(\boldsymbol{\beta})\overline{\boldsymbol{U}}_n(\boldsymbol{\beta})-\\ \qquad n\overline{\boldsymbol{U}}_n^{\text{T}}(\boldsymbol{\beta})\nabla C_n^{-1}(\boldsymbol{\beta})\overline{\boldsymbol{U}}_n(\boldsymbol{\beta})\\ \nabla^2Q_n(\boldsymbol{\beta})=2n\nabla\overline{\boldsymbol{U}}_n^{\text{T}}(\boldsymbol{\beta})C_n^{-1}(\boldsymbol{\beta})\nabla\overline{\boldsymbol{U}}_n(\boldsymbol{\beta})+O_p(1)\end{cases}$$

1.2　估计的渐近性质

设 $\boldsymbol{\beta}_0$ 为参数 $\boldsymbol{\beta}$ 的真值，$\boldsymbol{\beta}$ 所处的参数空间记为 $\boldsymbol{\Theta}$。首先给出一些假设条件[4]（本文中 $E_{\boldsymbol{\beta}_0}(\cdot)$ 均表示随机变量关于真值 $\boldsymbol{\beta}_0$ 的数学期望）：

1）参数空间 $\boldsymbol{\Theta}$ 是可识别的，若 $\boldsymbol{\beta}\neq\boldsymbol{\beta}_0$，则 $E_{\boldsymbol{\beta}_0}[U_1(\boldsymbol{\beta})]\neq 0$；

2）对任意 $\boldsymbol{\beta}\in\boldsymbol{\Theta}$，$E_{\boldsymbol{\beta}_0}[\boldsymbol{U}_1(\boldsymbol{\beta})]$ 存在且有限，同时关于 $\boldsymbol{\beta}$ 连续；

3）$\boldsymbol{\Theta}$ 是紧空间，且 $\boldsymbol{\beta}_0$ 为 $\boldsymbol{\Theta}$ 的一个内点；

4）存在 $\boldsymbol{\beta}_0$ 的某个邻域 O_0，当 $\boldsymbol{\beta}\in O_0$ 时，有

$$C_n(\boldsymbol{\beta})\xrightarrow{a.s.}E_{\boldsymbol{\beta}_0}[\boldsymbol{U}_1(\boldsymbol{\beta})\boldsymbol{U}_1^{\text{T}}(\boldsymbol{\beta})]\triangleq\boldsymbol{\Sigma}_{\boldsymbol{\beta}_0}(\boldsymbol{\beta})$$

其中，$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_0}(\boldsymbol{\beta})$ 为正定矩阵，且关于 $\boldsymbol{\beta}\in O_0$ 连续；

5）$\overline{\boldsymbol{U}}_n(\boldsymbol{\beta})$ 关于 $\boldsymbol{\beta}$ 的一阶导数存在且连续，且当 $\boldsymbol{\beta}^*\xrightarrow{P}\boldsymbol{\beta}_0$ 时，有

$$\frac{\partial\overline{\boldsymbol{U}}_n(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}}\xrightarrow{P}\boldsymbol{B}(\boldsymbol{\beta}_0)=E_{\boldsymbol{\beta}_0}\left[\frac{\partial\overline{\boldsymbol{U}}_n(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right]$$

基于条件 1）～4），根据文献[12]和文献[2]中相关定理的证明，可得定理 1。

定理 1　在条件 1）～4）下，通过式(3)得到的 $\hat{\boldsymbol{\beta}}_{\text{ERQIF}}$ 是存在的，且当 $n\to\infty$ 时，有 $\hat{\boldsymbol{\beta}}_{\text{ERQIF}}\xrightarrow{a.s.}\boldsymbol{\beta}_0$，即得到了 $\boldsymbol{\beta}_0$ 的相合估计 $\hat{\boldsymbol{\beta}}_{\text{ERQIF}}$。

定理 2　在条件 1）～5）下，若 $\overline{\boldsymbol{U}}_n(\boldsymbol{\beta})$ 和 $C_n(\boldsymbol{\beta})$ 的二阶导数有有限的均值和方差，则当 $n\to\infty$ 时，有

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{ERQIF}}-\boldsymbol{\beta}_0)\xrightarrow{D}N_p(0,\boldsymbol{J}^{-1}(\boldsymbol{\beta}_0))$$

即证明了所得估计的渐进正态性。其中,$\boldsymbol{J}(\boldsymbol{\beta}_0)=\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}_{\boldsymbol{\beta}_0}^{-1}(\boldsymbol{\beta}_0)\boldsymbol{B}(\boldsymbol{\beta}_0)$。

证明 因 $\hat{\boldsymbol{\beta}}_{\mathrm{ERQIF}}=\arg\min_{\beta} Q_n(\boldsymbol{\beta})$,则 $\nabla Q_n(\hat{\boldsymbol{\beta}}_{\mathrm{ERQIF}})=0$。由 Taylor 公式得到

$$\mathbf{0}=\nabla Q_n(\hat{\boldsymbol{\beta}}_{\mathrm{ERQIF}})=\nabla Q_n(\boldsymbol{\beta}_0)+\nabla^2 Q_n(\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_{\mathrm{ERQIF}}-\boldsymbol{\beta}_0)$$

其中 $\tilde{\boldsymbol{\beta}}$ 位于 $\hat{\boldsymbol{\beta}}_{\mathrm{ERQIF}}$ 与 $\boldsymbol{\beta}_0$ 之间。从而有

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{ERQIF}}-\boldsymbol{\beta}_0)=-\left(\frac{1}{2n}\nabla^2 Q_n(\tilde{\boldsymbol{\beta}})\right)^{-1}\frac{1}{2\sqrt{n}}\nabla Q_n(\boldsymbol{\beta}_0)$$

其中

$$\nabla Q_n(\boldsymbol{\beta}_0)=2n\nabla\overline{\boldsymbol{U}}_n^{\mathrm{T}}(\boldsymbol{\beta}_0)C_n^{-1}(\boldsymbol{\beta}_0)\overline{\boldsymbol{U}}_n(\boldsymbol{\beta}_0)+O_p(1)$$

$$\nabla^2 Q_n(\tilde{\boldsymbol{\beta}})=2n\nabla\overline{\boldsymbol{U}}_n^{\mathrm{T}}(\tilde{\boldsymbol{\beta}})C_n^{-1}(\tilde{\boldsymbol{\beta}})\nabla\overline{\boldsymbol{U}}_n(\tilde{\boldsymbol{\beta}})+O_p(1)$$

由条件 4)和 5)以及 $\sqrt{n}\overline{\boldsymbol{U}}_n(\boldsymbol{\beta}_0)\xrightarrow{D}N_{kp}(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}_0))$,可得

$$\frac{1}{2\sqrt{n}}\nabla Q_n(\boldsymbol{\beta}_0)=\nabla\overline{\boldsymbol{U}}_n^{\mathrm{T}}(\boldsymbol{\beta}_0)C_n^{-1}(\boldsymbol{\beta}_0)\sqrt{n}\overline{\boldsymbol{U}}_n(\boldsymbol{\beta}_0)+O_p(n^{-1/2})=\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}_{\boldsymbol{\beta}_0}^{-1}(\boldsymbol{\beta}_0)\sqrt{n}\overline{\boldsymbol{U}}_n(\boldsymbol{\beta}_0)+O_p(n^{-1/2})\xrightarrow{D}N_p(\mathbf{0},\boldsymbol{J}(\boldsymbol{\beta}_0))$$

$$\frac{1}{2n}\nabla^2 Q_n(\tilde{\boldsymbol{\beta}})=\nabla\overline{\boldsymbol{U}}_n^{\mathrm{T}}(\tilde{\boldsymbol{\beta}})C_n^{-1}(\tilde{\boldsymbol{\beta}})\nabla\overline{\boldsymbol{U}}_n(\tilde{\boldsymbol{\beta}})+O_p(1)\xrightarrow{P}\boldsymbol{J}(\boldsymbol{\beta}_0)$$

因此可得当 $n\to\infty$ 时,$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{ERQIF}}-\boldsymbol{\beta}_0)\xrightarrow{D}N_p(\mathbf{0},\boldsymbol{J}^{-1}(\boldsymbol{\beta}_0))$。

推论 1 在定理 2 的条件下,可得 $\hat{\boldsymbol{\beta}}_{\mathrm{ERQIF}}$ 渐近协方差的相合估计 $\hat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{ERQIF}})=(\hat{\boldsymbol{B}}_n^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_n^{-1}\hat{\boldsymbol{B}}_n)^{-1}$,其中

$$\hat{\boldsymbol{B}}_n=\sum_{i=1}^{n}(\hat{\boldsymbol{B}}_{i1},\cdots,\hat{\boldsymbol{B}}_{ik})^{\mathrm{T}}$$

$$\hat{\boldsymbol{\Sigma}}_n=\sum_{i=1}^{n}(\hat{\boldsymbol{\Sigma}}_{i1},\cdots,\hat{\boldsymbol{\Sigma}}_{ik})^{\mathrm{T}}$$

$$\hat{\boldsymbol{\Sigma}}_{is}^{\mathrm{T}}=\boldsymbol{D}_i^{\mathrm{T}}\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_s h_i^{\gamma}(\mu_i(\hat{\boldsymbol{\beta}}))h_i^{\gamma}(\mu_i(\hat{\boldsymbol{\beta}}))^{\mathrm{T}}\boldsymbol{M}_s\boldsymbol{A}_i^{-1/2}\boldsymbol{D}_i$$

$$\hat{\boldsymbol{B}}_{is}^{\mathrm{T}}=\boldsymbol{D}_i^{\mathrm{T}}\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_s\dot{h}_i^{\gamma}(\mu_i(\hat{\boldsymbol{\beta}}))\boldsymbol{D}_i(s=1,\cdots,k)$$

$$\dot{h}_i^{\gamma}(\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}))=\partial h_i^{\gamma}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))/\partial\boldsymbol{\mu}_i|_{\boldsymbol{\mu}_i=\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})}$$

$$\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}_{\mathrm{ERQIF}}$$

2 ERQIF 算法

2.1 尺度参数和调节参数

要通过式(3)得到 ERQIF 估计 $\hat{\boldsymbol{\beta}}_{\mathrm{ERQIF}}$,首先需要估计尺度参数 ϕ,并且对调节参数 γ 进行选择,γ 的选取对估计的有效性和稳健性有至关重要的影响。假设 $\tilde{\boldsymbol{\beta}}$ 为 $\boldsymbol{\beta}$ 当前的估计,通过绝对偏差中位获取 ϕ 的稳健估计

$$\hat{\phi}=\{1.483\mathrm{medi}\{|\hat{\xi}_{ij}-\mathrm{medi}(\hat{\xi}_{ij})|\}\}^2$$

其中 $\hat{\xi}_{ij}=\boldsymbol{A}_{ij}^{-1/2}(\boldsymbol{Y}_{ij}-\mu_{ij}(\tilde{\boldsymbol{\beta}}))$,$\boldsymbol{A}_{ij}$ 是矩阵 $\boldsymbol{A}_i$ 对角线上第 j 个元素。同时通过最小化 $\hat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})$ 的行列式[13]得到当前最优的 γ

$$\gamma_{\mathrm{opt}}=\arg\min_{\gamma}(\det(\hat{\mathrm{Cov}}(\tilde{\boldsymbol{\beta}})))$$

其中 $\hat{\mathrm{Cov}}(\tilde{\boldsymbol{\beta}})$ 由推论 1 给出。

2.2 估计算法

通过一个迭代算法得到 ERQIF 估计,具体步骤如下:

1)给定 $\boldsymbol{\beta}$ 的初始估计 $\hat{\boldsymbol{\beta}}^{(0)}$ 以及迭代收敛阈值 ε,并令 $k=0$;

2)利用当前估计 $\hat{\boldsymbol{\beta}}^{(k)}$ 来估计尺度参数 $\hat{\phi}^{(k)}$ 和调节参数 $\gamma_{\mathrm{opt}}^{(k)}$;

3)利用公式(4)迭代求解 $\hat{\boldsymbol{\beta}}^{(k+1)}$;

4)若 $\|\hat{\boldsymbol{\beta}}^{(k+1)}-\hat{\boldsymbol{\beta}}^{(k)}\|<\varepsilon$,则终止计算,并令 $\hat{\boldsymbol{\beta}}_{\mathrm{ERQIF}}=\hat{\boldsymbol{\beta}}^{(k+1)}$;否则令 $k\leftarrow k+1$,返回步骤 2)。

在数值计算中,利用独立工作相关矩阵的 GEE 方法[1]得到初始估计 $\hat{\boldsymbol{\beta}}^{(0)}$,并取 $\varepsilon=10^{-3}$,计算结果表明一般在迭代 50 步内即可收敛。

3 数值模拟

为了得到估计的有限样本性质,本文通过随机模拟考察数据出现离群点时估计的稳健性和工作相关矩阵被正确或错误指定对估计效率的影响。用本文提出的 ERQIF 方法与 GEE、ERGEE 方法分别计算 3 种参数估计的样本偏差(Bias)和标准差(SD)并进行比较。设样本容量 $n=100$,重复观测次数 $m=5$,模拟次数为 500。在模拟计算中,分别指定组内工作相关矩阵 $\boldsymbol{R}_i(\alpha)$ 为 Exch 和 AR(1)两种结构。

3.1 线性模型

当响应变量是连续型变量时,假设模型满足:

$$y_{ij}=\sum_{k=1}^{3}x_{ij}^{(k)}\boldsymbol{\beta}_{0k}+\varepsilon_{ij}(i=1,\cdots,n;j=1,\cdots,m) \tag{5}$$

其中,$\beta_{01}=0.7$,$\beta_{02}=0.7$,$\beta_{03}=-0.4$;$x_{ij}^{(k)}\sim N(0,1)$;$\mathrm{Cov}(x_{ij}^{(k)},x_{ij}^{(l)})=0.5^{|k-l|}$,$k,l=1,2,3$。随机误差 $\boldsymbol{\varepsilon}_i=(\varepsilon_{i1},\cdots,\varepsilon_{i5})^{\mathrm{T}}$ 的真实工作相关矩阵是 Exch 结构,相关系数 $\alpha=0.7$。

分别考虑两种情况:

①$\varepsilon_{ij} \sim N(0,1)$；

②$\varepsilon_{ij} \sim 0.9N(0,1) + 0.1N(0,100)$。

用 3 种方法计算得到两种情况下两种结构的估计值，结果分别如表 1、表 2 所示。

表 1　情况①下参数的估计结果

Table 1　Parameter estimation results for ①

$\boldsymbol{R}(\alpha)$	方法	Bias[a)]	SD[a)]	Bias[b)]
Exch	GEE	0.00138	0.14103	0.00027
	ERGEE	0.00141	0.12976	0.00030
	ERQIF	0.00102	0.10477	0.00072
AR(1)	GEE	0.00171	0.13864	0.00464
	ERGEE	0.00023	0.12640	0.00331
	ERQIF	0.00096	0.12080	0.00005
$\boldsymbol{R}(\alpha)$	方法	SD[b)]	Bias[c)]	SD[c)]
Exch	GEE	0.14411	0.00047	0.13546
	ERGEE	0.13888	0.00151	0.13167
	ERQIF	0.13526	0.00183	0.09885
AR(1)	GEE	0.14679	0.00238	0.14924
	ERGEE	0.15721	0.00215	0.11762
	ERQIF	0.11879	0.00012	0.10518

a—β_{01}；b—β_{02}；c—β_{03}。

表 2　情况②下参数的估计结果

Table 2　Parameter estimation results for ②

$\boldsymbol{R}(\alpha)$	方法	Bias[a)]	SD[a)]	Bias[b)]
Exch	GEE	0.00919	1.18311	0.00518
	ERGEE	0.00448	0.91688	0.00860
	ERQIF	0.00009	0.28698	0.00034
AR(1)	GEE	0.00208	1.08281	0.00070
	ERGEE	0.00284	0.93709	0.03906
	ERQIF	0.00029	0.29395	0.00078
$\boldsymbol{R}(\alpha)$	方法	SD[b)]	Bias[c)]	SD[c)]
Exch	GEE	0.74718	0.00213	1.02006
	ERGEE	0.87324	0.02064	0.73074
	ERQIF	0.33117	0.00005	0.47756
AR(1)	GEE	1.28386	0.00310	1.20267
	ERGEE	0.73802	0.05721	0.67969
	ERQIF	0.33555	0.00037	0.29530

a—β_{01}；b—β_{02}；c—β_{03}。

从表 1 可以看出，不管相关结构是否正确指定，3 种方法的 Bias 差异不大，说明 3 种方法均能得到参数的相合估计。由 GEE 和 ERGEE 方法相比较可知，ERQIF 方法的 SD 均较小，特别是当工作相关结构被错误指定时 SD 有较大降低，说明 ERQIF 方法可避免对讨厌参数的估计，同时提高估计的效率。

从表 2 可以看出，GEE 方法的 SD 值受离群点的影响很大，说明该方法不能得到参数的稳健估计。通过引入加权指数得分函数得到的 ERGEE 和 ERQIF 方法可以有效控制离群点的影响，从而保证了估计的稳健性。同时从 SD 的结果可以看出，不管相关结构是否正确指定，ERQIF 方法均优于 ERGEE 方法，其估计效率有了显著提高。

3.2　逻辑回归模型

当响应变量是 0–1 型变量时，假设模型满足：

$$\ln(\mu_{ij}/(1-\mu_{ij})) = x_{ij}\beta_0 (i=1,\cdots,n; j=1,\cdots,m) \tag{6}$$

其中 x_{ij} 独立地服从[0,1]上的均匀分布，$\beta_0 = 0.4$，$\boldsymbol{Y}_i$ 的真实相关矩阵 $\boldsymbol{R}_i(\alpha)$ 为可交换结构，且 $\alpha = 0.2$。在模拟计算中，分别指定 $\boldsymbol{R}_i(\alpha)$ 为 Exch 和 AR(1) 两种结构；参考文献[14]生成不同结构下的 y_{ij}；为了研究稳健性，对部分数据进行扰动，即将 $\boldsymbol{x}_{ij}(i=1,\cdots,10)$ 的每个值都减去 5，计算得到 3 种方法下参数的偏差和标准差如表 3 所示。

表 3　参数 β_0 的估计结果

Table 3　Parameter estimation results for β_0

$\boldsymbol{R}(\alpha)$	方法	Bias[a)]	SD[a)]	Bias[b)]	SD[b)]
Exch	GEE	0.03404	0.24458	0.04147	0.28801
	ERGEE	0.01701	0.04938	0.01370	0.09124
	ERQIF	0.00851	0.01198	0.01048	0.00159
AR(1)	GEE	0.07315	0.23402	0.08171	0.28940
	ERGEE	0.01613	0.05126	0.01979	0.04655
	ERQIF	0.00521	0.04131	0.00996	0.00511

a—数据无扰动；b—数据有扰动。

从表 3 可以看出，3.2 节和 3.1 节的模拟结果相似。在不同情况下，ERQIF 方法在 Bias 与 SD 上均比其他方法占优势，特别当数据有扰动时，ERGEE 与 ERQIF 方法比非稳健 GEE 方法的 Bias 有显著降低；且 ERQIF 较 ERGEE 的 SD 更小，且在工作相关结构被错误指定的情况下，ERQIF 也能得到更稳健的估计，估计效率更高。

总之，从 3.1 节和 3.2 节的模拟结果可以得出，当数据出现离群点或工作相关矩阵被错误指定时，ERQIF 方法均有更好的稳健性和有效性。

4 结束语

本文将加权指数得分函数和 QIF 方法相结合,得到了纵向数据下广义线性模型的 ERQIF 估计,证明了估计的渐近性质,并通过数值计算得到了稳健的模拟结果。由于本文仅研究了模型参数估计问题,未来可望进一步研究各种半参数模型的稳健估计和变量选择问题,并对估计的稳健性进行理论分析。

参考文献:

[1] Liang K Y, Zeger S L. Longitudinal data analysis using generalized linear models[J]. Biometrika, 1986, 73(1): 13-22.

[2] Qu A, Lindsay B G, Li B. Improving generalised estimating equations using quadratic inference functions[J]. Biometrika, 2000, 87(4): 823-836.

[3] Lian H, Liang H, Wang L. Generalized additive partial linear models for clustered data with diverging number of covariates using GEE[J]. Statistica Sinica, 2014, 24: 173-196.

[4] Qu A, Li R Z. Quadratic inference functions for varying-coefficient models with longitudinal data[J]. Biometrics, 2006, 62(2): 379-391.

[5] Tian R Q, Xue L G, Liu C L. Penalized quadratic inference functions for semiparametric varying coefficient partially linear models with longitudinal data[J]. Journal of Multivariate Analysis, 2014, 132: 94-110.

[6] Fan Y L, Qin G Y, Zhu Z Y. Variable selection in robust regression models for longitudinal data[J]. Journal of Multivariate Analysis, 2012, 109: 156-167.

[7] Zheng X Y, Fung W K, Zhu Z Y. Robust estimation in joint mean—covariance regression model for longitudinal data[J]. Annals of the Institute of Statistical Mathematics, 2013, 65(4): 617-638.

[8] 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.

[9] Wang X Q, Jiang Y L, Huang M, et al. Robust variable selection with exponential squared loss[J]. Journal of the American Statistical Association, 2013, 108(502): 632-643.

[10] Lv J, Yang H, Guo C H. An efficient and robust variable selection method for longitudinal generalized linear models[J]. Computational Statistics & Data Analysis, 2015, 82: 74-88.

[11] Xue L, Qu A N, Zhou J H. Consistent model selection for marginal generalized additive model for correlated data[J]. Journal of the American Statistical Association, 2010, 105(492): 1518-1530.

[12] Hansen L P. Large sample properties of generalized method of moments estimators[J]. Econometrica, 1982, 50(4): 1029-1054.

[13] Yao W X, Lindsay B G, Li R Z. Local modal regression[J]. Journal of Nonparametric Statistics, 2012, 24(3): 647-663.

[14] Oman S D. Easily simulated multivariate binary distributions with given positive and negative correlations[J]. Computational Statistics & Data Analysis, 2009, 53(4): 999-1005.

Robust quadratic inference functions for generalized linear models with longitudinal data

GUAN XiaoNi HUANG Bin*

(Faculty of Science, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: This paper presents an efficient and robust estimation method for generalized linear models with longitudinal data. By using a quadratic inferential function (QIF) to improve the weighted exponential score function, we can obtain an effective and robust quadratic inferential function (ERQIF). Under some regularity conditions, the resulting estimators are consistent and asymptotically normal distributed. Finally, simulation studies show that the proposed estimators have robust and efficient numerical results, even when many outliers are included and the working correlation matrix is misspecified.

Key words: longitudinal data; weighted exponential score function; quadratic inferential function (QIF); robust estimation

(责任编辑:汪 琴)