Statistics and Application 统计学与应用, 2018, 7(4), 367-375
Published Online August 2018 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2018.74043

Hans汉斯

A Robust Estimation Method for the Partial Linear Models with Longitudinal Data

Xiaofei Sun[1], Shaomin Li[2], Kangning Wang[1*]

[1]School of Statistics, Shandong Technology and Business University, Yantai Shandong
[2]School of Economics, Shandong University, Jinan Shandong
Email: *wkn1986@126.com

Received: Jul. 3rd, 2018; accepted: Jul. 23rd, 2018; published: Jul. 30th, 2018

Abstract

Based on the exponential squared loss function, we propose a robust estimation method for the parametric parts in the partial linear models. By using the kernel approximation and transformation, we first absorb the nonparametric part. Then by using the exponential squared loss function to deduce the influence of outliers, we propose robust generalized estimation equations and empirical likelihood ratio function for estimation and inference. Under some regularity conditions, we show that the resulting estimators are consistent and asymptotic normality. Simulation results also illustrate the robustness and efficiency advantages of our method.

Keywords

Longitudinal Data, Partial Linear Models, Robust Estimation, Empirical Likelihood

纵向数据部分线性模型的一种稳健估计方法

孙晓霏[1]，李劭珉[2]，王康宁[1*]

[1]山东工商学院统计学院，山东 烟台
[2]山东大学经济学院，山东 济南
Email: *wkn1986@126.com

收稿日期：2018年7月3日；录用日期：2018年7月23日；发布日期：2018年7月30日

摘 要

本文利用指数平方损失函数，针对纵向数据部分线性模型的参数部分提出了一种稳健的估计方法。首先通过核估计以及变换将非参部分消去，然后利用指数平方损失函数降低异常值的影响，构造出稳健的广义估计方程，最终推导出稳健的经验似然比函数，由此进行估计和统计推断。在一定正则条件下该估计

*通讯作者。

文章引用：孙晓霏，李劭珉，王康宁. 纵向数据部分线性模型的一种稳健估计方法[J]. 统计学与应用，2018，7(4): 367-375. DOI: 10.12677/sa.2018.74043

具有相合性以及渐近正态性，蒙特卡洛模拟显示本文的方法具有稳健性以及较高的效率。

关键词

纵向数据，部分线性模型，稳健估计，经验似然

1. 引言

部分线性模型是一类非常重要的半参数模型，比单纯的线性回归模型或非参数模型有更大的适应性以及更强的解释能力，在医学和计量经济学等领域有着广泛的应用。而在这些领域中常常用到的一种数据类型为纵向数据，即每个个体经过多次测量得到的数据集。考虑来自 n 个个体的数据，其中第 i 个个体有 m_i 次观测，$i=1,\cdots,n$，总的观测数为 $N=\sum_{i=1}^{n}m_i$。设 Y_{ij} 和 $\left(X_{ij},T_{ij}\right)$ 分别是是第 i 个个体第 j 次观测的响应变量和协变量，其中 X_{ij} 是 $p\times 1$ 维向量，Y_{ij} 和 T_{ij} 都是一维变量。纵向数据的部分线性模型为：

$$Y_{ij}=X_{ij}^{\mathrm{T}}\beta+g\left(T_{ij}\right)+\varepsilon_{ij},\ i=1,2,\cdots,n;\ j=1,2,\cdots,m_i \tag{1}$$

其中 β 是 $p\times 1$ 维未知参数向量，$g(\cdot)$ 是未知的基准函数，ε_{ij} 为随机误差项，假定 $E\left[\varepsilon_{ij}\mid T_{ij}\right]=0$，$E\left[\varepsilon_{ij}^2\mid T_{ij}\right]=\sigma_\varepsilon^2$，不失一般性，不妨假定 $T_{ij}\in[0,1]$。一般假定来自不同个体的观测是相互独立的(称之为组间独立)，而来自同一个个体的观测具有一定的相关性(称之为组内相关)。

Zeger 和 Diggle [1] (1994)最早对模型(1)进行了研究，利用后移算法(back-fitting)，结合核估计以及最小二乘估计将参数 β 和函数 $g(\cdot)$ 估计出来。Lin 和 Carroll [2] (2001)利用 profile 估计方程估计参数部分，并结合估计方程与核估计方法估计非参部分。He 等[3] (2002)利用 B-样条逼近非参部分，然后选择适当的损失函数用 M 估计方法估计参数 β 以及样条系数。Fan 和 Li [4] (2004)利用局部多项式逼近非参部分，在假定工作独立协方差矩阵下，提出两种估计方法：差分估计和 profile 最小二乘估计。Xue 和 Zhu [5] (2008)利用核估计以及差分法消去未知的基准函数 $g(\cdot)$，并利用经验似然(Owen, 1988) [6]方法估计 β 以及进行统计推断。以上方法都没有考虑纵向数据模型中的组内相关性以及估计的稳健性，而在对实际问题的研究中得到的数据常常存在异常值或与假定的分布不符，于是估计的稳健性变得尤为重要。

本文首先利用广义估计方程(Liang, 1986) [7]的思想，引入工作相关矩阵将组内相关性考虑进来，然后利用指数平方损失函数(Wang, 2013) [8]降低异常值的影响，并在估计方程中加 W 权重减轻杠杆值的影响，得到稳健的广义估计方程，并以此构造出稳健的经验似然函数，得到稳健的经验似然估计，并进行统计推断。

2. 稳健经验似然的构造

首先对(1)式两边取给定 T_{ij} 下的条件期望，得

$$E\left[Y_{ij}\mid T_{ij}\right]=E\left[X_{ij}^{\mathrm{T}}\mid T_{ij}\right]\beta+g\left(T_{ij}\right),$$

两式相减得

$$Y_{ij}-E\left[Y_{ij}\mid T_{ij}\right]=\left(X_{ij}-E\left[X_{ij}\mid T_{ij}\right]\right)^{\mathrm{T}}\beta+\varepsilon_{ij},$$

令 $Y_{ij}^{*}=Y_{ij}-E\left[Y_{ij}\mid T_{ij}\right]$，$X_{ij}^{*}=X_{ij}-E\left[X_{ij}\mid T_{ij}\right]$，则模型变为

$$Y_{ij}^{*}=X_{ij}^{*\mathrm{T}}\beta+\varepsilon_{ij},\ i=1,2,\cdots,n;\ j=1,2,\cdots,m_i。$$

为了稳健，本文引入指数平方损失函数 $\varphi_\gamma(r)=-\frac{2r}{\gamma}\exp\left(-\frac{r^2}{\gamma}\right)$，其中 $\gamma>0$ 为调节参数，当 γ 非常大时，$1-\exp\left(-r^2/\gamma\right)\approx r^2/\gamma$，于是得到的估计与最小二乘估计类似；当 γ 较小时，指数平方损失函数能够降低较大的残差对参数估计的影响，即选取较小的 γ 能够有效的控制异常值的影响，提高估计的稳健性，最优 γ 值的选取见第三节。

令 $Y_i^{*}=\left(Y_{i1}^{*},Y_{i2}^{*},\cdots,Y_{im_i}^{*}\right)^{\mathrm{T}}$，$X_i^{*}=\left(X_{i1}^{*},X_{i2}^{*},\cdots,X_{im_i}^{*}\right)^{\mathrm{T}}$，$\varepsilon_i=\left(\varepsilon_{i1},\varepsilon_{i2},\cdots,\varepsilon_{im_i}\right)^{\mathrm{T}}$，$r_i^{*}(\beta)=Y_i^{*}-X_i^{*}\beta$，$V_i=Var(\varepsilon_i)=\sigma_\varepsilon^2R_i(\rho)$，$R_i(\rho)$ 为工作相关矩阵，ρ 为未知的参数，注意在随后的矩条件 $E\left[Z(\beta_0)\right]=\mathbf{0}$ 中可将方差 σ_ε^2 消去，于是不妨直接假定 $V_i=R_i(\rho)$。构造辅助随机向量

$$Z_i(\beta)=X_i^{*\mathrm{T}}V_i^{-1}h_i^{\gamma}\left(r_i^{*}(\beta)\right),\ i=1,\cdots,n。$$

其中 $h_i^{\gamma}\left(r_i^{*}(\beta)\right)=C_i\left[\varphi_\gamma\left(r_i^{*}(\beta)\right)-E\left[\varphi_\gamma\left(r_i^{*}(\beta)\right)\right]\right]$，减去 φ_γ 的期望是为了确保 Fisher 一致性(He [9], 2005)。权重矩阵 $C_i=diag\left(c_{i1},\cdots,c_{im_i}\right)$ 用来约束协变量空间中杠杆点(leverage points)的影响，常用的权重是 Mallows-type 权函数

$$c_{ij}=c\left(x_{ij}\right)=\min\left\{1,\left\{\frac{b_x}{\left(x_{ij}-m_x\right)^{\mathrm{T}}S_x^{-1}\left(x_{ij}-m_x\right)}\right\}^{\frac{\xi}{2}}\right\},$$

其中 $\xi\geq1$，m_x 和 S_x 分别是对 x_{ij} 位置参数和尺度参数的稳健估计，b_x 是 $\chi^2(p)$ 的 95%分位数。

利用辅助向量 $Z_i(\beta)$ 可以定义 β 的稳健经验似然比函数，然而此函数不能直接用于估计 β 以及进行统计推断，因为 $Z_i(\beta)$ 中含有未知的函数 $E\left[Y_{ij}\mid T_{ij}\right]$ 和 $E\left[X_{ij}\mid T_{ij}\right]$ 以及未知的相关性参数 ρ，需要用各自的估计量替代。对于 ρ 可以用稳健的矩估计方法，例如，对于可交换工作相关结构，ρ 的估计为

$$\hat{\rho}=\frac{1}{nH^2}\sum_{i=1}^{n}\frac{1}{m_i(m_i-1)}\sum_{j\neq k}\varphi_\gamma\left(e_{ij}\right)\varphi_\gamma\left(e_{ik}\right),$$

对于一阶自回归工作相关结构，ρ 的估计为

$$\hat{\rho}=\frac{1}{nH^2}\sum_{i=1}^{n}\frac{1}{m_i-1}\sum_{j\leq m_i-1}\varphi_\gamma\left(e_{ij}\right)\varphi_\gamma\left(e_{i,j+1}\right),$$

其中 $H^2=E\left[\varphi_\gamma^2\left(e_{ij}\right)\right]$。

对于 $E\left[Y_{ij}\mid T_{ij}\right]$ 和 $E\left[X_{ij}\mid T_{ij}\right]$ 的估计可用非参数统计方法，本文选用核估计

$\hat{E}\left[Y_{ij}\mid T_{ij}=t\right]=\sum_{i=1}^{n}\sum_{j=1}^{m_i}W_{ij}(t)Y_{ij}$ 和 $\hat{E}\left[X_{ij}\mid T_{ij}=t\right]=\sum_{i=1}^{n}\sum_{j=1}^{m_i}W_{ij}(t)X_{ij}$，其中

$W_{ij}(t)=K_h\left(T_{ij}-t\right)\Big/\sum_{k=1}^{n}\sum_{l=1}^{m_k}K_h\left(T_{kl}-t\right)$，$K_h(\cdot)=K(\cdot/h)$，$K(\cdot)$ 是一个核函数。

用估计值分别替代 $Z_i(\beta)$ 中的 ρ、$E\left[Y_{ij}\mid T_{ij}\right]$ 和 $E\left[X_{ij}\mid T_{ij}\right]$，得到 $Z_i(\beta)$ 的估计量 $\hat{Z}_i(\beta)$，以此构建 β 的稳健似然比函数

$$L(\beta)=\max\left\{\prod_{i=1}^{n}np_i\,\middle|\,p_i\geq0,\sum_{i=1}^{n}p_i=1,\sum_{i=1}^{n}p_i\hat{Z}_i(\beta)=\mathbf{0}\right\},$$

利用拉格朗日乘子法可以得出，$L(\beta)$ 的最大值在 $p_i=\dfrac{1}{n\left(1+\lambda^{\mathrm{T}}\hat{Z}_i(\beta)\right)}$，$i=1,\cdots,n$ 处取到，其中 λ 满足

$$\sum_{i=1}^{n}\frac{\hat{Z}_i(\beta)}{1+\lambda^{\mathrm{T}}\hat{Z}_i(\beta)}=\mathbf{0},$$

于是 β 的对数经验似然比函数为

$$l(\beta)=-\sum_{i=1}^{n}\log\left(1+\lambda^{\mathrm{T}}\hat{Z}_i(\beta)\right)。$$

最大化似然比函数可以求出 β 的估计 $\hat{\beta}_{\mathrm{MELE}}$，根据 Xue [5]和 He [9]，在一定正则条件下，有

$$\sqrt{n}\left(\hat{\beta}_{MELE}-\beta_0\right)\xrightarrow{d}N\left(\mathbf{0},\Sigma_0^{-1}\Omega_0\Sigma_0^{-1}\right),$$

以及

$$-2l(\beta)\xrightarrow{d}\chi^2(p),$$

其中 $\chi^2(p)$ 表示自由度为 p 的卡方分布，于是本文估计具有相合性以及渐近正态性，且 β 的置信域为

$$I_\alpha=\left\{\beta\middle|-2l(\beta)\le\chi^2_{1-\alpha}(p)\right\},$$

其中 α 为显著性水平。

3. 算法

下面给出具体的算法：

Step 1：用核估计计算 $\hat{E}\left[Y_{ij}\,|\,T_{ij}\right]$ 和 $\hat{E}\left[X_{ij}\,|\,T_{ij}\right]$，对数据进行变换 $\hat{Y}_{ij}^{*}=Y_{ij}-\hat{E}\left[Y_{ij}\,|\,T_{ij}\right]$，$\hat{X}_{ij}^{*}=X_{ij}-\hat{E}\left[X_{ij}\,|\,T_{ij}\right]$；

Step 2：用 GEE 估计作为参数估计的初始值 $\beta^{(0)}$，令 t=1；

Step 3：在估计值为 $\beta^{(t)}$ 时，求出相关性参数的估计 $\hat{\rho}$，并代入工作相关矩阵得到 $\hat{V}_i=R_i(\hat{\rho})$。通过最小化 $\det\left(Cov\left(\beta^{(t)}\right)\right)$ 得到调节参数 γ 的估计 $\gamma_{opt}^{(t)}$，其中 $\det(\cdot)$ 表示行列式运算；

$$Cov\left(\beta^{(t)}\right)=\left[\hat{\Sigma}\left(\mu_i\left(\beta^{(t)}\right),\hat{\rho}\right)\right]^{-1}\hat{\Omega}\left(\mu_i\left(\beta^{(t)}\right),\hat{\rho}\right)\left[\hat{\Sigma}\left(\mu_i\left(\beta^{(t)}\right),\hat{\rho}\right)\right]^{-1},$$

$$\hat{\Sigma}\left(\mu_i\left(\beta^{(t)}\right),\hat{\rho}\right)=\sum_{i=1}^{n}X_i^{\mathrm{T}}\hat{V}_i^{-1}E\left[\frac{\partial h_i^{\gamma}\left(\mu_i\left(\beta^{(t)}\right)\right)}{\partial\mu_i\left(\beta^{(t)}\right)}\right]X_i,$$

$$\hat{\Omega}\left(\mu_i\left(\beta^{(t)}\right),\hat{\rho}\right)=\sum_{i=1}^{n}X_i^{\mathrm{T}}\hat{V}_i^{-1}\left[h_i^{\gamma}\left(\mu_i\left(\beta^{(t)}\right)\right)\right]\left[h_i^{\gamma}\left(\mu_i\left(\beta^{(t)}\right)\right)\right]^{\mathrm{T}}\hat{V}_i^{-1}X_i;$$

Step 4：最小化 $-2l(\beta,\hat{\rho})$ 得到新的估计值 $\beta^{(t+1)}$；

Step 5：令 $t=t+1$，迭代 Step 2~Step 3 直到收敛，得到最终的估计值 $\hat{\beta}_{MELE}$ 和 $\hat{\gamma}_{opt}$。

4. 模拟

本节通过蒙特卡洛模拟，比较本文方法与 Xue 提出方法的稳健性。模型设定为

$$y_{ij}=x_{ij}\beta+g\left(t_{ij}\right)+\varepsilon_{ij},\ i=1,2,\cdots,n;\ j=1,2,\cdots,m,$$

其中 $x_{ij}\sim Uniform(-1,1)$，$t_{ij}\sim Uniform(0,1)$，$g(t)=3\sin\left(\frac{\pi}{2}t\right)$，$(\varepsilon_{i1},\cdots,\varepsilon_{iT})^{\mathrm{T}}\sim N\left((0,\cdots,0)^{\mathrm{T}},\sigma^2R(\rho)\right)$，令 $\sigma^2=1$，$R(\rho)$ 分别取可交换结构(Exch)和一阶自相关结构(AR-1)，参数 ρ 选取 0.3 和 0.7 两个数值，取 $n=50$，100；$m=3$，10。估计时所用的工作相关性矩阵分别选取独立结构(Ind)、可交换结构(Exch)和一阶自相关结构(AR-1)。令数据以概率 δ 受到来自分布 $\chi^2(5)-5$ 的污染，考虑 $\delta=0.05,0.1$ 和 0.2 三种不同的污染程度。由 n、m、ρ 和 δ 的不同取值以及真实相关矩阵的不同结构共得到 24 种不同的情形，每种情形分别进行 1000 次模拟，计算出对应的偏(Bias)、均方误差(MSE)、覆盖率(CP)以及置信区间宽度(Width)，结果列在下文的表格和图中。

表 1 所列为四种估计的偏(Bias)与均方误差(MSE)的计算数值，其中 $\hat{\beta}(\mathrm{Xue})$ 表示用 Xue 的方法得到

Table 1. The bias and mean squared error of estimators
表 1. 估计的偏(Bias)与均方误差(MSE)

	δ	$\hat{\beta}$	Exch				AR−1			
			$\rho=0.3$		$\rho=0.7$		$\rho=0.3$		$\rho=0.7$	
			Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
$n=50$, $m=3$	0.05	$\hat{\beta}$(Xue)	−0.0515	0.0439	−0.0596	0.0430	−0.0749	0.0554	−0.0391	0.0484
		$\hat{\beta}$(Exch)	−0.0422	0.0242	−0.0470	0.0215	−0.0680	0.0395	−0.0402	0.0250
		$\hat{\beta}$(AR-1)	−0.0415	0.0244	−0.0448	0.0221	−0.0680	0.0395	−0.0372	0.0242
		$\hat{\beta}$(Ind)	−0.0443	0.0258	−0.0463	0.0290	−0.0681	0.0408	−0.0326	0.0311
	0.1	$\hat{\beta}$(Xue)	−0.0691	0.0596	−0.0559	0.0595	−0.0553	0.0574	−0.0458	0.0566
		$\hat{\beta}$(Exch)	−0.0566	0.0322	−0.0428	0.0296	−0.0427	0.0357	−0.0362	0.0225
		$\hat{\beta}$(AR-1)	−0.0576	0.0323	−0.0415	0.0298	−0.0416	0.0352	−0.0339	0.0216
		$\hat{\beta}$(Ind)	−0.0575	0.0331	−0.0428	0.0360	−0.0478	0.0370	−0.0436	0.0291
	0.2	$\hat{\beta}$(Xue)	−0.1047	0.0820	−0.0329	0.0746	−0.0640	0.0759	−0.0655	0.0707
		$\hat{\beta}$(Exch)	−0.0694	0.0335	−0.0359	0.0314	−0.0457	0.0313	−0.0396	0.0225
		$\hat{\beta}$(AR-1)	−0.0677	0.0328	−0.0396	0.0329	−0.0440	0.0296	−0.0362	0.0223
		$\hat{\beta}$(Ind)	−0.0697	0.0339	−0.0354	0.0396	−0.0490	0.0317	−0.0396	0.0276
$n=50$, $m=10$	0.05	$\hat{\beta}$(Xue)	−0.0291	0.0103	−0.0239	0.0093	−0.0102	0.0103	−0.0197	0.0102
		$\hat{\beta}$(Exch)	−0.0203	0.0053	−0.0192	0.0029	−0.0095	0.0059	−0.0143	0.0046
		$\hat{\beta}$(AR-1)	−0.0196	0.0057	−0.0191	0.0035	−0.0104	0.0056	−0.0189	0.0033
		$\hat{\beta}$(Ind)	−0.0209	0.0064	−0.0191	0.0057	−0.0102	0.0061	−0.0143	0.0060
	0.1	$\hat{\beta}$(Xue)	−0.0328	0.0149	−0.0173	0.0120	−0.0278	0.0127	−0.0196	0.0139
		$\hat{\beta}$(Exch)	−0.0255	0.0065	−0.0190	0.0044	−0.0180	0.0063	−0.0136	0.0052
		$\hat{\beta}$(AR-1)	−0.0242	0.0068	−0.0177	0.0053	−0.0180	0.0056	−0.0160	0.0041
		$\hat{\beta}$(Ind)	−0.0236	0.0074	−0.0160	0.0068	−0.0189	0.0064	−0.0132	0.0062
	0.2	$\hat{\beta}$(Xue)	−0.0237	0.0175	−0.0243	0.0190	−0.0195	0.0184	−0.0203	0.0188
		$\hat{\beta}$(Exch)	−0.0188	0.0063	−0.0189	0.0042	−0.0145	0.0075	−0.0144	0.0061
		$\hat{\beta}$(AR-1)	−0.0187	0.0068	−0.0164	0.0049	−0.0140	0.0074	−0.0142	0.0051
		$\hat{\beta}$(Ind)	−0.0183	0.0073	−0.0200	0.0060	−0.0134	0.0074	−0.0165	0.0070

Continued

$n=100$, $m=3$	0.05	$\hat{\beta}$(Xue)	−0.0388	0.0162	−0.0425	0.0188	−0.0271	0.0170	−0.0395	0.0196
		$\hat{\beta}$(Exch)	−0.0294	0.0101	−0.0265	0.0071	−0.0225	0.0098	−0.0300	0.0073
		$\hat{\beta}$(AR-1)	−0.0319	0.0103	−0.0269	0.0076	−0.0241	0.0098	−0.0279	0.0070
		$\hat{\beta}$(Ind)	−0.0318	0.0110	−0.0279	0.0106	−0.0242	0.0107	−0.0311	0.0114
	0.1	$\hat{\beta}$(Xue)	−0.0378	0.0224	−0.0326	0.0199	−0.0240	0.0221	−0.0408	0.0267
		$\hat{\beta}$(Exch)	−0.0238	0.0112	−0.0202	0.0064	−0.0250	0.0117	−0.0313	0.0120
		$\hat{\beta}$(AR-1)	−0.0221	0.0111	−0.0213	0.0062	−0.0243	0.0115	−0.0328	0.0117
		$\hat{\beta}$(Ind)	−0.0230	0.0115	−0.0260	0.0093	−0.0255	0.0118	−0.0296	0.0151
	0.2	$\hat{\beta}$(Xue)	−0.0572	0.0356	−0.0358	0.0274	−0.0198	0.0290	−0.0323	0.0345
		$\hat{\beta}$(Exch)	−0.0244	0.0121	−0.0243	0.0100	−0.0197	0.0109	−0.0261	0.0121
		$\hat{\beta}$(AR-1)	−0.0249	0.0122	−0.0219	0.0099	−0.0198	0.0108	−0.0240	0.0123
		$\hat{\beta}$(Ind)	−0.0261	0.0128	−0.0261	0.0137	−0.0208	0.0111	−0.0228	0.0143
$n=100$, $m=10$	0.05	$\hat{\beta}$(Xue)	−0.0138	0.0050	−0.0143	0.0054	−0.0167	0.0050	−0.0055	0.0053
		$\hat{\beta}$(Exch)	−0.0108	0.0022	−0.0074	0.0012	−0.0140	0.0028	−0.0116	0.0024
		$\hat{\beta}$(AR-1)	−0.0105	0.0026	−0.0067	0.0014	−0.0148	0.0027	−0.0099	0.0017
		$\hat{\beta}$(Ind)	−0.0106	0.0029	−0.0100	0.0027	−0.0136	0.0029	−0.0080	0.0032
	0.1	$\hat{\beta}$(Xue)	−0.0177	0.0068	−0.0148	0.0063	−0.0136	0.0058	−0.0173	0.0059
		$\hat{\beta}$(Exch)	−0.0096	0.0026	−0.0112	0.0014	−0.0111	0.0026	−0.0118	0.0024
		$\hat{\beta}$(AR-1)	−0.0101	0.0027	−0.0114	0.0017	−0.0119	0.0025	−0.0100	0.0018
		$\hat{\beta}$(Ind)	−0.0110	0.0030	−0.0145	0.0028	−0.0114	0.0027	−0.0130	0.0030
	0.2	$\hat{\beta}$(Xue)	−0.0145	0.0092	−0.0131	0.0111	−0.0168	0.0098	−0.0142	0.0087
		$\hat{\beta}$(Exch)	−0.0095	0.0025	−0.0103	0.0027	−0.0135	0.0041	−0.0092	0.0025
		$\hat{\beta}$(AR-1)	−0.0094	0.0026	−0.0104	0.0035	−0.0130	0.0037	−0.0072	0.0022
		$\hat{\beta}$(Ind)	−0.0096	0.0029	−0.0091	0.0042	−0.0127	0.0041	−0.0109	0.0031

的估计，$\hat{\beta}(\text{Exch})$、$\hat{\beta}(\text{AR-1})$ 和 $\hat{\beta}(\text{Ind})$ 表示本文方法利用不同工作相关性矩阵结构得到的估计，第一行的 Exch 和 AR-1 表示真实的相关性矩阵的结构。通过比较可以发现，在多数情形下，本文估计量(不论利用何种工作相关矩阵结构)的偏差要更加接近于 0；而在所有情形下，本文估计量的均方误差要远远小于 Xue 的估计，大多不到其一半，这表明本文构建的估计量有较强的稳健性。当污染率增加时，四种估计的均方误差也随之增加，这是由于污染数据的比例增加导致估计结果的波动增大，而相对于 Xue 的方法本文估计量的均方误差仍处于较低的水平。当 n 或 m 增大时，四种估计的均方误差都减小，这是由于样本量增大导致估计的方差减小。

图 1 表示真实的相关性矩阵结构为 Exch 时的不同情形下四种估计的均方误差，其中圆圈符号代表 Xue 的估计，三角形代表本文提出的估计，显然在所有情形下本文方法的均方误差都远小于 Xue 的方法。图 1 中的前 3 个三角表示在 $n=50$，$m=3$，$\rho=0.3$，污染率为 0.05 时分别利用 Exch、AR-1 和 Ind 三种工作矩阵结构得到的估计，可以发现工作矩阵用 Exch 结构得到的估计的均方误差最小，用 AR-1 结构得到的结果与之相近，用 Ind 结构(即不考虑组内相关性)得到的估计的均方误差最大。前 3 个圆圈表示对应情形下 Xue 的估计，由于 Xue 的估计没有用到工作相关矩阵，因此得到的估计相同，于是这三个圆圈处

于相同的高度。从圆圈的分布可以看出 3 个圆圈为一个阶梯，第 1~3 个阶梯所处情形的区别在于污染率分别为 0.05、0.1 和 0.2，于是均方误差呈现上升的趋势；前 3 个阶梯与随后的 3 个阶梯所处情形的区别在于相关性参数的不同，因此这两段阶梯形状和高度相似；前 6 个阶梯与随后的 6 个阶梯所处情形的区别在于 m 分别取 3 和 10，m 取 10 时样本量增加了大约两倍，估计的均方误差更小，于是第 7~12 个阶梯要明显低于前 6 个阶梯。对三角形的走势进行同样的分析也能得出类似的结果，而且在每个三角形的阶梯中，总是第 1 个三角形最低，第 3 个三角形最高，这表明当模型中存在组内相关性而在估计时忽略相关性(用 Ind 工作矩阵)得到的估计效率最低，而选取的工作矩阵结构为真实的结构时估计的效率最高。

表 2 列出的是 n = 50，m = 3 时不同情形下四种方法求出的置信水平为 95%的置信区间的覆盖率(CP)

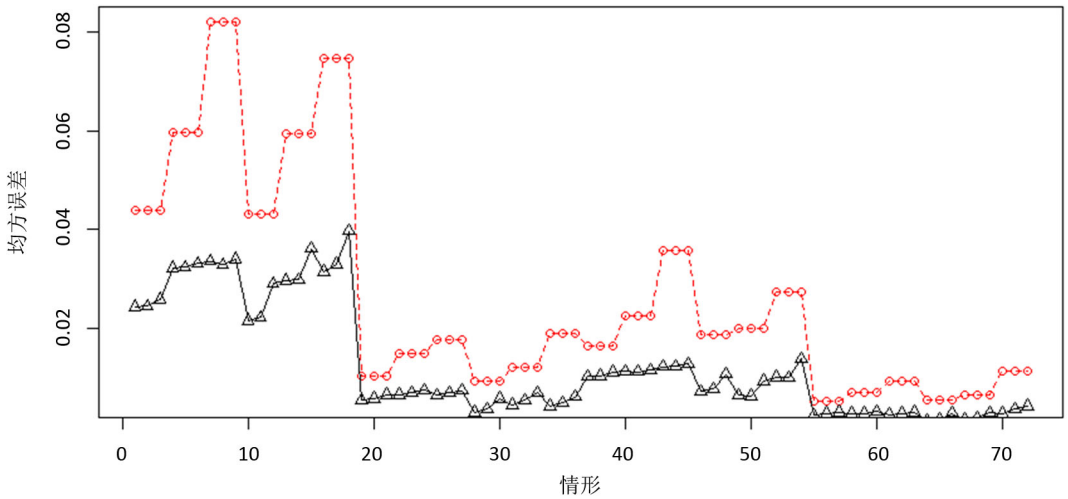

Figure 1. Mean squared error of four estimators under different conditions
图 1. 不同情形下四种估计的均方误差

Table 2. Coverage probabilities and width of the confidence interval
表 2. 置信区间的覆盖率(CP)和宽度(Width)

δ	$\hat{\beta}$	Exch				AR-1			
		$\rho = 0.3$		$\rho = 0.7$		$\rho = 0.3$		$\rho = 0.7$	
		CP	Width	CP	Width	CP	Width	CP	Width
0.05	$\hat{\beta}$(Xue)	91.6%	0.6858	89.5%	0.7033	90.7%	0.6940	90.9%	0.6895
	$\hat{\beta}$(Exch)	91.8%	0.6086	90.0%	0.5140	92.2%	0.6280	91.0%	0.5294
	$\hat{\beta}$(AR-1)	92.1%	0.6157	89.1%	0.5372	92.3%	0.6232	91.2%	0.5217
	$\hat{\beta}$(Ind)	91.9%	0.6279	91.1%	0.6438	91.6%	0.6413	92.4%	0.6361
0.1	$\hat{\beta}$(Xue)	90.9%	0.7808	89.4%	0.7872	88.6%	0.7810	91.0%	0.7944
	$\hat{\beta}$(Exch)	91.9%	0.6524	91.7%	0.5729	88.7%	0.6553	91.1%	0.5949
	$\hat{\beta}$(AR-1)	91.3%	0.6586	91.3%	0.5929	89.2%	0.6527	91.8%	0.5919
	$\hat{\beta}$(Ind)	91.5%	0.6710	90.4%	0.6727	89.3%	0.6649	91.7%	0.6704
0.2	$\hat{\beta}$(Xue)	90.7%	0.9642	89.7%	0.9621	91.3%	0.9636	90.8%	0.9579
	$\hat{\beta}$(Exch)	91.9%	0.7466	91.9%	0.6930	91.5%	0.7878	90.4%	0.7080
	$\hat{\beta}$(AR-1)	91.5%	0.7517	91.8%	0.7100	91.4%	0.7871	90.9%	0.7106
	$\hat{\beta}$(Ind)	92.0%	0.7577	92.0%	0.7565	91.3%	0.7861	90.5%	0.7578

和宽度(Width)，第一行的 Exch 和 AR-1 表示真实的相关性矩阵的结构。通过比较可以发现，在几乎所有情形下，与用 Xue 的方法求出的置信区间相比，本文方法(不论用何种工作相关矩阵结构)得到的置信区间的覆盖率更加接近真实的置信水平 95%，而且宽度更小，这表明用本文方法进行统计推断能够得到更加稳健的结果。通过比较利用三种不同的工作相关结构得到的结果发现，当模型中存在组内相关性而在统计推断时忽略相关性(用 Ind 工作矩阵)得到的结果最差，而选取的工作矩阵结构为真实的结构时统计推断的结果最好。从表 2 中还可以看出，污染率由 0.05 增大到 0.2 时，置信区间的宽度也相应增加，这是由于数据污染越严重，估计量的波动越大，达到相同的置信率所需要的置信区间也随之扩大。

5. 结论

本文利用指数平方损失函数，针对纵向数据部分线性模型的参数部分提出了一种稳健的估计方法。首先通过核估计以及变换将非参部分消去，然后利用广义估计方程(GEE)的思想，通过假定相关性结构将组内相关性考加入到参数估计过程中；同时为了提高估计的稳健性，引入指数平方损失函数降低异常值的影响，并利用 W 权重减轻杠杆值的影响，构造出稳健的广义估计方程，最终推导出稳健的经验似然比函数，通过最大化似然比函数可得到参数的稳健估计，在一定正则条件下该估计具有相合性以及渐近正态性，且似然比函数服从渐近卡方分布，由此可进行稳健的统计推断。

本文在生成的数据中加入一定比例的污染，考虑了不同样本容量、不同相关性结构以及不同的污染率，对 24 种情形分别进行蒙特卡洛模拟，并将本文提出的方法与 Xue 的方法进行比较。结果表明，多数情形下本文估计量(不论利用何种工作相关结构)的偏差要更加接近于 0，且在所有情形下的均方误差远小于 Xue 估计的均方误差，这表明本文构建的估计量有较强的稳健性。将所有情形得到的结果放在一张图上可以明显看出，若存在组内相关性而估计时忽略，估计的有效性最差；而工作相关矩阵用可交换结构(Exch)和一阶自相关结构(AR-1)得到的估计的差别相对较小；若假定的工作相关性矩阵的结构与真实结构相同，估计的效果最好。与用 Xue 方法求得的置信区间相比，不论工作相关性矩阵的结构是否准确，本文方法得到的置信区间有着更小的宽度以及更加接近真实置信水平的覆盖率，这表明用本文方法进行统计推断能够得到更加稳健的结果。

基金项目

本项目受国家自然科学基金(11231005; 71673171)和山东省自然科学基金(ZR2017BA002)资助。

参考文献

[1] Zeger, S.L. and Diggle, P.J. (1994) Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters. *Biometrics*, **50**, 689. https://doi.org/10.2307/2532783

[2] Lin, X. and Carroll, R.J. (2001) Semiparametric Regression for Clustered Data Using Generalized Estimating Equations. *Journal of the American Statistical Association*, **96**, 1045-1056. https://doi.org/10.1198/016214501753208708

[3] He, X. and Kim, M.O. (2002) On Marginal Estimation in a Semiparametric Model for Longitudinal Data with Time-Independent Covariates. *Metrika*, **55**, 67-74. https://doi.org/10.1007/s001840200187

[4] Fan, J. and Li, R. (2004) New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis. *Journal of the American Statistical Association*, **99**, 710-723. https://doi.org/10.1198/016214504000001060

[5] Xue, L.G. and Zhu, L.X. (2008) Empirical Likelihood-Based Inference in a Partially Linear Model for Longitudinal Data. *Science in China Series A*: *Mathematics*, **51**, 115-130. https://doi.org/10.1007/s11425-008-0020-4

[6] Owen, A.B. (1988) Empirical Likelihood Ratio Confidence Intervals for a Single Functional. *Biometrika*, **75**, 237-249. https://doi.org/10.1093/biomet/75.2.237

[7] Liang, K. and Zeger, S.L. (1986) Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, **73**, 13-22. https://doi.org/10.1093/biomet/73.1.13

[8] Wang, X., Jiang, Y., Huang, M., *et al*. (2013) Robust Variable Selection with Exponential Squared Loss. *Journal of the*

American Statistical Association, **108**, 632-643. https://doi.org/10.1080/01621459.2013.766613

[9] He, X., Fung, W.K. and Zhu, Z. (2005) Robust Estimation in Generalized Partial Linear Models for Clustered Data. *Journal of the American Statistical Association*, **100**, 1176-1184. https://doi.org/10.1198/016214505000000277