文章编号:1000-5862(2015)04-0360-05

纵向数据测量误差模型的 2 次统计推断

张 涛[1,3],章 溢[2]

(1. 中国社会科学院金融研究所,北京 100028;2. 江西师范大学计算机信息工程学院,江西 南昌 330022
3. 兴业银行博士后工作站,福建 福州 350001;)

摘要:考虑纵向数据的线性误差模型,其中协变量含有测量误差. 使用 2 次函数推断方法得到回归参数的估计,证明所得到的估计渐近地服从正态分布;对参数的假设检验问题,证明所得统计量渐近地服从 χ^2 分布,并通过数值模拟讨论方法的有限样本性质. 最后,该方法被用于 1 组艾滋病数据的实证分析中.

关键词:纵向数据;测量误差;QIF 方法

中图分类号:O 212 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2015.04.06

0 引言

数据的测量误差问题广泛存在于实际中. 如果忽略测量误差问题,直接对相应模型的参数进行估计,那么所得的估计统计量是有偏的,因而也是不相合的. 在独立数据的假定下,W. A. Fuller 等[1-2] 介绍线性回归模型和非线性回归模型下的参数估计问题.

对于纵向数据,也有大量文献研究混合模型和转移误差模型的统计推断问题. J. Buonaccorsi 等[3-4] 考虑了线性和非线性混合模型中的测量误差问题. W. Pan 等[5] 研究了广义转移误差模型的极大似然估计. 以上文献所得结论需要对模型的分布做一些特别的假定. Zhang Tao 等[6] 讨论了相关结构对参数估计的影响,但是没有考虑数据的测量误差问题.

在没有假定任何分布的条件下,对含有测量误差问题的纵向数据模型的统计推断往往因个体内部相关结构而变得复杂. 目前国内对这一领域问题的研究,往往建立在一种工作独立的基础上,即忽略纵向数据个体内部的相关结构[2]. 但是该假定会导致估计参数效率的损失. 因而,提高参数估计效率的一个方法就是在估计过程中考虑个体的相关结构. A. Qu 等[7] 提出了 2 次推断函数方法,把广义矩方法推广到纵向数据. Lai Peng 等[8-10] 从不同的角度推广了 2 次函数方法. 但是以上文献没有考虑测量误差问题. 陈广雷等[11-12] 考虑变系数模型的测量误差问题.

本文考虑纵向数据的线性模型,其中假设响应变量为连续,部分协变量含有测量误差. 在估计的过程中,通过 2 次推断函数方法考虑个体内部的相关结构. 与现有处理纵向数据测量误差问题的文献相比,本文是在没有假定任何分布的情况下考虑个体的相关结构.

1 模型和参数估计过程

令 $\boldsymbol{Y}_i$ 为 $n_i \times 1$ 维响应变量,$\boldsymbol{X}_i$,$\boldsymbol{Z}_i$ 分别为 $n_i \times p$ 和 $n_i \times q$ 维协变量. 假定个体之间的观测相互独立. 考虑如下模型 $\boldsymbol{Y}_i = \boldsymbol{X}_i\beta_1 + \boldsymbol{Z}_i\beta_2 + \boldsymbol{\varepsilon}_i$,$\boldsymbol{W}_i = \boldsymbol{Z}_i + \boldsymbol{U}_i$,其中 $\boldsymbol{\beta} = (\beta_1, \beta_2)$ 表示未知的回归参数向量. 假定 $E\boldsymbol{\varepsilon}_i = \boldsymbol{0}$,$E\boldsymbol{U}_i = \boldsymbol{0}$;$\boldsymbol{Z}_i$,$\boldsymbol{U}_i$ 和 $\boldsymbol{\varepsilon}_i$ 不相关. 随机误差 $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{in_i})'$ 的分布是未知的. $\forall i, j$,令 $K = \min(n_i, n_j)$,假定 $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{iK})'$ 和 $\boldsymbol{\varepsilon}_j = (\varepsilon_{j1}, \cdots, \varepsilon_{jK})'$ 有相同的协方差结构 $\boldsymbol{V}_K$.

在真实的协方差矩阵已知的情况下,广义估计方程为 $\sum_{i=1}^{n} \dot{\boldsymbol{\mu}}_i \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}$,其中 $\boldsymbol{\mu}_i = \boldsymbol{X}_i\beta_1 + \boldsymbol{Z}_i\beta_2$,$\dot{\boldsymbol{\mu}}_i$ 是 $n_i \times p$ 矩阵,它表示 $\boldsymbol{\mu}_i$ 对参数 $\boldsymbol{\beta}$ 的 1 阶导数,$\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\alpha)\boldsymbol{A}_i^{1/2}$,$\boldsymbol{A}_i$ 表示第 i 个体的协方差矩阵对角线元素所构成的矩阵,$\boldsymbol{R}_i(\alpha)$ 表示相关结构矩阵. 上式又可以表示为 $\sum_{i=1}^{n} (\boldsymbol{X}_i, \boldsymbol{Z}_i)' \boldsymbol{A}_i^{-1/2}\boldsymbol{R}_i^{-1}(\alpha) \cdot \boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta_1 - \boldsymbol{Z}_i\beta_2) = \boldsymbol{0}$.

收稿日期:2014-11-25

基金项目:国家自然科学基金(71361015),江西省自然科学基金(20142BAB201013)和江西师范大学青年成长基金(004796)资助项目.

作者简介:张 涛(1977-),男,江苏沛县人,讲师,博士,主要从事纵向数据,广义线性模型和时间序列的研究.

由于 Z_i 含有测量误差,提出修正的广义估计方程为

$$\sum_{i=1}^{n}\{(X_i,W_i)'A_i^{-1/2}R_i^{-1}(\alpha)A_i^{-1/2}(Y_i-X_i\beta_1-W_i\beta_2)-D_i\beta\}=0, \quad (1)$$

其中 $D_i=E[0_{n_i\times p},U_i)'A_i^{-1/2}R_i^{-1}(\alpha)A_i^{-1/2}(0_{n_i\times p},U_i)]$.

类似于文献[7],把 $R_i(\alpha)$ 分解为一系列基矩阵的线性组合

$$R_i^{-1}(\alpha)=\sum_{j=1}^{m}\alpha_jM_j, \quad (2)$$

其中 $\alpha_1,\cdots,\alpha_m$ 为未知常数,$M_1,\cdots,M_m$ 为已知矩阵. 一些已知常用的相关结构都包含在(2)式之中.

把(2)式代入(1)式,可以得到

$$\sum_{i=1}^{n}\{(X_i,W_i)'A_i^{-1/2}(\alpha_1M_1,\cdots,\alpha_mM_m)A_i^{-1/2}(Y_i-X_i\beta_1-W_i\beta_2)-(\alpha_1D_i^{(1)},\cdots,\alpha_mD_i^{(m)})\beta\}=0,$$

其中 $D_i^{(k)}=E[\xi'_iA_i^{-1/2}M_kA_i^{-1/2}\xi_i]$,$\xi_i=(0_{n_i\times p},U_i)$.

定义 $G_n(\beta)=n^{-1}\sum_{i=1}^{n}g_i(\beta)$,其中

$$g_i(\beta)=\begin{pmatrix}(X_i,W)'A_i^{-1/2}M_1A_i^{-1/2}(Y_i-X_i\beta_1-W_i\beta_2)-D_i^{(1)}\beta\\ \vdots\\ (X_i,W)'A_i^{-1/2}M_mA_i^{-1/2}(Y_i-X_i\beta_1-W_i\beta_2)-D_i^{(m)}\beta\end{pmatrix},$$

则 2 次推断函数定义为

$$Q_n(\beta)=G'_n(\beta)C_n^{-1}(\beta)G_n(\beta),$$

其中 $C_n(\beta)=n^{-1}\sum_{i=1}^{n}g_i(\beta)g'_i(\beta)$. 参数 β 的估计可以定义为 $\beta=\underset{\beta}{\operatorname{argmin}}Q_n(\beta)$.

2　渐近性质

为了得到参数的渐近性质,需要以下假设:

(S_1) 参数空间 Θ 是 $\mathbf{R}^{p+q}$ 的 1 个紧集,真实参数 β_0 是其一内点;

(S_2) 随机误差向量满足 $E\|\varepsilon_i\|^4<+\infty$;

(S_3) $E\|(X_i,Z_i)\|^4<+\infty$;

(S_4) 令

$$\lim_{x\to\infty}\frac{1}{n}\sum_{i=1}^{n}E\begin{pmatrix}(X_i,Z_i)'A_i^{-1/2}M_1A_i^{-1/2}(X_i,Z_i)\\ \vdots\\ (X_i,Z_i)'A_i^{-1/2}M_mA_i^{-1/2}(X_i,Z_i)\end{pmatrix}=J_0.$$

引理 1　假定(S_1) ~ (S_4) 成立,则

(i) 当 $n\to\infty$ 时,$\forall\beta\in\Theta$,$\dot{G}_n$ 依概率收敛到 $-J_0$,$G_n(\beta)$ 依概率收敛到 $J_0(\beta_0-\beta)$,其中 $\dot{G}_n(\beta)=\partial G_n(\beta)/\partial\beta$;

(ii) $G_n(\beta_0)$ 满足 Lyapunov 中心极限定理条件,即

$$n^{1/2}C_0^{-1/2}G_n(\beta_0)\xrightarrow{D}N(0,I_{m(p+q)}),$$

其中 C_0 为 $C_n(\beta)$ 在 β_0 处的极限,$I_{m(p+q)}$ 为 $m(p+q)$ 阶单位矩阵;

(iii) $\forall\beta\in\Theta$,$C_n(\beta)$ 一致收敛于某一 $C(\beta)$.

证　类似于文献[13]可以证明,详细过程略.

定理 1　在(S_1) ~ (S_4) 的假定下,$\hat{\beta}$ 依概率收敛到 β_0,且有

$$\sqrt{n}(\hat{\beta}-\beta_0)\xrightarrow{D}N(0,(J'_0C_0^{-1}J_0)^{-1}).$$

证　根据引理 1,有

$$\sqrt{n}G_n(\beta_0)\xrightarrow{D}N(0,C_0), \quad (3)$$

根据 $\hat{\beta}$ 的定义,显然有 $Q_n(\hat{\beta})\leqslant Q_n(\beta_0)$.

由(3)式得 $Q_n(\beta_0)=G'_n(\beta_0)C_n^{-1}(\beta_0)\cdot G_n(\beta_0)=O_p(n^{-1})=o_p(1)$,即 $Q_n(\hat{\beta})\xrightarrow{P}0$.

据文献[13]的引理 4 得 $\hat{\beta}\xrightarrow{P}\beta_0$,因为 $\hat{\beta}$ 满足 $Q_n(\beta)$ 最小. 因而有 $\dot{Q}_n(\hat{\beta})=0$. 根据 Taylor 展开式可得

$$0=\dot{Q}_n(\hat{\beta})=\dot{Q}_n(\hat{\beta}_0)+\ddot{Q}_n(\tilde{\beta})(\hat{\beta}-\beta_0),$$

其中 $\tilde{\beta}$ 介于 $\hat{\beta}$ 和 $\hat{\beta}_0$ 之间. 因此有

$$\hat{\beta}-\beta=-\ddot{Q}_n^{-1}(\tilde{\beta})\dot{Q}_n(\beta).$$

注意到 $\ddot{Q}_n=2\dot{G}'_nC_n^{-1}\dot{G}_n+o_p(1)$,结合引理 1 可得 $\ddot{Q}_n(\tilde{\beta})\xrightarrow{P}2J'_0C_0^{-1}J_0$,并经过简单计算 $\sqrt{n}(\hat{\beta}-\beta_0)=-\sqrt{n}(J'_0C_0^{-1}J_0)^{-1}(J'_0C_0^{-1}G_n(\beta_0))+o_p(1)$. 结合(3)式有

$$\sqrt{n}(\hat{\beta}-\beta_0)\xrightarrow{D}N(0,(J'_0C_0^{-1}J_0)^{-1}).$$

假定参数 $\beta=(\gamma,\delta)$,其中 γ 是 s 维感兴趣参数,δ 是 $(p+q-s)$ 维讨厌参数. 定义

$$\tilde{\delta}=\underset{\delta}{\operatorname{argmin}}Q_n(\gamma,\delta),\ (\hat{\gamma},\hat{\delta})=\underset{(\gamma,\delta)}{\operatorname{argmin}}Q_n(\gamma,\delta),$$

可以得到定理 2.

定理 2　假定(S_1) ~ (S_4) 成立,对于原假设 $H_0:\gamma=\gamma_0$,则

$$Q_n(\gamma_0,\tilde{\delta})-Q_n(\hat{\gamma},\hat{\delta})\xrightarrow{D}\chi^2(s).$$

证　类似于文献[7]中定理 1 的证明,详细过程略.

定理 2 提供了一种对感兴趣的回归参数的检验方法. 与定理 1 的渐近正态性结论相比,定理 2 的检验方法避免了估计参数 γ 的协方差矩阵.

3 数值模拟

将通过数值模拟来说明以上理论的有限样本性质. 在模拟中,产生50个个体,每个个体包含着10次观测. 另一方面,允许每个观测以0.2的概率随机缺失. 考虑如下模型:

$$\boldsymbol{y}_i = \beta_1 \boldsymbol{x}_i + \beta_2 z_i + \boldsymbol{\varepsilon}_i, \boldsymbol{w}_i = z_i + \boldsymbol{u}_i.$$

对模型参数假定$\beta_1 = \beta_2 = 1$,$\boldsymbol{\varepsilon}_i$服从10维正态分布$N(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{R}(\rho))$,$\boldsymbol{x}_i$和$z_i$来自多元正态分布$N(\boldsymbol{\mu}, \boldsymbol{I})$,其中$\boldsymbol{\mu} = (0.1, 0.2, \cdots, 1.0)'$,$\boldsymbol{R}(\rho)$表示可交换相关结构或者AR(1)相关结构,$\boldsymbol{U}_i$服从10维正态分布$N(\boldsymbol{0}, \sigma_u^2 \boldsymbol{I})$. 在模拟中,设定$\sigma_\varepsilon^2 = 1.0$和$\sigma_u^2 = 0.6^2$. 为了估计$\boldsymbol{D}^{(k)}$,对$W_{ij}$进行重复观测. 以上模拟进行5 000次,模拟结果如表1所示.

表1 真实相关结构为可交换的模拟结果

工作相关结构	相关参数ρ	参数1			参数2		
		偏差	标准差	标准误差	偏差	标准差	标准误差
忽略结构	0.3	0.034	0.053	0.052	−0.123	0.049	0.048
	0.5	0.035	0.057	0.055	−0.124	0.051	0.050
	0.7	0.033	0.061	0.058	−0.125	0.054	0.053
真实结构	0.3	−0.001	0.046	0.046	0.002	0.052	0.049
	0.5	0.000	0.042	0.042	0.002	0.047	0.045
	0.7	−0.002	0.038	0.036	0.001	0.041	0.040
独立	0.3	−0.001	0.053	0.052	0.002	0.058	0.056
	0.5	0.000	0.058	0.055	0.001	0.060	0.058
	0.7	−0.002	0.061	0.058	−0.001	0.063	0.061
可交换	0.3	−0.001	0.047	0.044	0.003	0.053	0.048
	0.5	0.000	0.043	0.040	0.002	0.048	0.044
	0.7	−0.002	0.038	0.035	0.001	0.042	0.038
AR(1)	0.3	−0.002	0.052	0.046	0.005	0.058	0.050
	0.5	−0.001	0.050	0.044	0.005	0.056	0.048
	0.7	−0.002	0.046	0.040	0.003	0.051	0.044

从表1可以得到结论:1)当测量误差完全被忽略时,对应估计的偏差比较大. 而考虑测量误差所得到的结果,偏差都比较小;2)在工作独立的假定下,随着相关参数ρ的增大,相应的标准差和标准误差增大,这表明相关性越强,工作对相关参数的估计效率越有重要的影响;3)当工作相关结构被正确指定时,所得到的估计最有效.

4 实证分析

本部分对1组艾滋病数据进行实证分析. 数据来自1984—1991年感染HIV病毒的283个个体. 数据包含患者感染HIV病毒时的年龄、抽烟情况、感染前血液中CD4细胞的百分比,以及随着观测时间变化感染HIV病毒后血液CD4细胞的百分比等变量. 本文的目的是研究年龄、吸烟的状态和感染HIV病毒前血液中CD4细胞的百分比等因素对感染后血液中CD4百分比的影响是否显著.

假定Y_{ij}表示CD4细胞的百分比,X_1、X_2和Z分别表示年龄、吸烟状态和感染艾滋病以前的CD4细胞的百分比. 考虑模型

$$Y_{ij} = X_{1,i}\beta_1 + X_{2,i}\beta_2 + Z_i\beta_3 + \gamma_0 + \sum_{i=1}^{3} \gamma_r t_r^{\ ij} + \varepsilon_{ij}.$$

文献[4]研究了CD4细胞的测量误差问题. 因此假定$W_i = Z_i + U_i$.

由于没有重复测量数据,类似于文献[4],取σ_u^2为W的方差的15%和30%. 因此,这里假定σ_u^2分别为0,10.536和21.072等3种情形,分别在工作独立、可交换和AR(1)相关结构下,表2提供了参数的估计和标准差. 结果显示在0.05的显著水平下,年龄和吸烟对CD4百分比的影响在统计学意义上是不显著的. 而CD4细胞的百分比和感染HIV病毒以前CD4细胞的百分比之间存在着强烈的正相关关系.

表 2　AID 数据的参数估计(标准误差)

情形	参　数	工作结构		
		独立	可交换	AR(1)
1	年龄	-0.072(0.078)	-0.044(0.066)	-0.014(0.066)
	吸烟	0.622(1.134)	0.515(1.054)	0.559(1.047)
	感染前 CD4 细胞	0.375(0.067)	0.374(0.064)	0.384(0.063)
2	年龄	-0.090(0.079)	-0.060(0.066)	-0.031(0.067)
	吸烟	0.414(1.144)	0.273(1.064)	0.320(1.057)
	感染前 CD4 细胞	0.448(0.082)	0.454(0.078)	0.469(0.078)
3	年龄	-0.116(0.081)	-0.082(0.069)	-0.060(0.070)
	吸烟	0.107(1.175)	-0.069(1.096)	-0.052(1.088)
	感染前 CD4 细胞	0.557(0.109)	0.576(0.103)	0.606(0.105)

根据定理 2,考虑模型参数的假设检验问题. 假定工作相关结构为可交换和 AR(1) ,$\sigma_u^2=0$,表 3 给出相应问题的检验结果. 情形 2 和情形 3 的结果类似,这里没有单独列出. 表 2 和表 3 显示,对假设检验问题 $\beta_1=0$,$\beta_2=0$ 和 $\beta_3=0$,定理 1 和定理 2 的 2 种方法得到的结论是一致的. 为了检验年龄和吸烟状态是否同时显著,在可交换相关结构的假定下,计算统计量的值为 0.727; 在 1 阶自相关工作结构的假定下,计算统计量的值为 0.293,对应的 P 值分别为 0.695 1 和 0.863 8. 这表明年龄和吸烟状态不是显著因素. 该结论与现有研究成果[15]是一致的.

表 3　对艾滋病数据参数的假设检验

	可交换		AR(1)	
原假设	统计量	P 值	统计量	P 值
$\beta_1=0$	0.336	0.562	0.029	0.864
$\beta_2=0$	0.270	0.603	0.229	0.632
$\beta_3=0$	23.319	0.000	21.881	0.000
$(\beta_1,\beta_2)=0$	0.727	0.695	0.293	0.864

5　结论

在前人的工作基础上,考虑了含有测量误差问题的纵向数据的统计推断. 给出了参数估计的大样本性质,并且解决了对感兴趣参数的假设检验问题. 模拟结果显示:

1) 直接忽略测量误差所得到的估计是有偏的;

2) 相关结构对参数估计的效率有重要的影响,当正确的相关结构被指定时,所得到的估计效率最高;

3) 对 1 组艾滋病数据进行了实证研究,结果表明,年龄和吸烟状态这 2 个因素对感染 HIV 病毒以后血液中 CD4 细胞百分比这一指标影响在统计学意义上是不显著的,而感染以前血液中 CD4 细胞所占百分比在统计学意义上是显著因素.

6　参考文献

[1] Fuller W A. Measurement error models [M]. New York: Wiley,1987.

[2] Carroll R J,Ruppert D,Stefanski L A. Measurement error in nonlinear models: A modern perspective [M]. New York: Chapman and Hall,2006.

[3] Buonaccorsi J,Demidenko E,Tosteson T. Estimation in longitudinal random effects models with measurement error [J]. Statistica Sinica,2000,10(2) : 885-903.

[4] Wang Naisyin,Lin Xihong,Gutierrez R G,et al. Bias analysis and SIMEX approach in generalized linear mixed measurement error models [J]. Journal of the American Statistical Association,1998,93(4) : 249-261.

[5] Pan Wenqin,Lin Xihong,Zeng Donglin. Structural inference in transition measurement error models for longitudinal data [J]. Biometrics,2006,62(6) : 402-412.

[6] Zhang Tao,Zhu Zhongyi. Efficient inference based on block empirical likelihood for longitudinal partially linear regression models [J]. Chinese Journal of Applied Probability and Statistics,2010,26(3) : 323-335.

[7] Qu Annie,Lindsay B G,Li Bing. Improving generalized estimating equations using quadratic inference functions [J]. Biometrika,2000,87(2) : 823-836.

[8] Lai Peng,Li Gaorong,Lian Hua. Quadratic inference functions for partially linear single-index models with longitudinal data [J]. Journal of Multivariate Analysis,2013,118 (8) : 115-127.

[9] Philip M W. A bias-corrected covariance estimator for improved inference when using an unstructured correlation

with quadratic inference functions [J]. Statistics & Probability Letters,2013,83(6):1553-1558.

[10] Tian Ruiqin,Xue Liugen,Liu Chunling. Penalized quadratic inference functions for semiparametric varying coefficient partially linear models with longitudinal data [J]. Journal of Multivariate Analysis,2014,132(5):94-110.

[11] 陈广雷. 变系数测量误差模型的 B-样条估计 [J]. 应用数学,2014,27(1):45-51.

[12] 徐修友,黄彬. 协变量含测量误差的变系数偏线性模型的变量选择问题研究 [J]. 北京化工大学学报:自然科学版,2013,40(6):325-333.

[13] Bai Yang,Zhu Zhongyi,Fung Wing. Partial linear models for longitudinal data based on quadratic inference functions [J]. Scandinavian Journal of Statistics,2008,35(1):104-118.

[14] Liang Hua,Wang Suojin,Carroll R J. Partially linear models with missing response variables and error-prone covariates [J]. Biometrika,2007,94(2):185-198.

[15] Qu Ainne,Li Runze. Quadratic inference functions for varying-coefficient models with longitudinal data [J]. Biometrics,2006,62(7):379-391.

The Quadratic Inference Functions in Measurement Error Model for Longitudinal Data

ZHANG Tao[1,3], ZHANG Yi[2]

(1. Institute of Finance and Banking,Chinese Academy of Social Sciences,Beijing 100028,China;
2. College of Computer and Information Engineering,Jiangxi Normal University,Nanchang Jiangxi 30022,China;
3. Postdoctoral Workstation of CIB,Fuzhou Fujian 350001,China)

Abstract: A linear model for longitudinal data with continuous responses and error-prone covariates via quadratic inference functions methods is considered. Asymptotic normality of the parameter estimators is established by quadratic inference functions. In order to testing interested parameter,the statistic that proposed asymptotically follows a chi-squared distribution. The finite-sample properties of the procedures are studied through Monte Carlo simulations. At last,an application to a longitudinal study is used to illustrate the procedure developed here.

Key words: longitudinal data; measurement error; QIF method

(责任编辑:曾剑锋)