

引用格式:李宏凯,肖松涛,欧阳应根,等.一类线性混合效应模型分位数估计的影响分析[J].北京化工大学学报(自然科学版),2021,48(3):106–113.

LI HongKai, XIAO SongTao, OUYANG YingGen, et al. Influence analysis for quantile estimation of a class of linear mixed models[J]. Journal of Beijing University of Chemical Technology (Natural Science), 2021, 48(3):106–113.

一类线性混合效应模型分位数估计的影响分析

李宏凯¹ 肖松涛² 欧阳应根² 李志强^{1*}

(1. 北京化工大学 数理学院, 北京 100029; 2. 中国原子能科学研究院 放射化学研究所, 北京 102413)

摘 要:当回归模型误差服从非对称或非正态分布时,尤其是在重尾分布或分布受污染的情况下,如何检测纵向数据中的异常值是数据分析中的一个重要问题。为了克服非正态分布模型误差的影响,采用稳健的分位数方法对一类线性混合效应模型进行参数估计,并分别基于数据删除模型和均值漂移模型构造强影响点的诊断度量和异常值的检验统计量,以有效地检测强影响点和异常值点。在识别强影响点时,为了减轻计算负担,利用光滑逼近的方法给出了数据删除模型参数的一步近似估计,并据此构造出基于损失函数的距离和 Cook 距离。为了能够识别异常值点,首先构造出检验异常值点的 Wald 统计量,然后基于数据删除模型和均值漂移模型的系数估计的等价性,利用 Bootstrap 抽样得到检验的拒绝域。数值模拟结果表明,本文所提的诊断度量和检验统计量都能够很好地判断出强影响点和异常值点。最后应用本文方法针对化学实验纵向数据进行了影响分析。

关键词:线性混合效应模型;分位数估计;强影响点;异常值;Bootstrap 抽样

中图分类号: O212 **DOI:** 10.13543/j.bhxbzr.2021.03.013

引 言

线性混合效应模型是处理纵向数据最主要的模型之一,Diggle 等^[1]首次基于线性混合模型详细地研究了针对纵向数据的统计分析方法。随着纵向数据研究的不断发展,线性混合模型在生物、医学、卫生、心理学、农学 and 经济学等各个领域都有了非常广泛的应用。

影响分析作为统计建模中非常重要的一部分,主要被应用于识别数据中的强影响点和异常值点。近年来,关于线性混合模型的影响分析问题已经成为了研究热点之一。例如,当总体误差服从正态分布时,Pan 等^[2]基于似然函数及其导出的 Q 函数研究了线性混合模型在不同的随机效应协方差结构下的影响诊断问题;当模型误差服从指数族分布时,

Xu 等^[3]基于数据删除模型对广义线性混合效应模型进行影响分析,基于似然函数关于随机效应的条件期望构造了 Q 函数,进而建立了广义 Cook 距离和似然距离等;Pinho 等^[4]考虑了广义线性混合模型不同参数的影响分析问题,并且基于固定效应和随机效应估计的联合影响构建了广义 Cook 距离。此外,孙慧慧等^[5]与 Tang 等^[6]还分别研究了线性混合模型和部分线性混合模型的局部影响分析问题。然而,在很多情况下纵向数据模型的误差可能服从非正态分布或者非对称分布,此时一般的估计精度很容易受到影响^[4]。分位数回归估计作为一种稳健的估计方法,不仅可以克服上述问题,并且能够获得数据的总体分布信息,尤其是当回归误差项服从重尾分布或其分布受到污染时,分位数回归估计具有更高的稳健性。

Koenker^[7]最早对混合效应模型的分位数估计方法进行了研究,他利用加权分位数和对随机效应进行 L1 惩罚的方式,给出了模型的损失函数,获得了模型系数和随机效应的分位数估计。Geraci 等^[8]利用 Monte Carlo expectation maximization (MCEM) 算法消除了随机效应的影响,并求解了线性混合模型系数的分位数估计。尽管分位数回归估计具有稳

收稿日期:2020-11-04

基金项目:国家自然科学基金(21790371);中央高校基本科研业务费专项资金(XK2020-03)

第一作者:男,1995 年生,硕士生

* 通信联系人

E-mail: lizq@mail.buct.edu.cn

健性,但是 Narula 等^[9]的研究表明,在分位数 τ 较大或者较小的情况下,估计结果也会受到极端异常点的影响,此外他们的数值模拟结果和实际数据分析也说明了针对最小绝对偏差(least absolute deviation, LAD)回归进行影响分析等的统计诊断技术对异常数据的检测是非常重要的。尽管目前已经有很多关于影响分析的研究,但由于计算及统计推断的复杂性,针对线性混合效应模型分位数估计的影响分析的文献还很少见到,因此基于线性混合效应模型分位数估计对强影响点的诊断度量和异常值的检测进行研究是十分必要的。

本文利用线性混合效应模型的分位数估计处理服从非对称、非正态,尤其是重尾或轻尾分布的纵向数据模型的影响分析问题。为了检验强影响点并减少计算量,通过对分位数回归的光滑逼近得到了数据删除模型参数的一步近似估计,并构造了检验强影响点的损失函数距离和 Cook 距离。在识别异常值时,首先证明了数据删除模型和均值漂移模型参数估计的等价性,然后基于参数等价性给出计算 Wald 统计量的方法,并利用 Bootstrap 方法得到异常值检验的拒绝域。

1 实验数据描述及线性混合模型

1.1 数据描述

为了对氨基羟基脒(HSC)水溶液及硝酸水溶液在 γ 射线作用下的辐解产物进行定量分析,考察不同辐解剂量、HSC 初始浓度以及硝酸浓度条件下主要辐解产物的生成量,分别配制浓度为 0.1、0.2、0.4、0.6、0.8、1.0 mol/L 的 HSC 溶液和浓度为 0.2、0.4、0.6、0.8、1.0、1.5 mol/L 的硝酸溶液,两两组合得到 36 种不同浓度配比的溶液,委托中国原子能科学研究院辐照中心,采用实验型钴源装置(^{60}Co 源装置),分别置于辐解剂量为 1×10^3 、 2×10^3 、 4×10^3 、 6×10^3 、 8×10^3 、 10×10^3 Gy 的条件下,进行三因素下的无重复实验,总共得到 216 个样品。

为了考察不同条件下辐解剂量(10^3 Gy)、HSC 初始浓度(mol/L)以及硝酸浓度(mol/L)对铵根离子浓度(mmol/L)变化情况的影响,首先对数据进行初步的相关分析和散点图对比,结果表明辐解剂量与铵根离子浓度大体呈线性关系,而单独的 HSC 初始浓度和硝酸浓度对铵根离子浓度的影响则无明显的趋势变化规律。因此,本文将辐解剂量作为重要的解释变量,而将硝酸浓度和 HSC 浓度的联合影响

作为铵根离子浓度的随机影响因素,建立线性混合效应模型,分别考察两个因素的影响大小。

1.2 线性混合模型及其分位数估计

设 y_{ij} 为第 i 组实验第 j 次观测记录的铵根离子浓度, $i = 1, 2, \dots, n$ 代表 HSC 初始浓度和硝酸浓度的不同取值的组合, $j = 1, 2, \dots, n_i$ 代表在这些不同取值组合下的观测数,总样本数 $N = \sum_{i=1}^n n_i$ 。假设 y_{ij} 服从如式(1)所示的一类线性混合效应模型

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i + \varepsilon_{ij} \quad (1)$$

式中, $\mathbf{x}_{ij} = (\mathbf{x}_{ij_1}, \mathbf{x}_{ij_2}, \dots, \mathbf{x}_{ij_p})^T$ 是 $p \times 1$ 维的与固定效应相关的重要回归变量, $\boldsymbol{\beta}$ 是 $p \times 1$ 维的固定效应向量, $b_i \sim N(0, \sigma_b^2 \mathbf{I})$ 是不可观测的随机效应, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2 \mathbf{I})$ 是相互独立的均值为零的随机误差, $\sigma_b^2 \mathbf{I}$ 和 $\sigma_\varepsilon^2 \mathbf{I}$ 分别是协方差,并且 ε_{ij} 与 b_i 相互独立。混合效应模型假定模型扰动项由两部分组成,一部分随观测时间同时也随个体而变化,另一部分只随个体而不随观测时间变化。引入随机效应可以使个体观测数据之间具有一定的相关性。由于随机效应 b_i 是彼此相互独立的,因此当将随机效应和模型误差合在一起看作是模型的扰动时,二者共同刻画了纵向数据的组间独立性和组内相关性。

由于很多化学实验数据通常服从非对称、非正态的概率分布,为了克服误差分布的影响,采用稳健的分位数估计方法。根据 Koenker^[7]的定义,在 τ 分位数下的线性混合模型具有如下形式

$$Q_{y_{ij}}(\tau | \mathbf{x}_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta}(\tau) + b_i, i = 1, 2, \dots, n, j = 1, 2, \dots, n_i \quad (2)$$

式中, $Q_{y_{ij}}(\cdot) = F_{y_{ij}}^{-1}(\cdot)$ 是 y_{ij} 的条件累积分布函数的逆,分位数 $\tau \in (0, 1)$, $\boldsymbol{\beta}(\tau)$ 是 τ 分位数下模型的固定效应, b_i 是不依赖分位数 τ 的随机效应。

为了解决估计随机效应时可能会面临的维数较高的问题, Koenker^[7]提出一种基于随机效应的 L1 惩罚方法,通过加权多个分位数下的信息来获取参数与随机效应的联合估计,其损失函数定义如下

$$L(\boldsymbol{\beta}, \mathbf{b}) = \sum_{k=1}^q \sum_{i=1}^n \sum_{j=1}^{n_i} w_k \rho_{\tau_k}(y_{ij} - b_i - \mathbf{x}_{ij}^T \boldsymbol{\beta}(\tau_k)) + \lambda \sum_{i=1}^n |b_i| \quad (3)$$

式中, $\rho_\tau(r) = r\tau I(r > 0) + r(\tau - 1)I(r < 0)$, q 是分位数个数, w_k 是第 k 个分位数的权重, λ 是惩罚因子。Koenker^[7]建议将惩罚因子 λ 设置为 $\lambda = \sigma_\varepsilon / \sigma_b$, 其中 σ_ε 和 σ_b 可通过计算软件进行求解,具体选取过程

详见第3节。

由文献[7]可知,由于随机效应服从零均值的正态分布,为了保证固定效应的估计量具有较好的统计性质,要求固定效应中必须包含截距项。通过极小化损失函数(式(3))获得的参数估计记为 $\hat{\eta} = (\hat{\beta}, \hat{b})^T$ 。为求解损失函数的极小化解,以往的学者们提出了不同的求解思想和算法。其中,Koenker^[7]将问题转化为线性规划问题,然后利用单纯形法或内点法等优化算法进行计算。针对式(3)的损失函数,R语言提供了 regression quantiles for panel data (RQPD)程序包用于求解纵向数据的分位数估计。

2 线性混合效应模型分位数估计的影响分析

实际数据中常常会存在一些异常数据,通常将对回归估计影响过大的数据点称为强影响点,而将偏离回归直线较远的点称为异常值点。有时强影响点也很有可能同时是异常值点。影响分析正是为了识别出这些数据点而发展起来的一个统计分支,影响分析中常用的诊断模型包括数据删除模型(case deletion model, CDM)和均值漂移模型(mean shift outlier model, MSOM)。

2.1 数据删除模型的影响分析

2.1.1 数据删除模型

数据删除模型主要通过依次删除每个数据点来求解参数,之后再对比各个参数估计,若数据中存在较强的影响点,那么当删除该点时,模型的参数估计相对于原估计就会有较大的变化。

假定删去第 (i_1, j_1) 个观测值时得到的模型为

$$y_{ij} = x_{ij}^T \beta + b_i + \varepsilon_{ij}, i = 1, 2, \dots, n, j = 1, 2, \dots, n_i, (i, j) \neq (i_1, j_1) \quad (4)$$

则模型(4)的参数估计记为 $\hat{\eta}_{[i_1 j_1]} = (\hat{\beta}_{[i_1 j_1]}, \hat{b}_{[i_1 j_1]})^T$ 。

2.1.2 参数一步(光滑)近似估计

模型(4)相较原模型(1)只是减少了1个数据,所以参数估计的方法及计算复杂度与原模型基本相同,但是当数据量很大时,由于需要逐一检验每个数据点,将会产生很大的计算负担。针对线性模型, Cook 等^[10]提出了利用似然函数的 Taylor 展开式构造参数近似估计的思想。

为了能够对不可导的损失函数(3)进行光滑逼近,以建立数据删除模型参数的一步近似估计,本文结合一般分位数参数估计的 minorize-maximization

(MM)算法^[11]和绝对值函数的光滑方法^[12],构造如下目标函数

$$M(r, b | \hat{r}) = \sum_{k=1}^q \sum_{i=1}^n \sum_{j=1}^{n_i} w_k \frac{1}{4} \left(\frac{r_{kij}^2}{\delta + |\hat{r}_{kij}|} + (4\tau_k - 2)r_{kij} + C_k \right) + \lambda \sum_{i=1}^n \mu \ln \left(e^{\frac{b_i}{\mu}} + e^{-\frac{b_i}{\mu}} \right) \quad (5)$$

式中, $r_{kij} = y_{ij} - x_{ij}^T \beta(\tau_k) - b_i$, $\delta > 0$, $\mu > 0$ 是接近零的正常数, C_k 是常数。对数据删除模型目标函数在 $\hat{\eta} = (\hat{\beta}, \hat{b})^T$ 处进行 Taylor 展开得

$$M_{[i_1 j_1]}(\eta_{[i_1 j_1]} | \hat{r}_{[i_1 j_1]}) = M_{[i_1 j_1]}(\hat{\eta} | \hat{r}_{[i_1 j_1]}) + (\eta_{[i_1 j_1]} - \hat{\eta})^T M'_{[i_1 j_1]}(\hat{\eta} | \hat{r}_{[i_1 j_1]}) + (\eta_{[i_1 j_1]} - \hat{\eta})^T \frac{M''_{[i_1 j_1]}(\hat{\eta} | \hat{r}_{[i_1 j_1]})}{2} (\eta_{[i_1 j_1]} - \hat{\eta}) + o(1) \quad (6)$$

对式(6)两端关于 $\eta_{[i_1 j_1]}$ 求导可得^[10]

$$\hat{\eta}_{[i_1 j_1]} \approx \hat{\eta} - [M''_{[i_1 j_1]}(\hat{\eta} | \hat{r}_{[i_1 j_1]})]^{-1} M'_{[i_1 j_1]}(\hat{\eta} | \hat{r}_{[i_1 j_1]}) \quad (7)$$

2.1.3 诊断度量

通常情况下,参数估计都是多维向量,不便于直接比较,所以需要通过构建诊断度量来衡量参数变化的大小。在普通的回归模型中,常用的诊断度量有似然距离和 Cook 距离等,本文将似然距离推广到基于损失函数的距离。

1) 根据损失函数(3),构建基于损失函数的距离

$$L_D = 2[L(\hat{\beta}, \hat{b}) - L_{[i_1 j_1]}(\hat{\beta}_{[i_1 j_1]}, \hat{b}_{[i_1 j_1]})] \quad (8)$$

式中, $L_{[i_1 j_1]}(\hat{\beta}_{[i_1 j_1]}, \hat{b}_{[i_1 j_1]})$ 是删去第 (i_1, j_1) 个观测值的数据删除模型的损失函数。该距离类似于似然距离,但是使用的目标函数不同。另外针对损失函数(3)中的权重选取,可以利用损失函数距离(8)分别研究数据在不同分位数下的情况。

2) Cook 距离定义为^[10]

$$L_{CD} = (\hat{\eta}_{[i_1 j_1]} - \hat{\eta})^T [\text{Var}(\hat{\eta})]^{-1} (\hat{\eta}_{[i_1 j_1]} - \hat{\eta}) / p \quad (9)$$

式中 p 为参数 η 的维数。由于没有对模型(2)的误差项作任何的分布假定,所以在中小样本的情况下很难求得 Cook 距离中的 $\text{Var}(\hat{\eta})$ 。为了解决此问题,本文在中小样本条件下采用 Bootstrap 方法来近似计算样本方差 $\text{Var}(\hat{\eta})$ 。文献[7]中介绍了在 τ 分位数下构建 Bootstrap 样本的抽样方法,但是由于本文采用的加权分位数可以同时用于多个分位数下的参数估计,因此本文抽样误差需要从 $\tau = 0.5$ 时的分位数下产生,具体操作步骤如下。

1) 利用容量为 N 的原始样本求解分位数参数估计 $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{b}})^T$ 。假设共有 q 个分位数, 则记 $\hat{\boldsymbol{\beta}}(\tau_{q/2})$ 为 $\tau = 0.5$ 分位数下的参数估计, 利用 $\hat{\boldsymbol{\varepsilon}}_{ij} = y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(\tau_{q/2}) - \hat{b}_i$ 生成残差集合 $\{\hat{\boldsymbol{\varepsilon}}_{ij}\}$ 。

2) 从残差集合 $\{\hat{\boldsymbol{\varepsilon}}_{ij}\}$ 中有放回地随机抽取一个容量为 N 的 Bootstrap 残差集合 $\{\hat{\boldsymbol{\varepsilon}}_{ij}^*\}$ 。

3) 利用 $y_{ij}^* = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(\tau_{q/2}) + \hat{b}_i + \hat{\boldsymbol{\varepsilon}}_{ij}^*$ 生成 Bootstrap 样本 $(\mathbf{x}, \mathbf{y}^*)_{ij}$ 。

4) 利用 Bootstrap 样本 $(\mathbf{x}, \mathbf{y}^*)_{ij}$ 求解多个分位数下的参数估计 $\hat{\boldsymbol{\eta}}_{\tau_{ijk}}^*$ 。

5) 重复步骤 2) 至步骤 4) 共 M 次, 记 τ_k 分位数下的参数估计 $\hat{\boldsymbol{\eta}}_{\tau_{km}}^* = (\hat{\boldsymbol{\eta}}_{\tau_{k1}}^*, \hat{\boldsymbol{\eta}}_{\tau_{k2}}^*, \dots, \hat{\boldsymbol{\eta}}_{\tau_{kM}}^*)$ 。

6) 计算 τ_k 分位数下参数的样本协方差阵

$$\text{Var}(\hat{\boldsymbol{\eta}}_{\tau_k}) \approx \frac{\sum_{m=1}^M (\hat{\boldsymbol{\eta}}_{\tau_{km}}^* - \bar{\boldsymbol{\eta}}_{\tau_k}) (\hat{\boldsymbol{\eta}}_{\tau_{km}}^* - \bar{\boldsymbol{\eta}}_{\tau_k})^T}{M-1}$$

式中, $\bar{\boldsymbol{\eta}}_{\tau_k} = \frac{\sum_{m=1}^M \hat{\boldsymbol{\eta}}_{\tau_{km}}^*}{M}$ 为 τ_k 分位数下的样本均值。

2.2 均值漂移模型的影响分析

2.2.1 均值漂移模型

对 (i_1, j_1) 个观测值添加漂移项

$$\begin{cases} y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i + \varepsilon_{ij}, \text{ 且 } (i, j) \neq (i_1, j_1) \\ y_{i_1 j_1} = \mathbf{x}_{i_1 j_1}^T \boldsymbol{\beta} + b_{i_1} + \gamma_{i_1 j_1} + \varepsilon_{i_1 j_1}, \text{ 当 } (i, j) = (i_1, j_1) \text{ 时} \end{cases} \quad (10)$$

式中 $\gamma_{i_1 j_1}$ 是一个漂移项。记模型 (10) 的参数估计为 $\hat{\boldsymbol{\eta}}_{|i_1 j_1|} = (\hat{\boldsymbol{\beta}}_{|i_1 j_1|}, \hat{\boldsymbol{b}}_{|i_1 j_1|}, \hat{\gamma}_{i_1 j_1})^T$ 。此时均值漂移模型在分位数 $\tau_k (k=1, 2, \dots, q)$ 下的加权损失函数为

$$\begin{aligned} L_{|i_1 j_1|} &= \sum_{k=1}^q \sum_{i \neq i_1}^n \sum_{j=1}^{n_i} w_k \rho_{\tau_k}(y_{ij} - b_i - \mathbf{x}_{ij}^T \boldsymbol{\beta}(\tau_k)) + \\ &\sum_{k=1}^q \sum_{i=1}^n \sum_{j \neq j_1}^{n_i} w_k \rho_{\tau_k}(y_{ij} - b_{i_1} - \mathbf{x}_{ij}^T \boldsymbol{\beta}(\tau_k)) + \lambda \sum_{i=1}^n |b_i| + \\ &\sum_{k=1}^q w_k \rho_{\tau_k}(y_{i_1 j_1} - b_{i_1} - \mathbf{x}_{i_1 j_1}^T \boldsymbol{\beta}(\tau_k) - \gamma_{i_1 j_1 k}) \end{aligned} \quad (11)$$

式中 $\gamma_{i_1 j_1 k}$ 是分位数 τ_k 下的漂移项。均值漂移模型主要用来对每个数据点下的漂移项逐一进行假设检验: $H_{0k}: \gamma_{i_1 j_1 k} = 0$ 或 $H_{1k}: \gamma_{i_1 j_1 k} \neq 0$, 其中 $k=1, 2, \dots, q$ 。若拒绝 H_{0k} , 则说明在 τ_k 分位点下该数据点为异常值点。

2.2.2 $\gamma_{i_1 j_1 k}$ 的估计及数据删除模型和均值漂移模型系数估计的等价性

在分位数 τ_k 下, 对式 (11) 关于 $\gamma_{i_1 j_1 k}$ 进行极小

化, 可得

$$\rho_{\tau_k}(y_{i_1 j_1} - \mathbf{x}_{i_1 j_1}^T \hat{\boldsymbol{\beta}}_{|i_1 j_1|}(\tau_k) - \hat{b}_{|i_1 j_1| i_1 j_1} - \hat{\gamma}_{i_1 j_1 k}) = 0 \quad (12)$$

其中 $\hat{b}_{|i_1 j_1| i_1 j_1}$ 表示均值漂移模型中随机效应向量的估计值 $\hat{\boldsymbol{b}}_{|i_1 j_1|}$ 的第 $i_1 j_1$ 个分量, 由于 $\rho_{\tau}(r) = 0$ 当且仅当 $r = 0$, 因此可得

$$\hat{\gamma}_{i_1 j_1 k} = y_{i_1 j_1} - \mathbf{x}_{i_1 j_1}^T \hat{\boldsymbol{\beta}}_{|i_1 j_1|}(\tau_k) - \hat{b}_{|i_1 j_1| i_1 j_1} \quad (13)$$

根据式 (3) 和式 (11) 可以看出数据删除模型和均值漂移模型的系数 $\boldsymbol{\beta}$ 的估计是等价的, 由此可得

$$\hat{\gamma}_{i_1 j_1 k} = y_{i_1 j_1} - \mathbf{x}_{i_1 j_1}^T \hat{\boldsymbol{\beta}}_{[i_1 j_1]}(\tau_k) - \hat{b}_{[i_1 j_1] i_1 j_1} \quad (14)$$

式中 $\hat{\boldsymbol{\beta}}_{[i_1 j_1]}(\tau_k)$ 和 $\hat{b}_{[i_1 j_1] i_1 j_1}$ 是数据删除模型的参数估计。因此, 可以利用式 (14) 构造出原假设 H_{0k} 的检验统计量。

2.2.3 诊断统计量

为了得到统计量的抽样分布, 韦博成等^[13]在小样本中针对线性回归模型的异常值检验问题, 采用方差分析的方法构造出检验统计量; 曾林蕊等^[14]在误差的正态分布假设下, 针对大样本的半参数广义线性模型, 利用惩罚对数似然函数构建了 Score 检验统计量, 给出了上述问题的拒绝域。本文在中小样本和非正态或非对称的模型误差假定下, 构造了原假设 H_{0k} 下的 Wald 统计量

$$W = \frac{\gamma \cdot \gamma}{\text{Var}(\gamma)} \Big|_{\gamma = \hat{\gamma}_{i_1 j_1 k}} \quad (15)$$

式中, $\text{Var}(\gamma)$ 是漂移项的方差。统计量 (15) 在大样本下近似服从卡方分布, 但是在中小样本下其分布难以确定。因此, 本文采取 Bootstrap 法来构造上述统计量的拒绝域, 过程可以分为两步: 首先对统计量中的 $\text{Var}(\hat{\gamma}_{i_1 j_1 k})$ 进行近似计算; 其次利用统计量的 Bootstrap 样本构造假设检验的拒绝域。具体步骤如下所述。

1) 计算 $\text{Var}(\hat{\gamma}_{i_1 j_1 k})$ 。与 2.1.3 节类似, 首先利用原样本在 $\tau = 0.5$ 下的参数估计生成残差集合 $\{\varepsilon_{ij}\}$, 然后在抽样中锁定残差集合的第 (i_1, j_1) 个样本不动, 对其余的 $N-1$ 个样本进行有放回地抽样, 再合在一起生成容量为 N 的 Bootstrap 残差集合 $\{\tilde{\varepsilon}_{ij}\}$, 最后根据 $\tilde{y}_{ij} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(\tau_{q/2}) + \hat{b}_i + \tilde{\varepsilon}_{ij}$ 生成 Bootstrap 样本 $(\mathbf{x}, \tilde{\mathbf{y}})_{i_1 j_1}$, 其中 $\hat{\boldsymbol{\beta}}(\tau_{q/2})$ 是分位数 $\tau = 0.5$ 下的系数估计, 并基于式 (14) 求解漂移项的估计, 计算估计 $\hat{\gamma}_{i_1 j_1 k}$ 的 Bootstrap 方差, 记为 $\text{Var}(\hat{\gamma}_{i_1 j_1 k})$ 。

2) 利用初始样本 (\mathbf{x}, \mathbf{y}) 求得 $\hat{\gamma}_{i_1 j_1 k}$, 结合步骤 1) 中的 $\text{Var}(\hat{\gamma}_{i_1 j_1 k})$, 根据式 (15) 计算检验统计量^[15],

记为 T_0 。

3) 对初始样本在 $\tau = 0.5$ 分位数下的误差项 $\{\varepsilon_{ij}\}$ 进行 M 次有放回地抽样,生成 Bootstrap 误差项 $\{\varepsilon_{ij}^*\}$,利用 $y_{ij}^* = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(\tau_{q/2}) + \hat{b}_i + \varepsilon_{ij}^*$ 获得 Bootstrap 样本 $(\mathbf{x}, \mathbf{y}^*)_{i_{\nu l}}$,再将该样本漂移项的估计结果代入式(15),得到检验统计量 $T_{\tau_{km}}^*, m = 1, 2, \dots, M, k = 1, 2, \dots, q$ 。

4) 将分位数 τ_k 下的 M 个统计量从小到大重新排序,设定检验水平 α ,则 $T_{\tau_{km}}^*$ 上的 α 分位点为第 $[M(1 - \alpha)] + 1$ 个,记为 $T_{\tau_{k\alpha}}^*$,其中 $[A]$ 表示对实数 A 向下取整。由此可以构造 H_{0k} 下的拒绝域为 $\{T_0 > T_{\tau_{k\alpha}}^*\}$ 。

3 数值模拟

在模拟中 y_{ij} 代表第 i 个个体的第 j 个观测值,根据如下模型生成数据。

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i + \varepsilon_{ij} \quad (16)$$

假设 $\mathbf{x}_{ij} = (1, \hat{\mathbf{x}}_{ij})^T$, $\hat{\mathbf{x}}_{ij}$ 服从标准正态分布 $N(0, 1)$, 随机效应 $b_i \sim N(0, 2)$, 参数 $\boldsymbol{\beta} = (1, 2)^T$, 模型误差 $\varepsilon_{ij} \sim \text{Exp}(0.5) - 2$, 其中 $\text{Exp}(\lambda)$ 是参数为 λ 的指数分布,数学期望为 $\frac{1}{\lambda}$ 。共 20 组数据,每组 10 个观测值。分位数取 $(0.1, 0.3, 0.5, 0.7, 0.9)$ 。

图 1 是 0.5 分位数下模型的系数估计随惩罚因子 λ 的变化曲线,在 λ 大于 7 时系数不再发生变化,说明此时惩罚因子过大,导致随机效应不再起作用。针对线性混合模型(16),利用 R 语言 lme4 包对其随机效应方差和误差方差进行估计,然后根据 Koenker^[7] 建议的方法,并结合图 1,在模拟中经过多次实验,最终选取 $\lambda = 2$ 。

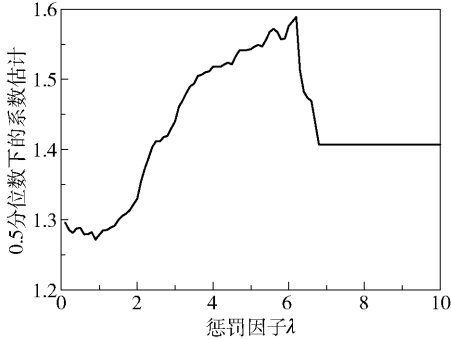


图 1 0.5 分位数下系数估计与 λ 的关系

Fig. 1 The relationship between coefficient estimation and λ in the 0.5 quantile

为了验证本文提出的数据删除模型参数的一步近似估计的准确性,表 1 给出了不同分位数下参数近似估计和直接计算的参数估计的对比情况,其中实验 1 ~ 5 分别为删去不同数据点所做的 5 次验证,评价指标设定为参数估计的相对偏差: $\text{bias}(\boldsymbol{\eta}) = \frac{\|\hat{\boldsymbol{\eta}}_0 - \hat{\boldsymbol{\eta}}_1\|_1}{\|\hat{\boldsymbol{\eta}}_0\|_1}$,其中 $\hat{\boldsymbol{\eta}}_0$ 是采用 R 语言 RQPD 包直接

逐一计算出的参数估计,作为真实估计, $\hat{\boldsymbol{\eta}}_1$ 是利用式(7)计算的 CDM 参数的近似估计结果, $\|\mathbf{A}\|_1$ 表示向量 \mathbf{A} 的一范数。可以看出 CDM 参数的近似估计结果与真实结果相接近,在样本量较大时,使用参数的近似估计可以减少计算时间。

表 1 CDM 参数近似估计与真实估计对比

Table 1 Comparison between approximate estimations and true values of CDM parameters

分位数	相对偏差				
	实验 1	实验 2	实验 3	实验 4	实验 5
0.1	6.25×10^{-4}	1.20×10^{-4}	1.65×10^{-5}	1.31×10^{-3}	1.79×10^{-5}
0.3	2.81×10^{-3}	2.58×10^{-3}	2.46×10^{-3}	2.03×10^{-3}	1.40×10^{-3}
0.5	6.91×10^{-3}	4.17×10^{-3}	8.51×10^{-6}	3.37×10^{-3}	1.94×10^{-3}
0.7	1.39×10^{-2}	2.51×10^{-3}	2.35×10^{-5}	1.34×10^{-3}	2.60×10^{-4}
0.9	1.61×10^{-3}	3.77×10^{-3}	1.62×10^{-5}	3.96×10^{-3}	5.97×10^{-4}

3.1 针对单个点的影响分析

首先对单个数据中是否是强影响点或异常值点进行验证,实验可以分为两部分:针对单个数据是否为强影响点的诊断,和是否为异常值点的诊断。针对由模型(16)所生成的模拟数据,对第 30 个点和第 50 个点增加扰动,分别为 $\mathbf{y}_{30} = \mathbf{y}_{30} + 6$ 和 $\mathbf{y}_{50} = \mathbf{y}_{50} - 6$,然后采用本文提出的基于损失函数的距离和 Cook 距离来检测强影响点。强影响点检验结果如图 2 所示。图 2(a) 从上到下分别为分位数等于 0.1、0.3、0.5、0.7、0.9 时的损失函数距离。可以看出损失函数距离可以很好地检验出第 30 个和第 50 个点为强影响点,并且在不同分位数下不同位置的强影响点的表现不同:低分位数利于检验位于数据下方的强影响点,高分位数则对位于数据上方的强影响点敏感。图 2(b) 是 Cook 距离的诊断结果,尽管结果的波动幅度与图 2(a) 不同,但是结论相同,可以相互验证,避免出现方法不准确的问题。

基于相同的模拟数据,对其进行异常值点检验,结果展示在图 3 中。图 3 中虚线上方是 Wald 统计量的拒绝域,可以明显看出,图 3(a) 中第 30 个和第

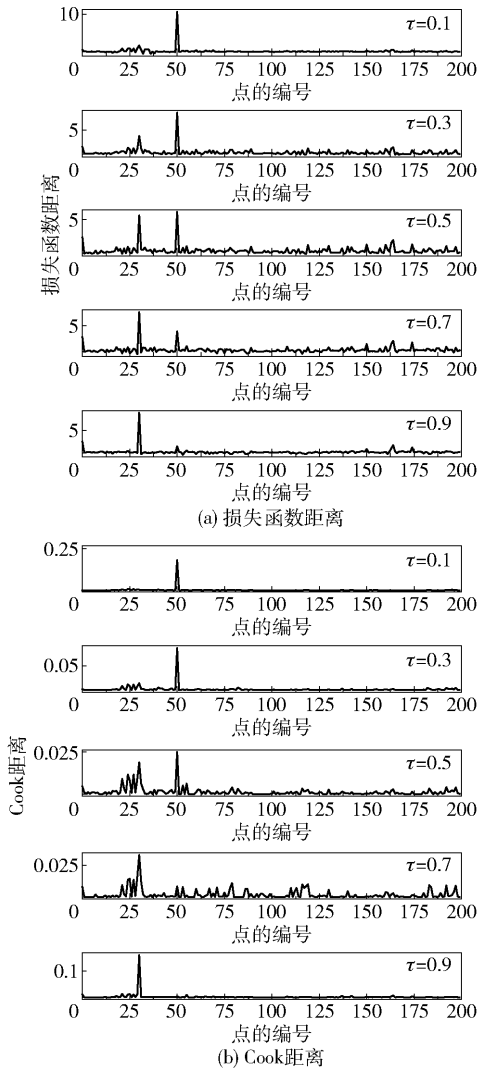


图 2 单个强影响点的影响分析

Fig. 2 Influence analysis of a single influential observation

50 个点位于拒绝域,图 3(b)中第 50 个点是异常值点,同时结合图 2 结果可知,异常值点与强影响点的检验结果相同,但是对比图 3(a)和(b)可以看出,低分位数下的结果对位于数据上方的异常值点更敏感。

3.2 针对同一个体一组点的影响分析

对属于同一个体的一组数据是否为强影响点或异常值点的情形进行检验,类似于 3.1 节实验,分为针对强影响点的诊断和针对异常值点的诊断。对初始模拟数据的第 3 组数据和第 14 组数据进行扰动： $y_{3.} = y_{3.} + 6$ 和 $y_{14.} = y_{14.} - 6$ 。篇幅所限,图 4 仅展示一组强影响点的损失函数距离的检验结果。从图 4 可以看出第 3 组和第 14 组数据的损失函数距离明显大于其他组,因此为强影响点,说明即使是在同一个体的一组数据都是强影响点的情况下,本文所提方法也具有良好的检验能力。

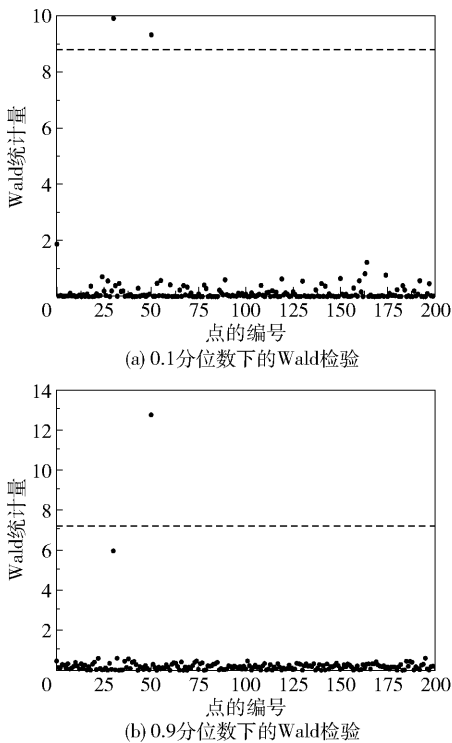


图 3 单个异常值点的影响分析

Fig. 3 Influence analysis of a single outlier

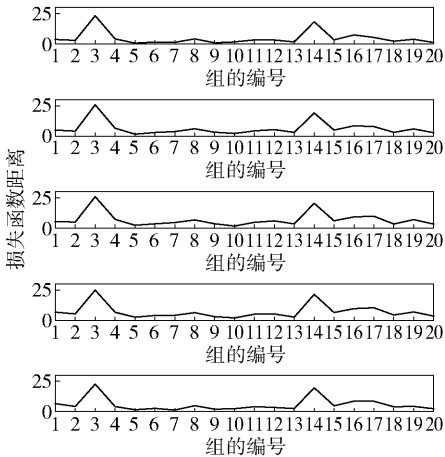


图 4 一组强影响点的损失函数距离

Fig. 4 Distance of loss function for a group of influence observations

4 实验验证与分析

实验数据共包含 216 个观测值,其中辐解剂量有 1、2、4、6、8、10(10^3 Gy)共 6 个水平,硝酸浓度也有 0.2、0.4、0.6、0.8、1.0、1.5 mol/L 共 6 个水平,HSC 浓度为 0.1、0.2、0.4、0.6、0.8、1.0 mol/L 共 6 个水平。根据 1.1 节数据描述的结果,以辐解剂量作为解释变量 X_{ij} ,硝酸浓度和 HSC 浓度的联合影响作为随机效应 μ_i ,共有 36 个取值。响应变量 c_{ij} 表示

为不同实验条件下生成的铵根离子浓度 (mmol/L)。为了对响应变量进行影响因素分析,可建立如下线性混合效应模型

$$c_{ij} = \beta_0 + \beta_1 X_{ij} + \mu_i + \varepsilon_{ij} \quad (17)$$

式中, $i=1, \dots, 36, j=1, \dots, 6$ 。基于模型 (17) 在 0.5 分位数下的参数估计作误差散点图,如图 5 所示,可以看出误差在零点上下波动,但是在上侧波动小而密集,在下侧波动大而疏松,可以明显得出数据总体上既不是对称分布,也不是正态分布,因此选择采用本文方法对其进行影响分析。

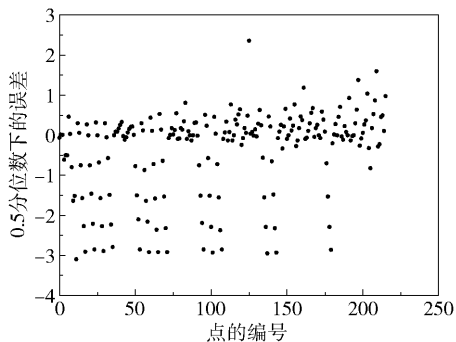
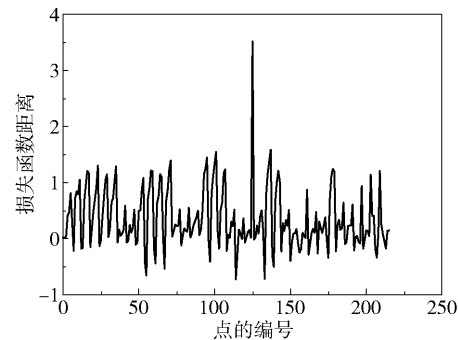


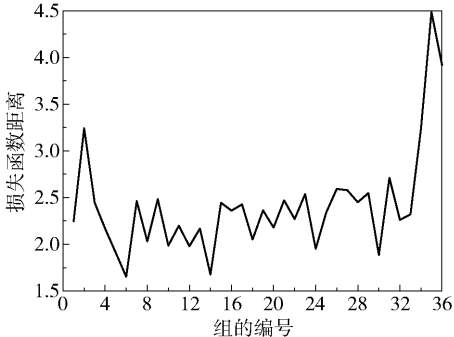
图 5 0.5 分位数下的误差散点图

Fig. 5 Scatter plot of error in the 0.5 quantile

由于在低分位数下检验结果不存在明显的强影响点,所以图 6 仅展示了 $\tau=0.9$ 时的检验结果。从



(a) 单个强影响点的损失函数距离



(b) 一组强影响点的损失函数距离

图 6 实证数据影响分析结果 ($\tau=0.9$)

Fig. 6 Influence analysis results for real data ($\tau=0.9$)

图 6 可以看出,在 0.9 的分位数下,第 125 个数据点和第 35 组数据相较于其他数据明显突出,结合图 5 可知二者均是位于数据上方的强影响点。

5 结束语

本文针对服从非对称、非正态,尤其是重尾或轻尾分布的纵向数据的影响分析问题,在中等样本下利用线性混合模型的分位数估计,分别构建检测强影响点的基于损失函数的距离和 Cook 距离,以及识别异常值点的 Wald 统计量。此外,为了减少计算量,给出数据删除模型参数的一步近似估计,并且利用 Bootstrap 方法求得统计量的拒绝域。模拟结果表明,本文所提的诊断度量和诊断统计量都可以很好地检验出数据中的强影响点和异常值点,实证分析结果也进一步证明了本文方法的实用性。

参考文献:

- [1] DIGGLE P J, HEAGERTY P, LIANG K Y, et al. Analysis of longitudinal data [M]. 2nd ed. Oxford: Oxford University Press, 2002.
- [2] PAN J X, FEI Y, FOSTER P J. Case-deletion diagnostics for linear mixed models [J]. Technometrics, 2014, 56(3): 269–281.
- [3] XU L, LEE S Y, POON W Y. Deletion measures for generalized linear mixed effects models [J]. Computational Statistics & Data Analysis, 2006, 51: 1131–1146.
- [4] PINHO L G B, NOBRE J S, SINGER J M. Cook's distance for generalized linear mixed models [J]. Computational Statistics & Data Analysis, 2015, 82: 126–136.
- [5] 孙慧慧, 林金官. 基于 M 估计的线性混合模型的局部影响分析 [J]. 应用概率统计, 2012, 28(2): 217–219, 221–223.
- [6] SUN H H, LIN J G. Local influence analysis of mixed effect linear models based on M-estimation [J]. Chinese Journal of Applied Probability, 2012, 28(2): 217–219, 221–223. (in Chinese)
- [7] TANG N S, DUAN X D. Bayesian influence analysis of generalized partial linear mixed models for longitudinal data [J]. Journal of Multivariate Analysis, 2014, 126: 86–99.
- [8] KOENKER R. Quantile regression [M]. Cambridge: Cambridge University Press, 2005.
- [9] GERACI M, BOTTAI M. Quantile regression for longitudinal data using the asymmetric Laplace distribution [J]. Biostatistics, 2007, 8(1): 140–154.

- [9] NARULA S C, WELLINGTON J F. Sensitivity analysis for predictor variables in the MSAE regression[J]. Computational Statistics & Data Analysis, 2002, 40: 355–373.
- [10] COOK R D, WEISBERG S. Residuals and influence in regression[M]//COX D R, HINKLEY D V. Monographs on statistics and applied probability. New York: Chapman and Hall, 1982.
- [11] HUNTER D R, LANGE K. Quantile regression via an MM algorithm [J]. Journal of Computational and Graphical Statistics, 2000, 9(1): 60–77.
- [12] 雍龙泉. 绝对值函数的一致光滑逼近函数[J]. 数学的实践与认识, 2015, 45(20): 250–255.
YONG L Q. Uniform smooth approximating functions for absolute value function[J]. Mathematics in Practice and Theory, 2015, 45(20): 250–255. (in Chinese)
- [13] 韦博成, 鲁国斌, 史建清. 统计诊断引论[M]. 南京: 东南大学出版社, 1991.
- WEI B C, LU G B, SHI J Q. Introduction to statistical diagnosis [M]. Nanjing: Southeast University Press, 1991. (in Chinese)
- [14] 曾林蕊, 朱仲义, 茆诗松. 半参数广义线性模型的影响分析与异常点检验[J]. 高校应用数学学报 A 辑, 2004, 19(3): 323–332.
ZENG L R, ZHU Z Y, MAO S S. Influence analysis and outlier tests for semiparametric generalized linear model [J]. Applied Mathematics—A Journal of Chinese Universities, Series A, 2004, 19(3): 323–332. (in Chinese)
- [15] 魏艳华, 王丙参, 邢永忠. 基于 Bootstrap 方法的回归分析的比较[J]. 统计与决策, 2016(3): 77–79.
WEI Y H, WANG B C, XING Y Z. Comparison of regression analysis based on Bootstrap method[J]. Statistics & Decision, 2016(3): 77–79. (in Chinese)

Influence analysis for quantile estimation of a class of linear mixed models

LI HongKai¹ XIAO SongTao² OUYANG YingGen² LI ZhiQiang^{1*}

(1. College of Mathematics and Science, Beijing University of Chemical Technology, Beijing 100029;

2. Institute of Radiochemistry, China Institute of Atomic Energy, Beijing 102413, China)

Abstract: How to detect outliers in longitudinal data when the model error obeys an asymmetric or non-normal distribution, especially under the condition of a heavy-tailed distribution or a contaminated distribution, is an important issue in data analysis. In order to overcome the influence of model errors with a non-normal distribution, a robust quantile method is adopted to estimate the parameters of the linear mixed model, and the data deletion model and the mean shift model used as a basis to construct the diagnostic metrics of influence observations and the test statistics of outliers. In order to reduce computation when testing the influential observations, a one-step approximate estimation of the parameters of the case deletion model is employed, and the distance is estimated based on the loss function and Cook's distance. To identify outliers, we establish Wald statistics, and then the equivalence of the parameter estimations of the case deletion model and the mean shift outlier model is confirmed and Bootstrap sampling is used to obtain the rejection region. A simulation study shows that the diagnostic measures and diagnostic statistics can accurately test the influential observations and outliers. Finally, we apply the method to study the influence analysis of longitudinal data in chemical experiments.

Key words: linear mixed model; quantile estimation; influential observations; outliers; Bootstrap sampling

(责任编辑:吴万玲)