

DOI :10.3876/j.issn.1000-1980.2009.04.026

基于 SA-MCMC 算法的非线性测量误差模型数据删除影响

徐 亮¹ 李 艳¹ 林金官¹ 夏乐天²

(1. 东南大学数学系 ,江苏 南京 210096 ;海海大学理学院 ,江苏 南京 210098)

摘要 :为了研究非线性测量误差模型强影响点的识别问题 ,首先将非线性测量误差模型中存在误差的不可观测的数据当作缺失数据 ,利用 SA-MCMC 算法求得模型参数的最大似然估计 ,然后用 Q 函数代替可观测数据的对数似然函数进行影响分析 ,得到了建立在 Q 函数基础上的广义 Cook 距离及其一步近似 ,最后通过算例说明了诊断统计量的有效性。
关键词 :缺失数据 ;MH 算法 ;SA-MCMC 算法 ; Q 函数 ;Cook 距离 ;数据删除
中图分类号 :O212.2 文献标识码 :A 文章编号 :1000-1980(2009)04-0492-03

非线性测量误差模型在生物育种、水利工程、工业质量控制和计量经济等领域得到了广泛的应用^[1-5]。统计诊断是近几十年来统计学迅速发展起来的一个新的分支学科。自 Cook^[6-7]提出统计模型诊断问题以来 ,模型诊断研究取得了长足的发展。Zhu 等^[8-10]对模型的诊断和分析提出了新的方法和诊断统计量。宗序平等^[11]利用截面似然的思想对非线性测量误差模型数据删除影响进行了分析。模型的诊断方法和估计方法是密切相关的。非线性测量误差模型的估计 ,通常采用 EM 算法^[12]或 SA-MCMC 算法^[13] ,其本质是先将测量误差视为假想的缺失数据 ,然后再把估计问题纳入带缺失数据的统计分析框架中。Zhu 等^[8]对带缺失数据的统计模型提出了基于 Q 函数的一般的数据删除影响分析方法。但是 ,对于测量误差模型 ,尤其是非线性测量误差模型 ,建立在 Q 函数基础上的 Cook 距离等诊断统计量还未见文献报道。为此 ,本文将用此方法对非线性测量误差模型进行数据删除影响分析。

1 非线性测量误差模型和 SA-MCMC 估计方法

根据文献 [14] 的定义 ,非线性测量误差模型可以写为

$$\begin{cases} Y = f(Z, \beta) + \varepsilon & \varepsilon \sim N(0, \sigma^2 I_n) \\ X = Z + \delta & \delta \sim N(0, I_n \otimes \Lambda) \end{cases}$$

其中 $Y = (y_1, \dots, y_n)^T$ $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$

式中 : Y —— n 维的可观测向量 ; X —— $n \times p$ 维的可观测矩阵 ; Z —— $n \times p$ 维的不可观测矩阵 ,只与 X 相差一个随机矩阵 δ ; $f(\cdot, \cdot)$ ——已知关于参数的非线性函数 ; I_n ——单位矩阵 ; σ^2 ——未知参数 ; β —— $p \times 1$ 未知参数 ; Λ ——已知正定矩阵 ; ε ——随机误差矩阵。这里 ε 和 δ 相互独立 ,并假定 Z 的第 i 行 z_i 是服从均值为 μ 、协方差阵为 Σ 的多元正态分布的 $N(\mu, \Sigma)$ 。如果用 x_i 和 z_i 分别表示 X 和 Z 的第 i 行的向量 , $\theta = (\beta^T, \sigma^2)^T$ 表示未知参数 (X, Y, Z) 为完全数据集 ,那么有关 (X, Y) 可观测数据的似然函数可以表示为

$$p_o \propto \prod_{i=1}^n \int p(y_i | x_i, z_i) \exp \left[-\frac{1}{2} (x_i - z_i)^T \Lambda^{-1} (x_i - z_i) - \frac{1}{2} (z_i - \mu)^T \Sigma^{-1} (z_i - \mu) \right] dz_i \tag{1}$$

其中 $p(y_i | x_i, z_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{e_i^2}{2\sigma^2} \right\}$ $e_i = y_i - f(z_i, \beta)$

记 $\pi(z_i | \theta, x_i, y_i)$ 为关于缺失数据 z_i 在已知参数和可观测数据下的条件分布 ,则有

收稿日期 :2008-09-09
基金项目 :江苏省自然科学基金(BK2008284) ;东南大学校基金(9207011430)
作者简介 :徐亮(1965—) ,男 ,江苏泰兴人 ,副教授 ,博士 ,主要从事统计诊断和统计计算研究。

$$\pi(z_i|\theta, x_i, y_i) \propto \exp\left\{-\frac{e_i^2}{2\sigma^2} - \frac{1}{2}(x_i - z_i)^T \Lambda^{-1}(x_i - z_i) - \frac{1}{2}(z_i - \mu)^T \Sigma^{-1}(z_i - \mu)\right\}$$

从式 (1) 可以看出, 在可观测数据的似然函数中, 积分很难求出. 因此, 很难通过它得到参数的估计, 同时还很难得到有关模型的诊断统计量, 因为那些统计量会牵涉到对数似然函数的一阶和二阶导数. 利用数据扩张的思想, 将 Z 看做缺失数据, 那么第 i 组完全数据的对数似然可以写为

$$l_{\theta}(\theta|x_i, y_i, z_i) \propto -\frac{1}{2}\ln(\sigma^2|\Lambda\Sigma|) - \frac{e_i^2}{2\sigma^2} - \frac{1}{2}(x_i - z_i)^T \Lambda^{-1}(x_i - z_i) - \frac{1}{2}(z_i - \mu)^T \Sigma^{-1}(z_i - \mu) \quad (2)$$

完全对数似然为

$$L_{\theta}(\theta|X, Y, Z) = \sum_{i=1}^n l_{\theta}(\theta|x_i, y_i, z_i) \quad (3)$$

采用 SA-MCMC 算法^[13]求参数的估计. 通过 MH 算法从给定观测数据关于缺失数据的条件分布中抽取样本, 然后再利用该样本通过蒙特卡罗方法计算对数似然函数的一阶和二阶导数阵的条件期望, 从而通过迭代可得到参数的最大似然估计. 记非线性测量误差模型参数 θ 的最大似然估计为 $\hat{\theta}$.

2 非线性测量误差模型的数据删除影响

模型的诊断分析方法主要有数据删除分析法和局部影响分析法 2 种. 本文讨论数据删除分析问题. 数据删除分析就是通过删除数据来研究数据点的作用. 删除第 i 个数据点后对应的量用带有下标 $[i]$ 的量来表示. 比如, 可以用 $Y_{[i]}$ 和 $X_{[i]}$ 表示删除第 i 个数据点后对应的可观测数据. 数据删除分析方法^[6]是基于观测数据的对数似然函数 $L_{\theta}(\theta|X, Y)$ 来分析删除数据对估计的影响的. 因为 $L_{\theta}(\theta|X, Y)$ 涉及高维积分问题, 直接采用该方法进行数据分析是困难的, 其相应的诊断统计也难以进行. 因此, 受 Zhu 等^[8]的启发, 基于 Q 函数

$$Q(\theta|\hat{\theta}) = E\{L_{\theta}(\theta|X, Y, Z)|X, Y, Z, \hat{\theta}\}$$

定义有关 $\hat{\theta}$ 和 $\hat{\theta}_{[i]}$ 的距离的度量

$$D_i = (\hat{\theta} - \hat{\theta}_{[i]})^T \{-\ddot{Q}(\hat{\theta})\}(\hat{\theta} - \hat{\theta}_{[i]}) \quad (4)$$

其中, 数学期望基于条件分布 $\pi(z_i|\theta, x_i, y_i)$ 进行计算; $\ddot{Q}(\hat{\theta}) = \partial^2 Q(\theta|\hat{\theta})/\partial\theta\partial\theta^T|_{\theta=\hat{\theta}}$. 该诊断统计量与 Cook 距离^[6]形式非常相似, 也称为广义 Cook 距离. D_i 越大, 表明第 i 个数据点的影响就越大.

上述式子, 可以通过 SA-MCMC 算法求得 $\hat{\theta}$ 和 $\hat{\theta}_{[i]}$. 但是, 对于 n 组数据的删除问题, 需要进行 n 次估计的计算, n 很大时, 计算量会很大. 因此, 可以采用一步近似 $\hat{\theta}_{[i]}^1$ 来代替 $\hat{\theta}_{[i]}$ 进行计算:

$$\hat{\theta}_{[i]}^1 = \hat{\theta} + \{-\ddot{Q}(\hat{\theta})\}^{-1}\ddot{Q}_{[i]}(\hat{\theta}|\hat{\theta}) \quad (5)$$

将式 (5) 代入式 (4) 得到广义 Cook 距离的一步近似

$$D_i^1 = (\hat{\theta} - \hat{\theta}_{[i]}^1)^T \{-\ddot{Q}(\hat{\theta})\}(\hat{\theta} - \hat{\theta}_{[i]}^1) = \hat{Q}_{[i]}^T \{\hat{\theta}|\hat{\theta}\} \{-\ddot{Q}(\hat{\theta})\}^{-1} \hat{Q}_{[i]}(\hat{\theta}|\hat{\theta}) \quad (6)$$

式 (5) 的近似精度不高. 对模型的影响分析而言, 当 $\hat{\theta} - \hat{\theta}_{[i]}^1$ 足够大时, 不必对 $\hat{\theta}_{[i]}^1$ 进行精确计算, 因此一步近似式 (6) 是可行的. 以下实例主要利用 D_i^1 进行数据删除影响分析.

3 算 例

Fuller^[14]根据某天然气可能储存地观测实验得到的超音速声波传播速度 y 和氮气体浓盐水饱和的百分比 x 共 12 组数据提出了非线性测量误差模型

$$y_i = \beta_0 + \beta_1\{\exp(\beta_2 z_i) - 1\}^2 + e_i$$
$$x_i = z_i + u_i \quad (i = 1, 2, \dots, 12)$$

其中 e_i 和 u_i 相互独立, 都服从 $N(0, \sigma^2)$. 利用 SA-MCMC 算法可得到未知参数的最大似然估计 $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2) = (1270.0, -52.0, -0.101, 2.541)$.

对这 12 组数据进行数据删除影响分析. 根据前面给出的公式计算得到广义 Cook 距离 D_i^1 ($i = 1, \dots, 12$), 如图 1 所

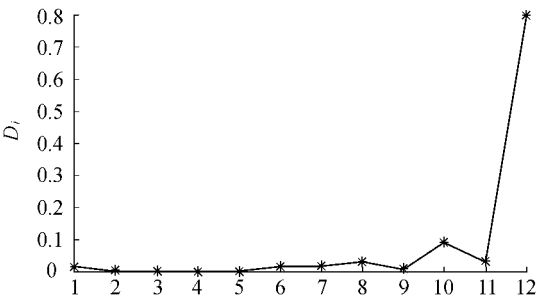


图 1 数据删除诊断量 D_i 散点图

Fig. 1 Distribution of case deletion of diagnostic measures D_i

示.从图1可以看出,第12组数据的广义Cook距离比其他11组数据的广义Cook距离大很多.这一结果表明,删除该组数据对模型参数估计的影响最大.因此,可以认为第12组数据点为强影响点.该组数据可能是一重要数据,也可能是由于实际记录时出现误差产生的,具体为哪种情况,需根据实际情况确定.

4 结 语

统计诊断是统计建模和数据分析的重要组成部分.在非线形测量误差模型分析中,由于观测数据似然函数的计算比较复杂,直接采用Cook方法比较困难.为此,本文利用Zhu等^[8]提出的数据删除影响分析方法,在 Q 函数的基础上对非线性测量误差模型进行了影响分析,导出了当前模型下的数据删除诊断量——广义Cook距离.实例表明,本文得到的广义Cook距离能有效地识别数据中的强影响点.

参考文献:

- [1] PRENTICE R L. Covariate measurement errors and parameter estimation in a failure time regression model[J]. *Biometrika*, 1982, 69(2): 331-342.
- [2] CARROLL R J, SPIEGELMAN C, LAN K K, et al. On error-in-variables for binary regression models[J]. *Biometrika*, 1984, 71(1): 19-76.
- [3] ARMSTRONG B. Measurement error in the generalized linear model[J]. *Comm Statist B*, 1985, 14(3): 529-544.
- [4] 贾仁甫, 陈守伦, 丁鑫. 需水量的线性测量误差模型[J]. *河海大学学报:自然科学版*, 2008, 36(4): 488-490.
- [5] STEFANSKI L A, CARROLL R J. Covariate measurement error in logistic regression[J]. *Annals of Statistics*, 1985, 13(4): 1335-1351.
- [6] COOK R D. Detection of influential observations in linear regression[J]. *Technometrics*, 1977, 19(1): 15-18.
- [7] COOK R D. Assessment of local influence[J]. *J Roy Statist Soc Ser B*, 1986, 48(1): 133-169.
- [8] ZHU H T, LEE S Y, WEI B C, et al. Case-delete measures for models with incomplete data[J]. *Biometrika*, 2001, 88(3): 727-737.
- [9] XU L, LEE S Y, POON W Y. Deletion measures for generalized linear mixed effects models[J]. *Comput Statist Data Anal*, 2006, 51(2): 1131-1146.
- [10] XU L, LEE S Y, POON W Y. Local influence analysis for ranking data[J]. *British Journal of Mathematical and Statistical Psychology*, 2008, 61(2): 133-161.
- [11] 宗序平, 孟国明. 非线性测量误差模型的影响分析[J]. *应用概率统计*, 2003, 19(1): 31-39.
- [12] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm (with discussion) [J]. *J Roy Statist Soc Ser B*, 1977, 39(1): 1-38.
- [13] GU M G, KONG F H. A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems[J]. *Proc Nat Acad Sci USA*, 1998, 95: 7205-7214.
- [14] FULLER W A. Measurement error models[M]. New York: Wiley, 1987.

Deletion influence of nonlinear measurement error models based on SA-MCMC algorithms

XU Liang¹, LI Yan¹, LIN Jin-guan¹, Xia Le-tian²

(1. Department of Mathematics, Southeast University, Nanjing 210096, China;

2. College of Science, Hohai University, Nanjing 210098, China)

Abstract: The deletion measures for nonlinear measurement error models were studied. The unobservable measurement errors were treated as the missing data. The maximum likelihood estimates were obtained using the stochastic approximation algorithms with the Markov Chain Monte Carlo (SA-MCMC) method. The logarithmic likelihood function of the observable data was replaced by the Q function. The Cook's distance and its one-step approximation were derived based on the Q function. The effectiveness of the diagnostic measures was validated with a real example.

Key words: missing data; MH algorithm; SA-MCMC algorithm; Q function; Cook's distance; case deletion