

DOI :10.3876/j.issn.1000-1980.2011.01.020

非线性分位点回归方法在大坝安全监测中的应用

龚晓雯^{1 2 3} 范 磊^{1 2 3}

(1.河海大学水文水资源与水利工程科学国家重点实验室 江苏 南京 210098 ;
2.河海大学水资源高效利用与工程安全国家工程研究中心 江苏 南京 210098 3.河海大学水利水电学院 江苏 南京 210098)

摘要 :基于分位点回归模型比最小二乘回归模型具有更强的统计分析能力 ,将参数估计的 MM 算法和统计诊断中的影响度量 MM 距离运用于大坝水平位移的统计模型 ,并给出了应用实例 .结果表明 ,分位点回归方法受异常点影响较小 ,训练及预报结果优于最小二乘法 ,将其应用于大坝水平位移安全监测 ,可提高预报模型的预测精度 ,能更好地对大坝运行性态进行分析 .
关键词 :非线性分位点回归 ;MM 算法 ;统计诊断 ;MM 距离 ;大坝安全监测
中图分类号 :TV698 文献标志码 :A 文章编号 :1000-1980(2011)01-0099-05

在分析大坝位移监测资料并建立安全监控模型时 ,一般采用最小二乘回归法建立统计模型^[1] .最小二乘回归法在理论和方法上均较为成熟 ,被广泛应用于曲线拟合、函数逼近、状态估计、时间序列建模和系统辨识等实际问题 .然而 ,最小二乘回归法对响应变量中的强影响点或异常点比较敏感 ,强影响点或异常点会使得统计推断的可靠性降低 .已有许多诊断方法可用来改进基本的最小二乘回归法 ,提高其对强影响点和异常点的识别和处理能力 .但改进的最小二乘回归法在处理异常点时具有较多的主观因素 ,剔除部分异常点会影响回归的结果 ,而且有些异常点在某些方面能反映特殊信息 ,不能随便剔除 .因此 ,有必要寻找一种能减少模型受异常点影响 ,在不剔除异常点的情况下可以得出较好估计结果的更稳健的估计方法 .分位点回归方法不仅具有对响应变量中强影响点或异常点稳健的优点 ,而且具有比最小二乘回归法更强大的统计分析能力 .因此 ,本文运用非线性分位点回归方法分析大坝位移监测资料并建立安全监控预报模型 .

1 分位点回归模型

定义 设任一实值随机变量 Y 的分布函数为 $F(y) = P(Y \leq y)$,对任意的 $0 < \tau < 1$,称 $y_\tau = F^{-1}(\tau) = \inf\{y : F(y) \geq \tau\}$ 为 $F(\cdot)$ 的 τ 分位点 .

由定义可知 ,总体分布 $F(\cdot)$ 的 $\tau(0 < \tau < 1)$ 分位点存在且唯一 .中位数 $\tau = \frac{1}{2}$,在所有的分位点中起着中心控制作用 ,在工程实际中具有重要的意义 .

设实值随机变量 Y 的分布函数为 $F(\cdot)$,当损失函数为 $\rho_\tau(\cdot) = \tau - I\{\tau < 0\}$ 或 $\rho_\tau(\cdot) = \frac{|t| + (2\tau - 1)t}{2}$ ($I\{\cdot\}$ 为示性函数)时 ,问题即变为求解 $E[\rho_\tau(Y - c)]$ 最小时的 c .

$$E[\rho_\tau(Y - c)] = \tau \int_c^{+\infty} (y - c) dF(\cdot) + (\tau - 1) \int_{-\infty}^c (y - c) dF(\cdot) \tag{1}$$

$F(c) = \tau$ 为方程 (1) 的解 .总体分布的 τ 分位点即由上面的优化问题导出 .

在不知道随机变量 Y 具体分布的情况下 ,要得出总体均值或总体分位数 ,必须进行参数估计 .最小二乘

回归法用样本均值 $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ 来估计总体均值 $\mu = E(\mu)$, 同样, 本文用样本 τ 分位点 $\hat{Y}_\tau$ 来估计总体 τ 分位点.

由于样本 τ 分位点是优化问题 $\min \sum_{i=1}^n \rho_\tau(Y_i - c)$ 的解, 当 τ 分位点与某个变量 x 有关, 即 $y_\tau = y_\tau(y|x) = f(x, \beta)$ 时, 称 $Q_\tau(y|x) \triangleq y_\tau = f(x, \beta)$ 为条件回归分位点函数, x 是与 y 相关的 n 维已知协变量, β 是 m 维未知参数. 因此, 可以通过优化问题 $\min_{\beta \in R^m} \sum_{i=1}^n \rho_\tau(y_i - f(x_i, \beta))$ 的求解来估计未知参数 β . 一般的非线性回归模型为

$$y_i = f(x_i, \beta) + \varepsilon_i \quad (2)$$

若 β^* 使 $S_\tau(\beta) = \sum_{i=1}^n \rho_\tau(y_i - f(x_i, \beta))$ 最小, 则 β^* 为未知参数 β 的 τ 分位点估计, 记为 $\hat{\beta}_\tau$. 函数 $S_\tau(\beta)$ 为目标函数. 使用回归分位点估计的回归模型(2)即为非线性回归模型.

2 MM 算法

由于分位点回归模型目标函数的复杂性, 本文采用计算非线性分位点的 MM (majorize-minimize) 迭代算法^[2], 首先寻找原目标函数的优化函数 (majorize), 然后将优化函数最小化 (minimize)^[3], 将一个较难的优化问题转化为一组较简单的优化问题.

假设要最小化的目标函数为 $U(\theta)$, 在寻找最小值点的过程中, 首先需构造一个优化函数 $Q(\theta|\theta^k)$, 使其满足

$$Q(\theta|\theta^k) = U(\theta^k) \quad Q(\theta|\theta^k) \geq U(\theta) \quad (3)$$

式中: θ^k ——当前的迭代值; $Q(\theta|\theta^k)$ —— $U(\theta)$ 在 θ^k 处的优化函数. 然后, 搜寻 θ^{k+1} 使得 $Q(\theta^{k+1}|\theta^k)$ 最小. 由 θ^{k+1} 以及式(3)得出

$$U(\theta^{k+1}) \leq Q(\theta^{k+1}|\theta^k) \leq Q(\theta^k|\theta^k) = U(\theta^k)$$

对于非线性分位点回归的目标函数 $S_\tau(\beta)$, $S_\tau^\varepsilon(\beta) = \sum_{i=1}^n \rho_\tau^\varepsilon(r_i)$ 为构造的一个处处可导的近似函数;

$Q_\tau^\varepsilon(\beta|\beta^{(k)}) = \sum_{i=1}^n \zeta_\tau^\varepsilon(r_{ij}|r_{ij}^{(k)})$ 为 $S_\tau^\varepsilon(\beta)$ 在 $\beta^{(k)}$ 处的优化函数. 其中, ε 是一个很小的正常数, $\rho_\tau^\varepsilon(t) = \rho_\tau(t) - \frac{\varepsilon}{2} \ln(\varepsilon + |t|)$, $t^{(k)}$ 代表某个 $r_{ij}^{(k)}$, $S_\tau^\varepsilon(t|t^{(k)}) = \frac{1}{4} \left(\frac{t^2}{\varepsilon + |t^{(k)}|} + (4\tau - 2)t + c \right)$, c 是使 $S_\tau^\varepsilon(t^{(k)}|t^{(k)}) = \rho_\tau^\varepsilon(t^{(k)})$ 的常数, $r_{ij}^{(k)} = r_{ij}(\beta^k) = y_{ij} - f(x_{ij}, \beta^{(k)})$, $j = 1, 2, \dots, n_i$, $i = 1, 2, \dots, N$.

在实际应用中, 为简化计算, 通常仅对 $Q_\tau^\varepsilon(\beta|\beta^{(k)})$ 用高斯-牛顿迭代法进行下降搜索.

$$\begin{aligned} \frac{\partial^2 Q_\tau^\varepsilon(\beta|\beta^{(k)})}{\partial \beta \partial \beta^T} &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^{n_i} \left[\frac{r_{ij}}{\varepsilon + |r_{ij}|} \frac{\partial r_{ij}(\beta)}{\partial \beta} \left(\frac{\partial r_{ij}(\beta)}{\partial \beta} \right)^T + \frac{r_{ij}}{\varepsilon + 2\tau - 1} \frac{\partial^2 r_{ij}(\beta)}{\partial \beta \partial \beta^T} \right] \approx \\ &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^{n_i} \frac{r_{ij}}{\varepsilon + |r_{ij}|} \frac{\partial r_{ij}(\beta)}{\partial \beta} \left(\frac{\partial r_{ij}(\beta)}{\partial \beta} \right)^T = \\ &= \frac{1}{2} V^T(\beta) W_\varepsilon(\beta^{(k)}) V(\beta) \quad (\text{当 } f(x_{ij}, \beta) = x_{ij}^T \beta \text{ 时}) \end{aligned} \quad (4)$$

$$\frac{\partial Q_\tau^\varepsilon(\beta|\beta^{(k)})}{\partial \beta} = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^{n_i} \left(\frac{r_{ij}}{\varepsilon + |r_{ij}|} + 2\tau - 1 \right) \frac{\partial r_{ij}(\beta)}{\partial \beta} = \frac{1}{2} V^T(\beta) V_\varepsilon(\beta) \quad (5)$$

其中: $V_\varepsilon(\beta) = \left(1 - 2\tau - \frac{r_{i1}(\beta)}{\varepsilon + |r_{i1}|}, 1 - 2\tau - \frac{r_{i2}(\beta)}{\varepsilon + |r_{i2}|}, \dots, 1 - 2\tau - \frac{r_{in_i}(\beta)}{\varepsilon + |r_{in_i}|} \right)^T$; $W_\varepsilon(\beta^{(k)})$ ——对角元为 $(\varepsilon + |r_{ij}|)^{-1}$ 的 $n \times n$ 对角阵.

由高斯-牛顿迭代法及式(4)(5)可知

$$\Delta_{\epsilon}^{(k)} = -[V^T(\beta^{(k)})W_{\epsilon}(\beta^{(k)})V(\beta^{(k)})]^{-1}V^T(\beta^{(k)})V_{\epsilon}(\beta^{(k)}) \quad \beta^{(k+1)} = \beta^{(k)} + \Delta_{\epsilon}^{(k)}$$

为了保证每次迭代都能减小优化函数 $Q_{\tau}^{\epsilon}(\beta|\beta^{(k)})$, 取 $\beta^{(k+1)} = \beta^{(k)} + \alpha^{(k)}\Delta_{\epsilon}^{(k)}$, $\alpha^{(k)} = \max\{2^{-\gamma} : Q_{\tau}^{\epsilon}(\beta^{(k)} + 2^{-\gamma}\Delta_{\epsilon}^{(k)}|\beta^{(k)}) < Q_{\tau}^{\epsilon}(\beta|\beta^{(k)}), \gamma \in N\}$.

MM 算法实际应用步骤见图 1.

3 非线性分位点回归模型的统计诊断

回归分位点估计是一类基本而重要的稳健估计,受响应变量中异常点的影响较小,但却可能对解释变量中的异常点(即高杠杆值点)敏感^[4].

在分位点回归模型中可以假定误差项服从非对称 Laplace 分布 $ALA_{\tau}(0, \sigma)$ ^[5-9].

文献[10]用一个取正值的函数 $m(\tau(\hat{\beta}), \tau(\hat{\beta}_i))$ 来度量删除第 i 个数据点对参数 β 估计的影响大小. 文献[11]基于某统计模型中的目标函数 $R(\cdot)$ 提出影响度量的拟似然距离 $QLD(\beta) = c |R(\hat{\beta}_{(i)}) - R(\hat{\beta})|$. 文献[12]中用了 Q 函数的方法.

根据以上方法,在分位点回归中,基于 MM 迭代算法中新构造的目标函数 $Q_{\tau}^{\epsilon}(\beta|\hat{\beta}_{\tau})$ 提出一种新的影响度量:MM 距离, $MLQ_{\tau}^{\epsilon} = Q_{\tau}^{\epsilon}(\hat{\beta}_{(i)}|\hat{\beta}_{\tau}) - Q_{\tau}^{\epsilon}(\hat{\beta}_{\tau}|\hat{\beta}_{\tau})$, $\hat{\beta}_{\tau}$ 和 $\hat{\beta}_{(i)}$ 分别是参数 β 的基于 MM 算法的 τ 分位点估计.

4 混凝土坝变形监测统计模型

混凝土坝的位移(δ)可分为水压分量(δ_H)、温度分量(δ_T)和时效分量(δ_{θ})3 个部分,其表达式详见文献[1, 13]. 根据本文所举例子,安全监控模型表达式为

$$\delta = \delta_H + \delta_T + \delta_{\theta} = a_0 + \sum_{i=1}^3 [a_{1i}(H_u^i - H_{u0}^i)] + \sum_{i=1}^3 [a_{2i}(H_d^i - H_{d0}^i)] + \sum_{i=1}^2 \left[b_{1i} \left(\sin \frac{2\pi it}{365} - \sin \frac{2\pi it_0}{365} \right) + b_{2i} \left(\cos \frac{2\pi it}{365} - \cos \frac{2\pi it_0}{365} \right) \right] + c_1(\theta - \theta_0) + c_2(\ln \theta - \ln \theta_0) \quad (6)$$

式中: H_u, H_{u0}, H_d, H_{d0} ——观测日、始测日对应的上、下游水头; a_{1i}, a_{2i} ——水压因子回归系数; b_{1i}, b_{2i} ——温度因子回归系数; c_1, c_2 ——时效因子回归系数; t ——位移监测日到起始监测日的累计天数; t_0 ——建模资料系列第一个监测日到始测日的累计天数; θ ——变形监测日至始测日的累计天数除以 100; θ_0 ——资料系列的第 1 个监测日到始测日的累计天数除以 100; a_0 ——常数项.

5 应 用

某大坝坝顶水平位移的观测采用视准线法,从右到左为 2~40 号坝段,共有 40 个水平位移测点,以 19 号厂房坝段 EX18 测点(埋设桩号 47.51,距坝轴线 14m,埋设高程 74m)水平位移监测资料为例,取 2002—2009 年之间的水平位移监测数据共 2 627 组,其中 2002—2007 年数据为训练用数据,2008—2009 年数据为预测用数据.同时与最小二乘法计算结果进行比较.训练结果见表 1、图 2 和图 3,预报结果见图 4,比较结果见表 2.

表 1 EX18 测点坝顶引张线水平位移不同分位点回归估计的比较

Table 1 Comparison of different quantile regression values of horizontal displacements of EX18 dam crest wire							
系数	a_0	a_{11}	a_{12}	a_{13}	a_{21}	a_{22}	a_{23}
$\tau = 0.5$	2.840 542	- 5.881 33	0.088 370 3	- 0.000 416 3	- 0.073 446 6	0.017 047 3	- 0.000 520 8
$\tau = 0.2$	2.476 292	- 68.791 22	1.150 905 0	- 0.006 385 9	- 0.106 742 0	0.015 504 1	- 0.000 416 1
$\tau = 0.9$	3.823 312	37.564 12	- 0.630 750 5	0.003 544 2	0.121 478 0	0.003 378 5	- 0.000 247 1
系数	b_{11}	b_{12}	b_{21}	b_{22}	c_1	c_2	
$\tau = 0.5$	- 1.457 331	0.076 062 0	1.932 913	0.018 399 2	0.138 399 9	- 0.060 947 44	
$\tau = 0.2$	- 1.166 958	0.103 468 6	2.048 882	0.132 104 0	0.140 673 8	- 0.718 918 20	
$\tau = 0.9$	- 1.721 295	- 0.047 886 8	1.809 810	- 0.051 629 2	0.133 249 4	- 0.697 551 30	

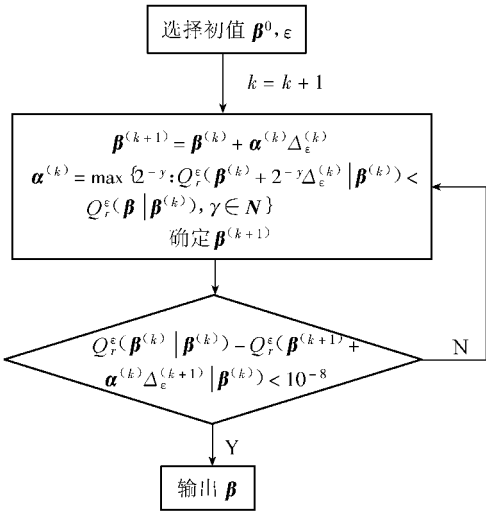


图 1 MM 算法实际应用步骤
Fig. 1 Procedure of MM algorithm

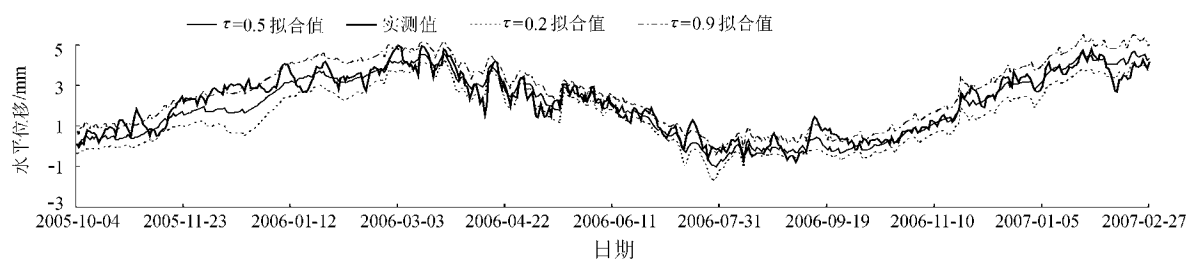


图2 EX18 测点水平位移不同分位点回归拟合值与实测值的比较

Fig. 2 Comparison between fitting and measured quantile regression value of horizontal displacements of EX18 dam crest wire

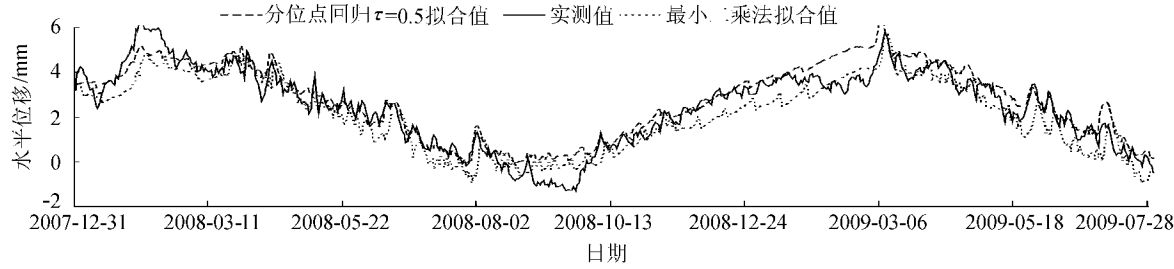


图3 EX18 测点水平位移中分位点回归、最小二乘回归拟合值与实测值的比较

Fig. 3 Comparison among nonlinear quantile regression , least square regression and measured data of horizontal displacements of EX18 dam crest wire

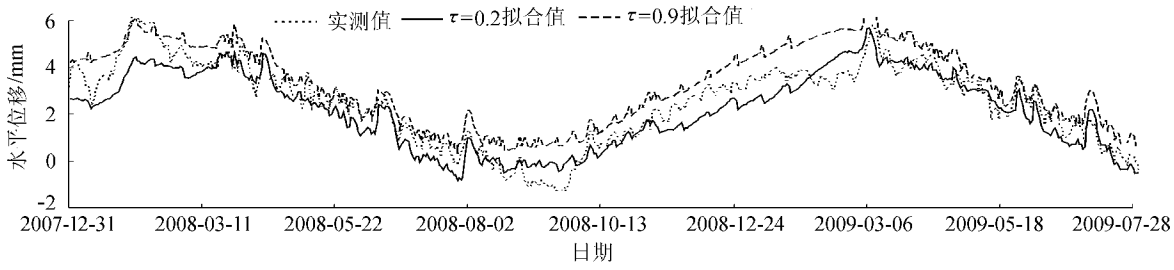


图4 EX18 测点水平位移分位点回归模型预报结果

Fig. 4 Predicted results of quantile regression values of horizontal displacements of EX18 dam crest wire

表2 删除 2007 年 8 月 4 日异常点前、后估计的变化

Table 2 Variation of values before and after deleting abnormal data on Aug. 4 , 2007

方 法		a_0	a_{11}	a_{12}	a_{13}	a_{21}	a_{22}	a_{23}
最小二乘法	删除前	5.850 700	- 20.121 00	0.329 600 0	- 0.001 776 2	0	0.008 886 2	- 0.000 283 79
	删除后	5.839 900	- 21.000 00	0.343 970 0	- 0.001 854 4	0	0.008 991 7	- 0.000 288 09
非线性分位点回归方法 ($\tau=0.5$)	删除前	2.840 542	- 5.881 330	0.088 370 3	- 0.000 416 3	- 0.073 446 6	0.017 047 3	- 0.000 520 80
	删除后	2.847 869	- 5.835 531	0.088 779 4	- 0.000 411 0	- 0.072 217 5	0.017 077 6	- 0.000 520 90
方 法		b_{11}	b_{12}	b_{21}	b_{22}	c_1	c_2	
最小二乘法	删除前	- 1.444 500	1.933 200 0	0.032 292	0	0.141 770 0	- 0.283 070 0	
	删除后	- 1.447 300	1.933 900 0	0.034 411	0	0.141 250 0	- 0.281 070 0	
非线性分位点回归方法 ($\tau=0.5$)	删除前	- 1.457 331	0.076 062 0	1.932 913	0.018 399 2	0.138 399 9	- 0.609 474 4	
	删除后	- 1.457 377	0.076 050 4	1.932 393	0.018 112 3	0.138 393 0	- 0.609 556 4	

由表 1 和图 2 的拟合结果可以看出 , $\tau=0.5$ 分位点(中分位点)与实测值拟合较好 ,且实测资料序列基本在 $\tau=0.2$ 与 $\tau=0.9$ 拟合曲线所包含的区间内 ,说明分位点拟合效果较好.

由图 3 的对比结果和图 4 的预报结果可以看出 ,中分位点的预报结果较最小二乘法预报结果好 ,且实测资料序列在 $\tau=0.2$ 与 $\tau=0.9$ 预报曲线所包含的区间内 ,分位点回归预报结果较好.

在研究数据点是否为强影响点或异常点时 ,最小二乘回归法中识别强影响点或异常点的统计诊断量都与残差密切相关 ,残差大则统计诊断量大 ;分位点回归方法用 MM 距离的数据删除模型(CDM)进行诊断.2

种方法诊断结果均表明 2007 年 8 月 4 日数据点为异常点.删除 2007 年 8 月 4 日数据点后对回归的影响结果见表 2.从表 2 可以看出,最小二乘参数估计所受的影响比分位点回归估计的大,回归分位点估计受响应变量中异常点的影响较小,但随解释变量中的异常点敏感,计算结果显示分位点回归方法具有很好的稳健性,受异常点影响较传统的最小二乘法要小.

6 结 论

分位点回归模型是一种重要的统计模型,比最小二乘回归模型具有更强的统计分析能力.回归分位点估计是一类基本的稳健估计,受数据异常点影响较小,并能给条件分布以更全面的统计描述;当误差项服从重尾分布或其分布受到污染时,它们比最小二乘估计的有效性更高.本文的应用结果表明,将分位点回归方法应用于大坝水平位移安全监测是合理、有效的.

参考文献:

[1] 吴中如.水工建筑物安全监控理论及其应用[M].北京:高等教育出版社,2003:10-62.

[2] HUNTER D R ,LANGE K. Quantile regression via an MM algorithm[J].Journal of Computational and Graphical Statistics ,2000 ,9(1): 60-77.

[3] 徐亮,周影辉,韦博成.基于 MM 算法的 LAD 回归影响分析[J].高校应用数学学报 A 辑,2007,22(1):48-58.(XU Liang , ZHOU Ying-hui ,WEI Bo-cheng. Influence analysis of LAD regression based on MM algorithm[J]. Applied Mathematics a Journal of Chinese Universities ,2007 ,22(1):48-58.(in Chinese))

[4] ADROVER J ,MARONNA R A ,YOUHAI V J. Robust regression quantiles[J]. Journal of Statistical Planning and Inference ,2004 ,122(1 -2):187-202.

[5] 林金官.非线性模型的异方差和变离差检验[D].南京:东南大学,2003.

[6] 孙瑞博.非线性 LAD 回归模型及其统计诊断[D].南京:东南大学,2003.

[7] 韦博成,林金官,吕庆哲.回归模型中异方差或变离差检验问题综述[J].应用概率统计,2003,19(2):210-220.(WEI Bo-cheng ,LING Jin-guan ,LU Qing-zhe. Developments of the tests for heteroscedasticity or varying dispersion in regression models[J]. Chinese Journal of Applied Probability and Statistics ,2003 ,19(2):210-220.(in Chinese))

[8] 王瑞瑞,林金官.基于固定效应的纵向数据分位点回归模型的参数估计及 CDM 和 MSOM 的等价性[J].应用数学,2009,22(3):610-615.(WANG Rui-rui ,LIN Jin-guan. Parameter estimation and equivalence between CDM and MSOM in quantile regression of longitudinal data models with fixed effect[J]. Mathematic Applicata ,2009 ,22(3):610-615.(in Chinese))

[9] BARRETO R A ,HUGHES A W. Under performers and over achievers : a quantile regression analysis of growth[J]. Economic Record , 2004 ,80(248):17-35.

[10] COOK R D ,WEISBERG S. Residuals and influence in regression[M]. New York :Chapman and Hall ,1982.

[11] SUN R B ,WEI B C. On influence assessment for LAD regressior[J]. Statistics & Probability Letters ,2004 ,67(2):97-110.

[12] ZHU H T ,LEE S Y ,WEI B C ,et al. Case-deletion measures for models with incomplete data[J]. Biometrika ,2001 ,88(3):727-737.

[13] 顾冲时,吴中如.大坝与坝基安全监控理论和方法及其应用[M].南京:河海大学出版社,2006:8-21.

Application of nonlinear quantile regression in dam safety monitoring

GONG Xiao-wen^{1,2,3} , FAN Lei^{1,2,3}

- (1. State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering , Hohai University , Nanjing 210098 , China ;
2. National Engineering Research Center of Water Resources Efficient Utilization and Engineering Safety , Hohai University , Nanjing 210098 , China ;
3. College of Water Conservancy and Hydropower , Hohai University , Nanjing 210098 , China)

Abstract : Based on the superiority of the quantile regression model to the least square regression model in statistical analysis , the MM algorithm of parameters and the MM distance for influence measure in the statistical diagnostics were introduced into the statistical model for horizontal displacement of dams . An application example was presented . The results show that the quantile regression method is less influenced by the abnormal data , and its training and predicted results are superior to those of the least square regression method . The proposed method is feasible for the horizontal displacement of dam safety monitoring and may improve the prediction accuracy of relevant models . It can be employed to analyze the operation behaviors of dams .

Key words : nonlinear quantile regression ; MM algorithm ; statistical diagnostics ; MM distance ; dam safety monitoring