

线性回归最小一乘法在水文流量预测中的应用

江秋阳 宋 飙 胡文全

(歙县水利局 安徽 歙县 245200)

摘要:简介线性回归最小一乘解的原理、绝对残差和最小原则,以实例说明在有异常点和无异常点情况下,线性回归最小一乘解与最小二乘解的异同,并分别用不同的最小一乘解方法求解回归方程,比较其最小绝对残差和。结果表明:最小一乘法的回归方程较之最小二乘法具有绝对残差和 $F_{\min}$ 最小的优越性,不会因特异点而使预报方程导致较大误差。在无异常点的情况下,通过均值点回归方程的最小一乘解和最小二乘解, $F_{\min}$ 值基本接近;在有异常点的情况下,最小二乘解的 $F_{\min}$ 值明显大于最小一乘解的 $F_{\min}$ 值。回归方程的最小一乘解算法,以不通过均值点而通过某一点的 $F_{\min}$ 值较小,其中,搜索试算法的 $F_{\min}$ 值最小,但该方法有选取初值的问题。

关键词:流量预测;最小一乘法;最小二乘法;绝对残差和

中图分类号:P333 文献标识码:B 文章编号:1006-764X(2004)07-0072-03

1 最小一乘法与最小二乘法

在水文分析、工程计算与预测预报中,线性回归模型应用比较广泛。回归模型参数大多采用最小二乘法估计,即

$$y = a + bx \tag{1}$$

式中: a 和 b 为待定参数。数理统计中,按照最小二乘原则确定 a 和 b ,即使 $\sum_{i=1}^n \Delta y_i^2$ 值最小。 $y = a + bx$ 为通过均值点的直线。最小二乘法要通过一系列计算方能求得参数 a 、 b 、 σ_x 、 σ_y 和 r 。但基于最小二乘法的回归分析有时很不理想,特别是样本数据较少,并含有个别异常点时,会使回归方程产生较大偏差,亦会因此而导致预测预报的失误。

采用绝对残差和最小原则的最小一乘法,其回归直线为中直线,不会因迁就特异点据而产生较大偏差。一般最小一乘法的目标函数为

$$F_{\min} = \sum_{i=1}^n |y_i - a - bx_i| \quad (i = 1, 2, \dots, n) \tag{2}$$

其中, F 与 b 的关系为分段线性函数,极值点在各分段的端点,并且每一端点对应于一个 b 值。即

$$b_i = y_i/x_i \quad (i = 1, 2, \dots, n) \tag{3}$$

将 b 值从小到大排列,即

$$b_1 \leq b_2 \leq b_3 \leq \dots \leq b_n \tag{4}$$

绘制 b 与 F 关系曲线(图 1)。由此可得 $F_{\min}$ 时

的 $b = b_r$ 值。此即式(2)中的最小一乘解。按式(4)顺序,找到式(3)中相应的 x_i 与 y_i 值,记为 (x_j, y_j) 。然后将式(2)右端分段,同时按 $b_j (j = 1, 2, \dots, n)$ 的顺序排列,即

$$|y_1 - bx_1|, |y_2 - bx_2|, \dots, |y_n - bx_n| \tag{5}$$

当 $b \leq b_1$ 时,式(5)各项为正值,有

$$F = \sum_{i=1}^n y_i - b \sum_{j=1}^n x_j$$

由此可见,在 F 与 b 的线性关系中,直线的斜率为 $S_0 = -\sum_{j=1}^n x_j$,见图 1 中曲线左端。

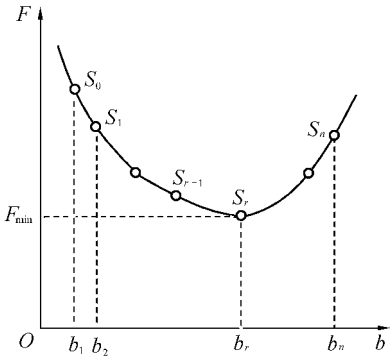


图 1 $F \sim b$ 相关曲线

当 $b_1 \leq b \leq b_2$ 时,式(5)的第 1 项 $y_1 - bx_1 < 0$,其余各项均为正值,于是有:

$$F = bx_1 - y_1 + \sum y_j - b \sum x_j$$

此段斜率为

作者简介:江秋阳(1967—),男,安徽黄山人,工程师,从事水资源管理工作。

$$S_1 = - \sum x_j + x_1 = S_0 + 2x_1$$

显然有：

$$|S_1| < |S_0|, S_2 = S_1 + 2x_2, |S_2| < |S_1|$$

并有： $b < b_1$ 时 $S_0 = - \sum |x_j| < 0$

$$b_1 < b < b_2 \text{ 时 } S_1 = S_0 + 2|x_1| < 0$$

$$b_{r-1} < b < b_r \text{ 时 } S_{r-1} = S_{r-2} + 2|x_r - 1| < 0$$

$$b_r < b < b_{r+1} \text{ 时 } S_r = S_{r-1} + 2|x_r| > 0$$

$$b > b_n \text{ 时 } S_n = S_{n-1} + 2|x_n| > 0$$

当分段直线的斜率有 $S_{r-1} < 0$ 变到 $S_r > 0$ 时，则出现 $F_{\min}$ 及相应 b_r 值(图 1)，此即为问题的解。

2 无异常点时线性回归方程的最小二乘解和最小一乘解

若已知某河流上游甲水文站的最大流量，可用回归方程来预报该河流下游乙水文站的最大流量。表 1 为某河流上下游甲、乙水文站的最大流量关系。 $y_{\text{下}}$ 为甲水文站的最大流量系列， $x_{\text{上}}$ 为乙水文站的相应最大流量系列。根据最小一乘法公式逐步求解，无异常点的回归方程为 $y = 1.66x - 79$ ， $F_{\min} = 354.27$ 。根据最小二乘法公式逐步求解(无异常点)得 $a = -59.1$ ， $b = 1.149$ ， $r = 0.991$ 。回归方程(均通过均值点(1151,1263))为

$$y = 1.149x - 59.1, F_{\min} = 356.895 \quad (6)$$

由上再计算 $x' = x - \bar{x}$ ， $y' = y - \bar{y}$ ，于是回归方程写成 $y' = b_j x'$ ，其中 $b_j = y'_i / x'_i$ ， x_j 为与 b_j 对应的值，从小到大排列。

表 1 某河流甲、乙水文站的最大流量系列

序号	$x_{\text{上}}/\text{m}^3$	$y_{\text{下}}/\text{m}^3$	序号	$x_{\text{上}}/\text{m}^3$	$y_{\text{下}}/\text{m}^3$
1	1590	1770	9	1380	1530
2	1170	1290	10	1080	1100
3	1500	1670	11	1110	1250
4	1230	1350	12	1090	1250
5	1010	1140	13	640	668
6	1290	1410	14	1090	1130
7	959	1060	15	1130	1230
8	996	1100	均值	1151	1263.20

由斜率计算式可得：

$$S_0 = - \sum |x_j| = -2508$$

$$S_1 = S_0 + 2|x_1| = -1486$$

$$S_2 = S_1 + 2|x_2| = -1102$$

$$S_3 = S_2 + 2|x_3| = -792$$

依次计算，当 $j = 10$ 时， S_j 变号，于是有 $b_r = b_{10} =$

1.166。通过均值点的回归方程为

$$y = 1.166x - 79, F_{\min} = 354.27 \quad (7)$$

比较式(6)和(7)可见， $F_{\min}$ 最小一乘解较最小二乘解小，且计算简捷方便。

3 有异常点时线性回归方程的最小一乘解和最小二乘解

上例中，序号 3(1500,1670)为正常点，而原来是因洪水期间防洪堤决口而出现的异常点据(1500,1490)。现仍按异常点据分别求最小一乘解与最小二乘解(图 2)。

a. 通过均值点的回归方程。由于最小一乘解与最小二乘解均通过均值点(1151,1263.2)，所以有最小一乘解为

$$y = 1.166x - 91.1, F_{\min} = 580.43 \quad (8)$$

最小二乘解为

$$y = 1.065x + 25.4, F_{\min} = 607.275 \quad (9)$$

可见，在有异常点情况下，两者回归方程截然不同。最小二乘解的 $F_{\min}$ 值明显大于最小一乘解的 $F_{\min}$ 值。

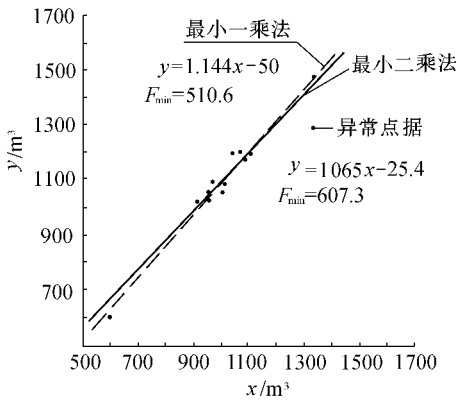


图 2 回归方程比较

b. 搜索试算法的最小一乘解。取 a 的初始值 a_0 ，得 $b_i = (y_i - a_0) / x_i$ ，进行与上例同样的计算，得到相应于最小 $F_{\min}$ 时的值，再设定若干个 a ，进行同样的计算可搜索到最终解，即最小 $F_{\min}$ 时的 b 值。回归方程为

$$y = 1.144x - 50, F_{\min} = 510.6 \quad (10)$$

c. 通过任一点的最小一乘解。目标函数： $F = \sum |y_i - a - bx_i|$ ，通过某一数值点(x_m, y_m)的回归方程为 $a = y_m - b_m x_m$ 。在本例中，先设(x_m, y_m) = (x_i, y_i)， $i = 1, 2, 3, \dots, n$ ，并计算 $x' = x_i - x_m$ ， $y' = y_i - y_m$ 。

用同样方法可求得 x', y' 的回归方程与 $F_{\min}$ 值。共进行了 $n = 15$ 次计算，取 $F_{\min}$ 最小的解。经计算比较，以通过点(1380,1530)的 $F_{\min}$ 值最小。回归方程为

$y = 1.142x - 45.96$, $F_{\min} = 522.33$ (11)

比较式 (6)~(11) 可见 , 以搜索试算法的 $F_{\min}$ 值最小 , 通过任一点的最小一乘解 $F_{\min}$ 值次之 , 而通过均值点的最小一乘解和最小二乘解 $F_{\min}$ 值均不为最小 .

4 结 论

比较计算结果可知 : ①在无异常点的情况下 , 通过均值点回归方程的最小一乘解和最小二乘解 , $F_{\min}$ 值基本接近 , 但最小一乘法较最小二乘法计算简捷方便 ; ②在有异常点的情况下 , 通过均值点回归方程的最小一乘解和最小二乘解 , 回归方程迥然不同 , 最小二乘解的 $F_{\min}$ 值明显大于最小一乘解的 $F_{\min}$ 值 ; ③回归方程的最小一乘解计算方法 , 以不通过均值点而通过某一点的 $F_{\min}$ 值较小 , 其中 , 搜索试算法的 $F_{\min}$ 值最小 , 但该方法有选取初值的问题 . 只

(上接第 53 页) 应将权重的选择与施工实践相结合 , 不断修正各个指标对应的权重 .

笔者采用前述的 AASHTO 付款系数计算方法 , 并结合专家调查法 (业主、施工单位和科研机构专家各占 1/3) 确定出 5 个关键的质量性能指标和相应的权重 , 计算了江苏省两条已投入使用的高速公路——公路 1、公路 2 的某个标段的付款系数 (计算过程略) , 见表 1 和表 2 .

表 1 公路 1 某标段中、下面层 P_{PF} 值计算结果

所在面层	P_{PF}				
	样本 1	样本 2	样本 3	样本 4	样本 5
下面层	1.010	1.024	1.001	1.011	1.007
中面层	0.986	1.038	0.933	0.898	0.892

表 2 公路 2 某标段中、下面层 P_{PF} 值计算汇总

所在面层	P_{PF}		
	样本 1	样本 2	样本 3
下面层	1.015	1.010	0.946
中面层	0.980	0.924	0.915

施工中连续 6 次试验的作为一个样本 , 即子样本大小 n 为 6 . 选用的质量性能指标为 : 级配、油石比、空隙率、压实度 , 权重选择为 0.3 0.2 0.3 0.2 . 由于施工数据不完整 , 两个路段选择的样本数量不相等 .

从表 1 和表 2 可以看出 : 下面层的 P_{PF} 基本大于 1 , 中面层的 P_{PF} 基本小于 1 , 说明下面层施工质量明显好于中面层的施工质量 . 公路 1 某标段下面层的第 2 个样本的 P_{PF} 远大于下面层其他样本的 P_{PF} , 同时 , 公路 2 某标段下面层的第 3 个样本的 P_{PF} 远小于下面层其他样本的 P_{PF} . 从目前这两条路段的使用情况来看 , 计算结果与实际情况相符 , 即计算出 P_{PF} 较高的路段 , 其使用情况也较好 , 反之亦然 . 这表明采用付款系数进行验收支付是合理可行的 .

有在确定通过均值点的回归方程后 , 进行搜索试算才较为方便 , 有时还会遇到多次搜索试算的问题 ; ④通过某一点 $(x_m, y_m) = (x_i, y_i)$ 回归方程的最小一乘解要进行多次运算 , 工作量较大 , 但可一次成功 , $F_{\min}$ 值也较小 ; ⑤最小一乘法的回归方程较之最小二乘法具有绝对残差和 $F_{\min}$ 最小的优越性 , 不会因特异点而使预报方程导致较大误差 . 因此 , 值得推广使用 .

(本文曾得到张善余老师的指导 , 特此致谢 .)

参考文献 :

[1] 金光炎 . 水文水资源随机分析 [M] . 北京 : 中国科技出版社 , 1993 . 119 .
 [2] 金光炎 . 线性回归与最小一乘解 [J] . 水文 , 1996 (3) : 1 ~ 6 .

(收稿日期 2004-04-28 编辑 : 马敏峰)

4 结 语

最终产品规范采用与质量直接相关的付款系数进行验收支付 , 更好地体现出“ 优质优价 ” 的思想 , 并成功实现了以经济杠杆来控制产品质量 , 使沥青路面施工管理更加规范、合理 . 我们要在综合分析国外最终产品规范的基础上 , 总结出其中科学合理的因素 , 结合具体试点项目 , 研究制定适合江苏省实际情况的最终产品规范 , 并分析其应用对工程造价、质量、施工以及管理所产生的影响 , 然后分阶段推广应用 , 进一步提高沥青路面施工水平和路面施工质量 .

参考文献 :

[1] Muench S T , Mahoney J P . A quantification and evaluation of WSDOT ' s hot mix asphalt concrete statistical acceptance specification [S] . Washington : Washington State Transportation Center (TRAC) . 2001 . 11 ~ 14 .
 [2] Butts N E , Ksaibati K . Asphalt pavement quality control/quality assurance programs in the United States [A] . In : Paper of TRB 2003 annual meeting [C] . Washington : Transportation Research Board 2002 . 4 ~ 14 .
 [3] Yeo R , Vuong B , Alderson A . Asphalt concrete statistics acceptance specification [S] . Wisconsin : Wisconsin Asphalt Pavement Association 2003 . Appendix A-III 56 ~ 58 , 59 ~ 60 .
 [4] Bohuslav T R . Material specification [S] . Texas : Texas Department of transportation . 2002 .
 [5] Emory J M , Duckert B . Field guide for the acceptance of hot mix and bridge deck waterproofing [M] . Ontario : Bituminous Section Materials Engineering and Research Office 2002 . 47 ~ 51 , Appendix C .
 [6] Bouldin M G , Elmore W E . End-result construction specification for hot mix asphalt [S] . Illinois : Department of Civil Engineering . 2003 . 28 ~ 29 .

(收稿日期 2004-05-27 编辑 : 骆超)