DOI:10.13656/j.cnki.gxkx.1998.03.008
广西科学 Guangxi Sciences 1998，5（3）：190～193

线性模型和广义线性模型中的一种强影响点的显著性检验法

Influential Point Test for Linear Model and Generalized Linear Model

龙 蓓　　林 路*

Long Bei　　Lin Lu

（桂林医学院数学教研室　桂林市乐群路 56 号　541001）

（Teaching and Research Section of Math.，Guilin Medical College，56 Lequnlu，Guilin，Guangxi，541001）

摘要　根据强影响点的实际意义，提出一种强影响点的显著性检验模型。解决了线性模型和广义线性模型的强影响点的显著性检验问题，其中的检验统计量分别是 F 检验统计量和Score检验统计量，实例表明此法较好.

关键词　统计诊断　线性模型　广义线性模型　强影响点

中图法分类号　O 212.1

Abstract　The influential point test for the linear model and the generalized linear model are discussed. We give the influential point test model and F statistic for the linear model and score statistic for the generalized linear model. Two numerical examples are given to illustrate our results.

Key words　statistical diagnostic，linear model，generalized linear model，influential point

影响分析是统计诊断中十分活跃的分支，其初期的最有实用价值的内容就是研究某些特定点对统计分析（如参数估计）的影响. 虽然线性模型中强影响点的显著性检验已有文章作过探讨，但对广义线性模型的强影响点的显著性检验问题却研究甚少. 作者根据强影响点的实际意义，提出了一种强影响点的显著性检验模型，解决了线性模型和广义线性模型的强影响点的显著性检验问题，其中的检验统计量分别是 F 检验统计量和Score检验统计量，实例表明此法较好.

1　检验模型

本文考察的线性模型为

$$Y = X\beta + \varepsilon,\quad \varepsilon \sim N(0, \sigma^2 I) \tag{1}$$

其中 $Y=(y_1, y_2, \cdots, y_n)$是观测向量，$X=(x_1, x_2, \cdots, x_n)^T$ 是 $n\times p$ 列满秩设计阵，$\beta=(\beta_1, \beta_2, \cdots, \beta_p)^T$ 为$p\times 1$ 未知参数，$\varepsilon=(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$ 为 $n\times 1$ 随机误差向量，β 在模型(1)中的最小二乘估计(最大似然估计)为 $\hat\beta=(\hat\beta_1,\hat\beta_2,\cdots,\hat\beta_p)^T=(X^TX)^{-1}X^TY$,

在模型(1)中，删除 $J=\{i_1, i_2, \cdots, i_k\}$ 中数据，$J\subset\{1, 2, \cdots, n\}$，得到数据删除模型.

$$Y(J) = X(J)\beta + \varepsilon(J),\ \varepsilon(J) \sim N(0, \sigma^2 I(J)) \tag{2}$$

其中 $Y(J)$, $X(J)$, $\varepsilon(J)$, $I(J)$ 分别是 Y, X, ε, I 删除第i_1行，第 i_2 行，…，第 i_k 行后得到的向量或矩阵，β 在模型(2)中的最小二乘估计(最大似然估计)为：

$$\hat\beta(J) = [\hat\beta(J)_1, \hat\beta(J)_2, \cdots, \hat\beta(J)_p]^T = (X^T(J)X(J))^{-1}X^T(J)Y(J).$$

本文还研究实用上较常用的典则联系的广义线性模型.

$$E(Y) = \mu,\quad \theta = G(\mu) = X\beta \tag{3}$$

其中 Y、X 和 β 与模型(1)中的表示式一致，$\mu=(\mu_1, \mu_2, \mu_3, \cdots, \mu_n)^T$, $\theta=(\theta_1, \theta_2, \cdots, \theta_n)^T=[g(\mu_1), g(\mu_2), \cdots, g(\mu_n)]^T$. y_i 服从指数族分布，其分布密度为：

$$p(y_i, \theta_i) = \exp\{\frac{\theta_i y_i - \psi(\theta_i)}{a(\varphi)} + c(y_i, \varphi)\},$$

$i=1,2, \cdots, n$. 其中 φ 为多余参数，因此假设 $a(\varphi)=1$, g 为一元严格增函数，即为 $\dot\psi$的反函数，记模型(3)中 β 的最大似然估计为$\hat\beta$, $\hat\beta$ 可由 $G-N$ 迭代公式求解.

1997-10-20 收稿。

*邵阳师范专科学校数学系，湖南邵阳，422000（Department of Math.，Shaoyang Teachers College，Shaoyang，Hunan，422000）

在模型(3)中，删除 $J=\{i_1, i_2, \cdots, i_k\}$ 中数据，得到其数据删除模型.

$$E(Y(J))=\mu(J),\ \theta(J)=G(J)(\mu)=X(J)\beta \tag{4}$$

其中 $Y(J)$, $\mu(J)$, $\theta(J)$, $G(J)$, $X(J)$ 分别是 Y, μ, θ, G, X 删除第 i_1 行，第 i_2 行，…，第 i_k 行后得到的向量或矩阵，模型(4)中 β 的最大似然估记记为 $\hat\beta(J)$.

记 $Y_J=(y_{i1}, y_{i2}, \cdots, y_{ik})^T$, $X_J=(x_{i1}, x_{i2}, \cdots, x_{ik})^T$，在影响分析中，常用 $\hat\beta$ 与 $\hat\beta(J)$ 的某种距离描述数据 (Y_J, X_J) 对回归分析(如参数估计)的影响. 例如，对参数估计来说，若 $\hat\beta(J)$ 与 $\hat\beta$ 的距离很大，则数据 (Y_J, X_J) 为强影响点；若 $\hat\beta(J)$ 与 $\hat\beta$ 相去不远，则 (Y_J, X_J) 不是强影响点，基于这种思想，本文考察原模型(1)(或(3))和数据删除模型(2)(或(4))中未知参数 β 的真值之间的差异[1]，若删除数据 (Y_J, X_J) 后，模型(2)(或(4))中的 β 的真值与模型(1)(或(3))中 β 的真值差异很大，则数据 (Y_J, X_J) 为强影响点，否则，(Y_J, X_J) 不是强影响点. 在模型(1)(或(3))中，用 $\hat\beta$ 近似地表示 β 的真值，我们就有如下想法：若模型(2)(或(4))中 β 的真值与 $\hat\beta$ 差异很大，则 (Y_J, X_J) 是强影响点，否则 (Y_J, X_J) 不是强影响点.

由上述思想，我们在数据删除模型(2)和(4)中提出如下假设检验：

$$H: \beta=\hat\beta \tag{5}$$

若原假设 H 被拒绝，则认为模型(2)(或(4))中的 β 的真值与模型(1)(或(3))中的 β 有显著差异，从而数据 (Y_J, X_J) 是强影响点.

2 检验统计量

先求线性模型中强影响点的检验统计量，显然，检验(5)是模型(2)中 β 的线性假设检验.

模型(2)的残差平方和为：

$$RSS(J)=\sum_{j\in J}(y_j-x_j^T\hat\beta(J))^2 \tag{6}$$

模型(2)中，在条件 $H: \beta=\hat\beta$ 下的残差平方和为：

$$RSS(J)_H=\sum_{j\in J}(y_j-x_j^T\hat\beta)^2 \tag{7}$$

记模型(1)的残差平方和为：

$$RSS=\sum_{j=1}(y_j-x_j^T\hat\beta)^2 \tag{8}$$

且记：

$\hat e_j=y_j-x_j^T\hat\beta$, $\hat e_J=(\hat e_{i1}, \hat e_{i2}, \cdots, \hat e_{ik})^T$, $P_J=X_J(X^TX)^{-1}X_J^T$.

我们可得到假设检验(5)的 F 检验统计量[3]：

$$\begin{aligned}F_H(J)&=\frac{1}{p}(RSS(J)_H-RSS(J))\Big/\frac{1}{n-p-k}RSS(J)\\&=\frac{n-p-k}{p}\cdot\frac{RSS-\hat e_J^T\hat e_J-(RSS-\hat e_J^T(I-P_J)^{-1}\hat e_J)}{RSS-\hat e_J^T(I-P_J)^{-1}\hat e_J}\\&=\frac{n-p-k}{p}\cdot\frac{\hat e_J^T((I-P_J)^{-1}-I)\hat e_J}{RSS-\hat e_J^T(I-P_J)^{-1}\hat e_J}\\&=\frac{n-p-k}{p}\cdot\frac{\hat e_J^T(I-P_J)^{-1}P_J\hat e_J}{RSS-\hat e_J^T(I-P_J)^{-1}\hat e_J},\end{aligned} \tag{9}$$

于是，我们有如下定理.

定理 1　在模型(2)中，若 $H: \beta=\hat\beta$ 为真，则：

$$F_H(J)=\frac{n-p-k}{p}\cdot\frac{\hat e_J^T(I-P_J)^{-1}P_J\hat e_J}{RSS-\hat e_J^T(I-P_J)^{-1}\hat e_J}\sim F(p, n-p-k) \tag{10}$$

其中，$F(p, n-p-k)$ 为自由度为 $(p, n-p-k)$ 的 F 分布.

由定理 1 知，给定显著性水平 α, $0<\alpha<1$，记 $F_\alpha(p, n-p-k)$ 是自由度为 $(p, n-p-k)$ 的 F 分布的上侧 α 分位点，当 $F_H(J)>F_\alpha(p, n-p-k)$ 时，不定式(5)，即认为 (Y_J, X_J) 是强影响点；不然就接受(5)，即认为 (Y_J, X_J) 不是强影响点.

由 $F_H(J)$ 的表达式(9)知，$F_H(J)$ 与 Cook 统计量和 $W-K$ 统计量等度量影响的统计量有相当的一致性[1]. 在本文后面的实例中，将再次证实这种一致性.

下面将定理 1 推广(或者说应用)到约束线性模型强影响点的显著性检验.

考察约束线性模型

$$\begin{cases}Y=X\beta+\varepsilon\\A^T\beta=0\end{cases}\quad \varepsilon\sim N(0, \sigma^2 I) \tag{11}$$

其中 A 为 $p\times q$ 列满秩矩阵，$q<p$，此模型的数据删除形式为

$$\begin{cases}Y(J)=X(J)\beta+\varepsilon(J)\\A^T\beta=0\end{cases}\quad \varepsilon(J)\sim N(0, \sigma^2 I(J)) \tag{12}$$

在模型(11)，(12)中，记约束 $A^T\beta=0$ 的解为 $\beta=Bt$，B 是 $p\times(p-g)$ 列满秩矩阵，满足 $A^TBt=0$，则分别得到与(11)和(12)等价的无约束线性模型[4]

$$Y=(XB)t+\varepsilon\quad \varepsilon\sim N(0, \sigma^2 I) \tag{13}$$

和　$Y(J)=(X(J)B)t+\varepsilon(J)$,

$$\varepsilon(J)\sim N(0, \sigma^2 I(J)) \tag{14}$$

若记(11)中 β 的最小二乘估计为 $\hat\beta_c$，残差为 $\hat e_j(c)=y_j-x_j^T\hat\beta_c$，(13)中 t 的最小二乘估计为 $\hat t$，要检验数据 (Y_J, X_J) 是否为模型(11)即模型(13)的强影响点，

只要对模型(14)检验如下假设:

$$H_c: t = \hat{t} \tag{15}$$

由式(9)知,检验(15)的 F 统计量为

$$F_{H_C}(J) = \frac{n-p+q-k}{p-q} \cdot \frac{\hat{e_J}(c)(I-P_J(c))^{-1}P_J(c)\hat{e_J}(c)}{RSS_c - \hat{e_J}(c)(I-P_J(c))^{-1}\hat{e_J}(c)} \tag{16}$$

其中 RSS_c 为(11)即(13)的残差平方和 $\sum_{j=1}^{n}(y_j - x_j^T\hat{\beta}_c)^2$, $\hat{e_J}(c) = (\hat{e_{i1}}(c), \hat{e_{i2}}(c), \cdots, \hat{e_{ik}}(c))^T$,由文献[3]知:

$$P_J(c) = X_J B(B^T(X^TX)B)^{-1}BX_J^T$$

$$= X_J((X^TX)^{-1} - (X^TX)^{-1}M(X^TX)^{-1})X_J^T$$

其中 $M = A(A^T(X^TX)^{-1}A)^{-1}A$,由此得到如下定理.

定理2 模型(12)(即模型(14)中,若式(15)为真,则:

$$F_{H_C}(J) = \frac{n-p+q-k}{p-q} \cdot \frac{\hat{e_J}(c)(I-P_J(c))^{-1}P_J(c)\hat{e_J}(c)}{RSS_c - \hat{e_J}(c)(I-P_J(c))^{-1}\hat{e_J}(c)} \sim F(p-q, n-p+q-k) \tag{17}$$

定理2表明,若 $F_{H_C}(J) > F_\alpha(p-q, n-p+q-k)$,则数据 (Y_J, X_J) 是模型(11)的强影响点;不然,(Y_J, X_J) 不是(11)的强影响点.

下面求广义线性模型中强影响点的检验统计量.

广义线性模型(4)(略去与参数无关的项后)的对数似数函数为:

$$L(\beta) = \sum_{j \in J}(y_j x_j^T\beta - \psi(x_j^T\beta)) \tag{18}$$

且: $\dot{L}(\beta) = X^T(J)S(J)$ (19)

其中 $S(J)$ 是由 $S_j, j \in \{1, 2, \cdots, n\} - J$ 组成的列向量, $S_j = y_j - \dot{\psi}(\theta_j) = y_j - \mu_j$

$$\ddot{L}(\beta) = -X^T(J)W(J)X(J) \tag{20}$$

其中 $W(J)$ 是 $W_j, j \in \{1, 2, \cdots, n\} - J$ 构成的对角阵, $W_j = V_{ar}(y_j) = \ddot{\psi}(\theta_j)$.

以下,若无特殊说明, $S(J)$, $W(J)$ 及 $S = (s_1, s_2, \cdots, s_n)^T$, $S_J = (s_{i1}, s_{i2}, \cdots, s_{ik})^T$, $W = \mathrm{diag}(w_1, w_2, \cdots, w_n)$, $W_J = \mathrm{diag}(w_{i1}, w_{i2}, \cdots, w_{ik})$ 等均表示在 $\hat{\beta}$ 处计值.我们知道[5],模型(4)对于假设检验(5)的Score检验统计量为:

$$SC_J = -(\dot{L}(\hat{\beta}))^T\ddot{L}^{-1}(\hat{\beta})\dot{L}(\hat{\beta}) \to \chi^2(p) \tag{21}$$

由式(19),(20)有:

$$SC_J = S^T(J)X(J)(X^T(J)W(J)X(J))^{-1}X^T(J)S(J) \tag{22}$$

由于在 $\hat{\beta}$ 处有 $X^TS = 0$(即 $\hat{\beta}$ 满足模型(3)的似然方程),故

$$X^T(J)S(J) = X^TS - X_J^TS_J = -X_J^TS_J \tag{23}$$

又由于

$$X^T(J)W(J)X(J) = X^TWX - X_J^TW_JX_J \tag{24}$$

由式(22),(23),(24)及矩阵和式求逆公式,得到

$$SC_J = S_J^TX_J(X^TWX)^{-1}X_J^T(I - W_JX_J(X^TWX)^{-1}X_J^T)^{-1}S_J \tag{25}$$

所以,有

定理3 在模型(4)中,若 $H: \beta = \hat{\beta}$ 为真,则Score检验统计量

$$SC_J = S_J^TX_J(X^TWX)^{-1}X_J^T(I - W_JX_J(X^TWX)^{-1}X_J^T)^{-1}S_J \to \chi^2(p) \tag{26}$$

定理3表明,若 $SC_J > \chi^2_\alpha(p)$,则认为数据 (Y_J, X_J) 是模型(3)中的强影响点;否则,数据 (Y_J, X_J) 不是模型(3)中的强影响点.

3 算例

下面的两个算例是分别应用定理1和定理3检验线性模型及广义线性模型中单个数据是否是强影响点,若逐个考察单个数据 (y_j, x_j^T) 的强影响点判别问题,式(9)为

$$F_H(j) = \frac{n-p-1}{p} \cdot \frac{P_{jj}e_j^2/(1-P_{jj})}{RSS - e_j^2/(1-P_{jj})}$$

其中, $P_{jj} = x_j^T(X^TX)^{-1}x_j$,为便于应用,将 $F_H(j)$ 用Cook统计量和 $W-K$ 统计量表示:

$$F_H(j) = \frac{(1-P_{jj})^{\sigma^2}}{\sigma^2(j)}D_j \tag{27}$$

$$F_H(j) = \frac{1-P_{jj}}{p}(W_{k_i})^2 \tag{28}$$

其中 $D_j = \dfrac{P_{jj}e_j^2}{P\sigma^2(1-P_{jj})^2}$ 是 (y_j, x_j^T) 的Cook统计量, $W-k_j = \dfrac{\sqrt{P_{jj}} \cdot e_j^2}{\sigma^2(j)(1-P_{jj})}$ 是 (y_j, x_j^T) 的 $W-K$ 统计量, $\sigma^2 = \dfrac{RSS}{n-p}$, $\sigma^2(j) = \dfrac{RSS(j)}{n-p-1}$,

对于单个数据 (y_j, x_j^T), 记 $h_{jj} = W_jx_j^T(X^TWX)^{-1}x_j$,式(25)为:

$$SC_J = \frac{h_{jj}s_j^2}{W_j(1-h_{jj})} = \frac{h_{jj}}{1-h_{jj}} \cdot rp_j \tag{29}$$

其中 rp_j 为 (y_j, x_j^T) 的Pearson残差: $rp_j = s_jw_j^{-\frac{1}{2}}$,

例1 在BOQ数据[5]中,已知 $n = 25$,对25组数据进行线性回归分析 $(p = 8)$,表1给出每个数据 (y_j, x_j^T) 的Cook统计量, $W-K$ 统计量. F 统计量的值是由式(28)算得的,该表说明,第23号点的 D_j 和 $W-k_j$ 都特别大,若用 D_j 和 $W-k_j$ 为标准,可以怀疑第23号点是强影响点[1],若用 F 统计量检验,只有第23号点的 F 统计量的值 $F_H(j) = 3.3839 > F_{0.05}(8, 16) = 2.59$,即强影响点的显著性检验表明,只有第23号点是强影响点,检

验结果与文献[2] 的影响分析一致.

表 1　BOQ 数据的影响度量

Table 1　BOQ data influence measurings

No.	P_{jj}	$W-k_j$	D_j	$F_H(j)$
1	0.257 3	－0.043 3	0.000	0.000 2
2	0.160 9	－0.031 8	0.000	0.000 1
3	0.161 4	－0.201 6	0.005	0.004 3
4	0.163 1	－0.077 8	0.001	0.000 6
5	0.147 5	－0.174 5	0.004	0.003 2
6	0.158 9	－0.247 9	0.008	0.006 5
7	0.182 9	0.356 3	0.016	0.013 0
8	0.359 1	－0.353 3	0.016	0.010 0
9	0.288 0	0.202 9	0.005	0.003 7
10	0.129 5	0.035 3	0.000	0.000 1
11	0.124 1	0.584 9	0.039	0.037 5
12	0.202 4	0.223 0	0.007	0.005 0
13	0.080 2	－0.053 7	0.000	0.000 3
14	0.096 9	－0.007 3	0.000	0.000
15	0.557 6	－2.828 2	0.761	0.442 3
16	0.402 4	－0.044 4	0.000	0.000 1
17	0.368 2	－1.016 2	0.123	0.081 6
18	0.446 5	1.128 6	0.154	0.088 1
19	0.088 6	0.310 6	0.012	0.011 0
20	0.366 3	－2.178 7	0.417	0.376 0
21	0.070 4	0.565 2	0.034	0.037 1
22	0.785 4	－3.071 5	1.079	0.253 1
23	0.988 5	－48.517 9	115.041	3.383 9
24	0.876 2	8.537 3	5.889	1.127 9
25	0.546 7	0.472 2	0.029	0.012 6

例 2　表 2 是一组人造的 Logistic 回归数据[2]，并给出 Pearson 残差和 Score 检验统计量，结果显示，只有第 11 号数据的 $S_{C_j}=2.061\ 3>\chi^2_{0.20}(1)=1.642$，即强影响点的显著性检验表明，只有第 11 号数据是强影响点，这结果与文献 [2] 中用其他统计量分析的结果是一致的.

表 2　Logistic 回归数据影响度量

Table 2　Influence measurings of Logistic regression data

No.	x_j	Logit (y_j/n_j)	rp_j	h_{jj}	Sc_j
1	1	0.5	－0.506	0.201 8	0.064 7
2	2	0.5	－0.615	0.211 5	0.101 5
3	3	0.8	－0.247	0.159 1	0.011 5
4	4	1.0	－0.055	0.122 7	0.000 4
5	5	1.3	0.248	0.101 1	0.006 9
6	6	1.3	0.157	0.093 0	0.002 5
7	7	1.6	0.417	0.097 0	0.018 7
8	8	1.8	0.535	0.111 8	0.036 0
9	9	2.1	0.715	0.136 0	0.080 4
10	10	2.1	0.642	0.168 4	0.083 5
11	17	1.0	－1.383	0.518 7	2.061 3

参考文献

1 林　路. 若干有偏估计的强影响点的显著性检验. 数学的实践与认识, 1997, 27 (3).

2 韦博成, 鲁国斌, 史建清. 统计诊断引论. 南京: 东南大学出版社, 1990.

3 陈希孺, 王松桂. 近代回归分析. 合肥: 安徽教育出版社, 1987.

4 虞克明, 王静龙. 约束线性模型异常值检验. 高校应用数学学报, 1994, 9 (4).

5 Myers R H. Classial and modern regression with application. Boston: Duxbury press, 1986.

6 Cox D R, Hinkley D V. Thoeretical statistics. London: Chapman and Hall, 1974.

（责任编辑：黎贞崇　邓大玉）

德国环境、生态、气候的优先研究领域

（1）实施新的能源研究计划——“第四 能源研究计划”。研究重点：提高煤和其他矿物能源的利用率并减少对环境的污染，可再生能源的利用，核能源研究以及能源的信息处理和分析系统。计划实施可使德国的二氧化碳排放量减少 35%。

（2）制定新的环境研究计划。研究重点：生态研究、环境技术研究和大气研究。计划实施可持续发展生态保护系统和集约型生产的环保技术，并建立相应的示范样板。

（3）建立一个通过人工智能并且实现了广泛联网的、具有最大灵活性和最佳效果的交通系统，减少资源消耗和环境负荷。

（摘自 中国科学院《科学发展报告》1997. P51）