第10卷 第3期
1990年8月

数学研究与评论
JOURNAL OF MATHEMATICAL
RESEARCH AND EXPOSITION

Vol.10 No.3
Aug. 1990

强影响的分布与稳定性度量

杨 虎

（重庆交通学院）

摘要 本文考虑了数据集在样本空间的影响分布，并且指出数据集仅对与之强相关的数据点有较大的影响，因而Cook距离 $D_i(X'X, c)=\frac{1}{c}\sum_{s=1}^{n}(y_s-y_s(i))^2$ 中这一部分数据处的影响较突出，我们考虑用 $\frac{1}{c}\sum_{s=1}^{n}|y_s-y_s(i)|$ 作为影响度量将更合理，同时它也比Cook距离具有更好的稳定性.

一、引 言

今天，影响分析(见[1])的研究领域已远远超出了线性回归的范围，在主成分分析[2]，判别分析[3]，时间序列分析[4]以及多元分析的一般参数估计问题[5]中都出现了一些工作.一些文献(如[1])并不认为影响分析是回归诊断的一部分，以上工作也说明了这一点,然而影响分析同残差分析，数据变换一样却是回归诊断所要研究的主要问题.

考虑多元线性回归模型

$$Y=X\beta+\varepsilon \tag{1}$$

其中$Y'=(y_1,y_2,\cdots,y_n)$为观测矩阵，$X'=(X_1,X_2,\cdots,X_n)$为设计阵，$\beta'=(\beta_1,\beta_2,\cdots,\beta_p)$为参数矩阵，$\varepsilon'=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)$为随机误差矩阵，这里$y_i,\beta_i,\varepsilon_i$为$m$维向量，$i$为不超过$n$的自然数.我们假定$\varepsilon_i, i=1,2,\cdots,n$互不相关，均值为零且协方差阵均为$\Sigma^2>0$(正定)，当$m=1$时记$\sigma=\Sigma$，这时$\varepsilon$的均值为零，协方差阵为$\sigma^2 I_n$($I_n$表示单位矩阵).$\beta$的最小二乘估计(记为LSE)为$\hat{\beta}=(X'X)^{-1}X'Y$，记$H=X(X'X)^{-1}X'$为帽子矩阵，其第$i$个对角元为$h_{ii}$，考虑剔除数据集$J$，它由$k$个数据$(X'_{i_s},y'_s), s=1,2,\cdots,k)$组成，从模型(1)中剔除这些数据后，得到如下模型

$$Y(J)=X(J)\beta+\varepsilon(J) \tag{2}$$

由模型(1)，(2)得到的Σ的LSE分别记为$\hat{\Sigma}$和$\hat{\Sigma}(J)$，(2)中β的LSE记为$\hat{\beta}(J)$，$\delta_J=Y_J-X_J\hat{\beta}$表示普通残差矩阵，这里$X_J,Y_J$表示$X,Y$中第$i_1,i_2,\cdots,i_k$行组成的矩阵.

$m=k=1$时，常用来判定第i组数据是否为强影响点的度量标准是Cook距离[6]

$$D_i(M,c)=(\hat{\beta}(i)-\hat{\beta})'M(\hat{\beta}(i)-\hat{\beta})/c \tag{3}$$

* 1988年10月19日收到.

$M>0$, $c>0$ 可适当选取，它和稳健统计中的一个重要概念——影响曲线有密切关系，事实上，Cook和Weisberg[7]证明了（3）正好是β的样本影响曲线$(n-1)(\hat{\beta}(i)-\hat{\beta})$的一种范数，从而使Cook距离得到一种理论支持．另一种影响度量是Andrews和Pregibon[8]提出的$A\ P$统计量

$$R_i=|Z'(i)Z(i)|/|Z'Z| \qquad (4)$$

这里$Z=(X,Y)$，$Z(i)=(X(i),Y(i))$，Draper和John[9]曾举例说明R_i很小的数据在Cook距离意义下未必是强影响的．此外，Welsch-kuh统计量[10]

$$W_i^2=(X_i'\hat{\beta}(i)-X_i'(\hat{\beta})^2/\hat{\sigma}^2(i)h_{ii} \qquad (5)$$

也被作为影响度量．从(5)式很容易使人认为W_i^2度量了第i个数据对自身预测的影响，而事实上$W_i^2=D(X'X,\hat{\sigma}^2(i))=(X\hat{\beta}-X\hat{\beta}(i))'(X\hat{\beta}-X\hat{\beta}(i))/\hat{\sigma}^2(i)$度量了第$i$个数据对各点预测影响的总和．（5）是一个很有意思的表达式．

其它诸如Johnson，Geisser的预测度量[11]，似然距离，Cook的局部影响度量[12]等等也可作为影响度量，这里不一一叙述．

二、主要结果

记$\|\cdot\|^2$为欧氏范数，$H_J=X_J(X'X)^{-1}X_J'$，$\hat{y}_j=X_j'\hat{\beta}$，$\hat{y}_j(J)=X_j'\hat{\beta}(J)$分别为从模型(1)和(2)中得到的$y_i$的预测值．

定义 我们用$\|\hat{y}_s-\hat{y}_s(J)\|\|\hat{\Sigma}(J)\|^{-1}$表示数据集$J$在样本空间中第$s$个点处的影响，记为$HP_J(s)$，则$J$的影响分布列为$\{HP_J(s):\ s=1,2,\cdots,n\}$．

定理1 当$k=1$时

$$HP_i(s)=\sqrt{h_{ss}h_{ii}^{-1}}HP_i(i)\,\rho_H(\hat{y}_j,\hat{y}_i) \qquad \forall s \qquad (6)$$

其中$\rho_H(\cdot,\cdot)$为Hotelling广义相关系数[13]．

证明 当矩阵A，C可逆时，我们有如下熟知的代数事实

$$(A+BCB')^{-1}=A^{-1}-A^{-1}B(B'A^{-1}B+C^{-1})^{-1}B'A^{-1} \qquad (7)$$

从而
$$(X'X-X_iX_i')^{-1}=(X'X)^{-1}+(X'X)^{-1}X_i(1-h_{ii})^{-1}X_i'(X'X)^{-1}$$

$$\begin{aligned}\hat{\beta}-\hat{\beta}(i)&=(X'X)^{-1}X'Y-(X'(i)X(i))^{-1}X'(i)Y(i)\\&=(X'X)^{-1}X'Y-[(X'X)^{-1}+(X'X)^{-1}X_i(1-h_{ii})^{-1}X_i'(X'X)^{-1}][X'Y-X_iy_i]\\&=(X'X)^{-1}X_iy_i+(X'X)^{-1}X_iy_i\frac{h_{ii}}{1-h_{ii}}-(X'X)^{-1}X_i(1-h_{ii})^{-1}\hat{y}_i\\&=(X'X)^{-1}X_i(1-h_{ii})^{-1}\delta_i \qquad (8)\end{aligned}$$

故

$$HP_i(s)=\|\hat{y}_s-\hat{y}_s(i)\|\|\hat{\Sigma}(i)\|^{-1}=|h_{si}|(1-h_{ii})^{-1}\|\delta_i\|\|\hat{\Sigma}(i)\|^{-1}，\text{而}$$

$$\rho_H^2(\hat{y}_s,\hat{y}_i)=\frac{|\mathrm{Cov}(\hat{y}_s,\hat{y}_i)\mathrm{Cov}^{-1}(\hat{y}_i,\hat{y}_i)\mathrm{Cov}(\hat{y}_i,\hat{y}_s)|}{|\mathrm{Cov}(\hat{y}_s,\hat{y}_s)|}$$

我们知道$\mathrm{Cov}(\hat{y}_s,\hat{y}_i)=\mathrm{Cov}(X_s'(X'X)^{-1}X'Y,\ X_i'(X'X)^{-1}X'Y)$，记$Y=(y^1,y^2,\cdots,y^m)$，易证$\mathrm{Cov}(y^l,y^l)=\sigma_{tl}I_m$，这里$(\sigma_{tl})_{m\times m}=\Sigma^2$，所以

$$\begin{aligned}\mathrm{Cov}(\hat{y}_s,\hat{y}_i)&=(\mathrm{Cov}(X_s'(X'X)^{-1}X'y^t,\ X_i'(X'X)^{-1}X'y^l))_{m\times m}\\&=(X_s'(X'X)^{-1}X'\sigma_{tl}I_mX(X'X)^{-1}X_i)_{m\times m}\end{aligned}$$

$$= h_{si}(\sigma_{ti})_{m\times m}$$
$$= h_{si}\Sigma^2$$

故 $\rho_H^2(\hat{y}_s,\hat{y}_i)=\dfrac{h_{si}^2}{h_{ii}h_{ss}}$，所以

$$HP_i(s)=\sqrt{h_{ss}h_{si}^{-1}}\ (1-h_{ii})^{-1}\|\delta_i\|\,\|\hat{\Sigma}(i)\|\,\rho_H(\hat{y}_s,\hat{y}_i)$$
$$=\sqrt{h_{ss}h_{ii}^{-1}}\ HP_i(i)\,\rho_H(\hat{y}_s,\hat{y}_i)$$

得证.

推论 当 $m=k=1$ 时，$HP_i(s)=\sqrt{h_{ss}}\,w_i\rho(\hat{y}_s,\hat{y}_i)$，$\rho(\cdot,\cdot)$为相关系数.

定理 2 当 $m=1$ 时

$$HP_J(s)<\frac{\sqrt{kh_{ss}\lambda_{\max}(H_J)}}{\hat{\sigma}(i)}\|(I_k-H_J)^{-1}\delta_J\|\,\rho_z(\hat{y}_s,\hat{y}_J)\tag{10}$$

其中 $\rho_z(\cdot,\cdot)$ 为复相关系数[14]，$\lambda(\cdot)$ 表示特征值.

证明 由(7)式可得

$$(X'X-X_J'X_J)^{-1}=(X'X)^{-1}+(X'X)^{-1}X_J'(I_k-H_J)^{-1}X_J(X'X)^{-1}$$

两边同时右乘 $X'Y=X'(J)Y(J)+X_J'Y_J$ 和 X_J' 分别得到

$$\hat{\beta}(J)+(X'(J)X(J))^{-1}X_J'Y_J=\hat{\beta}+(X'X)^{-1}X_J'(I_k-H_J)^{-1}\hat{Y}_J$$
$$(X'(J)X(J))^{-1}X_J'=(X'X)^{-1}X_J'(I_k+(I_k-H_J)^{-1}H_J)$$

注意到 $I_k+(I_k-H_J)^{-1}H_J=(I_k-H_J)^{-1}$，从以上两式我们容易推出

$$\hat{\beta}-\hat{\beta}(J)=(X'X)^{-1}X_J'(I_k-H_J)^{-1}\delta_J\tag{11}$$

从而 $HP_J(s)=|X_s'(X'X)^{-1}X_J'(I_k-H_J)^{-1}\delta_J|/\hat{\sigma}(J)$，而

$$\rho_z^2(X_s'\hat{\beta},X_J\hat{\beta})=\frac{\mathrm{Cov}(X_s'\hat{\beta},X_J\hat{\beta})\,\mathrm{Cov}^{-1}(X_J\hat{\beta},X_J\hat{\beta})\,\mathrm{Cov}(X_J\hat{\beta},X_s'\hat{\beta})}{\mathrm{Var}(X_s'(\hat{\beta})}$$
$$=X_s'(X'X)^{-1}X_J'H_J^{-1}X_J(X'X)^{-1}X_s/h_{ss}$$
$$=\lambda(H_J^{-1}X_J(X'X)^{-1}X_sX_s'(X'X)^{-1}X_J')/h_{ss}$$
$$>\lambda_{\min}(H_J^{-1})\lambda_{\min}(X_J(X'X)^{-1}X_sX_s'(X'X)^{-1}X_J')/h_{ss}$$
$$=X_s'(X'X)^{-1}X_J'X_J(X'X)^{-1}X_s/h_{ss}\lambda_{\max}(H_J)$$

故
$$HP_J^2(s)=\mathrm{tr}[(I_k-H_J)^{-1}\delta_J\delta_J'(I_k-H_J)^{-1}X_J(X'X)^{-1}X_sX_s'(X'X)^{-1}X_J']/\hat{\sigma}^2(J)$$
$$<k\lambda_{\max}[(I_k-H_J)^{-1}\delta_J\delta_J'(I_k-H_J)^{-1}]\lambda_{\max}[X_J(X'X)^{-1}X_sX_s'(X'X)^{-1}X_J']/\hat{\sigma}^2(J$$
$$=k\delta_J'(I_k-H_J)^{-2}\delta_JX_s'(X'X)^{-1}X_J'X_J(X'X)^{-1}X_s/\hat{\sigma}^2(J)$$
$$<kh_{ss}\lambda_{\max}(H_J)\|(I_k-H_J)^{-1}\delta_J\|^2\rho_z^2(\hat{y}_s,\hat{y}_J)/\hat{\sigma}^2(J)$$

两边取平方根，即所欲证.

以上定理说明数据集仅对与之强相关的数据点处有较大影响，因而 Cook 距离

$$D_i(X'X,c)=\frac{1}{c}\sum_{s=1}^{n}(\hat{y}_s-\hat{y}_s(i))^2\tag{12}$$

中这些数据点处的影响较突出，基于这种考虑我们定义

$$HP_i=\frac{1}{c}\sum_{s=1}^{n}|\hat{y}_s-\hat{y}_s(i)|\tag{13}$$

作为影响度量，它把第 i 个数据对各点预测值的影响平等地加以考虑，因此比 Cook 距离更好地说明了第 i 组数据对回归分析影响的真实情况，同时，从稳健的角度出发，(13) 比 (12)

式具有更好的稳定性.

取 $c=n\hat{\sigma}(i)$ 时，$HP_i=\frac{1}{n}\sum_{s=1}^{n}HP_i(s)$，即我们用影响分布的平均值作为强影响诊断统计量，而Cook距离通常是影响分布的总平方和.

当 $k=1$ 时，由(8)式可以得到

$$HP_i=\frac{1}{n}\sum_{s=1}^{n}|h_{si}|(1-h_{ii})^{-1}\|\delta_i\|/\hat{\sigma}(i)$$

$$=(\frac{1}{n}\sum_{s=1}^{n}|h_{si}|)\frac{1}{h_{ii}\hat{\sigma}(i)}\|\delta_i\|\frac{h_{ii}}{1-h_{ii}} \tag{14}$$

记 $\mu_i=\frac{1}{n}\sum_{s=1}^{n}|h_{si}|$ 表示 H 的第 i 行的绝对平均，$P_i(X'X)=\frac{h_{ii}}{1-h_{ii}}$ 为 X_i' 的势(参见[15]，[16]，[12])，故(14)即

$$HP_i=\frac{\mu_i}{h_{ii}\hat{\sigma}(i)}\|\delta_i\|P_i(X'X) \tag{15}$$

当 $m=k=1$ 时，$HP_i=\frac{\mu_i}{\sqrt{1-h_{ii}}}|r_i^*|$，这里 $r_i^*=\frac{\delta_i}{\hat{\sigma}(i)\sqrt{1-h_{ii}}}$ 为Jackknife残差.

在考虑数据集的影响时，我们同样定义影响度量 HP_J 为分布列 $\{HP_J(s),\ s=1,2,\cdots n\}$ 的平均值，当 $m=1$ 时，由(11)式可得

$$HP_J=\frac{1}{n\hat{\sigma}(J)}\sum_{s=1}^{n}\|\hat{y}_s-\hat{y}_s(J)\|$$

$$=\frac{1}{n\hat{\sigma}(J)}\sum_{s=1}^{n}|X_J'(X'X)^{-1}X_J'(I_k-H_J)^{-1}\delta_J| \tag{16}$$

现在我们来看 μ_i，类似于Jackknife残差，对多元线性模型，仍记 $r_i^*=\frac{\delta_i}{\hat{\sigma}(i)\sqrt{1-h_{ii}}}$，则(15)可写为

$$HP_i=\frac{\mu_i}{\sqrt{1-h_{ii}}}\|r_i^*\| \tag{17}$$

而

$$\mu_i=\frac{1}{n}\sum_{s=1}^{n}|h_{si}|=\frac{1}{n}\sum_{s=1}^{n}\rho_H(\hat{y}_s,\hat{y}_i)\sqrt{h_{ss}h_{ii}} \tag{18}$$

所以 $HP_i=(\frac{1}{n}\sum_{s=1}^{n}\rho_H(\hat{y}_s,\hat{y}_i)\sqrt{h_{ss}})\|r_i^*\|P_i^{\frac{1}{2}}(X'X)$，同Cook距离比较，这个分解中，除势和残差外，还多了一项，这一项取决于其它点的位置(由 h_{ss} 度量)及与 $\hat{y}_i$ 的相关程度. 又由Cauchy-Schwarz不等式

$$\mu_i=\frac{1}{n}\sum_{s=1}^{n}|h_{si}|<\frac{1}{n}\sqrt{\sum_{s=1}^{n}h_{si}^2 n}=\sqrt{\frac{h_{ii}}{n}} \tag{19}$$

从而 $HP_i<\frac{1}{\sqrt{n}}\|r_i^*\|P_i^{\frac{1}{2}}(X'X)$，可见 HP_i 还与样本容量 n 有关，当 n 较大时，特别对大样本情形，影响分析失去了作用，这时也就无所谓强影响点了.

类似的，由Cauchy-Schwarz不等式，从(16)式可得

$$HP_J < \frac{1}{\sqrt{n}\hat{\sigma}(J)}\delta_J'(I_k - H_J)^{-1}H_J(I_k - H_J)^{-1}\delta_J$$

$$= \frac{1}{\sqrt{n}} D_J^{\frac{1}{2}}(X'X, \hat{\sigma}(J))$$

$$= \frac{w_J}{\sqrt{n}} \tag{20}$$

这里 $D_J(X'X, \hat{\sigma}(J))$，$w_J^2$是诊断强影响集的Cook距离和Welsch-Kuh统计量．(20)式说明 HP_J仍与样本容量 n 有关，如果 n 太大，就连数据集的影响都是有限的．

综上所述，HP_J统计量是一种平均影响，因此我们可以根据实际制定统一的标准临界值，当 HP_J超过这个标准值，就认为 J 具有强影响，而Cook距离在度量强影响时是相对的，这从(12)式也可看出，它表示了各点影响的总平方和，随着 n 的增大，Cook距离也会增大．因而，对不同的样本容量，实际中Cook距离的临界值也应有所不同，为解决这个问题只有取 $c = np\hat{\sigma}$等等．但还有另外的问题：对于模型的扰动(或误差)，D_J不如 HP_J稳定，这从(12)和(13)式看是明显的，对于模型的扰动造成LSE的扰动来讲，D_J比 HP_J灵敏，因而受到的影响也就较大，即 D_J抗干扰的能力不如 HP_J，也就是说 HP_J更稳定．

参 考 文 献

[1] Weisberg, S., Technometrics, 25(1983), 240—244.

[2] Critchley, F., Biometrika, 72(1985), 627—636.

[3] Campbell, N. A., Appli.Statist. 27(1978), 251—258.

[4] Martin, R.D., Yohai, V.J., Ann. Statist., 14(1986), 781—818.

[5] Radhakrishan, R., Kshirsagar, A. M., Commu. Statist., A 10(1981), 515—529.

[6] Cook, R. D., Weisberg, S., Technometrics, 22(1980), 495—508.

[7] Cook, R. D., Weisberg, S., Residuals and Influence in Regression, Chapman-Hall, New York, 1982.

[8] Andrews, D. F., Pregibon, D., J.Ray. Statist. Soc. B, 40(1978), 85—93.

[9] Draper, N. R., John, J. A., Technometrics, 23(1981), 21—26.

[10] 王松桂，数理统计与管理，6(1985)，38—49 和 1(1986)，40—48.

[11] Johnson, W., Geisser, S., J. Bmer.Statist. Assoc., 76(1983), 137—144.

[12] Cook, R. D., J. Ray. Statist. Soc. B., 48(1986), 133—169.

[13] Hotelling, H., Biometrika, 36(1936), 321—377.

[14] 王松桂，线性模型的理论及其应用，安徽教育出版社，1987.

[15] Weisberg, S., Applied Linear Regression, New York, 1980.

[16] Huber, P., Robust Statistics, New York, Wiley, 1981.

Distribution and Stability Measure of Influence

Yang Hu

(Chongqing Communication Institute)

Abstract

In this paper, we propose the influence distribution of data sets in the sample space, and show that data sets only have bigger influence in the data case which is strong correlation with, so it's more prominent that the influence of the part data in Cook-distance $D_i(X'X,c)=\frac{1}{c}\sum_{s=1}^{n}(\hat{y}_s-\hat{y}_s(i))^2$.

We propose a new influence measure, namely, $\frac{1}{c}\sum_{s=1}^{n}|\hat{y}_s-\hat{y}_s(i)|$, it's more rational and robust than $D_i(X'X,c)$.