Advances in Applied Mathematics 应用数学进展, 2020, 9(11), 2004-2009
Published Online November 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.911232

Hans汉斯

基于改进两参数估计的影响点检测

陈　菊，李　荣

贵州民族大学数据科学与信息工程学院，贵州　贵阳
Email: 1257358419@qq.com

收稿日期：2020年11月1日；录用日期：2020年11月18日；发布日期：2020年11月25日

摘　要

在改进两参数估计下对单个数据删除模型进行研究，通过对比删除某个观测值前后估计量的变化程度来度量相应观测值的影响程度，并由近似删除公式得到删除某个数据点前后改进两参数估计量间的关系；同时，在前人的基础上推导得到DFFITS统计量和Cook统计量新的表达形式，并在实例中用两种统计量来识别影响点，验证其合理性。

关键词

改进两参数估计，数据删除模型，近似删除公式，Cook统计量，影响点

Influence Points Detection Based on Modified Two-Parameter Estimator

Ju Chen, Rong Li

School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang Guizhou
Email: 1257358419@qq.com

Received: Nov. 1st, 2020; accepted: Nov. 18th, 2020; published: Nov. 25th, 2020

Abstract

The single data deletion model is studied under the modified two-parameter estimator. The influence degree of the corresponding observation value is measured by comparing the change degree of the estimators before and after deleting a certain observation value, and the relationship between the improved two-parameter estimator value before and after deleting a certain data point is obtained by the approximate deletion formula; at the same time, new expressions of DFFITS statistics and Cook statistics are derived on the basis of predecessors, and two kinds of

文章引用: 陈菊, 李荣. 基于改进两参数估计的影响点检测[J]. 应用数学进展, 2020, 9(11): 2004-2009.
DOI: 10.12677/aam.2020.911232

statistics are used to identify the influence points in an example to verify their rationality.

Keywords

Modified Two-Parameter Estimator, Data Deletion Model, Approximate Deletion Formula, Cook Statistics, Influence Point

1. 引言

考虑一般线性回归模型：

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

其中 y 是 $n\times 1$ 的响应变量，X 是 $n\times p$ 的已知设计矩阵，β 为 $p\times 1$ 未知参数向量，ε 是均值为 0、协方差矩阵为 $\sigma^2 I_n$ 的 n 维随机误差向量，I_n 表示 n 阶单位矩阵。

回归诊断中的影响分析主要是研究观测值对回归模型中回归参数估计的影响。对线性模型(1)，第 i 个样本点对参数 β 估计的影响通常是指删除该样本点后模型参数 β 估计的变化情况，若删除该样本点后模型参数 β 估计的变化较大，则认为该样本点对模型参数 β 估计的影响较大。

记 $y_{(i)}$ 和 $X_{(i)}$ 分别表示从 y 和 X 中删除第 i 个样本值后的观测向量和设计矩阵，则模型(1)删除第 i 个样本值后可表示为：

$$y_{(i)} = X_{(i)}\beta + \varepsilon \tag{2}$$

第 i 个样本点对参数 β 估计的影响分析即对模型(1)和模型(2)所得参数 β 估计变化大小的比较分析。

考虑删除第 i 个样本点后参数 β 估计变化的总和，Cook 和 Weisberg [1]提出了以 Cook 距离，即

$$D_i = \frac{\left(\hat{\beta}-\hat{\beta}_{(i)}\right)'\left(X'X\right)\left(\hat{\beta}-\hat{\beta}_{(i)}\right)}{ps^2} = \left(\frac{e_i^2}{ps^2}\right)\left[\frac{h_{ii}}{\left(1-h_{ii}\right)^2}\right] \tag{3}$$

作为第 i 个样本点对最小二乘估计的影响度量。其中 e_i 表示残差 $e = y-\hat{y}$ 的第 i 个分量，h_{ii} 表示帽子矩阵 $H = X\left(X'X\right)^{-1}X'$ 主对角线上的第 i 个元素，$s^2 = e'e/n-p$ 为 σ^2 的估计。

Belsey [2]等基于删除第 i 个样本点后响应变量 y 的预测值变化情况，提出了以 DFFITS 统计量作为第 i 个样本点对最小二乘估计的影响度量，即

$$DFFITS_{(i)} = \frac{x_i\left[\hat{\beta}_{LS}-\hat{\beta}_{LS(i)}\right]}{SE\left(x_i\hat{\beta}_{LS}\right)} = \left[\frac{e_i}{s_{(i)}}\right]\left[\frac{h_{ii}^{1/2}}{1-h_{ii}}\right] \tag{4}$$

其中 $\hat{\beta}_{LS} = \left(X'X\right)^{-1}X'y$ 和 $\hat{\beta}_{LS(i)} = \left(X_{(i)}'X_{(i)}\right)^{-1}X_{(i)}'y_{(i)}$ 是分别由模型(1)和模型(2)所得的最小二乘估计。$SE\left(x_i\hat{\beta}_{LS}\right)$ 表示 $x_i\beta$ 标准误差的估计值，$s_{(i)}$ 表示模型(2)中 σ 的估计值。

Cook [3]基于置信椭球提出数据删除法以判断各个数据点对最小二乘估计的贡献。此种方法是通过对比删除某个观测值前后估计量的变化程度来度量相应观测值的影响程度。随后，Chatterjee 和 Hadi [4]不断地完善数据删除法，给出了判别异常点，高杠杆点和影响点的若干统计量。

当模型(1)存在复共线性时，最小二乘估计往往表现不稳定，此时再基于最小二乘估计进行影响分析显然不太合适。Belsey [2]等发现有偏估计下检测到的影响点不同于最小二乘估计下检测到的影响点。Walker 和 Birch [5]使用数据删除法检测了岭估计下的影响点，给出近似删除公式。Jahufer 和 Jianbao [6]研究了改进岭估计下用于度量影响大小的统计量随岭参数变化的情况。Jahufer [7]基于 Liu 估计给出用于度量影响大小的 DFFITS 和不同形式的 Cook 距离表达式。Ertas [8]等给出 Liu 估计和改进 Liu 估计的度量影响大小的统计量，并讨论了影响点的识别。Yasin 和 Murat [9]研究了两参数岭回归的影响诊断。Adewale 和 Kayode [10]通过对比删除某个数据点前后两参数估计的影响程度。

关于参数 β 的估计，考虑存在一个关于 β 的先验信息 b，一些学者结合其他有偏估计提出了一系列的改进估计，如 Swindel [11]提出的改进岭估计，Li 和 Yang [12]提出的改进 Liu 估计等。类似的，Adewale [13]结合先验信息 b 和两参数估计(Ozkale 和 Kachiranlar [14])提出了改进两参数估计(MTPE)

$$\hat{\beta}_{MTPE}=\left(X'X+kI\right)^{-1}\left(\left(X'X+kdI\right)\hat{\beta}_{LS}+k\left(1-d\right)b\right) \tag{5}$$

其中岭参数 $k>0$，Liu 参数 $0<d<1$。

针对线性模型存在复共线性的情形，考虑改进两参数估计可以视为其他许多有偏估计的推广，如当 $k=0$ 或 $d=1$ 时，为最小二乘估计；$d=0$ 和 $b=0$ 时，为岭估计；$d=0$ 时，为改进岭估计等，本文主要探讨样本点对改进两参数估计的影响。

近似删除公式与检验统计量

根据等式(4)，第 i 个样本点对 MTPE 的影响度量统计量 DFFITS 可写为：

$$DFFITS_{(i)}=\frac{x_i\left[\hat{\beta}_{MTPE}-\hat{\beta}_{MTPE(i)}\right]}{SE\left(x_i\hat{\beta}_{MTPE}\right)} \tag{6}$$

其中 $\hat{\beta}_{MTPE(i)}$ 表示模型(2)中未知参数 β 的改进两参数估计，分母是 $x_i\hat{\beta}_{MTPE}$ 的标准误差估计值。$SE\left(x_i\hat{\beta}_{MTPE}\right)=s\left(i\right)\sqrt{\sum_{j=1}^{n}h_{MTPEij}^2}$，其中 h_{MTPEij} 是 H 矩阵的第 ij (h_{ij})个元素。

由等式(3)，Cook 统计量可写成如下两个表达式：

$$D_i^{*}=\frac{1}{ps}\left[\hat{\beta}_{MTPE}-\hat{\beta}_{MTPE(i)}\right]'\left(X'X\right)\left[\hat{\beta}_{MTPE}-\hat{\beta}_{MTPE(i)}\right] \tag{7}$$

$$D_i^{**}=\frac{1}{ps^2}\left[\hat{\beta}_{MTPE}-\hat{\beta}_{MTPE(i)}\right]'\left(KN^{-1}X'XN^{-1}K\right)\left[\hat{\beta}_{MTPE}-\hat{\beta}_{MTPE(i)}\right] \tag{8}$$

其中 $K=X'X+kI$，$N=\left(X'X+kdI\right)+k\left(1-d\right)\left(X'X+kI\right)^{-1}\left(X'X+kdI\right)$，$D_i^{*}$ 是等式(3)的直接推广，D_i^{**} 是基于方差 $\mathrm{var}\left(\hat{\beta}_{MTPE}\right)=\sigma^2\left[K^{-1}N\left(X'X\right)^{-1}NK^{-1}\right]$ 给出。$\hat{\beta}_{MTPE}-\hat{\beta}_{MTPE(i)}$ 反映了第 i 组数据对回归系数 β_{MTPE} 的影响大小且是一个向量，不便比较大小，而 MTPE 不是比例不变的(X 矩阵没有第 i 行元素)，所以设计矩阵 X 须在计算之前重新缩放。因此，为了 $\hat{\beta}_{MTPE}-\hat{\beta}_{MTPE(i)}$ 能够比较大小，可通过近似删除公式实现。

根据模型(2)，$\hat{\beta}_{MTPE(i)}$ 可以写成：$\hat{\beta}_{MTPE(i)}=\left(X_{(i)}'X_{(i)}+kI\right)^{-1}\left[\left(X_{(i)}'X_{(i)}+kdI\right)\hat{\beta}_{LS(i)}+k\left(1-d\right)b\right]$。利用谢尔曼－莫里森－伍德伯里(SMW)定理(Rao [15])，$\hat{\beta}_{MTPE(i)}$ 可以近似为：

$$\hat{\beta}_{MTPE(i)}=\left(X'X+kI-x_i'x_i\right)^{-1}\left[\left(X'X+kdI-x_i'x_i\right)\hat{\beta}_{LS(i)}+k\left(1-d\right)b\right]$$

式中 $K=X'X+kI$。

$$\begin{aligned}\hat{\beta}_{MTPE(i)} &= \left(K - x_i'x_i\right)^{-1}\left[\left(X'X + kdI - x_i'x_i\right)\hat{\beta}_{LS(i)} + k\left(1-d\right)b\right] \\ &= \left(K^{-1} + \frac{K^{-1}x_i'x_iK^{-1}}{1-x_iK^{-1}x_i'}\right)\left[\left(X'X + kdI\right)\hat{\beta}_{LS(i)} - x_i'x_i\hat{\beta}_{LS(i)} + k\left(1-k\right)b\right] \\ &\cong \hat{\beta}_{MTPE} + \frac{K^{-1}x_i'}{1-m_{ii}}\left[\hat{y}_{MTPEi} - y_i + m_{ii}y_i - m_{ii}y_i\right] \\ &\cong \hat{\beta}_{MTPE} - \frac{e_{MTPEi}K^{-1}x_i'}{1-m_{ii}}\end{aligned} \tag{9}$$

根据等式(9)，等式(6)~(8)的近似形式可以写成：

$$DFFITS_{(i)} \cong \left[\frac{m_{ii}}{1-m_{ii}}\right]\frac{e_{MTPEi}}{SE\left(x_i\hat{\beta}_{MTPE}\right)} \tag{10}$$

$$D_i^* \cong \left[\frac{1}{ps^2}\right]\left[\frac{e_{MTPEi}}{1-m_{ii}}\right]^2 x_i\left(X'X+kI\right)^{-1}\left(X'X\right)\left(X'X+kI\right)x_i' \tag{11}$$

$$D_i^{**} \cong \left[\frac{1}{ps^2}\right]\left[\frac{e_{MTPEi}}{1-m_{ii}}\right]x_i\left(X'X+kI\right)^{-1}\left(X'X+kdI\right)\left(X'X\right)^{-1}\left(X'X+kdI\right)\left(X'X+kI\right)^{-1}x_i' \tag{12}$$

其中 $m_{ii} = x_iK^{-1}x_i'$。

2. 实证分析

案例数据来自文献 Longley [16]，回归模型(1)给出如下：

$$y = X\beta + \varepsilon$$

其中 $X = \left(x_1, x_2, x_3, x_4, x_5, x_6\right)$，$y$ 是总派生就业，x_1 是 GNP 隐含价格平减指数，x_2 是国民生产总值，x_3 是失业率，x_4 是武装力量的规模，x_5 是 14 岁及以上的非机构人口，x_6 是时间。$X'X$ 的条件数为 43，275 (Walker 和 Birch [5])。Hoerl 和 Kennard [17]提出岭参数的计算方法，并定义为 $K = \frac{s^2}{\hat{\beta}_{\max}^2}$。在本文中，$k$ 的值计算为 5.36488e−08，根据文献(Ullah [18]等，2013)取 $d = 0.9$，下面通过 k 和 d 的值分别计算 Cook 距离和 DFFITS，并通过它们来找出影响点。

Cook [3]使用数据删除法得到了最小二乘估计下的 Cook 统计量，将点 5、16、4、10 和 15 确定为影响点。Walker 和 Birch [5]用基于岭估计的数据删除法发现点 16、10、4、15 和 1 是影响点。Jahufer 和 Jianbao [6]在前人的基础上使用数据删除法得到了基于修正岭估计下的影响点，分别为 16、4、1、10 和 15 五个最有影响的观测值。Ullah [18]等计算当 d 值等于 0.9 时，liu 回归中影响点的顺序为 16、5、4、10 和 15。Yasin 和 Murat [9]通过基于两参数岭估计的影响点检测确定了观测值 16、10、6、1 和 4 为影响点。Adewale 和 Kayode [10]通过 DFFITS 准则确定了强影响点为 16、10、4、5 和 15，通过 D^* 检测到的 4、10、16、5 和 1 以及 D^{**} 检测到的 16、5、4、15 和 1 分别作为它们的五个最有影响的观测值。

表 1 显示，所提出的统计量 $DFFITS_{(i)}$ 识别出影响点与其他作者的相同，只是顺序不同。使用 D_i^* 和 D_i^{**} 检测出的影响点与 Cook [3]和 Ullah [18]等人的相同，只是顺序不同。

应用于 Hald 数据

实例数据来源于文献 Hald [19]，包括四个回归变量与十三个观测值，矩阵 $X'X$ 的条件数为 249.578 (Adewale 和 Kayode [10])，条件数表明，该模型具有较强的复共线性。Cook [3]、Yasin 和 Murat [9]使用

影响统计量检测影响点也用这一数据集。根据文献(Adewale 和 Kayode [10]) k 和 d 的值分别为 0.0076761 和 1.18495。Cook [3]的研究中将观察值 8、3、11、6 和 13 按此顺序作为影响点，Yasin 和 Murat [9]基于两参数岭估计中利用两种不同形式的 Cook 距离检测到的影响点分别为 8、11、10、3、6 和 8、11、10、6、13，通过 DFFITS 检测的最有影响的五个观测值分别是 8、11、6、10 和 13。本文提出的两种形式的 Cook 距离和 DFFITS 在实例 Hald 数据集中计算的结果如表 2 所示。

Table 1. The five most influential observed values detected by DFFITS and two versions of Cook distance (Longley)
表 1. DFFITS 和两个版本的 Cook 距离检测出的最有影响的五个观察值(Longley)

$DFFITS_{(i)}$		D_i^*		D_i^{**}	
case	value	case	value	case	value
5	0.4644	5	0.0386	16	0.0262
16	0.4070	16	0.0278	5	0.0193
4	0.3069	4	0.0159	4	0.0144
10	0.2983	10	0.0151	10	0.0132
15	0.2575	15	0.0111	15	0.0107

Table 2. The five most influential observed values detected by DFFITS and two versions of Cook distance (Hald)
表 2. DFFITS 和两个版本的 Cook 距离检测的最有影响的五个观察值(Hald)

$DFFITS_{(i)}$		D_i^*		D_i^{**}	
case	value	case	value	case	value
3	0.91514	8	0.00133	8	0.00142
8	0.58233	11	0.00055	3	0.00075
13	0.50369	3	0.00036	11	0.00060
11	0.34744	13	0.00031	13	0.00040
4	0.30572	4	0.00020	4	0.00022

表 2 结果显示，本文用 $DFFITS_{(i)}$， D_i^* 和 D_i^{**} 与 Yasin 和 Murat [9]用相同的统计量都确定 8、11 和 13 是影响点，只是顺序不同。与 Cook [3]研究中使用影响统计量检测出的影响点 3、8、11 和 13 相同，顺序不同。

3. 结束语

本文考虑了线性模型存在复共线性时影响点检测的问题。提出利用改进两参数估计进行诊断的新方法。利用 SMW 定理和改进两参数估计中的近似删除公式，得到了 DFFITS 和两种不同 Cook 距离的近似形式。用两个实例说明了这些影响度量统计量的性能。结果表明，所提出的影响度量方法在检测影响点方面与现有的方法有较强的竞争力。这些影响度量方法将会帮助从业者决定是否保留、删除或缩减有影响的数据点时，使用稳健估计在研究中确定。

参考文献

[1] Cook, R.D. and Weisberg, S. (1982) Residuals and Influence in Regression. Chapman and Hall, New York.

[2] Belsley, D.A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, New York. https://doi.org/10.1002/0471725153

[3] Cook, R.D. (1977) Detection of Influential Observation in Linear Regression. *Technometrics*, **19**, 15-18. https://doi.org/10.1080/00401706.1977.10489493

[4] Chatterjee, S. and Hadi, A.S. (1986) Influential Observations, High Leverage Points, and Outliers in Linear Regression. *Statistical Science*, **1**, 379-393. https://doi.org/10.1214/ss/1177013622

[5] Walker, E. and Birch, J.B. (1988) Influence Measures in Ridge Regression. *Technometrics*, **30**, 221-227. https://doi.org/10.1080/00401706.1988.10488370

[6] Jahufer, A. and Chen, J.B. (2009) Assessing Global Influential Observations in Modified Ridge Regression. *Statistics & Probability Letters*, **79**, 513-518. https://doi.org/10.1016/j.spl.2008.09.019

[7] Jahufer, A. (2013) Detecting Global Influential Observations in Liu Regression Model. *Open Journal of Statistics*, **3**, 5-11. https://doi.org/10.4236/ojs.2013.31002

[8] Ertas, H., Erisoglu, M. and Kaciranlar, S. (2013) Detecting Influential Observations in Liu and Modified Liu Estimators. *Journal of Applied Statistics*, **40**, 1735-1745. https://doi.org/10.1080/02664763.2013.794203

[9] Yasin, A. and Murat, E. (2016) Influence Diagnostics in Two-Parameter Ridge Regression. *Journal of Data Science*, **14**, 33-52.

[10] Lukman, A.F. and Ayinde, K. (2018) Detecting Influential Observations in Two-Parameter Liu-Ridge Estimator. *Journal of Data Science*, **16**, 207-218.

[11] Swindel, F.F. (1976) Good Ridge Estimators Based on Prior Information. *Communications in Statistics—Theory and Methods*, **5**, 1065-1075. https://doi.org/10.1080/03610927608827423

[12] Li, Y. and Yang, H. (2012) A New Liu-Type Estimator in Linear Regression Model. *Statistical Papers*, **53**, 427-437. https://doi.org/10.1007/s00362-010-0349-y

[13] Adewale, F., Lukman, A.F., Ayinde, K., Kun, S.S. and Adewuyi, E.T. (2019) A Modified New Two-Parameter Estimator in a Linear Regression Model. *Modelling and Simulation in Engineering*, **2019**, Article ID: 6342702. https://doi.org/10.1155/2019/6342702

[14] Ozkale, M.R. and Kaçiranlar, S. (2007) The Restricted and Unrestricted Two-Parameter Estimators. *Communications in Statistics—Theory and Methods*, **36**, 2707-2725. https://doi.org/10.1080/03610920701386877

[15] Rao, C.R. (1973) Linear Statistical Inference and Its Applications. *Biometrics*, **31**, 791. https://doi.org/10.2307/2529568

[16] Longley, J.W. (1967) An Appraisal of Least Squares Programs for Electronic Computer from the Point of View of the User. *Journal of American Statistical Association*, **62**, 819-841. https://doi.org/10.1080/01621459.1967.10500896

[17] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Non-Orthogonal Problems. *Technometrics*, **12**, 55-67. https://doi.org/10.1080/00401706.1970.10488634

[18] Ullah, M.A., Pasha, G.R. and Aslam, M. (2013) Assessing Influence on the Liu Estimates in Linear Regression Models. *Communications in Statistics—Theory and Methods*, **42**, 3100-3116. https://doi.org/10.1080/03610926.2011.620206

[19] Jowett, G.H. (1953) Statistical Theory with Engineering Applications. By A. Hald; Statistical Tables and Formulas. By A. Hald. *Journal of the Royal Statistical Society*, *Series A* (*General*), **116**, 87-88. https://doi.org/10.2307/2980953