文章编号：1002-1566(2017)04-0620-12
DOI: 10.13860/j.cnki.sltj.20170331-001

Logistic 回归模型的统计诊断

曾婕[1,2] 胡国治[2]

(1. 北京工业大学 应用数理学院，北京 100124；2. 合肥师范学院 数学与统计学院，安徽 合肥 230601)

摘要：统计诊断的主要任务就是通过诊断统计量检测已知观测数据在用既定模型拟合时的合理性，主要是找出数据当中的异常点或强影响点。本文主要研究 Logistic 回归模型的诊断统计量和诊断统计图。用牛顿迭代法给出 Logistic 回归模型的极大似然估计值，根据扰动模型得到传统的诊断统计量，结合残差、杠杆值和系数变化三者构造新的诊断统计量，绘制新的诊断统计图，通过模拟研究说明新的诊断统计量的有效性，最后用一个实际案例说明新的诊断方法的应用并进一步验证其优越性。

关键词：Logistic 回归模型；强影响点；扰动模型；诊断统计量；统计诊断图

中图分类号：O212 **文献标识码**：A

Statistical Diagnostics for Logistic Regression Model

ZENG Jie[1,2] HU Guo-zhi[2]

(1. College of Applied Science, Beijing University of Technology, Beijing 100124, China,
2. College of Mathematics and Statistics, Hefei Normal University, Anhui Hefei 230601, China)

Abstract: The main task of statistical diagnostics is using the diagnosis statistics to detect the rationality of fitting the data with established model, the most important is identifying outliers or influential points. In this paper, we focus on the diagnosis statistics and diagnosis graph of Logistic regression model. We discuss how to obtain the maximum likelihood estimates by Newton iterative method, then we can get the traditional diagnostic statistics based on perturbation model. In addition, we combine residuals, leverage value and the change in the value of the estimated coefficients to construct new diagnosis statistics and plot new diagnosis graph. Through simulation the effectiveness of the new diagnosis statistics can be verified. Finally, we cite an example to illustrate the application of new diagnostic method and to further validate its superiority.

Key words: logistic regression model, influential points, perturbation model, diagnosis statistics, diagnosis graph

0 引言

在二十世纪，Verhulst (1938)[1] 提出用 Logistic 函数作为增长曲线，有力地推动了人口统计学的发展。进入二十一世纪，Logistic 函数在经济学、医学等领域得到广泛应用，主要用于处理因变量为分类数据的情形。在实际中，数据可能会包含异常点和强影响点，而通常的极大似然估计法对于这些数据点非常敏感，使得拟合出的模型缺乏稳健性。因此在建立 Logistic 回归

收稿日期: 2015 年 12 月 22 日 **收到修改稿日期**: 2016 年 11 月 7 日
基金项目: 本文的研究受到合肥师范学院横向项目 (HX2016002) 的资助。

模型之后，需要利用统计诊断工具来检测模型的稳健性，确定模型的合理性。通过统计诊断，不仅可以检测出严重偏离既定模型的数据点即异常点，而且可以寻找出对统计诊断结果影响特别大的点，即强影响点。

统计诊断方法起源于线性回归模型，已有丰硕的成果。Cook (1977)[2] 提出了线性回归模型统计诊断方法，如 Cook (1979)[3] 统计量等。此后，大量学者尝试把此方法推广到非线性回归模型中。Pregibon (1981)[4] 利用对 Logistic 回归模型中对数似然函数加上扰动系数的方法寻找到强影响点，并把该方法推广到广义线性回归模型。为了直观地反映统计诊断过程，Landwehr 等 (1984)[5] 提出了一系列统计诊断图，如：Pearson 残差图、删除样本点对系数估计的影响图等。韦博成、林金官等 (2009)[6] 系统地总结了各种模型的统计诊断方法，其中包含对广义线性模型的统计诊断研究。本文总结现有 Logistic 回归模型统计诊断方法，并进一步进行研究。

1 Logistic 回归模型及其估计

1.1 模型简介

设随机变量 X 的分布函数为

$$F(x,\mu,\sigma)=\frac{1}{1+e^{-\frac{x-\mu}{\sigma}}},\qquad -\infty<x<\infty,$$

其中：$-\infty<\mu<\infty$，$\sigma>0$，则称 X 服从参数为 μ，σ 的一元 Logistic 分布，其中 μ 称为该分布的位置参数，σ 称为该分布的尺度参数，此时，X 的密度函数为

$$f(x,\mu,\sigma)=\frac{e^{-\frac{x-\mu}{\sigma}}}{\sigma(1+e^{-\frac{x-\mu}{\sigma}})^2},\qquad -\infty<x<\infty.$$

设变量 Y 为二分类的随机变量，$Y=1$、$Y=0$ 分别表示事件发生和不发生；利用二分类 Logistic 回归模型来刻画 Y 与协变量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_m)^T$ 之间的关系。二分类的 Logistic 回归模型的形式如下：

$$P(Y=1|\boldsymbol{x})=\frac{e^{g(\boldsymbol{x})}}{1+e^{g(\boldsymbol{x})}},\tag{1}$$

其中，

$$g(\boldsymbol{x})=\beta_1x_1+\beta_2x_2+\cdots+\beta_mx_m.$$

本文主要研究形如式 (1) 的二分类 Logistic 回归模型。

1.2 模型估计

设变量服从二项分布 $y\sim B(n,p)$。令 $\theta=\mathrm{logit}(p)=\log\frac{p}{1-p}$，则 y 的概率密度函数为

$$f(y;\theta)=\binom{n}{y}p^y(1-p)^{n-y}=\exp\big(y\theta-a(\theta)+b(y)\big),\qquad y=0,1,\cdots,n,$$

其中 $a(\theta)=n\log(1+e^\theta)$，$b(y)=\log\binom{n}{y}$，在本文中 $\log(\cdot)$ 都指 $\log_e(\cdot)$。相应的对数似然函数为 $l(\theta;y)$，

$$l(\theta;y)=y\theta-a(\theta)+b(y).$$

可以得到关于 θ 的对数似然方程

$$s(\theta;y)=\frac{\partial}{\partial\theta}l(\theta;y)=y-\dot{a}(\theta)=y-np.$$

若存在 N 个相互独立且服从二项分布的观测值 $y_i\sim B(n_i,p_i)$，这 N 个观测值的对数似然函数是

$$l(\boldsymbol{\theta};\boldsymbol{y})=\sum_{i=1}^{N}l(\theta_i,y_i)=\sum_{i=1}^{N}\big(y_i\theta_i-a(\theta_i)+b(y_i)\big). \tag{2}$$

若 Logistic 回归模型的数据中有 N 个协变类型，每个协变类型中因变量取值为 1 的次数 $y_i\sim B(n_i,p_i)$，又在 Logistic 回归模型中有如下关系：

$$\boldsymbol{\theta}=\operatorname{logit}(\boldsymbol{p})=\boldsymbol{X\beta},$$

其中 $\boldsymbol{X}=\{X_1,X_2,\cdots,X_m\}$。参照式 (2)，于是得到 m 维参数向量 $\boldsymbol{\beta}$ 的对数似然函数：

$$l(\boldsymbol{X\beta};\boldsymbol{Y})=\sum_{i=1}^{N}l(\boldsymbol{x}_i^T\boldsymbol{\beta};y_i)=\sum_{i=1}^{N}\big(y_i\boldsymbol{x}_i^T\boldsymbol{\beta}-a(\boldsymbol{x}_i^T\boldsymbol{\beta})+b(y_i)\big). \tag{3}$$

式 (3) 对应的极大似然方程是 $\partial l(\boldsymbol{X}\hat{\beta};\boldsymbol{Y})/\partial\hat{\beta}=\boldsymbol{0}$。即 β 的极大似然估计值满足下列方程：

$$\sum_{i=1}^{N}x_{ij}\big(y_i-\dot{a}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}})\big)=0,\qquad j=1,2,\cdots,m. \tag{4}$$

其中 a 上面有 k 个点代表 $\frac{\partial^k}{\partial\theta^k}a(\theta)$，记 $\boldsymbol{s}=\boldsymbol{y}-\dot{a}(\boldsymbol{x}_i\hat{\beta})=\boldsymbol{y}-n\hat{\boldsymbol{p}}$，则似然方程组式 (4) 可以写成矩阵形式

$$\boldsymbol{X}^T\boldsymbol{s}=\boldsymbol{X}^T(\boldsymbol{Y}-\hat{\boldsymbol{Y}})=\boldsymbol{0}. \tag{5}$$

式 (5) 虽然形式上与线性回归模型的似然方程一样，但是对 $\hat{\beta}$ 来说却不是线性的，故需使用 Newton 迭代法来估计 $\hat{\beta}$，得出 $-\partial\boldsymbol{X}^T\boldsymbol{s}/\partial\hat{\beta}=\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}$，其中 $\boldsymbol{V}=\operatorname{diag}\big(\ddot{a}(\boldsymbol{x}_i\hat{\beta})\big)=\operatorname{diag}\big(n_i\hat{p}_i(1-\hat{p}_i)\big)$，根据牛顿迭代法，得到迭代方程：

$$\beta^{(t+1)}=\beta^{(t)}+(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{s},\qquad t=0,\cdots,t_{(0)}. \tag{6}$$

上式中的 $\boldsymbol{V}$ 和 $\boldsymbol{s}$ 都是在 $\boldsymbol{\beta}^{(t)}$ 处估计出来的。式 (6) 在 $t=t_0$ 处收敛，得到 $\hat{\beta}=\beta^{(t_0)}$。同时可以计算出 $n_i\hat{p}_i$ 的值，记 $\hat{y}_i=n_i\hat{p}_i$，及 y_i 方差估计值 $v_{ii}=n_i\hat{p}_i(1-\hat{p}_i)$。

若记矩阵 $\boldsymbol{z}^{(t)}=\boldsymbol{X}\beta^{(t)}+\boldsymbol{V}^{-1}\boldsymbol{s}$，则

$$\begin{aligned}\beta^{(t+1)}&=\beta^{(t)}+(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{s}\\&=(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})\beta^{(t)}+(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{s}\\&=(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{X}\beta^{(t)}+\boldsymbol{V}^{-1}\boldsymbol{s})\\&=(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{z}^{(t)}.\end{aligned}$$

在收敛点处，$\boldsymbol{z}=\boldsymbol{X}\hat{\beta}+\boldsymbol{V}^{-1}\boldsymbol{s}$，所以得到 β 的极大似然估计值是

$$\hat{\beta}=(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{z}. \tag{7}$$

以上估计出 $\hat{\beta}$ 的方法叫作迭代再加权最小二乘法 (IRLS)。若假定响应变量是 $\boldsymbol{V}^{1/2}\boldsymbol{z}$，自变量为 $\boldsymbol{V}^{1/2}\boldsymbol{X}$，式 (7) 即为该线性回归模型的最小二乘估计。根据未知参数 β 估计的相似性，启发将线性回归模型的统计诊断方法扩展到 Logistic 回归模型的统计诊断。

2 Logistic 回归模型的统计诊断方法及相关诊断统计量

2.1 Logistic 回归模型的残差

与线性回归模型类似，残差和杠杆值这两个统计量也是对 Logistic 回归模型进行统计诊断的重要工具，但在线性回归模型中一个重要的假设是残差的方差不依赖于条件期望 $E(Y_i \mid X_i)$，而在 Logistic 回归模型中残差项的取值与因变量的取值有关，因此残差的方差是与 $E(Y_i \mid X_i)$ 有关的函数。

在线性回归模型中，残差的定义形式是唯一的，而对于 Logistic 回归模型，有两种残差定义的形式，一种是 Pearson 残差 [7]，另一种是 Deviance 残差 [7]。Pearson 残差 χ_i：

$$\chi_i = \frac{s_i}{\sqrt{v_{ii}}} = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p})}},$$

是观测概率与预测概率之差除以估计概率的二项分布标准差。由于每一个残差都被其标准误的近似估计所除，因此在模型正确且样本容量很大时，Pearson 残差应该近似服从标准正态分布。

Deviance 残差 d_i:

$$d_i = \pm\sqrt{2}\big(l(\hat{\theta}_i; y_i) - l(\boldsymbol{x}_i\hat{\beta}; y_i)\big)^{1/2} = \pm\sqrt{-2\big(y_i \ln\hat{p}_i + (1-y_i)\ln(1-\hat{p}_i)\big)};$$

上式中正负号取决于 $\hat{\theta}_i$ 与 $\boldsymbol{x}_i\hat{\beta}$ 的大小比较，如果因变量 $y_i = 1$，则有 $d_i = \sqrt{-2\ln\hat{p}_i}$；如果因变量 $y_i = 0$，则有 $d_i = -\sqrt{-2\ln(1-\hat{p}_i)}$。在大样本情况下，单个 Deviance 残差近似服从正态分布。当案例有较大的 d_i 值时说明模型不能较好地拟合此案例。因此，残差绝对值异常大的观测案例，便是一个异常值点。

2.2 Logistic 回归模型的帽子矩阵

Logistic 回归模型的帽子矩阵为

$$P = \boldsymbol{V}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{1/2}.$$

令 $\tilde{\boldsymbol{X}} = \boldsymbol{V}^{1/2}\boldsymbol{X}$，则 $P = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T$，与线性回归模型的帽子矩阵的形式一致，具有一切线性回归模型帽子矩阵的性质。主对角线上的元素 p_{jj}，$j = 1, \cdots, n$ 称为 Logistic 回归模型的杠杆值，它可以测量 x_j 到数据中心的距离，即第 j 个案例与其他案例离散程度。Logistic 回归中的杠杆值 p_{jj} 不仅取决于自变量的值，而且依赖于因变量的观测值。杠杆值的取值范围是 $[0, 1]$，其平均值是 m/N。当 $p_{jj} > 2m/N$ 时就认为是异常大，这个水平便是杠杆度平均值的 2 倍。同时得到投影矩阵

$$M = \boldsymbol{I} - P = \boldsymbol{I} - \boldsymbol{V}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{1/2}.$$

根据矩阵分块定理 [8]，可得到另一个有用的统计量 [4]：

$$\bar{P} = \boldsymbol{V}^{1/2}\bar{\boldsymbol{X}}(\bar{\boldsymbol{X}}^T\boldsymbol{V}\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}^T\boldsymbol{V}^{1/2},$$

其中 $\bar{\boldsymbol{X}}=[\boldsymbol{X};\boldsymbol{Z}]$。$\bar{p}_{jj}=p_{jj}+(\chi_j)^2/\chi^2$ 是 $\bar{P}$ 的对角元素 (证明见文 [9]), $0\leq\bar{p}_{jj}\leq 1$ 且 $ave(\bar{p}_{jj})=m/N$。所以，当 $\bar{p}_{jj}$ 接近于 1 时，$\bar{p}_{jj}$ 对应的那个观测值可能是由该模型拟合得不好的点 (较大的 χ_j^2)，或是离其他数据较远的点 (大的 p_{jj} 值)，或者包含此两种情形的点。于是，由该矩阵能产生一个有价值的诊断图是 $\frac{\chi_j^2}{\chi^2}$ 对 p_{jj} 的散点图，可由 $2ave(\bar{p}_{jj})$、$3ave(\bar{p}_{jj})$ 甚至是 $4ave(\bar{p}_{jj})$ 的等高线来寻找模型的强影响点。

2.3 基于扰动模型的诊断统计量

在前面引入的诊断统计量可以寻找出不能被模型拟合的点及对拟合有较大影响的点。但这些统计量并没有定量地刻画出特殊影响案例对模型拟合起的作用大小。在本节中，通过引入扰动模型来分别研究每个样本点对模型拟合的影响。

设一组服从二项分布的样本数据 $\{y_i:\ i=1,\cdots,N\}$，若样本包括 K 个协变量类型，且第 k 种协变类型中有 ω_k 个样本点，则这组样本数据的对数似然函数为

$$l(\theta;\boldsymbol{Y})=\sum_{k=1}^{K}\omega_k l(\theta;y_k)=\sum_{k=1}^{K}\omega_k\big(y_k-a(\theta)+b(y_k)\big).$$

若 $K=N$，那么对数似然函数为

$$l_\omega(\boldsymbol{X}\beta;\boldsymbol{Y})=\omega_i\sum_{i=1}^{N}l(\boldsymbol{x}_i\beta;y_i),\tag{8}$$

其中，$\omega_i=1$，$i=1,\cdots,N$。现定义

$$\omega_i=\begin{cases}\omega, & i=l,\\ 1, & \text{其他}.\end{cases}$$

其中，$0\leq\omega\leq 1$。称 ω 是第 l 个数据点对模型的扰动系数。令 $W=\mathrm{diag}(1,\cdots,\omega,\cdots,1)$，当 $\omega=1$ 时 W 是单位矩阵。通过式 (8) 得出 β 的极大似然估计值 $\hat{\beta}(\omega)$。极大似然方程组为

$$\sum_{i=1}^{N}x_{ij}\omega_i s_i=0,\qquad j=1,\cdots,m.\tag{9}$$

式 (9) 的矩阵形式是

$$\boldsymbol{X}^T W\boldsymbol{s}=\boldsymbol{0}.\tag{10}$$

l_ω 对 β_j 和 β_k 求混合偏导得到 $-\boldsymbol{X}^T\boldsymbol{V}^{1/2}W\boldsymbol{V}^{1/2}\boldsymbol{X}$。所以，利用牛顿迭代法得到迭代方程：

$$\beta^{(t+1)}(\omega)=\beta^{(t)}(\omega)+(\boldsymbol{X}^T\boldsymbol{V}^{1/2}W\boldsymbol{V}^{1/2}\boldsymbol{X})^{-1}\boldsymbol{X}^T W\boldsymbol{s}.\tag{11}$$

当 $\omega=1$ 时，W 是单位矩阵，此时式 (11) 化简为一般的 Logistic 回归模型求解 $\hat{\beta}=\hat{\beta}(1)$ 时的迭代式 (6)。为了得到式 (10) 的解，从 $\hat{\beta}=\hat{\beta}(1)$ 即式 (7) 开始，根据式 (11) 一步迭代，得到

$$\hat{\beta}^{(1)}(\omega)=(\boldsymbol{X}^T\boldsymbol{V}^{1/2}W\boldsymbol{V}^{1/2}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{1/2}W\boldsymbol{V}^{1/2}\boldsymbol{z}.\tag{12}$$

显然，式 (12) 等于设计矩阵是 $\boldsymbol{V}^{1/2}\boldsymbol{X}$，响应变量是 $\boldsymbol{V}^{1/2}\boldsymbol{z}$ 的线性扰动方程中的非迭代最小二乘解。

线性回归模型的方差加权扰动模型为

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2/\omega_i), \qquad i = 1, \cdots, N$$

其中 $\omega_i = \begin{cases} \omega, & i = l \\ 1, & \text{其他} \end{cases}$，且 $0 \le \omega \le 1$。模型的系数 $\beta(\omega)$ 的最小二乘估计值为

$$\hat{\beta}(\omega) = \hat{\beta} - \frac{1-\omega}{1+(\omega-1)p_{ll}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_l s_l.$$

其中 $\hat{\beta}$ 是一般线性回归模型 $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$，$\varepsilon_i \sim N(0, \sigma^2)$，$i = 1, \cdots, N$ 的解。

因为在 Logistic 回归模型中设计矩阵为 $\boldsymbol{V}^{1/2}\boldsymbol{X}$，所以代入上面定理的结论可得

$$\hat{\beta}^l(\omega) = \hat{\beta} - \frac{(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{x}_l^T s_l(1-\omega)}{\big(1-(1-\omega)p_{ll}\big)}.$$

$\omega = 1$ 时对应的是 Logistic 回归模型参数的估计值，故当 ω 逐渐减小至 0 的过程中，即使第 l 个观测值在拟合中起越来越小的作用。当 ω 逐渐减小时，系数估计值的变化较小，则表明第 l 个样本点对系数的估计作用较小，即对模型拟合影响较小。若 ω 的较小改变能使 $\hat{\beta}$ 产生较大的变化，那么第 l 个样本点就有可能是特殊影响案例。当 $\omega = 1$ 时，$\hat{\beta}^l(\omega) = \hat{\beta}$，而当 $\omega = 0$ 时即为删除第 l 个数据点后模型的系数估计值，此时

$$\hat{\beta}^l(0) = \hat{\beta} - \frac{(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{x}_l^T s_l}{1-p_{ll}},$$

于是删除第 l 个样本点后的系数估计的变化为：

$$\triangle\hat{\beta}_l = \hat{\beta} - \hat{\beta}^l(0) = \frac{(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{x}_l^T s_l}{1-p_{ll}},$$

因此可以得到 Logistic 回归模型的 Cook 统计量：

$$\triangle_l\hat{\beta} = (\triangle\hat{\beta}_l)^T(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})(\triangle\hat{\beta}_l). \tag{13}$$

Cook 统计量主要用于衡量删除第 l 个样本点后对模型估计系数的影响。

同理可得因删除第 l 个样本点导致的 Pearson 残差的减少量以及 D 残差的变化量：

$$\triangle_l\chi^2 = \frac{s_l^2}{1-p_{ll}}, \tag{14}$$

$$\triangle_l D = \frac{d_l^2}{1-p_{ll}}. \tag{15}$$

以上讨论的式 (13)、式 (14) 和式 (15) 式是 Logistic 回归模型的主要统计诊断量，借助这些统计量可以找到拟合不好的样本点 (大的 $\triangle_l\chi^2$ 值或大的 $\triangle_l D$ 值) 和对参数估计影响较大的样本点 (大的 $\triangle_l\hat{\beta}$ 值)。在实际分析中，通过绘制三个统计量的指标图来判定样本点的大致影响，同时由删除样本点前后模型的变化，最终确定强影响点。

2.4 修正的诊断统计量

根据式 (13)、式 (14) 和式 (15) 作出的诊断图初步判定强影响点，但判断效果不好，因为这三个图只分别讨论了各个案例对系数估计、Pearson 残差和 D 残差的影响，并没有将这三方

面的影响结合起来。因此，我们有必要找到更有效的 Logistic 回归模型的诊断统计量。考虑到删除第 l 个样本会对所有的 m 个系数都产生影响，所以我们设

$$\triangle_l^{(k)} = \frac{|\triangle \hat{\beta}_l^{(k)}|}{SE(\hat{\beta}^l(0)^{(k)})},$$

其中，$l = 1, \cdots, N$，$k = 1, \cdots, m$。$\triangle_l \hat{\beta}^{(k)}$ 表示的是 $\triangle_l \hat{\beta}$ 的第 k 个分量，而 $SE(\hat{\beta}^l(0)^{(k)})$ 表示 $\hat{\beta}^l(0)$ 的第 k 个分量的标准差。若令 $\triangle_l = \sum_{k=1}^{m} \triangle_l^{(k)}$，那么 $\triangle_l$ 衡量的实际上是第 l 个样本点对所有系数产生的整体影响。$\triangle_l$ 与 Cook 统计量同样是描述第 l 个样本点对模型系数的影响，都仅仅描述的是对模型拟合影响的一部分，试想我们能否以 $\omega_{1l} = p_{ll}/m$ 和 $\omega_{2l} = \chi_l^2/\chi^2$ 作为权重对 $\triangle_l$ 进行加权得到新的诊断统计量 $\triangle^{(1)}$ 和 $\triangle^{(2)}$，这样得到的新的统计量 $\triangle^{(1)}$ 结合了第 l 个样本点杠杆值的信息和对参数估计的信息，$\triangle^{(2)}$ 结合了第 l 个样本点残差的信息和对参数估计的信息，所以新的诊断统计量与 Cook 统计量相比含有更多有价值的信息，在诊断强影响点方面有更高的效率。得到的新的诊断统计量 $\triangle^{(1)}$ 和 $\triangle^{(2)}$，分别是：

$$\triangle^{(1)} = \omega_{1l}\triangle_l, \quad l = 1, \cdots, N; \qquad \triangle^{(2)} = \omega_{2l}\triangle_l, \quad l = 1, \cdots, N.$$

同时我们也可以想到将 $\triangle_l / \sum_{l=1}^{N} \triangle_l$ 作为权重对 $\triangle_l \chi^2$ 和 $\triangle_l D$ 进行加权处理，于是得到新的统计量：

$$\triangle^{(3)} = \frac{\triangle_l}{\sum_{l=1}^{N} \triangle_l} \triangle_l \chi^2, \qquad l = 1, \cdots, N;$$

$$\triangle^{(4)} = \frac{\triangle_l}{\sum_{l=1}^{N} \triangle_l} \triangle_l D, \qquad l = 1, \cdots, N.$$

$\triangle^{(3)}$、$\triangle^{(4)}$ 这两个新的统计量既结合了第 l 个样本对残差项的影响，又包含了第 l 个样本对系数估计产生的整体影响，大大提高了统计诊断的精度，能够更准确地找到数据当中的异常值点和强影响点。

由 2.2 节的帽子矩阵分析我们知道 $\bar{p}_{jj} = p_{jj} + \frac{\chi_j^2}{\chi^2}$，$\bar{p}_{jj}$ 综合了杠杆值以及残差的信息，根据这个统计量我们可以做出：$\frac{\chi_j^2}{\chi^2}$ 对 p_{jj} 的散点图，该图记为 $plot(\frac{\chi_j^2}{\chi^2}, p_{jj})$，这个图反应了样本的杠杆值和 Pearson 残差两方面的信息，根据这个图我们可以大致找出强影响点。于是我们想到利用 $plot(\frac{\chi_j^2}{\chi^2}, p_{jj})$ 和 $\triangle^{(1)}$、$\triangle^{(2)}$、$\triangle^{(3)}$、$\triangle^{(4)}$ 的指标散点图这五个诊断图来对 Logistic 回归模型当中的数据进行统计诊断应该会比传统的诊断图效率更高。

结合文章第 2 节我们讨论过的所有诊断统计方法和诊断统计量，可以总结出对 Logistic 回归模型做统计诊断的一般方法。首先根据已知的数据建立最终的主效应模型，然后计算出各个诊断统计量，为了更加直观地找出异常值点和强影响点，接着根据诊断统计量画出诊断统计指标图。一般情况下，$\triangle_l\chi^2$、$\triangle_l D$、$\triangle_l \hat{\beta}$ 的指标图是最常见的，结合之前改进的新的诊断统计量，还可以作出 $\frac{\chi_j^2}{\chi^2}$ 对 p_{ll}、$\triangle^{(1)}$、$\triangle^{(2)}$、$\triangle^{(3)}$ 和 $\triangle^{(4)}$ 这五幅更有效的诊断图。结合诊断图找出疑似的强影响点，根据模型删除疑似点前后的变化程度来分析该点对模型拟合的影响程度，从而最终确定疑似点是否为强影响点。

3 模拟分析

下面通过随机模拟 [10] 的方法来说明前面介绍的诊断图的有效性。考虑 Logistic 回归模型

$$\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

取 $\beta_0=-0.1$，$\beta_1=0.1$，$\beta_2=0.2$，$X_1\sim N(0,1)$，$X_2\sim N(0.5,1)$，根据给定模型及系数取值产生 200 个协变量及响应变量 y。现人工产生 2 个强影响点，将 56 号中 X_1 变成 -4，X_2 变成 3.8，128 号 X_1 变成 6.1，X_2 变成 -7.2。

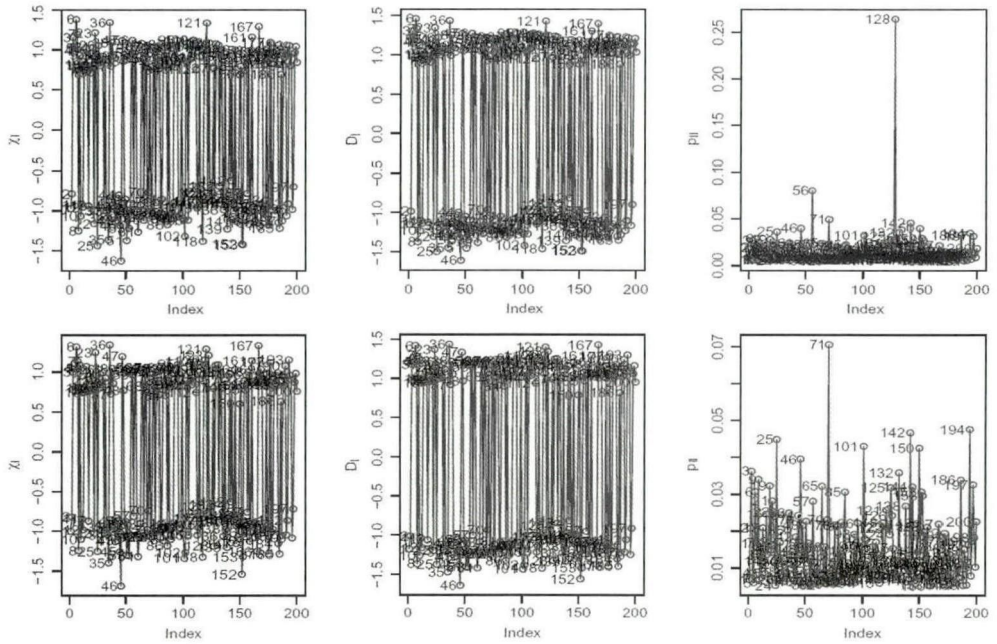

图 1 加入人工强影响点前后各样本点处 χ_l、D_l 及杠杆值 p_{ll} 变化图

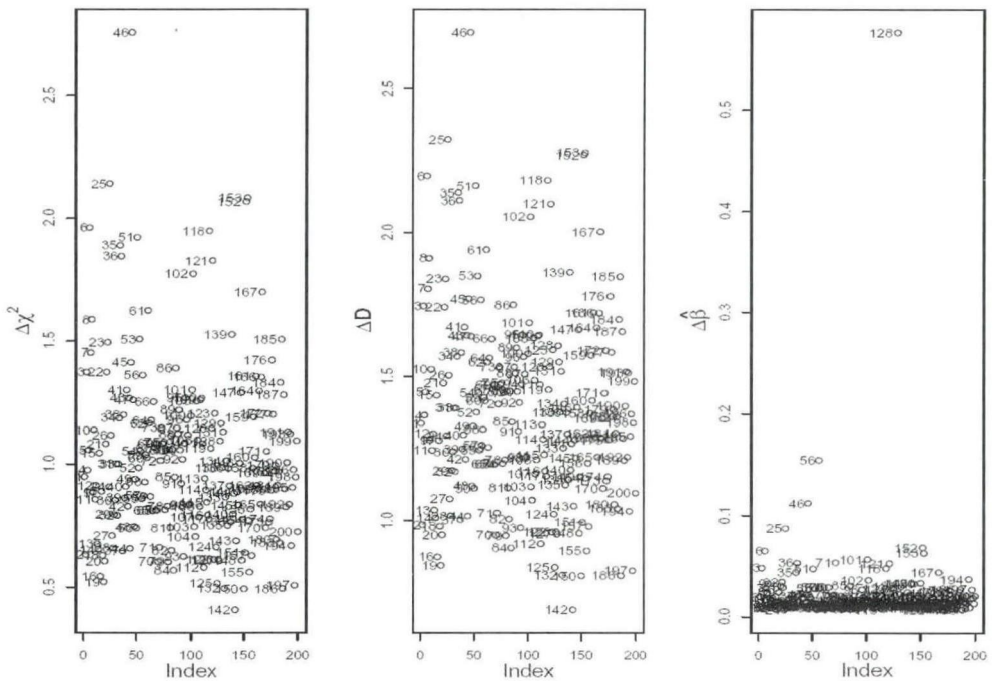

图 2 各样本点 $\triangle\chi^2$、$\triangle D$ 和 $\triangle\hat{\beta}$ 的指标散点图

在图 1 中，第一行的三个图表示引入两个强影响点后 χ_l、D_l 及杠杆值 p_{ll} 变化，第二行的三个图表示在原始数据下 χ_l、D_l 及杠杆值 p_{ll} 变化，可以看出 χ_l 图、D_l 图不能清晰诊断出

强影响点，而从 p_{ll} 变化图中可以直观地看出 56 号、128 号明显不同于其他样本点，且从原始数据中发现 71 号亦是一个强影响点。在图 2 中，$\triangle\chi^2$ 图、$\triangle D$ 图可以看出 46 号强影响点，从 $\triangle\hat{\beta}$ 中可以发现 128 号是强影响点。综合两张图发现诊断效果不高，现考虑用改进的诊断图来观察强影响点。

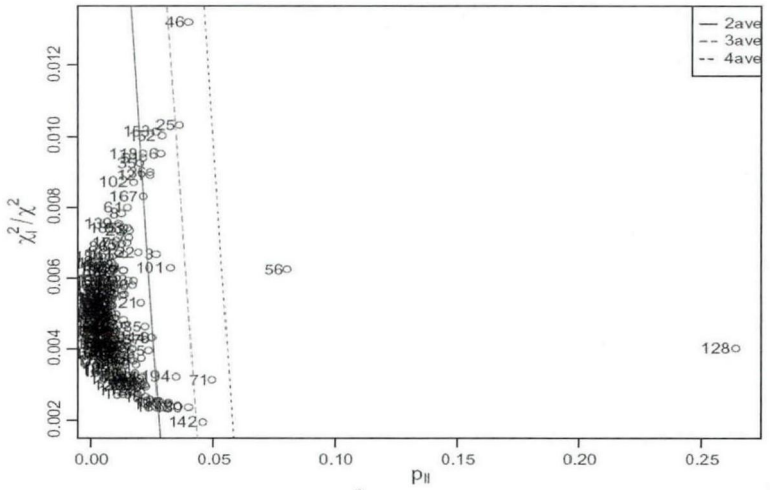

图 3 各样本点 $\frac{\chi_l^2}{\chi^2}$ 对 p_{ll} 的指标散点图

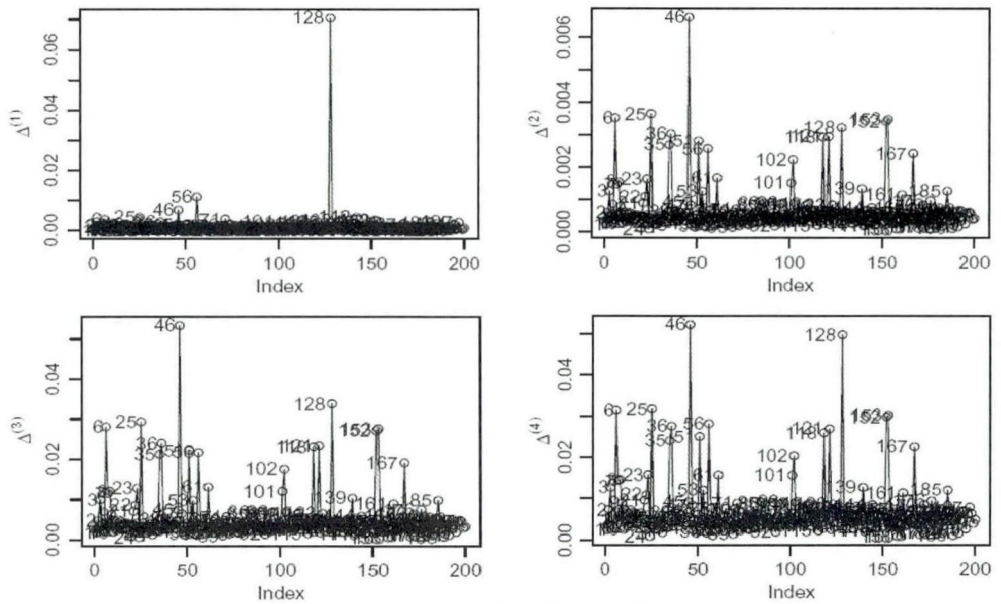

图 4 各样本点 $\triangle^{(1)}$、$\triangle^{(2)}$、$\triangle^{(3)}$ 和 $\triangle^{(4)}$ 指标散点图

根据 Logistic 回归模型中帽子矩阵部分内容可知 $\bar{p}_{jj} = p_{jj} + (\chi_j)^2/\chi^2$，于是可以画出 $\frac{\chi_j^2}{\chi^2}$ 对 p_{jj} 的指标散点图，为了更直观地找出 $\bar{p}_{jj}$ 值很大的点，在图中可用不同颜色的斜率为 -1 的虚线表示 $2ave(\bar{p}_{jj})$、$3ave(\bar{p}_{jj})$ 和 $4ave(\bar{p}_{jj})$ 等高线。一般我们关心的是超过 $3ave(\bar{p}_{jj})$ 的那些点，因为 $2ave(\bar{p}_{jj})$ 等高线以内的点认为是正常的点，介于 $2ave(\bar{p}_{jj})$ 和 $3ave(\bar{p}_{jj})$ 等高线之间的点认为是对模型影响程度较弱的点，而超过 $3ave(\bar{p}_{jj})$ 的点是对模型拟合影响程度较强的那些点，处于 $3ave(\bar{p}_{jj})$ 等高线右边越远位置的点认为是对模型影响越大的点，当然有可能是

对模型残差贡献比较大，也有可能是杠杆值比较大，两种情况下的点对模型拟合的影响都比较大。从图 3 中可以明显发现 25、46、71、142 号点是图形中位于 $3ave(\bar{p}_{jj})$ 等高线右边的点，尤其是 56、128 号点的取值更是超过了 $4ave(\bar{p}_{jj})$ 等高线。

图 4 是根据改造的诊断统计量画出的诊断统计图，从图中可看出 46、56、128 均为影响较大的点。因此结合图 3 和图 4 不仅能明显诊断出人工设定的两个强影响点 56 号和 128 号，还能找出数据中本身就存在的 46 号强影响点，所以通过模拟发现根据本文改进的诊断统计量作出的诊断图是更加有效的。

4 实例分析

4.1 Logistic 回归模型的建立

Baystate 医疗中心收集了 189 名孕妇的资料用以研究哪些变量会导致低出生体重 [11]。数据涉及到的指标分别为：(1) LBW 表示新生儿的体重是否为低出生体重，1 表示是低出生体重婴儿 (体重 <2500 克)，0 表示不是低出生体重婴儿 (体重 ≥ 2500 克)；(2) RACE 表示孕妇的种族；(3) AGE 表示孕妇的年龄；(4) LWT 表示妇女在最后一次月经期的体重。以 LBW 作为因变量，RACE、AGE 和 LWT 作为自变量建立 Logistic 回归模型，用 R 软件进行分析 [12]，得到最终的主效应模型为：

$$\text{logit}(\pi) = 0.756 - 0.533RACE + 0.037AGE - 0.011LWT.$$

该模型中的各个系数都是显著的，我们利用建立的主效应模型对样本中的数据进行统计诊断。

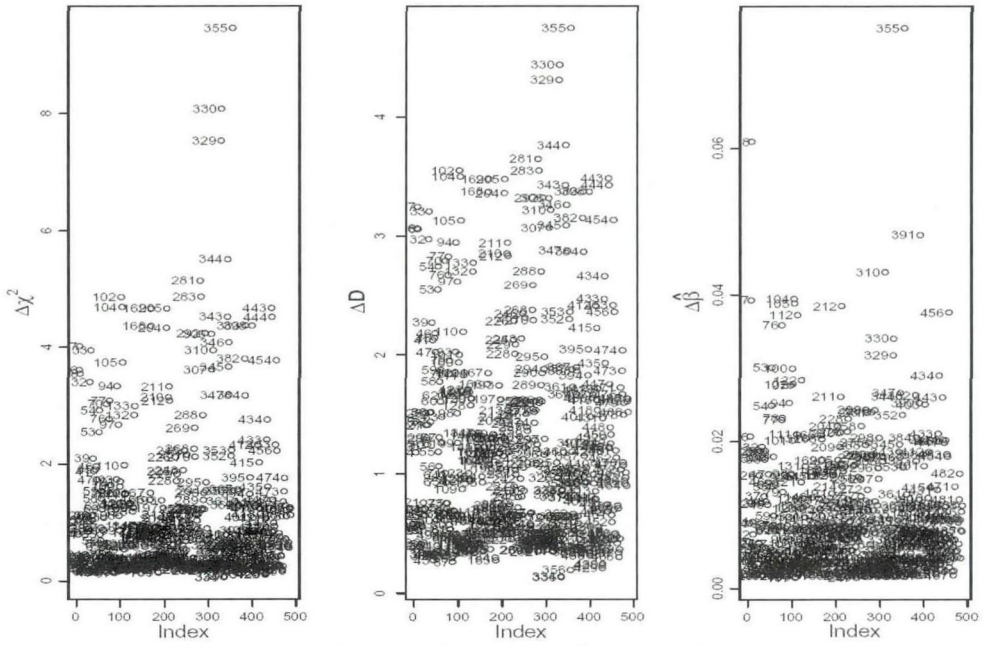

图 5 各样本点 $\triangle\chi^2$、$\triangle D$ 和 $\triangle\hat{\beta}$ 的指标散点图

4.2 模型的统计诊断分析

建立了最终的主效应模型之后，我们对模型中的数据进行统计诊断分析。不妨先试试用传统的诊断统计图进行分析会得出怎样的结论。我们对该例中涉及到的 189 个样本点进行编

号，分别计算出 189 个样本点对应的 $\triangle_l\hat{\beta}$、$\triangle_l\chi^2$ 和 $\triangle_l D$ 值，于是我们可以画出 $\triangle_l\hat{\beta}$、$\triangle_l\chi^2$ 和 $\triangle_l D$ 的指标图。

在图 5 中各个点旁边的数字就是各样本点的编号，根据之前的讨论我们知道 $\triangle\chi^2$、$\triangle D$ 和 $\triangle\hat{\beta}$ 的值越大的点就越有可能是强影响点。根据图 5 中的三个指标图的情况我们初步找出第 8、329、330、355 号点为疑似强影响点，因为这四个点不是对模型残差影响很大就是对系数的估计影响很大，对应的就是图中最顶端的几个点。这是根据传统的统计诊断图找出的几个疑似强影响点，下面我们针对该案例采用新的诊断统计图看看结果如何。

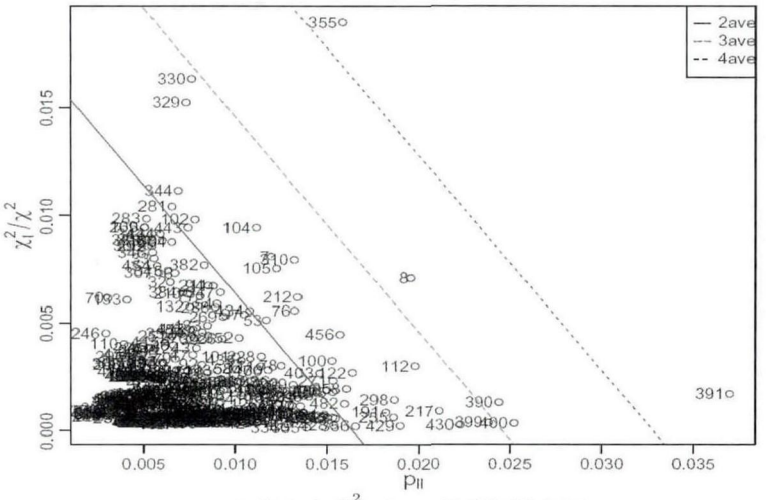

图 6 各样本点 $\frac{\chi_l^2}{\chi^2}$ 对 p_{ll} 的指标散点图

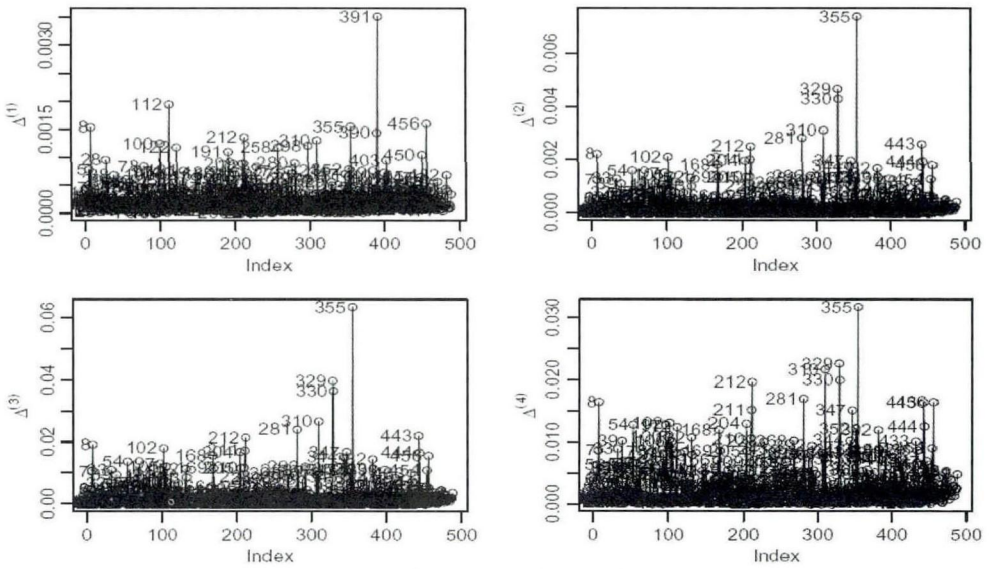

图 7 各样本点 $\triangle^{(1)}$、$\triangle^{(2)}$、$\triangle^{(3)}$ 和 $\triangle^{(4)}$ 指标散点图

图 6 是针对本案例当中的数据，画出的 $\frac{\chi_j^2}{\chi^2}$ 对 p_{jj} 的指标散点图。从图中看出 391、355、8 号点是图形中位于 $3ave(\bar{p}_{jj})$ 等高线右边的点，尤其是 391、355 号点的取值更是超过了 $4ave(\bar{p}_{jj})$ 等高线。其中 391 号点是杠杆值最大的点，355 号点是 Pearson 残差和杠杆值取值都较大的点，8 号点是杠杆值取值较大的点。这三个点都极有可能是模型的强影响点。刚才用传统的诊

断统计图找出的 330、329 号疑似强影响点在图 6 中虽然位于 $3ave(\bar{p}_{jj})$ 等高线以内，但是非常接近于 $3ave(\bar{p}_{jj})$ 等高线的位置，并且这两个点对应的 Pearson 残差值确实比较大，所以我们暂且也把这两个点列入疑似强影响点的行列，结合接下来要计算的 $\triangle^{(1)}$、$\triangle^{(2)}$、$\triangle^{(3)}$ 和 $\triangle^{(4)}$ 的取值情况再做定夺。

图 7 是根据各个样本点处 $\triangle^{(1)}$、$\triangle^{(2)}$、$\triangle^{(3)}$ 和 $\triangle^{(4)}$ 的取值画出的统计诊断图。结合这四幅图的情况，我们发现第 391、355、329、330 号样本点普遍在四个诊断统计量的取值相对于其他样本点来说都要大一些，而我们知道这四个统计量取值越大的点越有可能是强影响点，所以根据图 7 找出的强影响点是 391、355、329、330 号数据点。结合图 6 和图 7 发现数据当中的强影响点有 8、329、330、355、391 这五个样本点。找出的这五个强影响点对于模型拟合的影响是不一样的：8 号样本点对于模型估计系数的影响是第二大的；330 号和 329 号样本点对模型 Pearson 残差和 D 残差的影响程度分别位于所有样本点的第 2 和第 3 位；355 号样本点是 Pearson 残差和杠杆值取值都很大的点，不管是对模型估计系数还是模型残差的影响都是最大的；391 号样本点是所有数据点当中杠杆值最大的点，也就是说该点离其他数据点的距离是最远的。

我们对低出生体重的例子分别采用了传统的和改进的统计诊断图的方法，前者找出了 8、329、330、355 号这四个强影响点，后者找出了 8、329、330、355、391 这五个强影响点，并且事实证明 8 号点确实对模型系数估计有很大的影响。所以改进的统计诊断图有更高的效率，可以比传统的诊断统计图找出更多的强影响点。

5 总结

统计诊断是回归分析的一个重要内容。本文针对 Logistic 回归模型的统计诊断问题进行了深入的研究，既总结归纳了传统的 Logistic 回归模型的诊断统计量和诊断统计图，也提出了改进的诊断统计量和诊断统计图。根据模拟和实例分析，发现改进的诊断统计量确实比传统的诊断统计量效率更高，能够帮助我们更加方便准确有效地找出模型当中的强影响点。

[参考文献]

[1] Verhulst P J. Notice sur la lois que la population suit dans sons accroissement [J]. Corr. Math. Phys., 1938, 10: 113–121.

[2] Cook R D. Detection of influential observation in linear regression [J]. Technometrics, 1977, 19: 15–18.

[3] Cook R D. Influential observation in linear regression [J]. Journal of the American Statistical Association, 1979, 74: 169–174.

[4] Pregibon D. Logistic regression diagnostic [J]. The Annals of Statistics, 1981, 9(4): 705–724.

[5] Landwehr J M, Pregibon D, Shoemaker A C. Graphical methods for assessing logistic regression model [J]. Journal of the American Statistical Association, 1984, 79(285): 61–71.

[6] 韦博成, 林金官, 解锋昌. 统计诊断 [M]. 北京: 高等教育出版社, 2009: 169–197.

[7] 王济川, 郭志刚. Logistic 回归模型 — 方法与应用 [M]. 北京: 高等教育出版社, 2001: 195–196.

[8] Rao C R, Toutenburg H. Linear Model and Generalizations [M]. Berlin: Springer, 2008: 322–324.

[9] 谭宏卫, 曾婕. Logistic 回归模型的影响分析 [J]. 数理统计与管理, 2013, 32(3): 476–485.

[10] 戴琳送, 林金官. 广义泊松回归模型的统计诊断 [J]. 统计与决策, 2013, 21: 29–33.

[11] Hosmer D W, Lemeshow S. Applied Logistic Regression [M]. New York: John Wiley, 2000: 23–28. http://www-unix.oit.umass.edu/ statdata.

[12] 汤银才. R 软件与统计分析 [M]. 北京: 高等教育出版社, 2008: 265–271.