文章编号: 1002-1566 (2013) 03 - 0476 - 10

Logistic 回归模型的影响分析

谭宏卫[1] 曾捷[2]

(1. 贵州民族大学理学院, 贵州 贵阳 550025; 2. 北京工业大学应用数理学院, 北京 100124)

摘要: Logistic 回归模型的影响分析是 Logistic 回归诊断研究中的重要内容。常用的分析方法都是轮换地删除数据点后的逐步判断, 而这个判断的过程主要体现在模型的诊断图上。鉴于此, 通过构造诊断统计量来有效地开发诊断图成为影响分析的核心内容, 并由此能较为准确地探寻出模型的强影响点。本文通过对 Logistic 回归模型帽子矩阵的分解以及对轮换地删除数据点后的系数估计的相对变化量进行加权, 得出 Logistic 回归模型诊断图使其能比传统的诊断图更准确地判断出模型的强影响点。

关键词: Logistic 回归模型; 影响分析; 扰动分析; 诊断统计量; 诊断图

中图分类号: O212 **文献标识码**: A

Influence Analysis for Logistic Regression Model

TAN Hong-wei[1] ZENG Jie[2]

(1. School of Sciences, Guizhou University for Nationalities, Guizhou Guiyang 550025, China, 2. College of Applied Sciences, Beijing University of Technology, Beijing 100124, China)

Abstract: Influence analysis for logistic regression model is important content in the process of diagnosis study. The common method of analysis is stepwisely judgment when data points are alternately deleted and the judgment program is mainly reflected diagnosis charts of the model. In view of this, diagnosis charts of influence analysis are effectively developed through the construction of diagnostic statistics that is core content. At the same time, it can accurately find out the strong influential points of the model. This article has got diagnostic charts more accurate than conventional diagnostic charts for determining strong influential points of the model by decomposing the hat matrix for the logistic regression model and alternately deleting the data point estimates of the coefficients of relative changes weighting.

Key words: logistic regression model, influence analysis, perturbations analysis, diagnosis statistics, diagnosis graph

0 引言

近年来, Logistic 回归模型的应用相当广泛, 已经渗透到医学、经济学、生物学、犯罪心理学、工程技术学等领域。主要原因是: Logistic 回归模型是处理分类数据 (包括连续数据) 的有力工具, 且对解释变量几乎没有任何限制。本文第一部分特简要介绍 Logistic 回归模型。建立 Logistic 回归模型与其他传统模型一样, 主要有两个目的: (1) 用模型去挖掘数据所隐含的内在信息, 以及用模型去衡量解释变量与响应变量的相依关系; (2) 预测或为决策者提供某些

收稿日期: 2012 年 8 月 28 日 **收到修改稿日期**: 2013 年 1 月 2 日
基金项目: 2011 年贵州民族学院学生科研基金资助。

先验信息，作出较准确的决策。Logistic 回归模型要较准确地完成上述两个目的，模型的稳健性必不可少。而 Logistic 回归模型的影响分析正是通过一系列技术去探究影响模型稳健性的一些观测数据即强影响点。模型强影响点的主要特征是：数据点对模型的估计系数、拟合优度及残差产生较大的影响，进而影响模型的整体效果。

影响分析是回归诊断研究中的重要过程。Logistic 回归模型的诊断主要是在线性回归模型的诊断基础上发展和提出的。Cook (1977)[1] 提出了一系列的线性回归模型诊断方法，其中最为著名的是 Cook (1979)[2] 统计量。此后，许多学者直接将 Cook 的诊断思想移入非线性模型中，并取得良好效果。Cook 和 Weisberg (1980)[3] 提出一种泛型 (不管回归模型的类型) 的经验影响函数，这个函数只是理论上能达到寻找强影响点的目的，但很难实施，运算量过大，且效果并不太好。Pregibon (1981)[4] 在线性模型的基础上，利用扰动原理探究 Logistic 回归模型的强影响点；与此同时，Pregibon 利用扰动方程推出了 Logistic 回归模型的诊断统计量与 Cook 统计量形式上几乎一致；不仅如此，Landwehr 和 Pregibon 等 (1984)[5] 还提出了一系列有价值的诊断统计图。如指标图、杠杆值对 Pearson 残差图、删除数据的系数影响图等，使强影响点达到可视化的效果。近年来，对 Logistic 回归模型的诊断研究相对较少，Sugata (2008)[6] 考虑 Logistic 回归模型删除数据后对各个系数估计分量的绝对影响来探索强影响点，但其缺陷在于：轮换地删除数据点后并没有从整体上考虑其影响程度 [7−8]。韦博成，林金官等 (2009)[9] 讨论了广义线性模型的回归诊断，但 Logistic 回归模型的回归诊断还欠进一步地系统化。本文第二、三部分将阐述如何利用绝对加权的方法从整体上衡量删除强影响点后的影响程度，并用图形展示其功效。最后，文章的第四部分用《米其林指南》的数据进行实证研究来验证其优越性。

1 Logistic 回归模型及其估计

1.1 模型简介

设 y 为一个二分类反应变量，常用 $y=1$ 表示某研究事件发生 (或暴露)，用 $y=0$ 表示未发生 (或未暴露)；$\boldsymbol{x}=(x_1,x_2,\cdots,x_p)'$ 为相应的 p 维解释变量。考虑概率 $P(y=1|\boldsymbol{x})$，表示在给定 x 的条件下 $y=1$ 的概率，并记 $\pi(\boldsymbol{x})=P(y=1|\boldsymbol{x})$。对这个条件概率常用标准 Logistic 分布的分布函数来描述。设函数 $g(\boldsymbol{x})=\beta_0+\beta_1x_1+\cdots+\beta_px_p$，于是有

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

其中，$\boldsymbol{\beta}=(\beta_0,\beta_1,\cdots,\beta_p)'$，称 (1) 式为多元 Logistic 回归模型，$g(\boldsymbol{x})$ 为 Logistic 回归模型的 logit 转换，即有

$$g(\boldsymbol{x})=\operatorname{logit}\big(\pi(\boldsymbol{x})\big)=\log\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}.$$

若模型有 n 个观测数据 $\boldsymbol{x}_i=(x_{i1},x_{i2}\cdots,x_{ip})'$，$i=(1,2,\cdots,n)$；则

$$\boldsymbol{X}=\begin{pmatrix}1 & \boldsymbol{x}_1'\\ 1 & \boldsymbol{x}_2'\\ \cdots & \cdots\\ 1 & \boldsymbol{x}_n'\end{pmatrix}_{n\times(p+1)}$$

为 Logistic 回归模型的设计矩阵。

1.2 模型估计

由 y 是二分类反应变量，且在 $\boldsymbol{x}$ 处的概率为 $\pi(\boldsymbol{x})$，则 y 在 x 给定的条件下服从均值为 $\pi(\boldsymbol{x})$ 的两点分布。于是 $\boldsymbol{\beta}=(\beta_0,\beta_1,\cdots,\beta_p)'$ 的似然函数为

$$l(\boldsymbol{\beta})=\prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

其中 $\pi_i=P(y_i=1|\boldsymbol{x}_i)$ 表示在给定 $\boldsymbol{x}_i$ 的条件下 $y_i=1$ 的概率。令 $L(\boldsymbol{\beta})=\log l(\boldsymbol{\beta})$，用其建立似然方程

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}=\sum_{i=0}^{n}x_{ij}(y_i-\pi)=0, \tag{2}$$

其中 $j=1,\cdots,p$，$x_{0j}=1$，于是 (2) 式的矩阵表达式为

$$\boldsymbol{X}'\hat{\boldsymbol{e}}=0, \tag{3}$$

其中 $\hat{\boldsymbol{e}}=(\hat{e}_1,\hat{e}_2,\cdots,\hat{e}_n)'$，$\hat{e}_i=y_i-\hat{\pi}_i$，$i=1,\cdots,n$，且 $\hat{\pi}_i$ 是 π_i 的极大似然估计 (MLE)，并称 (3) 式为 Logistic 回归模型的正则方程。利用 Newton-Raphson 算法及重复加权的最小二乘估计得 (3) 式的数值解为

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}\boldsymbol{Z}.$$

其中 $\boldsymbol{Z}=\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{V}^{-1}\hat{\boldsymbol{e}}$，$\boldsymbol{V}=\mathrm{diag}\big(\hat{\pi}_1(1-\pi_1),\cdots,\hat{\pi}_n(1-\hat{\pi}_n)\big)$。

2 Logistic 回归模型的诊断统计量

Logistic 回归模型的残差及其杠杆值是回归诊断分析中的核心，几乎所有的诊断统计量都与这两个量有关。在线性模型中的学生化残差、Cook 统计量、Diffits 统计量等都与之密切相关。而在 Logistic 回归模型中的残差主要是 Pearson 残差和 Deviance 残差 [10]，对这两种残差的深刻理解与 Logistic 回归模型中的一个重要概念密切相关—协变类型；所谓协变类型是指模型中协变量不同值的特定组合。对于一个含有分类自变量的模型来说，数据所含的协变类型的数目是由每个分类自变量的取值所决定。

2.1 残差及帽子矩阵

2.1.1 Pearson 残差和 Deviance 残差

设 y_i 表示在第 i 个协变类型中 $y=1$ 发生的案例数，m_i 表示第 i 个协变类型的总案例数；由于 $y\sim B(1,\pi_i)$，于是有 $y_i\sim B(m_i,\pi_i)$。令

$$r(y_i,\pi_i)=\frac{y_i-m_i\pi_i}{\sqrt{m_i\pi_i(1-\pi_i)}}, \tag{4}$$

当模型满足 m- 渐近或 n- 渐近 [10] 时，$r(y_i,\pi_i)$ 是渐近服从 $N(0,1)$；而 $\hat{\pi}_i$ 是 π_i 的 MLE，将其代入 (4) 式，有

$$r_i=r(y_i,\pi_i)=\frac{y_i-m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}}.$$

称 r_i 为 Logistic 回归模型的 Pearson 残差。显然，当 Logistic 回归模型对数据拟合比较合理时，则 $\hat{\pi}_i$ 是 π_i 的最小方差无偏估计，于是 $r(y_i,\hat{\pi}_i)$ 也渐近服从 $N(0,1)$，统计量

$$\chi^2=\sum_{i=1}^{I}r_i^2. \tag{5}$$

渐近服从自由度为 $I-p-1$ 的 χ^2 分布，其中 I 为模型的协变类型总数。称 (5) 式为 Pearson χ^2 统计量。

设 l_f 表示拟合模型的似然值，l_s 表示相应的饱和模型的似然值，则偏差 (Deviance)

$$D=2(\log l_s-\log l_f).$$

于是，Deviance 残差就定义为 $d=\pm\sqrt{D}$，d 的 $\pm$ 与 $y_i-m_i\hat{\pi}_i$ 的 $\pm$ 是一致的。可以证明：第 i 个协变类型的 Deviance 残差为

$$d(y_i,\hat{\pi}_i)=\pm\Big(2\Big(y_i\log\frac{y_i}{m_i\hat{\pi}_i}+(m_i-y_i)\log\frac{m_i-y_i}{m_i(1-\hat{\pi}_i)}\Big)\Big)^{1/2}.$$

与此同时，可以验证 Deviance 残差与 Pearson 残差有相似的性质。

2.1.2 Logistic 回归模型的帽子矩阵及其分解

在线性模型中的帽子矩阵 $\mathbb{H}=\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$，投影矩阵为 $\mathbb{M}=I-\mathbb{H}$，其中 $\mathbb{X}$ 为线性模型的设计矩阵。而 Logistic 回归模型的帽子矩阵 [4] 为

$$H=\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{V}^{\frac{1}{2}}. \tag{6}$$

令 $\boldsymbol{X}=\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{X}$，则 $H=\widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'$ 为线性模型的帽子矩阵，其中 $\widetilde{\boldsymbol{X}}$ 为该模型下的设计矩阵。于是得出：(6) 式中的 H 同样具有线性模型的帽子矩阵 $\mathbb{H}$ 的一切性质，且主对角线上的元素 $h_{jj},j=1,\cdots,n$ 亦称为 Logistic 回归模型的杠杆值。

引理 设帽子矩阵 $H(\boldsymbol{X})=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$，若设计矩阵 $\widetilde{\boldsymbol{X}}$ 能按列分解为 $\boldsymbol{X}=(\boldsymbol{X}_1;\boldsymbol{X}_2)$，则 $H(\boldsymbol{X})=H(\boldsymbol{X}_1)+H(M(\boldsymbol{X}_1)\boldsymbol{X}_2)$，其中 $M(\boldsymbol{X}_1)=I-H(\boldsymbol{X}_1)$。

证明见 [11]。

定理 若 $\overline{H}=\boldsymbol{V}^{\frac{1}{2}}\overline{\boldsymbol{X}}(\overline{\boldsymbol{X}}'\boldsymbol{V}\overline{\boldsymbol{X}})^{-1}\overline{\boldsymbol{X}}\boldsymbol{V}^{\frac{1}{2}}$，$\overline{\boldsymbol{X}}=(\boldsymbol{X};\boldsymbol{Z})$，则有 $\overline{h}_{jj}=h_{jj}+\frac{r_j^2}{\chi^2}$，其中 $\overline{h}_{jj}$ 为 $\overline{H}$ 主对角线上的元素。

证明 设 $\overline{\mathbb{X}}=\boldsymbol{V}^{\frac{1}{2}}\overline{\boldsymbol{X}}$，则有 $\boldsymbol{V}^{\frac{1}{2}}\overline{\boldsymbol{X}}=(\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{X};\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{Z})=(\widetilde{\boldsymbol{X}};\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{Z})$，于是由引理得

$$\overline{H}=H(\overline{\mathbb{X}})=H(\widetilde{\boldsymbol{X}})+H\big(M(\widetilde{\boldsymbol{X}})\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{Z}\big).$$

即有

$$\overline{H}=H+H\big(M(\widetilde{\boldsymbol{X}})V^{\frac{1}{2}}\boldsymbol{Z}\big).$$

而由 $\boldsymbol{Z}=\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{V}^{-1}\hat{\boldsymbol{e}}$，得

$$\boldsymbol{Z}-\boldsymbol{X}\hat{\boldsymbol{\beta}}=\big(I-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}'V\big)\boldsymbol{Z}=\boldsymbol{V}^{\frac{1}{2}}M\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{Z}.$$

其中 $M=M(\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{X})=M(\widetilde{\boldsymbol{X}})$，于是 $\boldsymbol{V}^{\frac{1}{2}}\hat{\boldsymbol{e}}=M(\widetilde{\boldsymbol{X}})\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{Z}$；令 $\boldsymbol{r}=\boldsymbol{V}^{\frac{1}{2}}\hat{\boldsymbol{e}}$，即 $\boldsymbol{r}$ 为模型的 Pearson 残差向量，则有

$$H\big(M(\widetilde{\boldsymbol{X}})\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{Z}\big)=H(\boldsymbol{r})=\boldsymbol{r}(\boldsymbol{r}'r)^{-1}\boldsymbol{r}'.$$

于是得 $\overline{H}=H+H(\boldsymbol{r})$, 即有 $\overline{h}_{jj}=h_{jj}+\frac{r_j^2}{\chi^2}$。 □

由定理可知当 $\overline{h}_{jj}$ 较大时, 则第 j 个数据点有可能是一个强影响点; 于是由定理能产生一个有价值的诊断图是 h_{jj} 对 $\frac{r_j^2}{\chi^2}$ 的散点图, 设 $\overline{h}_{jj}$ 的平均值用 $ave(\overline{h}_{jj})$ 表示, 显然 $ave(\overline{h}_{jj})=\frac{p+2}{n}$。可由 $2ave(\overline{h}_{jj})$, 和 $3ave(\overline{h}_{jj})$ 甚至是 $4ave(\overline{h}_{jj})$ 的等高线来寻找模型的强影响点。

2.2 Logistic 回归模型的扰动分析及诊断统计量

在诊断分析中, 当数据点发生扰动变化时, 其整个系统应发生怎样的变化; 现考虑第 j 个数据点对模型的影响程度, 则定义

$$w_i=\begin{cases} w, & i=j, \\ 1, & \text{其他}, \end{cases} \tag{7}$$

其中 $i=1,\cdots,n$; $w\in[0,1]$, 称 (7) 式中的 w 为第 j 个数据点对模型的扰动系数 [4]。令 $W=\mathrm{diag}(1,\cdots,w,\cdots,1)$, 即当 $w=1$ 时 W 是单位矩阵; 则 Logistic 回归模型相应的正则方程为: $\boldsymbol{X}'\boldsymbol{W}\hat{\boldsymbol{e}}=0$, 同样由 Newton-Raphson 算法得模型系数估计的数值解为

$$\hat{\boldsymbol{\beta}}_j(w)=(\boldsymbol{X}'\boldsymbol{V}^{\frac{1}{2}}W\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{\frac{1}{2}}W\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{Z}, \tag{8}$$

$\hat{\boldsymbol{\beta}}_j(w)$ 表示在第 j 个数据发生扰动时模型相应的系数估计, 而由数据点未发生扰动时的系数估计 $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}\boldsymbol{Z}$, 则有

$$\hat{\boldsymbol{\beta}}_j(w)=\hat{\boldsymbol{\beta}}-\frac{(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}x_j\hat{e}_j(1-w)}{1-(1-w)h_{jj}}, \tag{9}$$

对 (9) 式的证明见 [4]。当 w 在 0 到 1 之间变化时, 由 (8) 式及 (9) 式都可得出第 j 个数据扰动下的模型系数估计, 显然当 $w=1$ 时 $\hat{\boldsymbol{\beta}}_j(w)=\hat{\boldsymbol{\beta}}$, 而当 $w=0$ 时即为删除第 j 个数据点后模型的系数估计为

$$\hat{\boldsymbol{\beta}}_j(0)=\hat{\boldsymbol{\beta}}-\frac{(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}x_j\hat{e}_j}{1-h_{jj}},$$

于是有

$$\Delta\hat{\boldsymbol{\beta}}_j=\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_j(0)=\frac{(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}x_j\hat{e}_j}{1-h_{jj}},$$

$\Delta\hat{\boldsymbol{\beta}}_j$ 表示去掉第 j 个数据点后的系数估计变化量, 将 $\Delta\hat{\boldsymbol{\beta}}_j$ 作二次型变换得

$$\Delta_j\hat{\beta}=\Delta\hat{\boldsymbol{\beta}}_j'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})\Delta\hat{\boldsymbol{\beta}}_j. \tag{10}$$

$\Delta_j\hat{\beta}$ 称为 Logistic 回归模型的 Cook 统计量, 它衡量的是删除第 j 个数据点后对模型估计系数的影响; 通过扰动系数对拟合值影响进而对似然值产生影响, 由此亦可得出删除第 j 个点后对拟合值 $\hat{\pi}_j$ 产生影响的两个诊断统计量分别为

$$\Delta\chi_j^2=\frac{r_j^2}{1-h_{jj}}, \tag{11}$$

$$\Delta D_j=\frac{d_j^2}{1-h_{jj}},\qquad d_j=d(y_j,\hat{\pi}_j). \tag{12}$$

上述的 (10), (11), (12) 式是 Logistic 回归模型主要的诊断统计量 [10]。在实际的诊断过程中, 并不是通过三个统计量的渐近分布来诊断强影响点的, 而是利用这三个统计量作诊断图判定第 j 个点的大致影响能力, 再从数理的角度进行比较分析, 最终确定模型的强影响点。

3 系统化的诊断方法

在实际应用中, 探寻 Logistic 回归模型的强影响点并不是从纯理论的角度出发去分析, 更多的是利用诊断图去直观的判定模型的强影响点, 然后利用删除该点后的比较分析最终确定模型的强影响点。鉴于此, 针对具体的模型去挖掘有价值的诊断图也是诊断分析中的重要方法。

3.1 Logistic 回归模型的诊断图

Logistic 回归模型的强影响点主要影响的是该数据点的残差和杠杆值, 进而影响其估计值 $\hat{y}$ 和估计系数 $\hat{\beta}$; 显然, Logistic 回归模型的诊断统计量都是以两者为中心所建立的如 (10), (11), (12) 式。于是根据数据影响方向的不同, 模型主要分为三类诊断统计量: (1) r_j, d_j, h_{jj} 为基本统计量; (2) 影响模型拟合值的统计量 $\Delta\chi_j^2$, ΔD_j; (3) 影响模型估计系数的统计量 $\Delta_j\hat{\beta}$。在基本统计量中, h_{jj} 表示第 j 个样本点到样本空间中心的一种比例距离, 即 $0 \le h_{jj} \le 1$。显然当 r_j, d_j, h_{jj} 都较大时 (一般大于其 2 倍均值), 第 j 个点有可能是一个强影响点。同理 $\Delta\chi_j^2$, ΔD_j, $\Delta_j\hat{\beta}$ 的大小同样可以度量其影响程度, 即三个量越大, 越有理由说明第 j 个点是强影响点, 于是能作出一些常用的诊断图:

(1) $\Delta\chi_j^2$, ΔD_j, $\Delta_j\hat{\beta}$ 的指标图;

(2) $\Delta\chi_j^2$, ΔD_j, $\Delta_j\hat{\beta}$ 对 $\hat{y}_j$ 的散点图;

(3) $\Delta\chi_j^2$, ΔD_j, $\Delta_j\hat{\beta}$ 对 h_{jj} 的散点图。

虽然这些常用的诊断图都能初步的判定强影响点, 但由三个平方型统计量得出的诊断图判定功效并不高, 初判也不够精确。因此, 探索 Logistic 回归模型的有效诊断图成为必然。

由帽子矩阵分解定理有 $\overline{h}_{jj} = h_{jj} + \frac{r_j^2}{\chi^2}$, 这个统计量综合了杠杆值及残差的信息, 由此作出诊断图: $\frac{r_j^2}{\chi^2}$ 对 h_{jj} 的散点图记为 $plot(\frac{r_j^2}{\chi^2}, h_{jj})$, 再利用 $2ave(\overline{h}_{jj})$, $3ave(\overline{h}_{jj})$ 和 $4ave(\overline{h}_{jj})$ 的等高线去确定强影响点。这个图较为全面的反映了数据的信息, 并且主要的诊断方向是拟合值 $\hat{y}_j$。针对系数估计方向上的诊断, 设

$$\Delta_j^{(k)} = \frac{|\Delta\hat{\boldsymbol{\beta}}_j^{(k)}}{SE(\hat{\boldsymbol{\beta}}_j(0)(k))},$$

其中 $j = 1, \cdots, n$; $k = 1, \cdots, p+1$。$\Delta\hat{\boldsymbol{\beta}}_j^{(k)}$ 表示的是 $\Delta\hat{\boldsymbol{\beta}}_j$ 的第 k 个分量, 而 $SE(\hat{\boldsymbol{\beta}}_j(0)^{(k)})$ 表示 $\hat{\boldsymbol{\beta}}_j(0)$ 的第 k 个分量估计的标准差; 则令 $\Delta_j = \sum_{k=1}^{p+1}\Delta_j^{(k)}$, 实际上 Δ_j 衡量的是第 j 个点对估计系数产生的整体影响, 于是通过对 Δ_j 加权来弥补 (10) 式中广义 Cook 统计量 $\Delta_j\hat{\boldsymbol{\beta}}$ 作诊断图所致的缺陷。选取 $w_{1j} = \frac{h_{jj}}{p+1}$ 和 $w_{2j} = \frac{r_j^2}{\chi^2}$ 为权, 于是得诊断统计量为

$$\Delta^{(1)} = \sum_{j=1}^{n} w_{1j}\Delta_j, \qquad \Delta^{(2)} = \sum_{j=1}^{n} w_{2j}\Delta_j.$$

而由 $\Delta^{(1)}$, $\Delta^{(2)}$ 得出的指标图, 也能比较准确的找到 Logistic 回归模型的强影响点; 于此, 在 Logistic 回归模型中, 利用 $plot(\frac{r_j^2}{\chi^2}, h_{jj})$ 及 $\Delta^{(1)}$, $\Delta^{(2)}$ 的指标图来综合判定强影响点优于传统意义上的诊断图。最后, 本文将在第四个部分中用实例来验证这个问题。

3.2 Logistic 回归模型的诊断步骤

(1) 建立 Logistic 回归模型的最终主效应模型 [10];

(2) 利用软件计算数据的协变类型, 进而求出 $\overline{h}_{jj}$, $ave(\overline{h}_{jj})$, $\Delta^{(1)}$, $\Delta^{(2)}$;

(3) 作诊断图 $plot(\frac{r_j^2}{\chi^2}, h_{jj})$, $plot(\Delta^{(1)})$, $plot(\Delta^{(2)})$, 并由图形初判模型的强影响点;

(4) 通过删除初判的强影响点, 来分析该点对模型的具体影响, 并建立比较分析表, 最终确定 Logistic 回归模型的强影响点。

4 实证分析及结论

4.1 数据说明及最终主效应模型的建立

MichelinNY[12] 是 2006 年由《米其林指南》对纽约各大餐馆的评级数据, 共 164 家参评餐馆。其数据的主要评定指标分别为

- Inmichelin 表示所评餐馆是否进入《米其林指南》, 用 Yes 和 No 分别表示进入与未进入;
- Food 为顾客评定的饮食质量指标;
- Decor 为餐馆的内外环境指标;
- Service 为餐馆的服务质量指标;
- Price 为用餐价格。

本节拟用 R 软件进行分析, 用 y 代替 Inmichelin 表示响应变量, 且用 $y=1$ 表示所评餐馆进入《米其林指南》, 否则 $y=0$。其余的指标为解释变量建立 Logistic 回归模型。经过反复的回归, 最终得到的主效应模型为

$$\begin{aligned}\text{logit}(\pi) = \quad & -63.76436 + 0.64274Food + 1.50597Decor + 1.12633Service \\ & +7.29827\log(Price) - 0.07613Decor \times Service.\end{aligned}$$

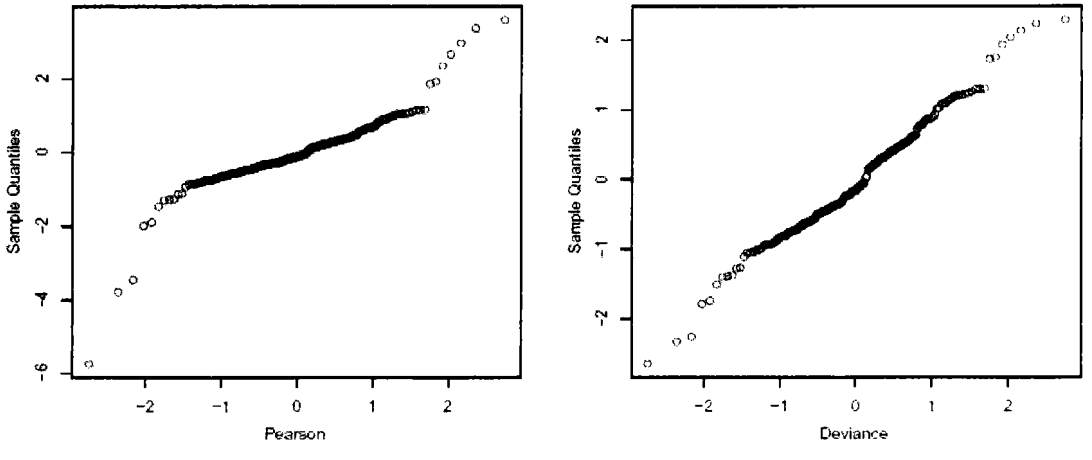

图 1 Pearson 残差和 Deviance 残差的 QQ 图

为了进一步说明利用 Logistic 回归模型拟合 MichelinNY 数据的合理性, 可以通过 Pearson 残差及 Deviance 残差的 QQ 图来初判其拟合优度。当然在此并不讨论拟合优度问题, 只是为了在有效模型的基础上, 讨论影响模型稳健性的一些强影响点。从图 1 可以看出, 两种残差大

部分点都集中于 $(-2,2)$ 之间，且大部分点大致分布于一条直线上。这说明用 Logistic 最终主效应模型拟合 MichelinNY 数据是比较合理的。

4.2 模型的影响分析

由 (10)、(11)、(12) 式得 $\Delta\chi_j^2$, ΔD_j, $\Delta_j\hat{\beta}$, $j=1,\cdots,n$ 即为删除第 j 个数据点所得到的三种平方型变化量；于是得到 164 维的向量 $\Delta\chi^2$, ΔD, $\Delta\hat{\beta}$, 实际上向量的维数是 163 维，这是因为 MichelinNY 数据的协变类型是 163 个，即几乎一个点就是一个协变类型。于是协变类型的个数对诊断分析几乎不会产生影响 (若协变类型数较少时，回归诊断要以协变类型为单位进行分析)。

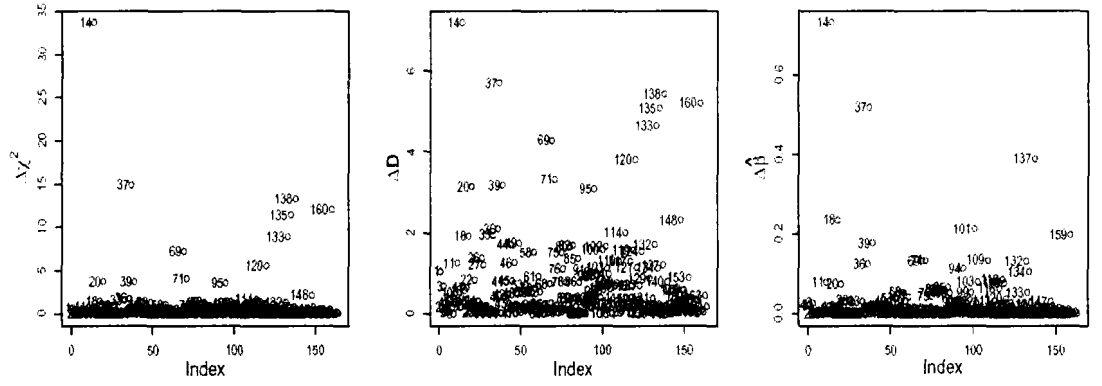

图 2 $\Delta\chi^2$, ΔD, $\Delta\hat{\beta}$ 的指标图

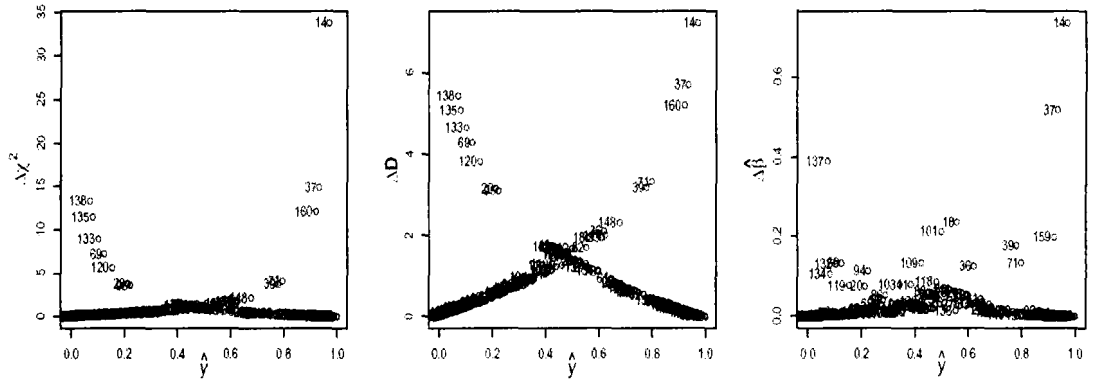

图 3 $\Delta\chi^2$, ΔD, $\Delta\hat{\beta}$ 对 $\hat{y}$ 的散点图

如图 2、图 3 分别是三个量的指标图及与 $\hat{y}$ 的散点图；很显然，当 $\Delta\chi_j^2$, ΔD_j, $\Delta_j\hat{\beta}$ 越大时就越能说明第 j 个点有可能是个强影响点，图中的标号表示样本点的编号。组图 3 的三幅图上的点都很有规律性，这都是由于 y 的分布所致；结合图 2、图 3 得出结论：第 14 号，第 37 号很有可能是强影响点。实际上，在初判中还有可能多选几个变化值较大的点，但这样没有必要，当第 37 号个体代入模型中分析变化较大时才往下选取。

由分解定理可知 $\overline{h}_{jj}=h_{jj}+\frac{r_j^2}{\chi^2}$，利用 R 软件计算出 h_{jj} 和 $\frac{r_j^2}{\chi^2}$，于是得 $plot(\frac{r_j^2}{\chi^2},h_{jj})$。如图 4 所示，图中分别用 $2ave(\overline{h}_{jj})$, $3ave(\overline{h}_{jj})$, $4ave(\overline{h}_{jj})$ 的等高线将可能的强影响点划分区域，$2ave(\overline{h}_{jj})$ 以内为正常点；$2ave(\overline{h}_{jj})$ 到 $3ave(\overline{h}_{jj})$ 之间的点是相对较弱的一类影响点；$3ave(\overline{h}_{jj})$ 到 $4ave(\overline{h}_{jj})$ 之间的点是中等程度的影响点；$4ave(\overline{h}_{jj})$ 之外是强影响点。于是结合诊断图 4 初判模型可能的强影响点是：14 号，37 号，86 号，104 号，110 号，18 号，102 号，11 号；实际

上后 6 类主要是由于杠杆值过大所引起的，而 11 号的杠杆值是最大的。而通过计算诊断统计量 $\Delta^{(1)}$, $\Delta^{(2)}$ 得到图 5 分别为 $plot(\Delta^{(1)})$, $plot(\Delta^{(2)})$；由于 $\Delta^{(1)}$, $\Delta^{(2)}$ 使用的权不一样，于是图 5 中影响点的类型也不一样。而由图 5 得到的影响点是：11 号，14 号，37 号；最后再综合图 4 和图 5 得出整个模型的强影响点为 11 号，14 号，37 号，且由表 1 亦可得出：14 号数据点并没有对系数产生显著性的影响，但是对模型残差的影响是次大的。而 37 号个体对模型残差影响较小，但对系数的影响最大，即删除第 37 号数据点后尤其是变量 $Service, Decor$ 和 $Decor \times Service$ 的显著性明显提高；最后，11 号个体对残差的影响是最大的，进而对估计值的影响也是最大的。因此，结合表 1，图 4 和图 5 得知上述的结论是合理的。

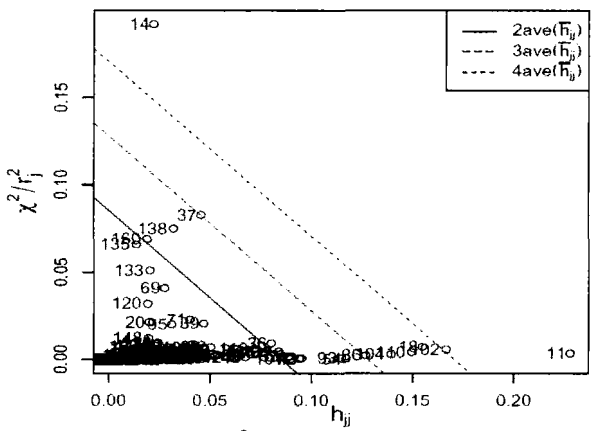

图 4 $\frac{r_j^2}{\chi^2}$ 对 h_{jj} 的散点图

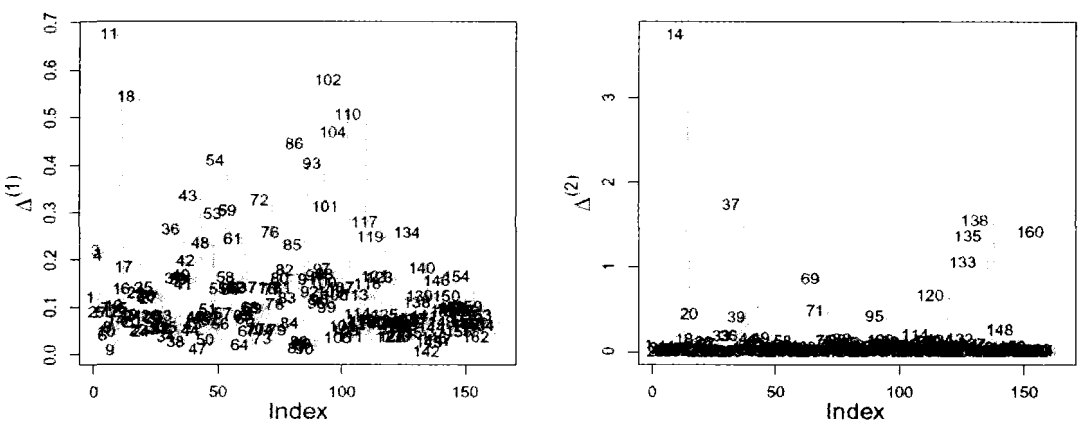

图 5 $\Delta^{(1)}$, $\Delta^{(2)}$ 的指标图

表 1 Logistic 回归模型的影响分析表

模型参数	观测数据	删除 14 号	删除 37 号	删除 11 号
Intercept	−63.76 (***)	−65.27 (***)	−70.20 (**)	−59.13 (***)
Food	0.64 (***)	0.69 (***)	0.76 (***)	0.62 (***)
Decor	1.50 (**)	1.38 (**)	1.76 (***)	1.30 (*)
Service	1.13 (*)	1.02 (*)	1.26 (**)	0.97 (.)
log (Price)	7.30 (***)	8.05 (***)	7.53 (***)	6.915 (***)
Decor:Service	−0.08 (**)	−.071 (**)	−.0.09 (***)	−0.067 (*)
模型残差	131.23	123.33	125.08	128.66

注：括号中的符号表示估计参数的显著性水平如:'***':0.001,'**':0.01,'*':0.05,'.':0.1,' ':1。

4.3 结论

本文利用两类方法来探寻 Logistic 模型的强影响点; 由 MichelinNY 数据的分析得知: 改进的诊断图得出的强影响点 (11 号, 14 号, 37 号) 比传统的诊断图得出的强影响点 (14 号, 37 号) 更精确. 于是, 上述方法为 Logistic 回归模型的影响分析提供了一个更有效的诊断方法。

[参考文献]

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

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

[3] Cook R D, Weisberg S. Characterizations of an empirical influence function for detecting influential cases in regression [J]. Technometrics, 1980, 22: 495–508.

[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 models [J]. Journal of the American Statistical Association, 1984, 79(385): 61–71.

[6] Sugata S R, Guria S. Diagnostics in logistic regression models [J]. Journal of the Korean Statistical Society, 2008, 37: 89–94.

[7] 徐宇航, 丁邦俊. 删失数据下几种两样本检验的功效研究 [J]. 数理统计与管理, 2011, (1): 68–73.

[8] 魏章进, 唐月玲, 隋广军. 热带气旋登陆概率的 Logistic 模拟 [J]. 数理统计与管理, 2012, (3): 13–21.

[9] 韦博成, 林金官, 解锋昌. 统计诊断 [M]. 北京: 高等教育出版社, 2009: 169–194.

[10] Hosmer D W, Lemeshow S. Applied Logistic Regression [M]. New York: John Wiley, 2000: 145–186.

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

[12] Sheather S J. A Modern Approach to Regression with R [M]. Berlin: Springer, 2009: 277-278. http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0- 387-09607-0?changeHeader