应 用 数 学
MATHEMATICA APPLICATA
2019, 32(4): 739-746

基于Pena距离的双重广义线性模型的统计诊断

邢伊琦, 吴刘仓, 聂兴锋
(昆明理工大学理学院, 云南 昆明 650093)

摘要： 本文主要研究双重广义线性模型, 考虑基于数据删除模型的参数估计和统计诊断, 比较删除模型与未删除模型相应的诊断统计量之间的变化. 首次提出基于双重广义线性模型下的Pena距离. 通过一些模拟研究以及实例分析, 比较不同诊断统计量判别异常点或强影响点的差异, 研究结果表明本文提出的理论和方法是行之有效的.

关键词： 双重广义线性模型;扩展拟似然;伪似然; Pena距离; 统计诊断
中图分类号： O212.1 **AMS(2000)主题分类：** 62J20
文献标识码： A **文章编号：** 1001-9847(2019)04-0739-08

1. 引言

经典的线性模型可以描述许多研究领域的现象, 但是, 在生物、医学、保险等领域的研究人员发现, 经典的线性模型会遇到很多方法上的困难, 在此基础上, 广义线性模型也就应运而生了. 广义线性模型是经典线性模型的推广. 1972年Nelder和Wedderburn首先提出广义线性模型[1], McCullagh和Nelder在1983年出版了系统论述此专题的专著《Generalized Linear Models》. 在一些经济领域和工业产品质量改进中, 对均值和方差同时建模是非常有必要的, 所以双重广义线性模型近年来引起了许多学者的关注. 1984年, Pregibon在文章中首先提出了对散度参数建模的广义线性模型[2]; WANG和ZHANG[3] 研究了双重广义线性模型中仅均值模型的变量选择; WU和LI[4]研究了逆高斯分布下联合均值和散度模型的变量选择; 陈海露[5]研究了双重广义线性模型的参数估计与变量选择; 吴刘仓等[6]研究了基于Box-Cox变换下联合均值与散度广义线性模型的极大似然估计; 胡江等[7]研究了基于Pena距离的广义线性回归模型的影响分析等.

在响应变量分布未知但已知其前两阶矩存在的情况下, Wedderburn提出了拟似然方法来进行参数估计. 拟似然的方法是假定总体前两阶矩阵存在, 然后通过对其对数似然方程求极值得到参数的估计值. 陈希孺[8]在广义线性模型中对拟似然方法有详细的阐述; 吴刘仓等[9]研究了缺失数据下双重广义线性模型的参数估计; 袁巧莉等[10]研究了混合双重广义线性模型的参数估计等.

我们知道, 统计诊断在数据分析中占有举足轻重的地位, 主要目的就是找出数据中的异常点或强影响点, 常用的统计量有似然距离、Cook距离等, Pena距离[11]这一诊断统计量是美国统计学教授Daniel Pena在2005年首次提出的, 并对其在线性模型上的影响分析做了详细的研

* **收稿日期：** 2018-06-28
基金项目： 国家自然科学基金项目 (11861041, 11261025)
通讯作者： 吴刘仓, 男, 汉族, 云南人, 教授, 研究方向：应用统计.

究, 这种方法与之前的方法有较大差别, 之前的方法是研究删除一(组)点,对回归分析的影响以及对预测值的影响, 或者是某个(组)样本点的微小扰动对参数估计的影响或是对模型预测的影响. 而Pena距离这一统计量则是研究的是样本中的某一个(组)点受其余各个(组)点的影响, 简单来说, 就是样本中各点删除后, 对某一特定的点的回归值或预测值的影响. 本文基于Pena距离, 采用伪似然和扩展拟似然的方法估计参数, 并通过数据删除模型的参数估计和统计诊断, 比较了删除模型和未删除模型对应的统计量之间的差异. 通过Monte Carlo模拟验证, 本文提出方法的有效性. 最后, 通过实例研究, 表明本文所提出的模型和方法是实用可行的.

2. 双重广义线性模型下的极大似然估计

我们先给出双重广义线性模型(Double Generalized Linear Model, DGLM)为:

$$\begin{cases} \mu_i = E(y_i|x,z), \\ \mathrm{Var}(y_i|x,z) = \phi_i V(\mu_i), \\ g(\mu_i) = x_i^{\mathrm{T}}\beta, \\ h(\phi_i) = z_i^{\mathrm{T}}\gamma, \\ i = 1,2,\cdots,n, \end{cases} \tag{2.1}$$

其中y_i为被解释变量, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^{\mathrm{T}}$, $z_i = (z_{i1}, z_{i2}, \ldots, z_{ip})^{\mathrm{T}}$为解释变量, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$为均值模型中的未知参数,$\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p)^{\mathrm{T}}$为散度模型中的未知参数.$x_i$, z_i两个解释变量可能完全相同, 部分相同或者完全不同, 但相同的解释变量在均值模型和方差模型中有不同的影响方式. $g(\mu_i) = x_i^{\mathrm{T}}\beta$是均值模型, $h(\phi_i) = z_i^{\mathrm{T}}\gamma$是散度模型, $g(\cdot)$和$h(\cdot)$是联系函数, $V(\cdot)$是关于均值的方差函数.

定理2.1 对于模型(2.1), 其Pena距离为:

$$S_i = \frac{s_i^{\mathrm{T}}s_i}{p\mathrm{Var}(\hat{y}_i)} = \frac{1}{ph_{ii}\hat{\sigma}^2}\sum_{j=1}^{n}\frac{h_{ij}^2\hat{e}_j}{(1-h_{jj})^2} = \frac{1}{ph_{ii}}\sum_{j=1}^{n}\frac{h_{ij}^2\hat{r}_j^2}{(1-h_{jj})},$$

其中$\hat{r}_j = \frac{\hat{e}_j}{\hat{\sigma}\sqrt{1-h_{jj}}}$为第$j$个学生化残差(标准化残差).

证 根据文[11], 我们定义Pena距离如下:

$$S_i = \frac{s_i^{\mathrm{T}}s_i}{p\mathrm{Var}(\hat{y}_i)},$$

其中$s_i = (\hat{y}_i - y_{i(1)}, \hat{y}_i - y_{i(2)}, \ldots, \hat{y}_i - y_{i(n)})$, $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$. 由韦博成等[12]统计诊断知, β的极大似然估计为$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$. 对于模型(2.1)，我们将联系函数$g(y_i)$在$y_i = \mu_i$处泰勒展开, 忽略二次项和更高次项, 得:

$$g(y_i) \approx g(\mu_i) + g'(\mu_i)(y_i - \mu_i) = \eta_i + (y_i - \mu_i)\frac{\partial\eta}{\partial\mu_i}.$$

对上式两边求方差，有：

$$\mathrm{Var}(g(\mu_i)) \approx (g'(\mu_i))^2\mathrm{Var}(y_i) = (\frac{\partial\eta}{\partial\mu_i})^2\phi_i V(\mu_i).$$

由$\hat{\beta}_i = \hat{\beta} - \frac{(X^{\mathrm{T}}X)^{-1}x_i\hat{e}_i}{1-h_{ij}}$, $\hat{h}_{ij} = {x_i}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_j$, $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, 其中$H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$是一个帽子矩阵, 且有$H^2 = H$, $H^{\mathrm{T}} = H$.

记$\hat{y}_i$为y_i的拟合值, $y_{i\hat{(j)}}$为删除第j个点后y_i的拟合值. 让$g(\hat{y}_i)$, $g(y_{i\hat{(j)}})$分别在$\hat{y}_i = \mu_i$, $y_{i\hat{(j)}} = \mu_i$处展开, 有：

$$g(\hat{y}_i) \approx g(\mu_i) + (\hat{y}_i - \mu_i)\frac{\partial\eta}{\partial\mu_i}, \quad g(y_{i\hat{(j)}}) \approx g(\mu_i) + (y_{i\hat{(j)}} - \mu_i)\frac{\partial\eta}{\partial\mu_i}.$$

上述两式相减, 得：

$$\hat{y_i} - y_{i\hat{(j)}} = (g(\hat{y_i}) - g(y_{i\hat{(j)}})(\frac{\partial \eta}{\partial \mu_i})^{-1}.$$

又由$g(\hat{y_i}) - g(y_{i\hat{(j)}}) = {x_i}^{\mathrm{T}}(\hat{\beta} - \hat{\beta_{(j)}})$, 有：

$$\hat{y_i} - y_{i\hat{(j)}} = \frac{x_j^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_j\hat{e_j}}{1 - \hat{h_{jj}}}(\frac{\partial \eta}{\partial \mu_i})^{-1}, j = 1, \ldots, n.$$

令$s_i = (\hat{y_i} - y_{i(1)}, \hat{y_i} - y_{i(2)}, \ldots, \hat{y_i} - y_{i(n)})$, 则有：

$$s_i^{\mathrm{T}}s_i = \sum_{j=1}^{n} \frac{\hat{h_{ij}}^{2}}{\hat{e}_j}^{2} (1 - h_{jj})^2(\frac{\partial \eta}{\partial \mu_i})^{-2}.$$

又$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$, 故$\mathrm{Var}(\hat{\beta}) = (X^{\mathrm{T}}X)^{-1}X\mathrm{Var}(Y)X(X^{\mathrm{T}}X)^{-1} = \sigma^2(X^{\mathrm{T}}X)^{-1}$, 而

$$\mathrm{Var}(g(y_i)) = \mathrm{Var}({x_i}^{\mathrm{T}}\hat{\beta}) = {x_i}^{\mathrm{T}}\mathrm{Var}(\hat{\beta})x_i = {x_i}^{\mathrm{T}}\sigma^2(X^{\mathrm{T}}X)^{-1}x_i = \sigma^2\hat{h_{ii}},$$

$$\mathrm{Var}(g(y_i)) \approx \mathrm{Var}(\hat{h_i} + (\hat{y_i} - \mu_i)\frac{\partial \eta}{\partial \mu_i}) = \mathrm{Var}(\hat{y_i}\frac{\partial \eta}{\partial \mu_i}) = (\frac{\partial \eta}{\partial \mu_i})^2\mathrm{Var}(\hat{y_i}),$$

故

$$\sigma^2\hat{h_{ii}} = (\frac{\partial \eta}{\partial \mu_i})^2\mathrm{Var}(\hat{y_i}), \quad \mathrm{Var}(\hat{y_i}) = \sigma^2\hat{h_{ii}}(\frac{\partial \eta}{\partial \mu_i})^{-2}.$$

故Pena距离为：

$$S_i = \frac{s_i^{\mathrm{T}}s_i}{p\mathrm{Var}(\hat{y_i})} = \frac{1}{ph_{ii}\hat{\sigma}^2}\sum_{j=1}^{n}\frac{h_{ij}^2\hat{e}_j}{(1 - h_{jj})^2} = \frac{1}{ph_{ii}}\sum_{j=1}^{n}\frac{h_{ij}^2\hat{r}_j^2}{(1 - h_{jj})},$$

其中$\hat{r}_j = \frac{\hat{e}_j}{\hat{\sigma}\sqrt{1-h_{jj}}}$为第$j$个学生化残差(标准化残差).

3. 双重广义线性模型的统计诊断

数据删除是统计诊断中最常用的也是最基本的方法之一, 比较删除第i个点前后模型参数估计量之间的差异, 能得出一些的结论. 用这些结论, 我们能评价我们估计方法的好坏, 详细内容参考韦博成[12]等的文献或书刊. 模型(2.1)的数据删除模型可表示为：

$$\begin{cases} \mu_j = \mathrm{E}(y_j|x, z), \\ \mathrm{Var}(y_j|x, z) = \phi_j V(\mu_j), \\ g(\mu_j) = x_j^{\mathrm{T}}\beta, \\ h(\phi_j) = z_j^{\mathrm{T}}\gamma, \\ j = 1, 2, \cdots, n, j \neq n. \end{cases} \tag{3.1}$$

对于未删除模型(2.1)和删除模型(3.1), 为检验第i个数据点在整个数据集中是否为异常点或强影响点, 可通过比较删除第i个点前后统计推断结果的变化, 看出这个点是否为异常点或者强影响点, 而统计推断结果的变化可由统计推断量来得到.

对于一组随机变量$y_1, y_2, \cdots, y_n$的分布是未知的, 但知道其期望和方差存在, 期望我们用$\mathrm{E}(y)$表示, 那么方差为:

$$\sigma_y^2 = \mathrm{Var}(y) = \mathrm{E}[(y - \mathrm{E}(y))^2] = \mathrm{E}(y^2) - \mathrm{E}^2(y).$$

从式中, 我们知道$\mathrm{E}(y)$是一阶原点矩, $\mathrm{Var}(y)$是二阶中心矩, 我们也可以认为方差是二阶原点矩减去期望的平方. 对于本文选用的模型, 我们选用扩展拟似然算法和伪似然算法来进行我们的参数估计. 本文采用的扩展拟似然函数(EQL)Q^+为：

$$Q^+(\mu, \phi|y) = \sum_{i=1}^{n}\log\{\frac{1}{\sqrt{2\pi\phi_i V(y_i)}}\exp(\frac{d_i}{2\phi_i})\} = -\frac{1}{2}\sum_{i=1}^{n}\log(2\pi\phi_i V(y_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{d_i}{\phi_i}, \tag{3.2}$$

其中π表示圆周率常数, $d_i = 2\int_{\mu_i}^{y_i} \frac{y_i - t}{V(t)}\mathrm{d}t$, 记$\rho_k(t) = (\frac{\mathrm{d}^k}{\mathrm{d}t^k})g^{-1}(t)$, $q_k(t) = (\frac{\mathrm{d}^k}{\mathrm{d}t^k})h^{-1}(t)$, $k = 1, 2$. 利用扩展拟似然函数Q^+对均值参数和散度参数进行计算, 有：

$$\frac{\partial Q^+}{\partial \beta} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i V(\mu_i)} \rho_1(x_i^{\mathrm{T}}\beta)x_i,$$

$$\frac{\partial Q^+}{\partial \gamma} = -\frac{1}{2}\sum_{i=1}^{n}(\frac{1}{\phi_i} - \frac{d_i}{\phi_i^2})q_1(z_i^{\mathrm{T}}\gamma)z_i,$$

$$\frac{\partial^2 Q^+}{\partial \beta \partial \gamma^{\mathrm{T}}} = -\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i^2 V(\mu_i)} \rho_1(x_i^{\mathrm{T}}\beta)q_1(z_i^{\mathrm{T}}\gamma)x_i z_i^{\mathrm{T}},$$

$$\frac{\partial^2 Q^+}{\partial \beta \partial \beta^{\mathrm{T}}} = -\sum_{i=1}^{n} \frac{V(\mu_i) + (y_i - \mu_i)V^{'}(\mu_i)}{\phi_i V^2(\mu_i)} \rho_1^2(x_i^{\mathrm{T}}\beta)x_i x_i^{\mathrm{T}} + \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\phi_i V(\mu_i)} \rho_2(x_i^{\mathrm{T}}\beta)x_i x_i^{\mathrm{T}},$$

$$\frac{\partial^2 Q^+}{\partial \gamma \partial \gamma^{\mathrm{T}}} = -\frac{1}{2}\sum_{i=1}^{n}(-\frac{1}{\phi_i^2} + \frac{2d_i}{\phi_i^3})q_1^2(z_i^{\mathrm{T}}\gamma)z_i z_i^{\mathrm{T}} - \frac{1}{2}\sum_{i=1}^{n}(\frac{1}{\phi_i} - \frac{d_i}{\phi_i^2})q_2(z_i^{\mathrm{T}}\gamma)z_i z_i^{\mathrm{T}}.$$

本文中采用的伪似然函数(PL)Q_p为:

$$Q_p(\mu, \phi|y) = \sum_{i=1}^{n} \log\{2\pi\phi_i V(\mu_i)\exp(\frac{X_i^2}{\phi_i})\},$$

其中$X_i^2 = \frac{(y_i - \mu_i)^2}{V(\mu_i)}$, 类似于扩展拟似然函数, 记$\rho_k(t) = (\frac{\mathrm{d}^k}{\mathrm{d}t^k})g^{-1}(t)$, $q_k(t) = (\frac{\mathrm{d}^k}{\mathrm{d}t^k})h^{-1}(t)$, $k = 1, 2$. 对伪似然函数Q_p计算均值参数以及散度参数，有：

$$\frac{\partial Q_p}{\partial \beta} = \sum_{i=1}^{n}\{\frac{V^{'}(\mu_i)}{V(\mu_i)} - \frac{2(y_i - \mu_i)V(\mu_i) + (y_i - \mu_i)^2 V^{'}(\mu_i)}{\phi_i V^2(\mu_i)}\}\rho_1(x_i^{\mathrm{T}}\beta)x_i,$$

$$\frac{\partial Q_p}{\partial \gamma} = \sum_{i=1}^{n}(\frac{1}{\phi_i} - \frac{X_i^2}{\phi_i^2})q_1(z_i^{\mathrm{T}}\gamma)z_i,$$

$$\frac{\partial^2 Q_p}{\partial \beta \partial \gamma^{\mathrm{T}}} = -\sum_{i=1}^{n} \frac{2(y_i - \mu_i)V(\mu_i) + (y_i - \mu_i)^2 V^{'}(\mu_i)}{\phi_i^2 V^2(\mu_i)} \rho_1(x_i^{\mathrm{T}}\beta)q_1(z_i^{\mathrm{T}}\gamma)x_i z_i^{\mathrm{T}},$$

$$\begin{aligned}\frac{\partial^2 Q_p}{\partial \beta \partial \beta^{\mathrm{T}}} =& \sum_{i=1}^{n}\{\frac{V^{''}(\mu_i)V(\mu_i) - (V(\mu_i))^2}{V^2(\mu_i)} + \frac{2V(\mu_i) + 2(y_i - \mu_i)V^{'}(\mu_i)}{\phi_i V^2(\mu_i)}\}\rho_1^2(x_i^{\mathrm{T}}\beta)x_i x_i^{\mathrm{T}} \\ &- \sum_{i=1}^{n} \frac{[-2(y_i - \mu_i)V^{'}(\mu_i) + (y_i - \mu_i)^2 V^{''}(\mu_i)]V(\mu_i) - 2(y_i - \mu_i)^2(V^{'}(\mu_i))^2}{\phi_i V^3(\mu_i)} \rho_1^2(x_i^{\mathrm{T}}\beta)x_i x_i^{\mathrm{T}} \\ &+ \sum_{i=1}^{n}\{\frac{V^{'}(\mu_i)}{V(\mu_i)} - \frac{2(y_i - \mu_i)V(\mu_i) + (y_i - \mu_i)^2 V^{'}(\mu_i)}{\phi_i V^2(\mu_i)}\}\rho_2(x_i^{\mathrm{T}}\beta)x_i x_i^{\mathrm{T}},\end{aligned}$$

$$\frac{\partial^2 Q_p}{\partial \gamma \partial \gamma^{\mathrm{T}}} = \sum_{i=1}^{n}(-\frac{1}{\phi_i^2} + \frac{2X_i^2}{\phi_i^3})q_1^2(z_i^{\mathrm{T}}\gamma)z_i z_i^{\mathrm{T}} + \sum_{i=1}^{n}(\frac{1}{\phi_i} - \frac{X_i^2}{\phi_i^2})q_2(z_i^{\mathrm{T}}\gamma)z_i z_i^{\mathrm{T}}.$$

用Gauss-Newton迭代法可得到参数极大似然估计的估计值. 设未删除模型的参数估计值用$\hat{\beta}, \hat{\gamma}$表示, 则$\hat{\theta} = (\hat{\beta}^{\mathrm{T}}, \hat{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. 删除模型的参数估计值用$\hat{\beta}_i, \hat{\gamma}_i$表示, 则有$\hat{\theta}_i = (\hat{\beta^{\mathrm{T}}}_i, \hat{\gamma^{\mathrm{T}}}_i)^{\mathrm{T}}$.

由Gauss-Newton迭代法可得到参数极大似然估计的估计值$\hat{\theta}$和$\hat{\theta}_i$, 但如果解释变量的维数为二维或者高于二维，这时参数$\hat{\theta}$ 和$\hat{\theta}_i$均为向量, 难以比较大小. 这时我们就可以用Cook距离来刻画参数的变化, Cook距离定义如下:

$$CD_i = \frac{(\hat{\theta} - \hat{\theta}_i)^{\mathrm{T}} H^{\mathrm{T}} H(\hat{\theta} - \hat{\theta}_i)}{p\hat{\sigma}^2},$$

其中$H=(x^{\mathrm{T}},z^{\mathrm{T}})^{\mathrm{T}}$为解释变量, p为对应解释变量的维数, $\hat{\sigma}^2$为未删除模型方差的估计值.

在分析具体数据时, 先计算出各点的Cook距离, 通过画散点图, 找出其中特别大的D_i, 对应数据点可能就是异常点或强影响点.

Pena距离与Cook距离相比较，前者研究的是删除一个(组)点后对估计值或预测值的影响, 而后者则研究的是样本中的某一点受其余各点的影响, 简单的说, 就是研究样本中各点删除后, 对某一特定的点的估计值或预测值的影响, Pena距离定义如下：

$$S_i=\frac{(\hat{\theta}_i-\hat{\theta}_{i(j)})^{\mathrm{T}}H^{\mathrm{T}}H(\hat{\theta}_i-\hat{\theta}_{i(j)})}{p\hat{\sigma}^2},$$

其中$H=X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$称为帽子矩阵, p为相应解释变量的维数, $\hat{\sigma}^2$为删除第i个点后模型方差的估计值. $\hat{\theta}_{i(j)}$是删除第j个点后第i个点的参数估计值. 具体分析时, 同样是先算出删除各点后某一点的S_i, 画出散点图, 其中S_i较大的就可能是异常点或强影响点.

4. Monte Carlo模拟

为了比较Pena距离和Cook距离的诊断效果, 本文我们采用Extra-Poisson模型进行模拟. Extra-Poisson 模型如下：

$$\begin{cases} y_i|m_i\sim \mathrm{Poisson}(m_i),\\ m_i\sim \mathrm{Gamma}(\nu_i,\alpha_i),\\ \mathrm{E}(y_i)=\mu_i=\nu_i\alpha_i,\\ \mathrm{Var}(y_i)=\nu_i\alpha_i+\nu_i\alpha_i^2=\mu_i(1+\alpha_i)=\mu_i\phi_i,\\ \mu_i=\exp(x_i^{\mathrm{T}}\beta),\\ \phi_i=\exp(z_i^{\mathrm{T}}\gamma),\\ i=1,2,\cdots,n; j=1,2. \end{cases} \tag{4.1}$$

根据模型(4.1)产生模拟数据, 其中y_i是根据双重广义回归模型产生的相互独立的响应变量, 解释变量x_i和z_i相互独立产生于U(0,1). 给定β_0和γ_0的真值分别为$\beta_0=(0,1,1)^{\mathrm{T}},\gamma_0=(0,1,1)^{\mathrm{T}}$. 将第170号, 190 号样本点的被解释变量的值做改变, 即从样本点中人为的制造两个异常点, 然后应用本文研究的Pena距离以及Cook 距离进行诊断, 根据异常点的诊断情况来判断本文提出的方法是否行之有效. 模拟结果如图1-4所示：

从图中我们可以看出, 无论是用PL或者EQL方法, 第170号点以及190号点均被诊断出来了, 这说明本文提出的方法是可行并且有效的, 下面用具体的实例进一步说明.

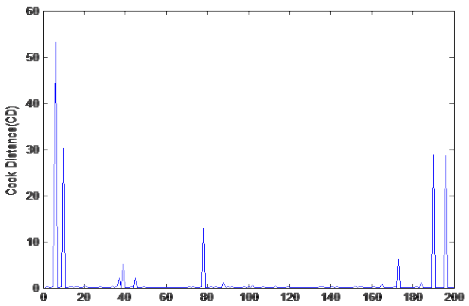

图1　PL的Cook距离CD散点图

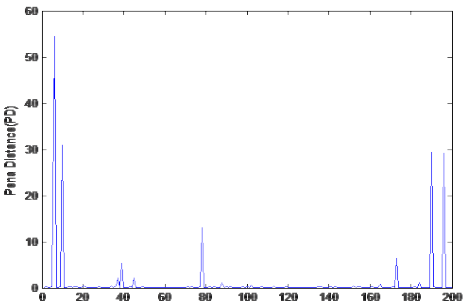

图2　PL的Pena距离PD散点图

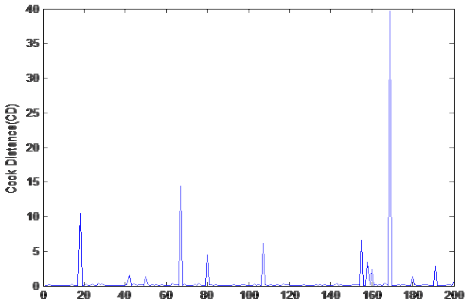

图3 EQL的Cook距离CD散点图

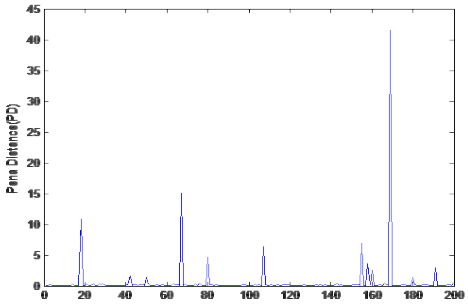

图4 EQL的Pena距离PD散点图

5. 实例分析

这里我们用一组检验某种工业用发动机性能试验的数据, 该试验使用的原料是柴油和从有机原料中通过蒸馏产生的气体的混合物, 在各种不同的速度x(计量单位：百转/分钟)下，测量发动机的马力y .

表 1 发动机性能数据

NO.	x	y	NO.	x	y	NO.	x	y
1	22	64.03	9	19	58	17	10.5	32.05
2	20	62.27	10	21	63.21	18	13	39.68
3	18	54.94	11	22	64.03	19	15	45.79
4	16	48.84	12	20	59.63	20	17	51.57
5	14	43.73	13	18	52.9	21	19	56.65
6	12	37.45	24	16	48.84	22	21	62.61
7	15	46.85	15	14	42.74	23	23	65.31
8	17	51.17	16	12	36.63	24	24	63.89

我们建立模型

$$\begin{cases} y_i \sim f(u_i, {\sigma_i}^2), \\ \mu_i = \exp(x_i^{\mathrm{T}}\beta), \\ \phi_i = \exp(z_j^{\mathrm{T}}\gamma), \\ i = 1, 2, \cdots, 21. \end{cases} \tag{5.1}$$

利用伪似然和扩展拟似然的估计方法得到的参数做统计诊断，得到图5-8的结果.

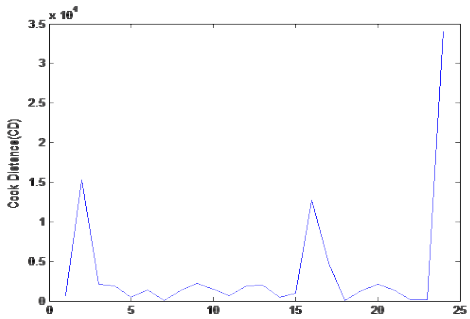

图5 发动机性能数据PL的Cook距离CD散点图

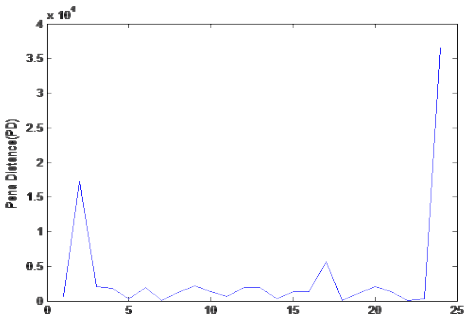

图6 发动机性能数据PL的Pena距离PD散点图

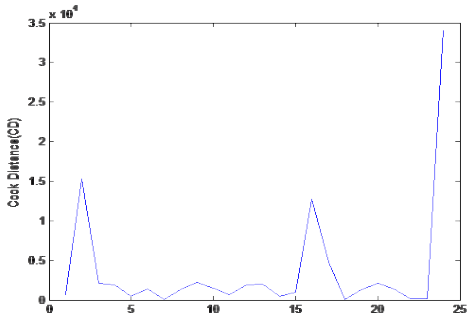

图7　发动机性能数据EQL的Cook距离CD散点图

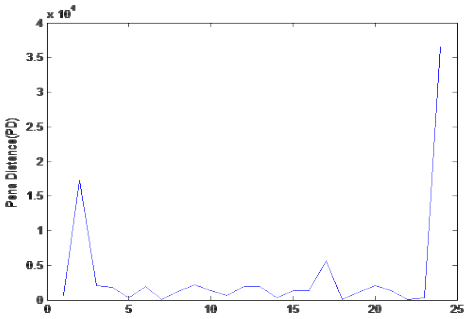

图8　发动机性能数据EQL的Pena距离PD散点图

从图5-8的结果可以看出, 用伪似然和扩展拟似然的方法估计参数做统计诊断的效果大致相同. 我们以图5和图6用伪似然的方法估计的参数做的统计诊断为例, 由图5可知第2号点、16号点和24号点可能为异常点或强影响点, 由图6可知第2号点和24号点可能为异常点或强影响点. 由文[12]中的例5.4 可知, 第2、24号点为异常点或强影响点, 比起Cook距离, Pena距离很好地诊断出了这个点.

6. 结论

本文针对分布未知但其一阶矩和二阶矩存在的随机变量, 建立了双重广义线性模型, 运用扩展拟似然和伪似然方法进行参数估计再用Pena距离和Cook距离进行统计诊断, 得到在一定的条件下Pena距离优于Cook距离的结论. 最后, 通过Monte Carlo模拟和实例研究的结果验证, 说明了本文所提出的模型与方法的有效性和实用性.

参考文献:

[1] NELDER J A, WEDDERBURN R W M. Generalized linear models[J]. Journal of the Royal Statistical Society: Series A, 1972, 135(3): 370-384.

[2] PREGIBON D. Review: P. McCullagh, J. A. Nelder, Generalized linear models[J]. The Annals of Statistics, 1984, 12: 1589-1596.

[3] WANG Darong, ZHANG Zhongzhan. Variable selectiong in joint generalized linear models[J]. Chinese Journal of Applied Probability and Statistics, 2009, 25(3): 245-256.

[4] WU Liucang, LI Huiqiong. Variable selectiong for joint mean and dispersion models of the inverse Gaussian distribution[J]. Metrika, 2012, 75: 795-808.

[5] 陈海露. 双重广义线性模型的参数估计与变量选择[D]. 北京：北京工业大学, 2011.

[6] 吴刘仓, 黄丽, 戴琳. Box-Cox变换下联合均值与散度模型的极大似然估计[J]. 统计与信息论坛, 2012, 27(5): 3-8.

[7] 胡江, 林金官, 赵彦勇. 基于Pena距离的广义线性回归模型的影响分析[J]. 应用数学, 2017, 30(3): 539-546.

[8] 陈希孺. 广义线性模型(六)[J]. 数理统计与管理, 2003, 22(2)：55-64.

[9] 吴刘仓, 邱贻涛, 詹金龙. 缺失数据下双重广义线性模型的参数估计[J]. 应用数学, 2014, 27(4): 714-724.

[10] 袁巧莉, 吴刘仓, 戴琳. 混合双重广义线性模型的参数估计[J]. 高校应用数学学报, 2017, 32(3): 267-176.

[11] PENA D. A new staistic for influence in linear regression[J]. Technometrics, 2005, 47(1): 1-12.

[12] 韦博成, 林金官, 解锋昌. 统计诊断[M]. 高等教育出版社, 2009.

Statistical Diagnostics for Double Generalized Linear Model Based on the Pena Distance

XING Yiqi, WU Liucang, NIE Xingfeng

(*Faculty of Science, Kunming University of Science and Technology, Kunming 650093, China*)

Abstract: This paper mainly studies the double generalized linear model, considers the parameters estimation and statistical diagnosis based on the data deletion model, and compares the changes between the corresponding diagnostic statistic of the deleted model and the non-deleted model. For the first time, the Pena distance based on the double generalized linear model is proposed. Through some Monte Carlo simulation studies and a real data analysis, the differences in different diagnosis statistics for discriminating between abnormal or strong influence points were compared. The results show the proposed theories and methods in this paper are effective.

Key words: Double generalized linear model; Extended quasi-likelihood; Pseudo-likelihood; Pena distance; Statistical diagnostics