应用概率统计　第二十九卷
第五期　2013年10月

Chinese Journal of Applied Probability
and Statistics Vol.29 No.5 Oct. 2013

删失线性模型的经验似然诊断 *

丁先文　　徐　亮

(江苏理工学院数理学院, 常州, 213001)　　(东南大学数学系, 南京, 211189)

摘　要

基于经验似然方法对删失线性模型进行统计诊断. 首先介绍基于经验似然方法的线性模型的统计诊断; 其次将删失线性模型转换为线性模型, 并对转换后的模型进行统计诊断; 最后通过模拟计算和实例分析说明了诊断方法的有效性.

关键词: 经验似然, 模型转换, 统计诊断, 有效性.

学科分类号: O212.2.

§1. 引　　言

在回归分析中, 对回归模型的统计诊断主要涉及异常点的识别, 强影响点分析等, 常用的诊断方法有数据删除方法, 局部影响分析方法和残差分析方法等, 近年来这些方法又得到了许多研究者的研究和推广, 如Belsley等(1980), Christensen等(1992), Critchley等(2001)等. Zhu和Lee (2003)研究了广义线性混合模型的局部影响分析. Zhu和Lee (2001)及Zhu等(2001)对带缺失数据的统计模型提出了基于Q函数的一般的影响分析方法. 该方法已被广泛应用于各种各样的统计模型, 如Lee和Xu (2004), Xu等(2006). Zhu等(2007)提出了一种以相应的统计模型的观测似然函数为基础的扰动选择方法和局部影响度量.

基于经验似然的统计分析是近年来统计学的热点研究课题. Owen (1988, 1991, 2001)介绍并发展了经验似然方法, 提出了经验似然比统计量, 对经验似然方法作了深入的研究. Owen (1988, 1991, 2001)首先将经验似然方法应用到线性模型, Kolaczyk (1994), Chen和Cui (2003)等将其推广到了广义线性模型, Qin和Lawless (1994)将经验似然推广到半参数模型中, Kitamura (2001)将经验似然方法应用到了经济模型的研究中. 经验似然方法同其它统计方法相比有许多突出的优点, 如用经验似然构造的置信区间具有域保持性, 变换不变性以及置信域的形状完全由数据决定, 此外还有Bartlett纠偏性及无需构造枢轴量等优点, 更为重要的是经验似然方法在构造置信域时可以避免估计量的渐近方差的估计, Hall和La Scala (1990)详细地讨论了经验似然的各种优点, 并与Bootstrap方法作了比较. 经验似然推断在总体均值推断、线性模型推断、分位数推断、估计方程推断及利用辅助信息进行推断等几种重要统计推断中有着广泛的应用. Zhu等(2008)将经验似然方法应用到统

*国家自然科学基金(NSFC11171065)、江苏省自然科学基金(NSFJSBK2011058)、江苏理工学院博士启动基金(KYY11052)和江苏理工学院研究基金(JG13022)资助.

本文2012年1月20日收到, 2012年7月8日收到修改稿.

计诊断中, 提出了基于经验似然的几个诊断统计量. 徐亮等(2011)将基于经验似然的统计诊断推广至部分线性模型, 对删失线性模型的统计诊断还未见文献, 本文基于经验似然将Zhu等(2008)的结果推广至删失线性模型.

设有线性回归模型

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\theta}+\boldsymbol{\varepsilon}, \tag{1.1}$$

其中$\boldsymbol{Y}=(y_1,\ldots,y_n)^T$, $\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_n)^T$, $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_p)^T$, $\boldsymbol{X}$为$n\times p$阶列满秩设计阵, $\boldsymbol{X}_i$为$\boldsymbol{X}$的第i行, $\mathsf{E}(\boldsymbol{\varepsilon})=0$, $\mathsf{Var}(\boldsymbol{\varepsilon})=\sigma^2\boldsymbol{I}$, 这是常见的完全数据回归模型, 然而在某些情况下, y_i可能被某个随机变量c_i所截掉, 从而导致y_i不能完全被观测到, 即观测到的数据不是$(y_i,\boldsymbol{X}_i^T)$, 而是$(z_i,\delta_i,\boldsymbol{X}_i^T)$, 其中$z_i=\min(y_i,c_i)$, $\delta_i=\boldsymbol{I}\{y_i\le c_i\}$, $i=1,\ldots,n$, $\boldsymbol{I}\{A\}$为集合A的指示函数, 并假设删失变量c_i是独立同分布的, 其共同分布为G. 上述模型就称为右删失线性模型. 由此可知, δ_i用来指示第i个观测值是y_i还是c_i, 若$\delta_i=1$, 则表示y_i, 若$\delta_i=0$, 则表示c_i. 在许多实际问题中, 诸如抽样调查, 生存分析, 可靠性寿命试验, 医药追踪试验中会产生一些数据删失, 因此许多研究者对删失数据的统计分析进行了广泛和深入的研究. 本文主要用经验似然方法来求右删失线性模型的各种诊断统计量, 对该模型进行统计诊断.

§2. 经验似然方法介绍

设$\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$为独立同分布样本, $\boldsymbol{x}_i\in\boldsymbol{R}^d$ $(i=1,\cdots,n)$, 共同分布为F, F中含有未知参数$\boldsymbol{\theta}=(\theta_1,\ldots,\theta_p)^T\in\boldsymbol{\Theta}$, 其中$T$代表转置(下同), F的具体形式也是未知的. 假设有估计函数$\boldsymbol{g}(\boldsymbol{x},\boldsymbol{\theta})=(g_1(\boldsymbol{x},\boldsymbol{\theta}),\ldots,g_r(\boldsymbol{x},\boldsymbol{\theta}))^T$, 满足

$$\mathsf{E}_F\{\boldsymbol{g}(\boldsymbol{x},\boldsymbol{\theta}_0)\}=0,\qquad \boldsymbol{\theta}_0\in\boldsymbol{\Theta}, \tag{2.1}$$

这里要求$r\ge p$, 其中E_F表示在分布F下求期望.

在估计方程(2.1)成立的条件下, 由经验似然方法构造如下的经验似然比函数

$$L_{\mathrm{E}}(\boldsymbol{\theta})=\sup\Big\{\prod_{i=1}^n np_i\Big|\sum_{i=1}^n p_i=1,p_i>0,\sum_{i=1}^n p_i\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})=0\Big\}. \tag{2.2}$$

由Qin和Lawless (1994)和Owen (2001)可知, 当$p_i(\boldsymbol{\theta})=n^{-1}\{1+\boldsymbol{t}_n(\boldsymbol{\theta})^T\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})\}^{-1}$时, (2.2)取得唯一值$L_{\mathrm{E}}(\boldsymbol{\theta})=\prod\limits_{i=1}^n\{1+\boldsymbol{t}_n(\boldsymbol{\theta})^T\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})\}^{-1}$, 其中$\boldsymbol{t}_n(\boldsymbol{\theta})\in\boldsymbol{R}^r$, 是方程$\sum\limits_{i=1}^n\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})\{1+\boldsymbol{t}_n^T\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})\}^{-1}=0$的解, 由$L_{\mathrm{E}}(\boldsymbol{\theta})$就可以对参数$\boldsymbol{\theta}$进行统计推断了.

类似于参数似然函数的情形, 定义对数经验似然函数$l_{\mathrm{E}}(\boldsymbol{\theta})$如下

$$l_{\mathrm{E}}(\boldsymbol{\theta})=\sum_{i=1}^n l_{\mathrm{E},i}(\boldsymbol{\theta})=-\sum_{i=1}^n\log\{1+\boldsymbol{t}_n(\boldsymbol{\theta})^T\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})\},$$

其中$l_{\mathrm{E},i}(\boldsymbol{\theta})=-\log\{1+\boldsymbol{t}_n(\boldsymbol{\theta})^T\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})\}$, 为了得到参数$\boldsymbol{\theta}$的估计值, 求解使$l_{\mathrm{E}}(\boldsymbol{\theta})$达最大值的 $\widehat{\boldsymbol{\theta}}$, 即有

$$l_{\mathrm{E}}(\widehat{\boldsymbol{\theta}})=\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}(l_{\mathrm{E}}(\boldsymbol{\theta})),$$

并将$\widehat{\boldsymbol{\theta}}$定义为参数$\boldsymbol{\theta}$的极大经验似然估计(MELE). 现定义

$$Q_n(\boldsymbol{t},\boldsymbol{\theta})=n^{-1}l_{\mathrm{E}}(\boldsymbol{t},\boldsymbol{\theta})=n^{-1}\sum_{i=1}^{n}l_i(\boldsymbol{t},\boldsymbol{\theta})=-n^{-1}\sum_{i=1}^{n}\log(1+\boldsymbol{t}^T\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})),$$

其中$l_i(\boldsymbol{t},\boldsymbol{\theta})=-\log(1+\boldsymbol{t}^T\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta}))$, 可通过如下方程组来求得$\widehat{\boldsymbol{\theta}}$和$\widehat{\boldsymbol{t}}=\widehat{\boldsymbol{t}}_n(\widehat{\boldsymbol{\theta}})$

$$\begin{cases}Q_{1,n}(\boldsymbol{t},\boldsymbol{\theta})=\partial_{\boldsymbol{t}}Q_n(\boldsymbol{t},\boldsymbol{\theta})=-n^{-1}\sum\limits_{i=1}^{n}\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})\{1+\boldsymbol{t}_n^T\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})\}^{-1}=0,\\ Q_{2,n}(\boldsymbol{t},\boldsymbol{\theta})=\partial_{\boldsymbol{\theta}}Q_n(\boldsymbol{t},\boldsymbol{\theta})=-n^{-1}\sum\limits_{i=1}^{n}\partial_{\boldsymbol{\theta}}\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})\boldsymbol{t}_n\{1+\boldsymbol{t}_n^T\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})\}^{-1}=0.\end{cases}\tag{2.3}$$

Qin和Lawless (1994)还证明了$\widehat{\boldsymbol{\theta}}$和$\widehat{\boldsymbol{t}}=\widehat{\boldsymbol{t}}_n(\widehat{\boldsymbol{\theta}})$的渐近性质如下

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{L}N(\boldsymbol{0},\boldsymbol{C}_{\boldsymbol{\theta}}),\qquad\sqrt{n}(\widehat{\boldsymbol{t}}-\boldsymbol{0})\xrightarrow{L}N(\boldsymbol{0},\boldsymbol{C}_{\boldsymbol{t}}),$$

其中$\xrightarrow{L}$表示依分布收敛, $\boldsymbol{C}_{\boldsymbol{\theta}}$和$\boldsymbol{C}_{\boldsymbol{t}}$为协方差矩阵且有

$$\boldsymbol{C}_{\boldsymbol{\theta}}=(S_{21}S_{11}^{-1}S_{12})^{-1},\qquad\boldsymbol{C}_{\boldsymbol{t}}=S_{11}^{-1}-S_{11}^{-1}S_{12}S_{22.1}^{-1}S_{21}S_{11}^{-1}.$$

现记$S_{22.1}=-S_{21}S_{11}^{-1}S_{12}$和矩阵$\boldsymbol{S}$:

$$\boldsymbol{S}=\boldsymbol{S}(0,\boldsymbol{\theta}_0)=\begin{pmatrix}S_{11}&S_{12}\\S_{21}&S_{22}\end{pmatrix}=\begin{pmatrix}\mathsf{E}_F(\boldsymbol{g}^{\otimes2})&-\mathsf{E}_F(\partial_{\boldsymbol{\theta}}\boldsymbol{g})^T\\-\mathsf{E}_F(\partial_{\boldsymbol{\theta}}\boldsymbol{g})&0\end{pmatrix}_{(0,\boldsymbol{\theta}_0)},$$

其中$\boldsymbol{g}=\boldsymbol{g}(\boldsymbol{x},\boldsymbol{\theta}_0)$, $\partial_{\boldsymbol{\theta}}\boldsymbol{g}=\partial\boldsymbol{g}(\boldsymbol{x},\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ (下同), 对任意的向量$\boldsymbol{a}$, $\boldsymbol{a}^{\otimes2}=\boldsymbol{a}\boldsymbol{a}^T$.

§3. 删失线性模型的诊断分析

3.1 模型转换与估计函数

对于删失线性模型, 无论单独使用非删失数据, 还是将删失数据简单的看作是真实数据, 都是不合理的. 换言之, 删失数据既不能丢掉也不能直接使用, 关键在于寻求一种方法尽量提取其中的信息. 而且基于完全数据提出的一系列统计方法也不能直接应用于删失数据, 一种自然的想法就是将删失数据模型通过某种方法转换成另外一种完全数据模型. Koul等(1981)针对删失数据模型提出了一种数据转换方法, 成功的解决了从删失数据模型到完全数据模型的转换问题. 首先进行数据转换, 令

$$y_{iG}=\frac{\delta_i z_i}{1-G(z_i)},\qquad i=1,\ldots,n,\tag{3.1}$$

即

$$y_{iG}=\begin{cases}\dfrac{y_i}{1-G(y_i)},&\delta_i=1;\\0,&\delta_i=0,\end{cases}\tag{3.2}$$

其中G是删失变量c_i的分布函数, 这时得到了新的数据集$(\boldsymbol{X}_i^T, y_{iG})$, $i=1,\ldots,n$, 并且可以证明$\mathsf{E}(y_{iG}|\boldsymbol{X}_i)=\boldsymbol{X}_i^T\boldsymbol{\theta}$, 这样我们就可以构造新的线性模型如下

$$\boldsymbol{Y}_G=\boldsymbol{X}\boldsymbol{\theta}+\boldsymbol{\varepsilon}, \tag{3.3}$$

其中$\boldsymbol{Y}_G=(y_{1G},\ldots,y_{nG})^T$, $\boldsymbol{X}$和$\boldsymbol{\theta}$的定义同(1.1), $\boldsymbol{\varepsilon}$期望仍为0, 但其方差有可能和(1.1)不同. 如果G已知就可以直接进行数据转换, 而事实上在很多情况下G的分布是未知的, 这时就需要寻找G的一个估计式来代替它. Susarla和Van Ryzin (1980)给出了G的一个如下形式的估计式$\widehat{G}_1$:

$$\widehat{G}_1(t)=\prod_{i=1}^{n}\left\{\frac{1+N^+(z_i)}{2+N^+(z_i)}\right\}^{\boldsymbol{I}\{z_i\le t,\delta_i=0\}},$$

其中$N^+(z_i)=\sum\limits_{j=1}^{n}\boldsymbol{I}\{z_j>z_i\}$, $\boldsymbol{I}\{A\}$为集合A的指示函数. 他们还证明了在某些条件下$\widehat{G}_1(t)$是$G(t)$的相合估计.

另外, Koul等(1981)又提出了G的K-M估计式$\widehat{G}_2$, 定义如下

$$\widehat{G}_2(t)=1-\prod_{i=1}^{n}\left(\frac{n-i}{n-i-1}\right)^{\boldsymbol{I}\{z_i\le t,\delta_i=0\}},$$

其中$\boldsymbol{I}\{A\}$为集合A的指示函数.

可见G的两种估计式都只与实际观测值$(z_i,\delta_i,\boldsymbol{X}_i^T)$有关, 因此通过实际观测数据就可以找到$G$的估计式$\widehat{G}_1(t)$和$\widehat{G}_2(t)$, 接下来只需将(3.2)式中的$G$用$\widehat{G}_1(t)$或$\widehat{G}_2(t)$代替就可以实现数据的转换了.

由(3.3)有$\mathsf{E}(y_{iG}-\boldsymbol{X}_i^T\boldsymbol{\theta}|\boldsymbol{X}_i)=0$, 由此可得$\mathsf{E}\{\boldsymbol{X}_i(y_{iG}-\boldsymbol{X}_i^T\boldsymbol{\theta})\}=0$, 因此在删失线性模型中可取估计函数为

$$\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})=\boldsymbol{X}_i(y_{iG}-\boldsymbol{X}_i^T\boldsymbol{\theta}), \tag{3.4}$$

其中$\boldsymbol{x}_i=(x_{i,1},x_{i,2})=(y_i,\boldsymbol{X}_i^T)$, 即此时观测值由两部分组成, 且有$r=p$, r为估计函数$\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})$的维数, 经验似然比函数为$L_{\mathrm{E}}(\boldsymbol{\theta})=\prod\limits_{i=1}^{n}\{1+\boldsymbol{t}_n(\boldsymbol{\theta})^T\boldsymbol{X}_i(y_{iG}-\boldsymbol{X}_i^T\boldsymbol{\theta})\}^{-1}$, $\boldsymbol{t}_n(\boldsymbol{\theta})\in\boldsymbol{R}^r$, 是方程$\sum\limits_{i=1}^{n}\boldsymbol{X}_i(y_{iG}-\boldsymbol{X}_i^T\boldsymbol{\theta})\{1+\boldsymbol{t}_n^T\boldsymbol{X}_i(y_{iG}-\boldsymbol{X}_i^T\boldsymbol{\theta})\}^{-1}=0$的解.

3.2　模型的数据删除度量

为了研究第i组数据点$(y_i,\boldsymbol{x}_i^T)$的影响, 通常考虑去掉$(y_i,\boldsymbol{x}_i^T)$的删除模型. 受Zhu等(2008)启发, 对应于参数似然的Cook距离, 可得经验Cook距离

$$\mathrm{ECD}_i(M)=(\widehat{\boldsymbol{\theta}}(i)-\widehat{\boldsymbol{\theta}})^TM(\widehat{\boldsymbol{\theta}}(i)-\widehat{\boldsymbol{\theta}}),$$

其中$\widehat{\boldsymbol{\theta}}(i)$为去掉第$i$组数据点$(y_i,\boldsymbol{x}_i^T)$后参数$\boldsymbol{\theta}$的估计值, M可取为某一正定矩阵, 例如$M=\partial^2_{\boldsymbol{\theta}}l_{\mathrm{E}}(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$, 同时, 也可得到经验似然距离

$$\mathrm{ELD}_i(M)=2\{l_{\mathrm{E}}(\widehat{\boldsymbol{\theta}})-l_{\mathrm{E}}(\widehat{\boldsymbol{\theta}}(i))\}.$$

若某个$\mathrm{ECD}_i(M)$或$\mathrm{ELD}_i(M)$较大时, 则可以认为第i组数据点为强影响点. 受Zhu等(2008)启发, 对删失线性模型, 可得$\widehat{\boldsymbol{\theta}}(i)$和$\widehat{\boldsymbol{t}}(i)$的一步近似结果为

$$\begin{aligned}\widehat{\boldsymbol{\theta}}(i) &= \widehat{\boldsymbol{\theta}} - n^{-1}S_{22.1}^{-1}S_{21}S_{11}^{-1}\boldsymbol{g}(\boldsymbol{x}_i,\widehat{\boldsymbol{\theta}})\cdot\{1+o_p(1)\},\\ \widehat{\boldsymbol{t}}(i) &= \widehat{\boldsymbol{t}} - n^{-1}(S_{11}^{-1}+S_{11}^{-1}S_{12}S_{22.1}^{-1}S_{11}^{-1})\boldsymbol{g}(\boldsymbol{x}_i,\widehat{\boldsymbol{\theta}})\cdot\{1+o_p(1)\}.\end{aligned}\tag{3.5}$$

这里$S_{11},S_{12},S_{22},S_{22.1}$如前定义, $\boldsymbol{g}(\boldsymbol{x}_i,\widehat{\boldsymbol{\theta}})=\boldsymbol{X}_i(y_{iG}-\boldsymbol{X}_i^T\widehat{\boldsymbol{\theta}})$.

3.3 模型的局部影响分析

设$\boldsymbol{\omega}=(\omega_1,\ldots,\omega_n)^T$为描述扰动因素的向量, 其定义域$\Omega$称为扰动空间, 它是$\mathbb{R}^n$上的某一开集. 受扰动后的对数经验似然函数定义为

$$l_{\mathrm{E}}(\boldsymbol{\theta}|\boldsymbol{\omega})=\sum_{i=1}^{n}\omega_i l_{\mathrm{E},i}(\boldsymbol{\theta}),$$

并且$\exists\,\boldsymbol{\omega}_0\in\Omega$, 使得$\forall\,\boldsymbol{\theta}$, 有$l_{\mathrm{E}}(\boldsymbol{\theta}|\boldsymbol{\omega}_0)=l_{\mathrm{E}}(\boldsymbol{\theta})$, 即当$\boldsymbol{\omega}=\boldsymbol{\omega}_0$时, 模型未受到扰动. 用$\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})$分别表示相应于原模型和扰动模型参数$\boldsymbol{\theta}$的极大经验似然估计, 且有$\widehat{\boldsymbol{\theta}}=\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}_0)$. 由以上假设, 经验似然距离为(Zhu等(2008))

$$\mathrm{LD}_{\mathrm{E}}(\boldsymbol{\omega})=2\{l_{\mathrm{E}}(\widehat{\boldsymbol{\theta}})-l_{\mathrm{E}}(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}))\}.$$

现考虑空间Ω中过$\boldsymbol{\omega}_0$以$\boldsymbol{h}$为方向的一条直线$\boldsymbol{\omega}(a)=\boldsymbol{\omega}_0+a\boldsymbol{h}$, 其中$\boldsymbol{\omega}(0)=\boldsymbol{\omega}_0$, $\boldsymbol{h}$为单位方向向量, a为实参数. 影响图$\boldsymbol{\pi}:\eta(\boldsymbol{\omega})=(\boldsymbol{\omega}^T,\mathrm{LD}_{\mathrm{E}}(\boldsymbol{\omega}))^T$在$\boldsymbol{\omega}=\boldsymbol{\omega}_0$处沿方向$\boldsymbol{h}$的曲率为

$$C_{\boldsymbol{h}}(\boldsymbol{\omega}_0)=\boldsymbol{h}^T H_{\mathrm{LD}_{\mathrm{E}}(\boldsymbol{\omega}_0)}\boldsymbol{h},$$

这里

$$H_{\mathrm{LD}_{\mathrm{E}}(\boldsymbol{\omega}_0)}=-2\frac{\partial^2\mathrm{LD}_{\mathrm{E}}(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}))}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}\Big|_{\boldsymbol{\omega}_0}=2\Delta^T\{-\partial_{\boldsymbol{\theta}}^2 l_{\mathrm{E}}(\boldsymbol{\theta})\}^{-1}\Delta\big|_{\boldsymbol{\omega}_0,\widehat{\boldsymbol{\theta}}},$$

其中$\Delta=\partial^2\mathrm{LD}_{\mathrm{E}}(\boldsymbol{\theta},\boldsymbol{\omega})/\partial\boldsymbol{\theta}\partial\boldsymbol{\omega}$为$p\times n$阶矩阵, 其第$(k,i)$个元素为$\partial_{\boldsymbol{\theta}_k}l_{\mathrm{E},i}(\boldsymbol{\theta})$. 影响曲率$C_{\boldsymbol{h}}(\boldsymbol{\omega}_0)$表示影响图$\boldsymbol{\pi}$在$\boldsymbol{\omega}_0$处沿方向$\boldsymbol{h}$的变化率, 它反映了模型对于$\boldsymbol{\omega}$沿$\boldsymbol{h}$方向扰动的敏感程度. 下面考虑基于影响曲率$C_{\boldsymbol{h}}(\boldsymbol{\omega}_0)$的局部影响度量. 根据对称矩阵的谱分解原理, 可以得到$H_{\mathrm{LD}_{\mathrm{E}}(\boldsymbol{\omega}_0)}$的谱分解如下

$$H_{\mathrm{LD}_{\mathrm{E}}(\boldsymbol{\omega}_0)}=\sum_{k=1}^{n}\lambda_k\boldsymbol{v}_k\boldsymbol{v}_k^T,$$

其中, λ_k为$H_{\mathrm{LD}_{\mathrm{E}}(\boldsymbol{\omega}_0)}$的特征值, $\boldsymbol{v}_k$为对应于λ_k的特征向量, 且有$\lambda_1\geq\cdots\geq\lambda_p=\cdots=\lambda_n=0$, $\{\boldsymbol{v}_k=(v_{k1},\ldots,v_{kn})^T:k=1,\ldots,n\}$为标准正交基, 即$H_{\mathrm{LD}_{\mathrm{E}}(\boldsymbol{\omega}_0)}\boldsymbol{v}_k=\lambda_k\boldsymbol{v}_k$. 由以上假设, 得到了局部影响分析的影响度量, 即对应于λ_1的特征向量$\boldsymbol{v}_1$为最大影响曲率方向. 据Zhu和Zhang (2004), 我们研究了影响度量

$$C_{\boldsymbol{e}_i}=\sum_{m=1}^{p}\lambda_m v_{mi}^2,\qquad i=1,\ldots,n,$$

指出了若某个$C_{\boldsymbol{e}_i}$较大($\boldsymbol{e}_i$指第i个分量为1, 其余分量为0的n维向量), 则对应的第i组数据点$(\boldsymbol{x}_i^T, y_i)$即为强影响点. 受Zhu等(2008)启发, 可得删失线性模型的诊断统计量的关系式

$$\begin{aligned} C_{\boldsymbol{e}_i} &= 2\mathrm{ECD}_i\{1+o_p(1)\} = 2\mathrm{ELD}_i\{1+o_p(1)\} = -2n^{-1}\Delta_i^T S_{22.1}^{-1}\Delta_i + o_p(1), \\ \sum_{i=1}^n C_{\boldsymbol{e}_i} &= 2\sum_{i=1}^n \mathrm{ECD}_i\{1+o_p(1)\} = 2\sum_{i=1}^n \mathrm{ELD}_i\{1+o_p(1)\} = 2p + o_p(1), \end{aligned} \tag{3.6}$$

其中, $\Delta_i = \partial_{\boldsymbol{\theta}} l_{\mathrm{E},i}(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}) = S_{21}S_{11}^{-1}\boldsymbol{g}(\boldsymbol{x_i}, \widehat{\boldsymbol{\theta}}) + o_p(1)$, $\boldsymbol{g}(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}) = \boldsymbol{X}_i(y_{iG} - \boldsymbol{X}_i^T\widehat{\boldsymbol{\theta}})$. 由此可知若某个$\mathrm{ECD}_i$或$C_{\boldsymbol{e}_i}$较大, 那么就可以认为第$i$个观测点$x_i$为强影响点.

3.4　模型的伪残差分析

在(2.1)式成立的条件下, 对每组观测数据的伪残差为

$$\boldsymbol{R}_i = (R_{i,1}, \ldots, R_{i,r})^T = \boldsymbol{g}(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}), \qquad i = 1, \ldots, n.$$

显然$\mathsf{E}_F(\boldsymbol{R}_i) \to 0$, $n \to \infty$. 进一步地, 受Zhu等(2008)启发, 研究了删失线性模型的标准化伪残差, 记$(\sigma_1^2, \ldots, \sigma_r^2) = \mathrm{diag}\{\mathsf{E}_F(\boldsymbol{g}^{\otimes 2})\}$, 其估计量为$(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_r^2)$, 标准化伪残差为

$$\boldsymbol{R}_i^s = (R_{i,1}^s, \ldots, R_{i,r}^s)^T = \Big(\frac{g_1(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})}{\widehat{\sigma}_1^2}, \ldots, \frac{g_r(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})}{\widehat{\sigma}_r^2}\Big)^T. \tag{3.7}$$

对估计量$(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_r^2)$, 有下面的近似公式

$$\begin{aligned} \mathsf{E}_F[g_k(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})] &\approx -n^{-1}\mathsf{E}_F\{\partial_{\boldsymbol{\theta}} g_k(\boldsymbol{x}_i)^T S_{22.1}^{-1}S_{21}S_{11}^{-1}\boldsymbol{g}(\boldsymbol{x}_i)\} - \frac{1}{n}\mathrm{tr}[\mathsf{E}_F\{\partial_{\boldsymbol{\theta}}^2 g_k(\boldsymbol{x}_i)\}S_{22.1}^{-1}]; \\ \widehat{\sigma}_k^2 &\approx \mathsf{Var}_F\{g_k(\boldsymbol{x}_i)\} - 2n^{-1}\mathsf{E}_F\{g_k(\boldsymbol{x}_i)\partial_{\boldsymbol{\theta}} g_k(\boldsymbol{x}_i)^T S_{22.1}^{-1}S_{21}S_{11}^{-1}\boldsymbol{g}(\boldsymbol{x}_i)\} \\ &\quad - n^{-1}\mathsf{E}_F\{\partial_{\boldsymbol{\theta}} g_k(\boldsymbol{x}_i)^T S_{22.1}^{-1}\partial_{\boldsymbol{\theta}} g_k(\boldsymbol{x}_i)\}, \end{aligned}$$

这里, $\boldsymbol{g}(\boldsymbol{x}_i) = \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$, $g_k(\boldsymbol{x}_i) = g_k(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$, $k = 1, \ldots, r$.

若对某个j, $|R_{i,j}^s|$较大(比如$|R_{i,j}^s| > 3$), 则可以认为第i组数据点为异常点. 以上各式中$\boldsymbol{X}_{ik}$为$\boldsymbol{X}_i$的第k个分量, 求期望之处用样本均值来代替, 例如

$$\mathsf{E}_F\{\partial_{\boldsymbol{\theta}} g_k(\boldsymbol{x}_i)^T S_{22.1}^{-1}\partial_{\boldsymbol{\theta}} g_k(\boldsymbol{x}_i)\} = \frac{1}{n}\sum_{i=1}^n \{\partial_{\boldsymbol{\theta}} g_k(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})^T S_{22.1}^{-1}\partial_{\boldsymbol{\theta}} g_k(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})\}$$

S_{ij}用$S_{nij}(\widehat{\boldsymbol{t}}, \widehat{\boldsymbol{\theta}})$代替且各式中均有$S_{22.1} = -S_{21}S_{11}^{-1}S_{12}$. 通过若干计算, 得到删失线性模型的计算公式如下

$$\begin{aligned} S_{n11} &= \partial_{\boldsymbol{t}} Q_{1,n} = \frac{1}{n}\sum_{i=1}^n \frac{\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})^T}{(1+\boldsymbol{t}^T\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}))^2}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}, \boldsymbol{t}=\widehat{\boldsymbol{t}}}; \\ S_{n12} &= \partial_{\boldsymbol{\theta}} Q_{1,n} = \frac{1}{n}\sum_{i=1}^n \frac{\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})\boldsymbol{t}^T\partial_{\boldsymbol{\theta}}\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}) - \partial_{\boldsymbol{\theta}}\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})(1+\boldsymbol{t}^T\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}))}{(1+\boldsymbol{t}^T\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}))^2}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}, \boldsymbol{t}=\widehat{\boldsymbol{t}}}; \\ S_{n22} &= \partial_{\boldsymbol{\theta}} Q_{2,n} = \frac{1}{n}\sum_{i=1}^n \frac{\partial_{\boldsymbol{\theta}}^T\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})\boldsymbol{t}\boldsymbol{t}^T\partial_{\boldsymbol{\theta}}\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})}{(1+\boldsymbol{t}^T\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}))^2}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}, \boldsymbol{t}=\widehat{\boldsymbol{t}}}, \end{aligned}$$

其中, $\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})=\boldsymbol{X}_i(y_{iG}-\boldsymbol{X}_i^T\boldsymbol{\theta})$, $\partial_{\boldsymbol{\theta}}\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{\theta})=-\boldsymbol{X}_i\boldsymbol{X}_i^T$. 至此我们就解决了删失线性模型的统计诊断问题, 它有以下几个关键步骤:

1. 对删失线性模型数据进行转换, 转换为一个关于全数据线性模型.
2. 由全数据线性模型求参数$\boldsymbol{\theta}$的极大经验似然估计$\widehat{\boldsymbol{\theta}}$(MELE).
3. 根据(3.5)、(3.6)、(3.7)计算各诊断统计量.
4. 由统计量散点图判定异常点.

§4. 模拟计算及实例分析

4.1 模拟数据分析

考虑下面二维线性回归模型

$$y_i=5+10x_i+\boldsymbol{e}_i,\qquad i=1,\ldots,150, \tag{4.1}$$

其中x_i产生于均匀分布$U[0,2]$, $\boldsymbol{e}_i\sim\chi^2_{20}-20$, 由模拟值$x_i$, $\boldsymbol{e}_i$和(4.1)式可求得y_i的模拟值, 令删失变量c_i服从指数分布$E(\lambda)$, 其中$\lambda=1/0.012$(此时删失率约为16%). 由此得到上述删失线性模型的观测数据$(\boldsymbol{X}_i^T,z_i,\delta_i)$如下

$$\boldsymbol{X}_i^T=(1,x_i),\qquad z_i=\min(y_i,c_i),\qquad \delta_i=\boldsymbol{I}\{y_i\le c_i\}.$$

令$y[50]=y[50]+30$, $y[100]=y[100]-30$, 此时50号点和100号点变为了异常点. 首先利用(3.1)将上述删失数据集$(\boldsymbol{X}_i^T,z_i,\delta_i)$转换为服从另一个线性模型的全数据集$(\boldsymbol{X}_i^T,y_{iG})$, 其中$\boldsymbol{X}_i^T=(1,x_i)$, $y_{iG}=\delta_i z_i/(1-G(z_i))$. 设全数据集$(\boldsymbol{X}_i^T,y_{iG})$服从如下线性模型

$$y_{iG}=\theta_1+\theta_2x_i+\boldsymbol{e}_i,$$

利用经验似然方法求出上述线性模型参数$\boldsymbol{\theta}$的MELE $\widehat{\boldsymbol{\theta}}$, 然后分别利用有关诊断统计量的公式计算出该模型下的诊断统计量$C_{\boldsymbol{e}_i}$和标准伪残差$R_{i,1}$, 散点图如下所示.

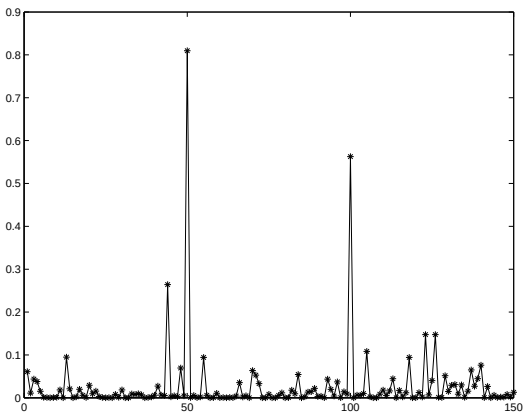

图1 $C_{\boldsymbol{e}_i}$的散点图

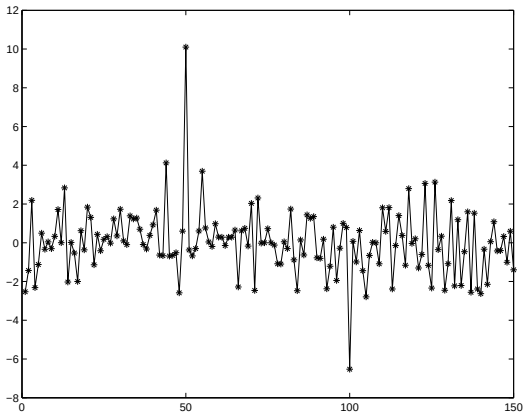

图2 标准化伪残差$R_{i,1}$

图1为$C_{\boldsymbol{e}_i}$的散点图, 由图1可以明显看出第50号点和第100号点为异常点. 图2为标准化伪残差$R_{i,1}$的散点图, 由图2可以明显看出第50号点和第100号点为异常点. 重复上述过程100次, 观察第50号点和第100号点分别能够被检验出来为异常点的次数. 得到在100次的模拟中第50号点被检验出了70次, 第100号点被检验出了78次, 后者被检出异常点的概率比前者要大, 这是因为在构造异常点时给50号点加了30, 这样由于50号点变大它被删失变量截掉的可能性也就变大, 而一旦它被删失变量截掉它就可能不再为异常点了, 而由于100号点减掉30它被删失变量截掉的可能性更小了, 因此它是异常点的概率就大于50号点为异常点的概率. 由上面的计算结果可以认为该方法是有效的. 尝试改变异常点设置的显著性, 给出检测为异常点的频率, 将50号点分别增加5、10、15进行100次模拟, 统计出检测为异常点的频率, 分别为0.05, 0.22, 0.45, 数值偏小, 说明效果不明显. 因为模拟中$\boldsymbol{e}_i \sim \chi^2_{20} - 20$, 标准差为$\sigma = \sqrt{40}$, 而增加量5, 10小于$2\sigma$, 因而被检测为异常点的可能性很小, 增加量15略大于2σ, 被检测为异常点的可能性变大, 当改变量为30时, 大于2σ, 被检测为异常点的次数增加, 模拟结果与判断标准[5]相一致.

4.2　实例分析

1967年10月1日Stanford大学启动了一项研究人工心脏移植手术存活时长的项目, 该项目截止于1974年4月1日, 在此期间共有69位病人接受了心脏移植手术, 研究人员对他们术后的存活时间进行了跟踪调查. 病人术后生存时间主要与2个独立的因素有关: 1. T5排异水平; 2. 病人接受心脏移植手术时的年龄. T5排异水平是一个度量心脏和接受体之间互斥程度的量, 一般认为当T5水平低于1.0时接受体与心脏结合情况会比较好, 当T5水平高于1.0时接受情况就比较差. 有四位病人的T5水平没有被测量到, 所以在对模型进行统计推断时我们忽略掉他们. 而年龄也与生存时间有较大关系, 一般年龄越大生存时间越短, 反之, 年龄越小生存时间越长. 数据见表1. 第二列表示病人术后生存时间; 第三列δ为研究项目结束时病人是否死亡, 1为死亡, 0为存活; 第四列x_1为$T5$水平值; 第五列x_2为病人接受手术时的年龄. 其中直到该研究项目截止时仍存活的有24位病人, 因此该24位病人的术后生存时间可看做是右截尾数据, 又因为他们接受手术的时间不同且有一定的随机性, 故可认为该组数据为随机右截尾数据.

首先建立线性模型(这里只考虑上面列出的两个主要因素)

$$y = \theta_1 + x_1\theta_2 + x_2\theta_3 + \varepsilon,$$

其中y为病人术后生存时间, x_1和x_2分别为病人T5水平和接受手术时的年龄. 而实际观测到的数据为(z_i, x_{1i}, x_{2i}), 其中$z_i = \min(y_i, c_i)$, y_i为病人实际存活时间, c_i为随机截尾值, 设其分布为G, x_{1i}, x_{2i}分别为第i个病人的T5水平和接受手术时的年龄. 类似于上例, 通过数据转换将删失数据集转换为一组服从另一个线性模型的全数据集, 然后根据(2.3)式求得全数据线性模型参数的MELE为$\widehat{\boldsymbol{\theta}} = (-0.8679, -0.1050, 0.0558)^T$, 进而求出相应的诊断统计量, 局部影响$C_{\boldsymbol{e}_i}$, 伪残差$R_{i,3}$, 散点图如图3和图4所示.

表1 心脏移植生存时间数据

序号	生存时间	δ	x_1	x_2	序号	生存时间	δ	x_1	x_2
1	15	1	1.11	54.3	34	838	0	0.19	41.6
2	3	1	1.66	40.4	35	65	1	0.66	49.1
3	624	1	1.32	50.0	36	815	0	1.93	32.7
4	46	1	0.61	42.5	37	551	1	0.12	48.9
5	127	1	0.36	48.0	38	66	1	1.12	51.3
6	64	1	1.89	54.6	39	228	1	1.02	19.7
7	1350	1	0.87	54.1	40	65	1	1.68	45.2
8	280	1	1.12	49.5	41	660	0	1.2	48
9	23	1	2.05	56.9	42	25	1	1.68	53
10	10	1	2.76	55.3	43	589	0	0.97	47.5
11	1024	1	1.13	43.4	44	592	0	1.46	26.7
12	39	1	1.38	42.8	45	63	1	2.16	56.4
13	730	1	0.96	58.4	46	12	1	0.61	29.2
14	136	1	1.62	52.0	47	499	0	1.7	52.2
15	1775	0	1.06	33.3	48	305	0	0.81	49.3
16	1	1	0.47	54.2	49	29	1	1.08	54
17	836	1	1.58	45.0	50	456	0	1.41	46.5
18	60	1	0.91	49	51	439	0	1.94	52.9
19	1536	0	0.96	58.4	52	48	1	3.05	53.4
20	1549	0	0.38	40.6	53	297	1	0.6	42.8
21	54	1	2.09	49	54	389	0	1.44	48.9
22	47	1	0.87	61.5	55	50	1	2.25	46.4
23	1	1	0.87	41.5	56	339	0	0.68	54.4
24	1367	0	0.75	48.6	57	68	1	1.33	51.4
25	1264	0	0.98	45.5	58	26	1	0.85	52.5
26	44	1	0	36.2	59	30	0	0.16	45.8
27	994	1	0.81	48.6	60	237	0	0.33	47.8
28	51	1	1.36	47.2	61	161	1	1.2	43.8
29	1106	0	1.35	36.8	62	167	0	0.46	26.7
30	253	1	1.08	48.8	63	110	0	1.78	23.7
31	51	1	1.51	52.5	64	13	0	0.77	28.9
32	875	0	0.98	38.9	65	1	0	0.67	35.2
33	322	1	1.82	48.1					

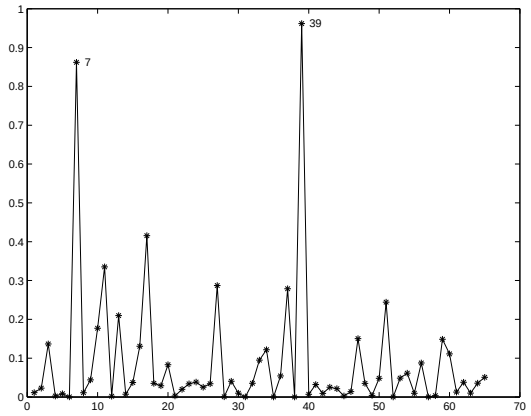

图3　$C_{\boldsymbol{e}_i}$的散点图

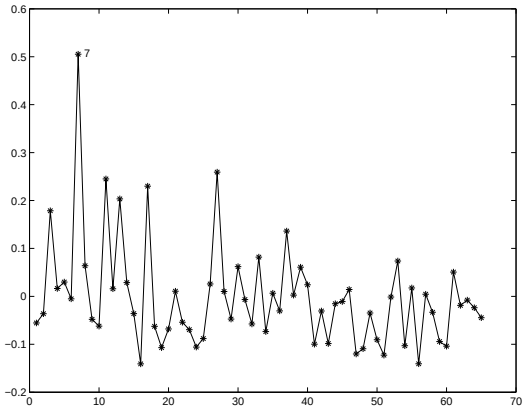

图4　标准化伪残差$R_{i,3}$

图3为$C_{\boldsymbol{e}_i}$的散点图, 从图3可以看出第7号点和39号点对应的统计量值明显比其它点大得多, 图4为标准化伪残差$R_{i,3}$的散点图, 由图4可以看出，第7号点的伪残差值比较大而第39号点的却不明显, 但是综合图3和图4我们还是可以认为第7号点和第39号点为异常点. 从表1中的数据可以看出第7号病人T5水平不是很低且年龄较大但其生存时间很长, 而第39号病人年龄很小且T5水平不是很高但其生存时间很短, 从数据出发和用诊断统计量得出的结论是一致的, 故有理由认为该诊断方法是有效的.

致谢　作者非常感谢审稿人提出的宝贵意见, 同时感谢李立宁对本文所做的部分数值模拟和计算工作.

参 考 文 献

[1] Belsley, D.A., Kuh, E. and Welsh, R.E., *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley, New York, 1980.

[2] Christensen, R., Pearson, L.M. and Johnson,W., Case-deletion diagnostics for mixed models, *Technometrics*, **34**(1992), 38–45.

[3] Critchley, F., Atkinson, R,A., Lu, G.B. and Biazi, E., Influence analysis based on the case sensitivity function, *Journal of the Royal Statistical Society: Series B*, **63**(2001), 307–323.

[4] Zhu, H.T. and Lee, S.Y., Local influence for generalized linear mixed models, *Canadian Journal of Statistics*, **31**(2003), 293–309.

[5] Zhu, H.T. and Lee, S.Y., Local influence for incomplete data models, *Journal of the Royal Statistical Society: Series B*, **63**(2001), 111–126.

[6] Zhu, H.T., Lee, S.Y., Wei, B.C. and Zhou, J.L., Case-deletion measures for models with incomplete data, *Biometrika*, **88**(2001), 727–737.

[7] Lee, S.Y. and Xu, L., Influence analysis of nonlinear mixed-effects models, *Computational Statistics & Data Analysis*, **45**(2004), 321–341.

[8] Xu, L., Lee, S.Y. and Poon, W.Y., Deletion measures for generalized linear mixed effects models, *Computational Statistics & Data Analysis*, **51**(2006), 1131–1146.

[9] Zhu, H.T., Ibrahim, J.G., Lee, S.Y. and Zhang, H.P., Perturbation selection and influence measures in local influence analysis, *The Annals of Statistics*, **35**(2007), 2565–2588.

[10] Owen, A., Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, **75**(1988), 237–249.

[11] Owen, A., Empirical likelihood for linear models, *The Annals of Statistics*, **19**(1991), 1725–1747.

[12] Owen, A., *Empirical Likelihood*, Chapman and Hall, New York, 2001.

[13] Kolaczyk, E.D., Empirical likelihood for generalized linear models, *Statistica Sinica*, **4**(1994), 199–218.

[14] Chen, S.X. and Cui, H.J., An extended empirical likelihood for generalized linear models, *Statistica Sinica*, **13**(2003), 69–81.

[15] Qin, J. and Lawless, J., Empirical likelihood and general estimating equations, *The Annals of Statistics*, **22**(1994), 300–325.

[16] Kitamura, Y., Asymptotic optimality of empirical likelihood for testing moment restrictions, *Econometrica*, **69**(2001), 1661–1672.

[17] Hall, P. and La Scala, B., Methodology and algorithms of empirical likelihood, *International Statisticl Review*, **58**(1990), 109–127.

[18] Zhu, H.T., Ibrahim, J.G., Tang N.S. and Zhang, H.P., Diagnostic measures for empirical likelihood of general estimating equations, *Biometrika*, **95**(2008), 489–507.

[19] 徐亮, 丁先文, 林金官, 基于经验似然的部分线性模型的统计诊断, 应用概率统计, **27**(2011), 91–102.

[20] Koul, H., Susarla, V. and Van Ryzin, J., Regression analysis with randomly right-censored data, *The Annals of Statistics*, **9**(1981), 1276–1288.

[21] Susarla, V. and Van Ryzin, J., Large sample theory for an estimator of the mean survival time from censored samples, *The Annals of Statistics*, **8**(1980), 1002–1016.

[22] Zhu, H.T. and Zhang, H.P., A diagnostic procedure based on local influence, *Biometrika*, **91**(2004), 579–589.

Diagnostic Measures for Censored Linear Models based on Empirical Likelihood Method

Ding Xianwen

(*School of Mathematical and Physics, Jiangsu University of Technology, Changzhou, 213001*)

Xu Liang

(*Department of Mathematics, Southeast University, Nanjing, 211189*)

In this paper, the diagnostic measures for censored linear models are studied based on the empirical likelihood method. First, the diagnostic measures for linear models are studied; Then, the censored linear models are converted to linear models, and the diagnostic measures for converted models are studied; Last, simulation studies and real data analysis are given to illustrate the validity of statistical diagnostic measures.

Keywords: Empirical likelihood, model convert, statistical diagnosis, validity.

AMS Subject Classification: 62J20.