第 36 卷 第 6 期 应 用 数 学 学 报 Vol. 36 No. 6
2013 年 11 月 ACTA MATHEMATICAE APPLICATAE SINICA Nov., 2013

基于模糊贴近度的模糊线性回归模型的统计诊断 *

张爱武

(盐城师范学院数学科学学院, 盐城 224002)

(E-mail: zaw_017@163.com)

摘 要 针对输入为实数、输出为模糊数的模糊线性回归模型, 讨论了基于数据删除的模型参数估计, 建立了检验观测数据中强影响点或异常点的统计诊断量, 并给出了运用该诊断量进行检验的一般步骤. 实例表明该统计诊断量是有效的.

关键词 模糊线性回归模型; 数据删除模型; 模糊贴近度

MR(2000) 主题分类 62G68

中图分类 O212.2; O159

1 引言

回归分析是分析因变量与一个或多个自变量之间相关关系的一种重要而全面的方法, 它在工程科学、社会科学、经济与金融领域都有着非常广泛的应用. 传统的回归分析方法往往要求自变量和因变量是精确的数据, 然而, 在实际问题中这些量往往不是精确的数据, 即用模糊数据来描述的. 例如观测的量是用语言来描述的, 如大的、重的、或近似等于 3 等等. 对这些问题的分析仅依靠传统的回归分析不能完全得到人们想要的结果, 因为一些指标的度量是模糊的. 借助于 Zadeh[1] 提出的模糊集理论, 研究者建立了不同的模糊回归分析模型及相应的解决问题的方法. 首先研究这个课题的是 Tanaka[2], Diamond[3] 提出了估计模糊回归系数的最小二乘法 (FLS), 其 FLS 估计与传统的 LS 估计相类似, 随后, Savic 和 Pedrycz[4] 将 FLS 与线性规划结合起来, 建立了模糊回归分析模型的两步法; 最近, Chang[5] 对模糊回归方法进行了比较, 总结了模糊回归的三种方法, 即最小模糊准则, 最小二乘拟合准则和区间回归分析方法. 模糊回归模型的主要目的是建立最小误差的好的模型, 根据我们定义的误差的多少, 这种方法可以分为两

本文 2011 年 8 月 22 日收到. 2013 年 9 月 12 日收到修改稿.

* 国家自然科学基金 (11171065) 和江苏省自然科学基金 (BK2011058) 资助项目.

类：一类是可能性方法 [2], 就是将模糊回归问题转化为目标函数是模糊回归系数的边宽总和最小的带约束条件的线性规划模型问题，另一类是最小二乘方法 [3], 就是通过两个模糊数之间的距离，使得偏差的平方和达到最小，从而达到确定回归系数的目的.

统计诊断就是对统计推断方法解决问题的全过程进行诊断，影响分析是统计诊断中十分重要的分支. 目前对建立在确定性数据基础上的线性回归模型统计诊断的研究已非常深入，理论上较为成熟，应用上也最为成功 [6,7]. 目前对模糊线性回归分析模型的诊断方法已有一些结果 [8−12], 但上述文献处理的方法都是采用的可能性方法，即将目标函数的最小化转化为一个线性规划问题来研究的，而当样本容量较大时，转化而成的线性规划问题的变量和附加变量会急剧增加，从而使得求解线性规划问题的工作量和计算量都非常大，同时这些方法都涉及到有关临界值的选取，这本身就是一个难点 [13]. 本文采用最小二乘的方法，针对应用非常广泛的输入为精确数 (实数) 、输出为模糊数的模糊线性回归分析模型，讨论了基于数据删除的模糊线性回归模型的参数估计，建立了检验观测数据中强影响点或异常点的统计诊断量 – 模糊贴近度，通过实例说明了该统计量的有效性.

2　模糊回归模型及参数估计

2.1　模糊数与模糊线性回归模型

定义 2.1[1]　实数域 R 上的模糊集 $\widetilde{A}$ 称为一个模糊数，若满足：

(1)　存在 $x_0 \in R$, 使得 $A(x_0) = 1$;

(2)　任意 $\alpha \in [0,1]$, $\widetilde{A}_\alpha = \{x | A(x) \geq \alpha\} = [\underline{A}_\alpha, \overline{A}_\alpha]$ 是一个闭区间.

这里 $A(x)$ 为 $\widetilde{A}$ 的隶属函数.

R 上的全体模糊数记为 $\widetilde{R}$.

若模糊数 $\widetilde{A}$ 的隶属函数为

$$A(x) = \begin{cases} 1 - \dfrac{m-x}{\alpha}, & m - \alpha \leq x \leq m, \ \ \alpha > 0, \\ 1 - \dfrac{x-m}{\beta}, & m \leq x \leq m + \beta, \ \ \beta > 0, \\ 0, & \text{其他}, \end{cases}$$

称 $\widetilde{A}$ 为三角模糊数，记为 (m, α, β). 特别地，当 $\alpha = \beta$ 时，称 $\widetilde{A}$ 为对称的三角模糊数，记为 (m, α).

假定 $(\widetilde{x_i}, \widetilde{y_i})$ $(i = 1, 2, \cdots, n)$ 是一组模糊输入和输出的观测数据，其中 $\widetilde{x_i}, \widetilde{y_i} \in \widetilde{R}$. 模糊线性回归模型就可以表示为

$$\widetilde{y} = \widetilde{\beta_0} + \widetilde{\beta_1}\widetilde{x}, \widetilde{\beta_0}, \widetilde{\beta_1} \in \widetilde{R}. \tag{2.1}$$

由最小二乘原理，我们的目的就是要确定模糊参数 $\widetilde{\beta_0}$ 和 $\widetilde{\beta_1}$, 使其在距离 d 的意义下关于模糊数据 $(\widetilde{x_i}, \widetilde{y_i})$ $(i = 1, 2, \cdots, n)$ 的误差平方和最小，即

$$\min \phi(\widetilde{\beta_0}, \widetilde{\beta_1}) = \sum_{i=1}^{n} d^2(\widetilde{y_i}, \widetilde{\beta_0} + \widetilde{\beta_1}\widetilde{x_i}).$$

两个模糊数 $\widetilde{a}=(a^-(\lambda),a^+(\lambda)),\ \widetilde{b}=(b^-(\lambda),b^+(\lambda))$ 的距离 [3] 为

$$d(\widetilde{a},\widetilde{b})=\Big(\int_0^1 f(\lambda)d^2(a_\lambda,b_\lambda)\,\mathrm{d}\lambda\Big)^{\frac{1}{2}},\tag{2.2}$$

其中 $d^2(a_\lambda,b_\lambda)=(a^-(\lambda)-b^-(\lambda))^2+(a^+(\lambda)-b^+(\lambda))^2, f(\lambda)$ 为 $[0,1]$ 上的单调增函数，且 $f(0)=0,\ \int_0^1 f(\lambda)\,\mathrm{d}\lambda=\frac{1}{2}$. 于是就转化为求方程组

$$\begin{cases}\delta_{\widetilde{\beta}_0}\phi(\widetilde{\beta}_0,\widetilde{\beta}_1)=0,\\ \delta_{\widetilde{\beta}_1}\phi(\widetilde{\beta}_0,\widetilde{\beta}_1)=0\end{cases}\tag{2.3}$$

在 $\widetilde{R}$ 上的解.

2.2 模糊线性回归模型的参数估计

为了讨论问题的方便，我们仅对输入为精确数，输出为模糊数的线性回归模型的情形进行讨论.

考虑模糊线性回归模型

$$\widetilde{y}=\widetilde{\beta_0}+\widetilde{\beta_1}x.$$

下面对参数的不同形式给出参数的估计.

2.2.1 $\beta_0\in R,\ \widetilde{\beta}_1\in\widetilde{R}$

引理 2.1[14] 设 $\widetilde{y}=\beta_0+\widetilde{\beta_1}\widetilde{x},(x_i,\widetilde{y_i}),\ i=1,2,\cdots,n$ 是一组观测数据，则

$$\begin{aligned}&\widehat{\beta}_0=\frac{\displaystyle\int_0^1\Big[\sum_{i=1}^n x_i^2\cdot\sum_{i=1}^n\overline{y}_i(\lambda)-\sum_{i=1}^n x_i\cdot\sum_{i=1}^n x_i\cdot\overline{y}_i(\lambda)\Big]\,\mathrm{d}\lambda}{\displaystyle n\sum_{i=1}^n x_i^2-\Big(\sum_{i=1}^n x_i\Big)^2},\\ &\widehat{\beta}_1^-(\lambda)=\frac{\displaystyle\sum_{i=1}^n x_i\cdot y_i^-(\lambda)-\widehat{\beta}_0\sum_{i=1}^n x_i}{\displaystyle\sum_{i=1}^n x_i^2},\qquad \widehat{\beta}_1^+(\lambda)=\frac{\displaystyle\sum_{i=1}^n x_i\cdot y_i^+(\lambda)-\widehat{\beta}_0\sum_{i=1}^n x_i}{\displaystyle\sum_{i=1}^n x_i^2},\end{aligned}\tag{2.4}$$

其中 $\overline{y}_i(\lambda)=\frac{1}{2}[y_i^-(\lambda)+y_i^+(\lambda)]$.

推论 2.2 设 $\widetilde{y}=\beta_0+\widetilde{\beta_1}x,(x_i,\widetilde{y_i}),\ i=1,2,\cdots,n$ 是一组观测数据，当观测数据 $\widetilde{y_i}=(y_i,l_i,r_i),\ i=1,2,\cdots,n$ 为三角模糊数，则

$$\begin{aligned}\widehat{\beta}_0 &= \frac{\sum\limits_{i=1}^{n} x_i^2 \cdot \sum\limits_{i=1}^{n} y_i - \sum\limits_{i=1}^{n} x_i \cdot \sum\limits_{i=1}^{n} x_i y_i - \frac{1}{4}\sum\limits_{i=1}^{n} x_i^2 (l_i - r_i) + \sum\limits_{i=1}^{n} x_i \cdot \sum\limits_{i=1}^{n} x_i (l_i - r_i)}{n\sum\limits_{i=1}^{n} x_i^2 - \big(\sum\limits_{i=1}^{n} x_i\big)^2},\\ \widehat{\widetilde{\beta}}_1 &= \left(\frac{\sum\limits_{i=1}^{n} x_i y_i - \widetilde{\beta}_0 \sum\limits_{i=1}^{n} x_i}{\sum\limits_{i=1}^{n} x_i^2}, \frac{\sum\limits_{i=1}^{n} x_i l_i}{\sum\limits_{i=1}^{n} x_i^2}, \frac{\sum\limits_{i=1}^{n} x_i r_i}{\sum\limits_{i=1}^{n} x_i^2} \right).\end{aligned} \tag{2.5}$$

特别地，当观测数据 $\widetilde{y} = (y_i, \mu_i)$ 为对称的三角模糊数时，则有

$$\widehat{\beta}_0 = \frac{\sum\limits_{i=1}^{n} x_i^2 \cdot \sum\limits_{i=1}^{n} y_i - \sum\limits_{i=1}^{n} x_i \cdot \sum\limits_{i=1}^{n} x_i y_i}{n\sum\limits_{i=1}^{n} x_i^2 - \big(\sum\limits_{i=1}^{n} x_i\big)^2}, \qquad \widehat{\widetilde{\beta}}_1 = \left(\frac{\sum\limits_{i=1}^{n} x_i y_i - \widehat{\beta}_0 \sum\limits_{i=1}^{n} x_i}{\sum\limits_{i=1}^{n} x_i^2}, \frac{\sum\limits_{i=1}^{n} x_i \mu_i}{\sum\limits_{i=1}^{n} x_i^2} \right), \tag{2.6}$$

其中 $\widehat{\widetilde{\beta}}_1$ 是对称的三角模糊数.

2.2.2 $\widetilde{\beta}_0 \in \widetilde{R},\ \beta_1 \in R$

引理 2.3[14] 设 $\widetilde{y} = \widetilde{\beta}_0 + \beta_1 x,\ (x_i, \widetilde{y}_i)\ i = 1, 2, \cdots, n$ 是一组观测数据，则

$$\begin{aligned}\widehat{\beta}_0^-(\lambda) &= \frac{\sum\limits_{i=1}^{n} y_i^-(\lambda) - \widehat{\beta}_1 \sum\limits_{i=1}^{n} x_i}{n}, \qquad \widehat{\beta}_0^+(\lambda) = \frac{\sum\limits_{i=1}^{n} y_i^+(\lambda) - \widehat{\beta}_1 \sum\limits_{i=1}^{n} x_i}{n},\\ \widehat{\beta}_1 &= \frac{\sum\limits_{i=1}^{n} \int_0^1 \big(nx_i - \sum\limits_{i=1}^{n} x_i\big)(y_i^-(\lambda) + y_i^+(\lambda))\,\mathrm{d}\lambda}{2\big(n\sum\limits_{i=1}^{n} x_i^2 - \big(\sum\limits_{i=1}^{n} x_i\big)^2\big)}.\end{aligned} \tag{2.7}$$

推论 2.4 设 $\widetilde{y} = \widetilde{\beta}_0 + \beta_1 x, (x_i, \widetilde{y}_i),\ i = 1, 2, \cdots, n$ 是一组观测数据，当观测数据 $\widetilde{y}_i = (y_i, l_i, r_i)$ 为三角模糊数时，则

$$\begin{aligned}\widehat{\widetilde{\beta}}_0 &= \left(\frac{\sum\limits_{i=1}^{n} y_i - \widehat{\beta}_1 \sum\limits_{i=1}^{n} x_i}{n}, \frac{\sum\limits_{i=1}^{n} l_i}{n}, \frac{\sum\limits_{i=1}^{n} r_i}{n} \right),\\ \widehat{\beta}_1 &= \frac{\sum\limits_{i=1}^{n} \big(nx_i - \sum\limits_{i=1}^{n} x_i\big)\, y_i - \frac{1}{4}\sum\limits_{i=1}^{n} \big(nx_i - \sum\limits_{i=1}^{n} x_i\big)(l_i - r_i)}{n\sum\limits_{i=1}^{n} x_i^2 - \big(\sum\limits_{i=1}^{n} x_i\big)^2},\end{aligned} \tag{2.8}$$

其中 $\widehat{\widetilde{\beta}}_0$ 为三角模糊数.

特别地，当观测数据 $\widetilde{y}_i=(y_i,\mu_i)$ 为对称的三角模糊数时，则

$$\widehat{\widetilde{\beta}}_0=\left(\frac{\sum\limits_{i=1}^{n}y_i-\widehat{\beta}_1\sum\limits_{i=1}^{n}x_i}{n},\frac{\sum\limits_{i=1}^{n}\mu_i}{n}\right),\qquad \widehat{\beta}_1=\frac{\sum\limits_{i=1}^{n}\big(nx_i-\sum\limits_{i=1}^{n}x_i\big)y_i}{n\sum\limits_{i=1}^{n}x_i^2-\big(\sum\limits_{i=1}^{n}x_i\big)^2},\tag{2.9}$$

其中 $\widehat{\widetilde{\beta}}_0$ 为对称的三角模糊数.

3 数据删除模糊线性回归模型的参数估计及统计诊断

3.1 基于数据删除的模糊线性回归模型

对于模糊线性回归模型 (2.1), 为了评价第 i 个数据点 $(\widetilde{x}_i,\widetilde{y}_i)$ 在回归分析中的作用与影响，可通过比较第 i 个数据点 $(\widetilde{x}_i,\widetilde{y}_i)$ 删除前后统计推断结果的变化，来检测这个点是否为异常点或强影响点. 删除第 i 个数据点以后的模型称为数据删除模型，其形式为

$$\widetilde{y}_{[i]}=\widetilde{\beta}_{0[i]}+\widetilde{\beta}_{1[i]}\widetilde{x},\qquad \text{其中 }\ \widetilde{\beta}_{0[i]},\widetilde{\beta}_{1[i]}\in\widetilde{R}.\tag{3.1}$$

差值 $\widetilde{y}_{[i]}-\widetilde{y}$ 就是第 i 个数据点对模型影响大小的一种度量，差值越大，影响就越大.

3.2 基于数据删除的模糊线性回归模型的参数估计

下面我们讨论模型 (3.1) 的参数估计问题，仅考虑输入为实数，输出为模糊数的情形.

由模型 (3.1) 及式 (2.5) 容易得到：

定理 3.1 设 $\widetilde{y}_{[i]}=\beta_{0[i]}+\widetilde{\beta}_{1[i]}x$, 其中 $\beta_{0[i]}\in R$, $\widetilde{\beta}_{1[i]}\in\widetilde{R}$, $(x_k,\widetilde{y}_k)$, $k\neq i$, $k=1,2,\cdots,n$ 是一组观测数据， $\widetilde{y}_k=(y_k,l_k,r_k)$ 为三角模糊数，则

$$\widehat{\beta}_{0[i]}=\frac{\sum\limits_{k\neq i}x_k^2\cdot\sum\limits_{k\neq i}y_k-\sum\limits_{k\neq i}x_k\cdot\sum\limits_{k\neq i}x_ky_k-\frac{1}{4}\sum\limits_{k\neq i}x_k^2(l_k-r_k)+\sum\limits_{k\neq i}x_k\cdot\sum\limits_{k\neq i}x_k(l_k-r_k)}{(n-1)\sum\limits_{k\neq i}x_k^2-\big(\sum\limits_{k\neq i}x_k\big)^2},$$

$$\widehat{\widetilde{\beta}}_{1[i]}=\left(\frac{\sum\limits_{k\neq i}x_ky_k-\widehat{\beta}_{0[i]}\sum\limits_{k\neq i}x_k}{\sum\limits_{k\neq i}x_k^2},\frac{\sum\limits_{k\neq i}x_kl_k}{\sum\limits_{k\neq i}x_k^2},\frac{\sum\limits_{k\neq i}x_kr_k}{\sum\limits_{k\neq i}x_k^2}\right).\tag{3.2}$$

推论 3.2 设 $\widetilde{y}_{[i]}=\beta_{0[i]}+\widetilde{\beta}_{1[i]}x$, 其中 $\beta_{0[i]}\in R$, $\widetilde{\beta}_{1[i]}\in\widetilde{R}$, $(x_k,\widetilde{y}_k)$, $k\neq i$, $k=1,2,\cdots,n$

是一组观测数据，当 $\widetilde{y}_k=(y_k,\mu_k,\mu_k)$ 为对称的三角模糊数，则

$$\begin{aligned}\widehat{\beta}_{0[i]}&=\frac{\sum\limits_{k\neq i}x_k^2\cdot\sum\limits_{k\neq i}y_k-\sum\limits_{k\neq i}x_k\cdot\sum\limits_{k\neq i}x_ky_k}{(n-1)\sum\limits_{k\neq i}x_k^2-\big(\sum\limits_{k\neq i}x_k\big)^2},\\ \widehat{\widetilde{\beta}}_{1[i]}&=\left(\frac{\sum\limits_{k\neq i}x_ky_k-\widehat{\beta}_{0[i]}\sum\limits_{k\neq i}x_k}{\sum\limits_{k\neq i}x_k^2},\frac{\sum\limits_{k\neq i}x_k\mu_k}{\sum\limits_{k\neq i}x_k^2},\frac{\sum\limits_{k\neq i}x_k\mu_k}{\sum\limits_{k\neq i}x_k^2}\right).\end{aligned}\tag{3.3}$$

由模型 (3.1) 及式 (2.7) 可以得到：

定理 3.3 设 $\widetilde{y}_{[i]}=\widehat{\beta}_{0[i]}+\beta_{1[i]}x$, 其中 $\beta_{0[i]}\in\widetilde{R}$, $\beta_{1[i]}\in R$, $(x_k,\widetilde{y}_k)$, $k\neq i$, $k=1,2,\cdots,n$ 是一组观测数据，$\widetilde{y}_k=(y_k,l_k,r_k)$ 为三角模糊数，则

$$\begin{aligned}\widehat{\widetilde{\beta}}_{0[i]}&=\left(\frac{\sum\limits_{k\neq i}y_k-\widehat{\beta}_{1[i]}\sum\limits_{k\neq i}x_k}{n-1},\frac{\sum\limits_{k\neq i}l_k}{n-1},\frac{\sum\limits_{k\neq i}r_k}{n-1}\right),\\ \widehat{\beta}_{1[i]}&=\frac{\sum\limits_{k\neq i}\big((n-1)x_k-\sum\limits_{k\neq i}x_k\big)y_k-\frac{1}{4}\sum\limits_{k\neq i}\big((n-1)x_k-\sum\limits_{k\neq i}x_k\big)(l_k-r_k)}{(n-1)\sum\limits_{k\neq i}x_k^2-\big(\sum\limits_{k\neq i}x_k\big)^2}.\end{aligned}\tag{3.4}$$

推论 3.4 设 $\widetilde{y}_{[i]}=\widetilde{\beta}_{0[i]}+\beta_{1[i]}x$, 其中 $\widetilde{\beta}_{0[i]}\in\widetilde{R}$, $\beta_{1[i]}\in R$, $(x_k,\widetilde{y}_k)$, $k\neq i$, $k=1,2,\cdots,n$ 是一组观测数据，当 $\widetilde{y}_k=(y_k,\mu_k,\mu_k)$ 为对称的三角模糊数，则

$$\begin{aligned}\widehat{\widetilde{\beta}}_{0[i]}&=\left(\frac{\sum\limits_{k\neq i}y_k-\widehat{\beta}_{1[i]}\sum\limits_{k\neq i}x_k}{n-1},\frac{\sum\limits_{k\neq i}\mu_k}{n-1},\frac{\sum\limits_{k\neq i}\mu_k}{n-1}\right),\\ \widehat{\beta}_{1[i]}&=\frac{\sum\limits_{k\neq i}\big((n-1)x_k-\sum\limits_{k\neq i}x_k\big)y_k}{(n-1)\sum\limits_{k\neq i}x_k^2-\big(\sum\limits_{k\neq i}x_k\big)^2}.\end{aligned}\tag{3.5}$$

4　基于模糊贴近度的模糊线性回归模型的统计诊断

由于 $\widetilde{y}_{[i]}$ 与 $\widetilde{y}$ 为两个模糊数，不便于比较，必须选择一个合适的距离，以便定量地比较影响的大小．在这里通过模糊数之间的距离来建立两个模糊回归方程之间的模糊贴近度来研究模糊线性回归模型的统计诊断.

在 (2.2) 式定义的两个模糊数距离的意义下，两个三角模糊数之间的距离可以定义为：

定义 4.1 设 $\widetilde{X}=(x,\xi,\eta)$, $\widetilde{Y}=(y,\lambda,\mu)$ 为两个三角模糊数，则称

$$d(\widetilde{X},\widetilde{Y})=\big([x-y+(\xi-\lambda)]^2+[x-y-(\eta-\mu)]^2+(x-y)^2\big)^{\frac{1}{2}} \tag{4.1}$$

为 $\widetilde{X}$ 与 $\widetilde{Y}$ 之间的距离.

特别地，当 $\widetilde{X}$ 与 $\widetilde{Y}$ 为对称的三角模糊数时，

$$d(\widetilde{X},\widetilde{Y})=\big(3(x-y)^2+2(\xi-\lambda)^2\big)^{\frac{1}{2}}. \tag{4.2}$$

有了模糊数的距离，下面给出模糊贴近度的公理化定义及计算公式：

定义 4.2[15] 设 $\widetilde{X}$ 和 $\widetilde{Y}$ 是两个模糊数. 定义 $S=S(\widetilde{X},\widetilde{Y})$, 若 S 满足：

(1) $0\le S\le 1$;

(2) 对于 $\widetilde{X}=\widetilde{Y}$, $S=1$;

(3) $S(\widetilde{X},\widetilde{Y})=S(\widetilde{Y},\widetilde{X})$;

(4) 当且仅当 $\widetilde{X}\cap\widetilde{Y}=\emptyset$ 时， $S(\widetilde{X},\widetilde{Y})=0$;

(5) 当 $\widetilde{X}\subset\widetilde{Y}\subset\widetilde{Z}$ 时，有 $S(\widetilde{X},\widetilde{Y})\ge S(\widetilde{X},\widetilde{Z})$,

则称 S 为 $\widetilde{X},\widetilde{Y}$ 的贴近度，即 $\widetilde{X}$ 与 $\widetilde{Y}$ 的接近程度.

从定义可知，贴近度是刻划了两个模糊集接近程度的一种度量. $S(\widetilde{X},\widetilde{Y})$ 值越大，表示 $\widetilde{X}$ 与 $\widetilde{Y}$ 越贴近. $S(\widetilde{X},\widetilde{Y})=1$, 表示 $\widetilde{X}$ 与 $\widetilde{Y}$ 完全相同， $S(\widetilde{X},\widetilde{Y})=0$, 表示 $\widetilde{X}$ 与 $\widetilde{Y}$ 完全不一致.

模糊量之间的贴近度计算方法很多，为实现的可靠性和方便性，定义基于距离度量的贴近度计算方法：

定义 4.3 设 $\widetilde{X}=(x,\xi,\eta)$, $\widetilde{Y}=(y,\lambda,\mu)$ 为两个三角模糊数. 称

$$S(\widetilde{X},\widetilde{Y})=\frac{d(\widetilde{X},\widetilde{Y})}{1+d(\widetilde{X},\widetilde{Y})} \tag{4.3}$$

为 $\widetilde{X}$ 与 $\widetilde{Y}$ 之间的贴近度. 式中 $d(\widetilde{X},\widetilde{Y})$ 由 (4.1) 式给出.

将两个模糊数的贴近度的计算公式加以改造，可以定义两个模糊回归方程的贴近度.

定义 4.4 设 $\widetilde{y}_i,\widetilde{y}'$ 是两个模糊回归方程对应于 x_i $(i=1,2,\cdots,n)$ 的预测值，称

$$S(L_1,L_2)=\frac{1}{n-2}\sum_{i=1}^{n}S(\widetilde{y}_i,\widetilde{y}') \tag{4.4}$$

为模糊回归方程 L_1 与 L_2 的贴近度. 式中 $S(\widetilde{y}_i,\widetilde{y}')$ 为模糊数 $\widetilde{y}_i,\widetilde{y}'$ 的贴近度.

在定义 4.4 中，分母用 $n-2$, 这是由于根据极值问题求得参数 $\widetilde{\beta}_0,\widetilde{\beta}_1$ 之后，再由回归方程去估计回归值，它已满足了两个条件，因而失去了两个自由度.

当 $S(L_1,L_2)$ 的值较大时，说明模糊回归方程 L_1 与 L_2 比较接近，反之，说明 L_1 与 L_2 误差较大.

在数据删除的模糊线性回归模型中，从直观上看，当删除第 i 个数据点后得到的模糊线性回归方程 L_i 与原回归方程 L 的贴近度较大时，说明第 i 个数据点为正常点；当

删除第 i 个数据点后得到的模糊线性回归方程 L_i 与原回归方程 L 的贴近度较小时，说明第 i 个数据点可能为异常点或强影响点.

为了从理论上判断删除的第 i 个数据点是否为异常点或强影响点，我们可以引进第 i 个数据点的中心模糊线性回归方程及模糊贴近度的相对偏差的概念.

定义 4.5 设样本集为 $\{(x_i,\widetilde{y}_i)\mid \widetilde{y}_i=(y_i,l_i,r_i),\ i=1,2,\cdots,n\}$, 称 $(\overline{x},\overline{\widetilde{y}})$ 为样本集的重心坐标，其中 $\overline{x}=\frac{1}{n}\sum\limits_{i=1}^{n}x_i$, $\overline{\widetilde{y}}=\big(\frac{1}{n}\sum\limits_{i-1}^{n}y_i,\frac{1}{n}\sum\limits_{i=1}^{n}l_i,\frac{1}{n}\sum\limits_{i=1}^{n}r_i\big)$. 将样本集中的第 i 个数据点 $(x_i,\widetilde{y}_i)$ 换成样本的重心坐标 $(\overline{x},\overline{\widetilde{y}})$, 这些数据满足的模糊线性回归方程 $\overline{L}_i$ $(i=1,2,\cdots,n)$ 称为第 i 个数据点的中心模糊线性回归方程.

定义 4.6 称

$$D_i=\frac{\big|S(L,L_i)-S(L,\overline{L}_i)\big|}{S(L,\overline{L}_i)},\qquad i=1,2,\cdots,n \tag{4.5}$$

为第 i 个数据点贴近度的相对偏差，其中 $S(L,\overline{L}_i)$ 为原回归方程与第 i 个数据点的中心模糊线性回归方程的贴近度.

一般认为，若相对偏差 $D_i>5\%$, 则第 i 个数据点就是对参数 $\widetilde{\beta}_0,\widetilde{\beta}_1$ 的估计影响特别大的点，因而可能是异常点或强影响点.

在具体作数据分析时，我们可以按下列步骤来操作：

第 1 步　计算出原模糊线性回归方程 L;

第 2 步　分别计算出删除第 i 个数据点后得到的模糊线性回归方程 L_i $(i=1,2,\cdots,n)$ 及原模糊线性回归方程 L 与模糊线性回归方程 L_i 的模糊贴近度 $S(L,L_i)$ $(i=1,2,\cdots,n)$;

第 3 步　分别求出第 i 个数据点的中心模糊线性回归方程及 $\overline{L}_i$ 与 L 的模糊贴近度 $S(L,\overline{L}_i)$ $(i=1,2,\cdots,n)$;

第 4 步　分别计算出每个数据点的相对偏差 D_i $(i=1,2,\cdots,n)$;

第 5 步　若相对偏差 $D_i>5\%$, 一般可认为第 i 个数据点可能是异常点或强影响点.

当然在诊断强影响点或异常点时，还必须考虑强影响点或异常点的实际意义以及具体数据的实际背景等.

5　实例分析

某信息公司近年来对产品销售量进行调查，得到的数据资料如表 1, 其中销售量 $\widetilde{y}_i$ 为三角模糊数. 应用前面提出的各种模型，给出问题的模糊线性回模型的分析结果，并讨论数据中的异常点.

表 1　产品销售情况 (数据来自 [14])

$\overline{x}_i$	$\widetilde{y}_i=(y_i,l_i,u_i)$	$\overline{x}_i$	$\widetilde{y}_i=(y_i,l_i,u_i)$
1987	(230,2,1)	1992	(257,2,1)
1988	(236,3,2)	1993	(262,3,2)
1989	(241,2,3)	1994	(276,3,3)
1990	(246,1,2)	1995	(281,2,3)
1991	(252,2,2)	1996	(286,1,2)

为了方便起见，记 $x_i=\overline{x}_i-1986$, 设 $\widetilde{y}=\beta_0+\widetilde{\beta}_1x$, 由 (2.5), 可以求得：

$$\beta_0=221.1685,\qquad \widetilde{\beta}_1=(6.4357,0.2935,0.3195),$$

即回归方程为

$$\widetilde{y}=221.1685+(6.4357,0.2935,0.3195)x.$$

利用 (3.3), (4.3) 及 (4.4) 式可求得删除第 i 个数据点后的回归方程及删除第 i 个数据点后的回归方程与原回归方程的贴近度 (见表 2), 第 i 个数据点的中心模糊线性回归方程及第 i 个数据点的中心模糊线性回归方程与原回归方程的贴近度 (见表 3).

表 2 删除数据点后的回归方程及删除数据点后的回归方程与原回归方程的贴近度

删除第 i 个数据点	删除数据点后的回归方程	贴近度
0	$\widetilde{y}=221.1685+(6.4357,0.2935,0.3195)x$	
1	$\widetilde{y}=219.6208+(6.6471,0.2891,0.3177)x$	0.9185
2	$\widetilde{y}=220.1532+(6.5666,0.2802,0.3123)x$	0.9304
3	$\widetilde{y}=221.2511+(6.4201,0.2846,0.3032)x$	0.9449
4	$\widetilde{y}=221.6806+(6.3748,0.2954,0.3117)x$	0.9487
5	$\widetilde{y}=221.3905+(6.4235,0.2861,0.3139)x$	0.9450
6	$\widetilde{y}=221.0054+(6.5064,0.2894,0.3352)x$	0.9371
7	$\widetilde{y}=220.7240+(6.5871,0.2738,0.3244)x$	0.9209
8	$\widetilde{y}=221.4294+(6.3141,0.2773,0.3084)x$	0.9430
9	$\widetilde{y}=222.1109+(6.2365,0.3125,0.3158)x$	0.9490
10	$\widetilde{y}=222.0352+(6.2822,0.3614,0.6140)x$	0.9469

表 3 中心回归方程及中心回归方程与原回归方程的贴近度

用样本重心替换第 i 个数据点	用样本重心替换第 i 个数据点后的回归方程	贴近度
0	$\widetilde{y}=221.1685+(6.4357,0.2935,0.3195)x$	
1	$\widetilde{y}=219.6694+(6.6470,0.2958,0.3244)x$	0.8156
2	$\widetilde{y}=213.4109+(7.5315,0.2883,0.3174)x$	0.9445
3	$\widetilde{y}=221.2526+(6.4217,0.2918,0.3090)x$	0.9449
4	$\widetilde{y}=221.6635+(6.3766,0.3019,0.3170)x$	0.9486
5	$\widetilde{y}=221.3697+(6.4262,0.2935,0.3192)x$	0.9450
6	$\widetilde{y}=220.9683+(6.5104,0.2968,0.3390)x$	0.9373
7	$\widetilde{y}=220.6715+(6.5909,0.2827,0.3291)x$	0.9229
8	$\widetilde{y}=221.4565+(6.3185,0.2863,0.3147)x$	0.9384
9	$\widetilde{y}=222.1296+(6.2384,0.3188,0.3218)x$	0.9490
10	$\widetilde{y}=222.0415+(6.2831,0.3634,0.3634)x$	0.9468

从表 2 和表 3 可以看出，删除第一个数据点后的回归方程与原回归方程的贴近度及第一个数据点中心回归方程与原回归方程的贴近度都是最小的，从而第一个数据点可能是强影响点或异常点

根据 (4.5) 式可以计算出删除第 i 个数据点后的回归方程与原回归方程的贴近度和第 i 个数据点中心回归方程与原回归方程的贴近度的相对偏差 (见表 4). 从表 4 的数据

我们可知，相对偏差 $D_1 = 12.6156\% > 5\%$, 从而第一个数据点可能是强影响点或异常点.

表 4　$S(L, \overline{L}_i)$ 与 $S(L, L_i)$ $(i = 1, 2, \cdots, n)$ 的相对偏差 (%)

数据点	相对偏差 D_i	数据点	相对偏差 D_i
1	12.6156	6	0.0008
2	1.4929	7	0.2198
3	0.0040	8	0.4980
4	0.0010	9	0.0002
5	0.0040	10	0.0092

综上，通过删除数据点后的回归方程与原回归方程的贴近度，中心回归方程与原回归方程的贴近度以及删除第 i 个数据点后的回归方程与原回归方程的贴近度和第 i 个数据点中心回归方程与原回归方程的贴近度的相对偏差三个方面，可以得出第一个数据点是强影响点或异常点，从而说明本文提出的用模糊贴近度的方法来检验模糊数据集中强影响点或异常点是有效的.

当数据量较大时，特别是每个数据点的预测值与观测值之间的残差都不大时，直观上就很难判断哪些点可能是强影响点或异常点. 如图 1 就是删除第一个数据后，观测数据 (样本数据) 与预测数据的对比图，从图 1 不能判断出第一个数据点是否是强影响点或异常点. 我们就需要采取一定的方法来加以区别，本文所提出的基于数据删除的模糊贴近度的方法就能较好地解决上述问题. 如实例中通过计算相应的模糊贴近度或相对偏差 (表 2 或表 4) 可以判断出第 1 号数据点可能是强影响点.

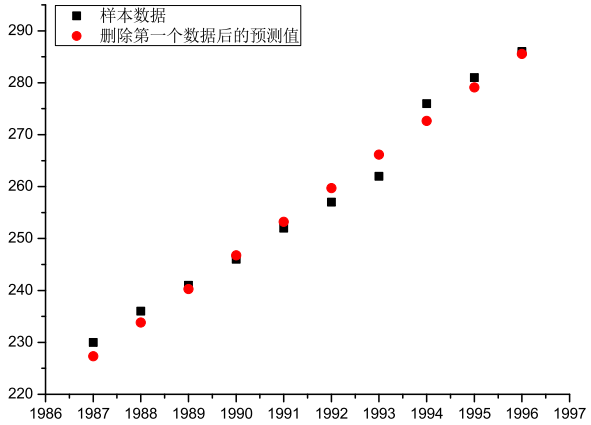

图 1

6　结论与进一步研究

基于数据删除法，我们得到了模糊线性回归模型在数据删除情况下的参数估计，同

时构造了检验观测数据中强影响点或异常点的模糊贴近度这一重要的诊断统计量，并提出了检验模糊观测数据中强影响点或异常点的一般步骤，通过对实际数据的研究，识别出其中的强影响点，进而表明本文所得统计量是有效的．此外，本文提出的基于数据删除的模糊贴近度的方法可以很容易编制成相应的软件包，以便在使用时直接调用，因此，更便于进行大规模的观测数据的诊断与挖掘．

本文仅对一元模糊线性回归模型 (输入为精确数) 进行了讨论，由于多元模糊线性回归模型的复杂性，本文的结论并不能直接推广到多元模糊线性回归模型的情形，但对多元模糊线性回归模型的情形另文将作进一步研究．

参 考 文 献

[1] Zadeh L A. The Concept of Linguistic Variable and its Application to Approximate Reasoning. *Inform Sci.*, 1975, 8: 199–244, 301–357

[2] Thanaka H, Uejina S, Asai K. Linear Regression Analysis with Fuzzy Model. *IEEE Trans Systems Man Cybernetics*, 1982, 12: 903–907

[3] Diamond P. Fuzzy Least Squares. *Information Science*, 1988, 46: 141–157

[4] Savic D A, Pedrycz W. Evaluation of Fuzzy Linear Regression Models. *Fuzzy sets and Systems*, 1988, 46: 141–157

[5] Chang Y H O, Ayyub B M. Fuzzy Regression Methods-a Comparative Assessment. *Fuzzy sets and Systems*, 2001, 119: 187–203

[6] Cook R D, Weisberg S. Residual and Influence in Regression. New York: Chapman and Hall, 1982

[7] 韦博成，林金官，解锋昌． 统计诊断． 北京：高等教育出版社， 2009: 19–44
(Wei B C, Lin J G, Xie F C． Statistic Diagnostic． Beijing: Higher Education Press, 2009, 19–44)

[8] Chen Y S. Outliers Detection and Confidence Interval Modification in Fuzzy Regression. *Fuzzy sets and Systems*, 2001, 119: 259–272

[9] Hung W L, Yang M S. An Omission Approach for Detecting Outliers in Fuzzy Regression Models. *Fuzzy Sets and Systems*, 2006, 157: 3109–3122

[10] Bang Y S. Robust Fuzzy Linear Regression Based on M-estimators. *J. App1. Math. and Computing*, 2005, 18: 591–601

[11] Peters G. Fuzzy Linear Regression with Fuzzy Interval. *Fuzzy sets and Systems*, 1994, 63: 45–55

[12] Sanchez J D A, Gomez A T. Estimating a Fuzzy Term Strcture of Interest Rates Using Fuzzy Linear Regression. *European J. Oper. Res.*, 2004, 154: 804–818

[13] Pierpaola D U, Tommaso G. A Least-squares Approach to Fuzzy Linear Regression Analysis. *Computational Statistics & Data Analysis*, 2003, 4: 427–440

[14] 曾文艺，李洪兴，施煜． 模糊线性回归模型 (1)． 北京师范大学学报 (自然科学版), 2006, 42: 120–125
(Zeng W Y, Li H X, Shi Y. Fuzzy Linear Regession Model(1). *Journal of Beijing Normal University* (Natural Science), 2006, 42: 120–125)

[15] 李洪兴，汪培庄． 模糊数学． 北京：国防工业出版社， 1993, 89–102

(Li H X, Wang P Z. Fuzzy Mathematics. National Defence Industry Press, 1993, 89–102)

Statistical Diagnostics of Fuzzy Linear Regression Model Based on Fuzzy Approach Degree

ZHANG Aiwu

(*School of Mathematics, Yancheng Teachers University, Yancheng* 224002)

(*E-mail: zaw_017@163.com*)

Abstract For fuzzy linear regression model with real input and fuzzy output, model parameter estimation is discussed based on data deletion. Statistical diagnostic capacity is established used to test strong influential points or outliers in the data. And the general procedure is given to examine with this statistical diagnostic capacity. Example suggests that this statistical diagnostic capacity is effective.

Key words fuzzy linear regression model; case deletion model; fuzzy approach degree

MR(2000) Subject Classification 62G68

Chinese Library Classification O212.2; O159