2012 年 10 月 北京航空航天大学学报 October 2012
第38卷 第10期 Journal of Beijing University of Aeronautics and Astronautics Vol. 38 No. 10

基于全信息的正态分布型数据的线性回归分析

王惠文 李 楠

(北京航空航天大学 经济管理学院, 北京 100191)

摘 要: 针对正态的分布型符号数据,提出一种新的线性回归分析方法. 以体现正态的分布型数据的全部原始信息为出发点,给出正态的分布型变量的一阶矩、二阶原点矩、二阶混合原点矩的定义和计算原则. 在此基础上,定义针对正态的分布型数据的线性回归模型以及残差信息的平方和,推导最小二乘回归系数. 仿真实验证明了该方法所得回归模型在解释能力和预测能力上的有效性以及相对于"中心法"的优越性. 给出的正态分布型变量数字特征的定义和计算原则为将其他经典的多元统计方法推广到分布型数据奠定了基础.

关 键 词: 正态的分布型符号数据; 数字特征; 全部信息; 线性回归

中图分类号: O 212

文献标识码: A **文 章 编 号**: 1001-5965(2012)10-1275-05

Linear regression analysis for normal distribution-valued data based on complete information

Wang Huiwen Li Nan

(School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: In light of normal distribution-valued symbolic data, a new method for building linear regression model was proposed. To reflect all the original information of the normal distribution-valued data, definition and calculation principle were proposed for first moments, second original moment and second mixed original moment of normal distribution-valued variables. On this basis, linear regression model for normal distribution-valued data and the sum of squares of the residual information were defined and least-squares regression coefficients were derived. Simulation results show that the explanatory power and predictive ability of the regression model derived by the proposed method are effective and outperform the "Centre Method". The definition and calculation principle of numerical characteristics laid the foundation for extending the other classical multivariate statistical method to distribution-valued data.

Key words: normal distribution-valued symbolic data; numerical characteristics; complete information; linear regression

随着计算机和信息技术的快速发展,产生数据和搜集数据的能力在日益提高,数据的极大丰富带来了对高效的数据分析技术的需求. 符号数据分析[1]为分析大规模数据供了新的思路,其处理的数据表中的每个单元可能是一个多值集合、一个区间、一个直方图或者一个连续型分布. 本文将聚焦研究一种普遍存在的分布型符号数据,即正态的分布型符号数据.

线性回归分析是一种重要的统计分析方法,目前,对符号数据的线性回归分析方法的研究主要集中于区间型符号数据和直方图型符号数据. 对于区间型符号数据,目前已有"中心法"[2]、"顶点法"[3]、"中心-半长法"[4]等. 这些方法主要的不足在于总是试图利用"局部信息"来代替"整体

收稿日期: 2012-01-06; **网络出版时间**: 2012-04-12 13:33

网络出版地址: www.cnki.net/kcms/detail/11.2625.V.20120412.1333.007.html

基金项目: 国家自然科学基金资助项目(71031001,70771004,70821061)

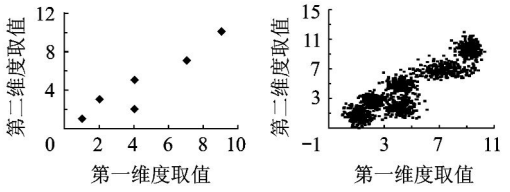
作者简介: 王惠文(1957 -),女,辽宁大连人,教授,wanghw@vip.sina.com.

信息”,因此对原始数据矩阵信息的使用不充分,难以获得准确的模型结果. 对于直方图型符号数据,文献[5]给出直方图数据的方差和协方差的定义,将经典的线性回归分析直接推广到直方图数据的线性回归分析. 文献[6]基于直方图数据之间的 Wasserstein 距离,建立与直方图相对应的分位数函数之间的线性回归模型.

本文针对正态的分布型符号数据,提出一种新的线性回归分析方法. 从方法论的特点来看,本文将摒弃在已有研究中所使用的局部信息替代的思路,而是以体现符号数据表中的全部信息为出发点,推导正态的分布型变量的数字特征的定义和计算原则,并在此基础上推导线性回归模型. 此外,本文将通过仿真实验验证本文提出的方法的正确性;同时通过与只采用均值信息的“中心法”相比较,说明新模型的优越性.

1 方法原理

本节首先来说明普通数据与正态的分布型数据的区别. 图 1a 是一个具有 6 个样本点的散点图,它对应着图 1c 表示的普通数据矩阵;图 1b 是一个具有 1 500 个样本点的散点图,如果用普通的数据矩阵表示,它将是一个 1 500 × 2 维的数据矩阵. 然而,通过对 1 500 个样本点的分布进行观察可以发现,很多样本点可以概括为一类,而当每一类样本点在 2 个维度上都服从或近似服从正态分布时,就可以用正态的分布型数据矩阵(图 1d)来大致表示这 1 500 个样本点的数据集合,从而大大地降低样本容量,并同时保留了原始数据的信息.



a 6 个样本点的散点图　　b 1500 个样本点的散点图

$$\begin{pmatrix}1 & 1\\ 2 & 3\\ 4 & 2\\ 4 & 5\\ 7 & 7\\ 9 & 10\end{pmatrix} \qquad \begin{pmatrix}N(1,0.5^2) & N(1,0.7^2)\\ N(2,0.4^2) & N(3,0.5^2)\\ N(4,0.6^2) & N(2,0.7^2)\\ N(4,0.5^2) & N(5,0.6^2)\\ N(7,0.6^2) & N(7,0.5^2)\\ N(9,1^2) & N(10,0.8^2)\end{pmatrix}$$

c 普通数据矩阵　　d 正态的分布型数据矩阵

图 1 普通数据与正态的分布型数据的区别

下面给出正态的分布型数据矩阵的一般形式. 已知一个 $n \times p$ 维的符号数据矩阵 $\boldsymbol{X}_{n\times p}$,它的每个元素 x_{ij} 都是一个数据集合,其分布形式为 $N(\mu_{ij},\sigma_{ij}^2)$,这里记为 $x_{ij} \sim N(\mu_{ij},\sigma_{ij}^2)$. 其中 μ_{ij} 和 σ_{ij}^2 可以是已知的总体分布参数,也可以由数据集合的样本均值和样本方差作为它们的估计. 这种数据矩阵为正态的分布型数据矩阵,记作

$$\boldsymbol{X}_{n\times p} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}_{n\times p} \tag{1}$$

其中,$\boldsymbol{X}_j = (x_{1j},x_{2j},\cdots,x_{nj})^{\mathrm{T}}, j = 1,2,\cdots,p$,为 n 个观测对象在第 j 个指标上取值的向量;$\boldsymbol{O}_i' = (x_{i1},x_{i2},\cdots,x_{ip}), i = 1,2,\cdots,n$,则是第 i 个观测对象在 p 个指标变量上取值的向量. 此外,假设数据集合 x_{ij} 和 x_{ik} 相互独立,即对于同一个观测对象,其各个指标变量之间是相互独立的.

变量的数字特征在线性回归模型的推导过程中起着重要的作用,为了将经典的线性回归方法推广到正态的分布型数据,本文将首先定义正态的分布型变量的数字特征.

1.1 正态的分布型变量的数字特征

定义 1 对于有 n 个观测对象的正态的分布型变量 $\boldsymbol{X}_j$,定义其一阶原点矩 $E(\boldsymbol{X}_j) \in \mathbf{R}$:

$$E(\boldsymbol{X}_j) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} E(x_{ij}) = \frac{1}{n}\sum_{i=1}^{n}\mu_{ij} \tag{2}$$

定义 2 对于有 n 个观测对象的任意 2 个正态的分布型变量 $\boldsymbol{X}_j$ 和 $\boldsymbol{X}_k (j \neq k)$,定义其二阶混合原点矩 $E(\boldsymbol{X}_j \cdot \boldsymbol{X}_k) \in \mathbf{R}$:

$$E(\boldsymbol{X}_j \cdot \boldsymbol{X}_k) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} E(x_{ij}x_{ik}) = \frac{1}{n}\sum_{i=1}^{n} E(x_{ij})E(x_{ik}) = \frac{1}{n}\sum_{i=1}^{n}\mu_{ij}\mu_{ik} \tag{3}$$

定义 3 对于有 n 个观测对象的正态的分布型变量 $\boldsymbol{X}_j$,定义其二阶原点矩 $E(\boldsymbol{X}_j^2) \in \mathbf{R}$:

$$E(\boldsymbol{X}_j^2) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} E(x_{ij}^2) = \frac{1}{n}\sum_{i=1}^{n}(\sigma_{ij}^2 + \mu_{ij}^2) \tag{4}$$

在此基础上,容易证明有如下定理成立.

定理 1 对于正态的分布型变量 $\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}$,定义 2 给出的二阶混合原点矩有如下运算性质:

1) $E(\boldsymbol{X} \cdot \boldsymbol{Y}) = E(\boldsymbol{Y} \cdot \boldsymbol{X})$;

2) $E((\boldsymbol{X}+\boldsymbol{Y}) \cdot \boldsymbol{Z}) = E(\boldsymbol{X} \cdot \boldsymbol{Z}) + E(\boldsymbol{Y} \cdot \boldsymbol{Z})$;

3) $E(\alpha\boldsymbol{X} \cdot \boldsymbol{Y}) = \alpha E(\boldsymbol{X} \cdot \boldsymbol{Y}), \alpha \in \mathbf{R}$.

定理 2 对于正态的分布型变量 $\boldsymbol{X},\boldsymbol{Y}$,定义 3 给出的二阶原点矩有如下运算性质:

1) $E(\boldsymbol{Y}^2) \geqslant 0$;

2) $E((\alpha \boldsymbol{X})^2)=\alpha^2 E(\boldsymbol{X}^2),\alpha\in\mathbf{R}$;

3) $E((\boldsymbol{X}+\boldsymbol{Y})^2)=E(\boldsymbol{X}^2)+2E(\boldsymbol{X}\cdot\boldsymbol{Y})+E(\boldsymbol{Y}^2)$.

1.2　正态的分布型数据的线性回归算法

设第 i 个观测对象在自变量上的取值 $\boldsymbol{O}_i'=(x_{i1},x_{i2},\cdots,x_{ip})$,假设 $x_{i1},x_{i2},\cdots,x_{ip}$ 之间相互独立;如果 $\beta_1,\beta_2,\cdots,\beta_p$ 是 p 个属于实数域的未知参数,有 $y_i=\sum_{j=1}^{p}\beta_j x_{ij}$,则根据正态分布的线性组合性质可知,$y_i\sim N\left(\sum_{j=1}^{p}\beta_j\mu_{ij},\sum_{j=1}^{p}\beta_j^2\sigma_{ij}^2\right)$.

设正态的分布型变量 $\boldsymbol{Y}$ 与 $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_p$ 之间的多元线性回归模型为

$$\boldsymbol{Y}=\beta_0\boldsymbol{X}_0+\beta_1\boldsymbol{X}_1+\beta_2\boldsymbol{X}_2+\cdots+\beta_p\boldsymbol{X}_p+\boldsymbol{\varepsilon}\quad(5)$$

其中,$\boldsymbol{X}_0$ 是一个 n 维列向量,其每个元素都是均值为 1,方差为 0 的正态分布;$\beta_0,\beta_1,\cdots,\beta_p$ 是 $p+1$ 个属于实数域的未知参数;$\boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)^{\mathrm{T}}$ 是随机误差向量,$\varepsilon_i\sim N(0,\sigma^2),i=1,2,\cdots,n$.

类似于经典的线性回归模型,寻找参数 $\beta_0,\beta_1,\cdots,\beta_p$ 的最小二乘估计量 $\hat{\boldsymbol{\beta}}=(\hat\beta_0,\hat\beta_1,\cdots,\hat\beta_p)^{\mathrm{T}}$. 由于 $y_i-\sum_{j=1}^{p}\hat\beta_j x_{ij}$ 为一个正态的分布型数据,而不是一个单值型数据,因此,本文采用正态的分布型变量 $\boldsymbol{Y}-\sum_{j=0}^{p}\hat\beta_j\boldsymbol{X}_j$ 的二阶矩来定义线性回归模型的残差信息平方和,即

$$I_{\mathrm{SSE}}=E\left(\boldsymbol{Y}-\sum_{j=0}^{p}\hat\beta_j\boldsymbol{X}_j\right)^2=E(\boldsymbol{Y}^2)-2\sum_{j=0}^{p}\hat\beta_j E(\boldsymbol{Y}\cdot\boldsymbol{X}_j)+\sum_{j=0}^{p}\hat\beta_j^2E(\boldsymbol{X}_j^2)+2\sum_{j=0}^{p}\sum_{\substack{k=0,\\k\neq j}}^{p}\hat\beta_j\hat\beta_k E(\boldsymbol{X}_j,\boldsymbol{X}_k)\quad(6)$$

对式(6)的未知参数求偏导数,并令这些偏导数等于 0,可得 $\hat\beta_0,\hat\beta_1,\cdots,\hat\beta_p$ 满足如下方程组:

$$\frac{\partial I_{\mathrm{SSE}}}{\partial\hat\beta_j}=-2E(\boldsymbol{Y}\cdot\boldsymbol{X}_j)+2\hat\beta_j E(\boldsymbol{X}_j^2)+2\sum_{\substack{k=0,\\k\neq j}}^{p}\hat\beta_k E(\boldsymbol{X}_j\cdot\boldsymbol{X}_k)=0\qquad j=0,1,2,\cdots,p\quad(7)$$

式(7)可表示为如下矩阵形式:

$$\begin{pmatrix}E(\boldsymbol{X}_0^2)&E(\boldsymbol{X}_0,\boldsymbol{X}_1)&\cdots&E(\boldsymbol{X}_0,\boldsymbol{X}_p)\\E(\boldsymbol{X}_1,\boldsymbol{X}_0)&E(\boldsymbol{X}_1^2)&\cdots&E(\boldsymbol{X}_1,\boldsymbol{X}_p)\\\vdots&\vdots&&\vdots\\E(\boldsymbol{X}_p,\boldsymbol{X}_0)&E(\boldsymbol{X}_p,\boldsymbol{X}_1)&\cdots&E(\boldsymbol{X}_p^2)\end{pmatrix}\begin{pmatrix}\hat\beta_0\\\hat\beta_1\\\vdots\\\hat\beta_p\end{pmatrix}=\begin{pmatrix}E(\boldsymbol{X}_0\cdot\boldsymbol{Y})\\E(\boldsymbol{X}_1\cdot\boldsymbol{Y})\\\vdots\\E(\boldsymbol{X}_p\cdot\boldsymbol{Y})\end{pmatrix}\quad(8)$$

在获得正态的分布型数据矩阵后,根据定义 1 ~ 定义 3 可以算出式(8)中的每一个元素,显然,它们都是实数域中的数据,记

$$\boldsymbol{V}=\begin{pmatrix}E(\boldsymbol{X}_0^2)&E(\boldsymbol{X}_0,\boldsymbol{X}_1)&\cdots&E(\boldsymbol{X}_0,\boldsymbol{X}_p)\\E(\boldsymbol{X}_1,\boldsymbol{X}_0)&E(\boldsymbol{X}_1^2)&\cdots&E(\boldsymbol{X}_1,\boldsymbol{X}_p)\\\vdots&\vdots&&\vdots\\E(\boldsymbol{X}_p,\boldsymbol{X}_0)&E(\boldsymbol{X}_p,\boldsymbol{X}_1)&\cdots&E(\boldsymbol{X}_p^2)\end{pmatrix}$$

$$\boldsymbol{W}=\begin{pmatrix}E(\boldsymbol{X}_0\cdot\boldsymbol{Y})\\E(\boldsymbol{X}_1\cdot\boldsymbol{Y})\\\vdots\\E(\boldsymbol{X}_p\cdot\boldsymbol{Y})\end{pmatrix}$$

则最小二乘估计量 $\hat{\boldsymbol{\beta}}=(\hat\beta_0,\hat\beta_1,\cdots,\hat\beta_p)^{\mathrm{T}}=\boldsymbol{V}^{-1}\boldsymbol{W}$.

2　仿真实验

本节通过仿真实验来验证本文提出的方法的有效性以及相对于“中心法”的优越性. 由于符号数据往往产生于对原始的大规模单值型数据的概括,那么采用符号数据方法对数据进行分析得到的结果应该与用经典的统计分析方法得到的结果一致,而不应该扭曲原始大规模数据的信息. 因此,本节的实验思路是对于给定的正态的分布型数据矩阵,采用仿真的方式来还原原始的大规模单值型数据矩阵,之后以经典最小二乘回归的回归系数为标杆来评价模型的解释能力,以对单值型数据点的预测效果来评价模型的预测能力.

2.1　实验数据

实验中将分别讨论一元线性回归模型和三元线性回归模型. 另外,实验中生成的每个正态的分布型数据矩阵将包含 n 个观测对象,前 n_1 个用来建立回归模型,后 n_2($n=n_1+n_2$)个用来评价回归模型的预测效果. 生成实验数据的方法如下:

1) 为了生成正态的分布型数据矩阵,首先随机生成解释变量 $\boldsymbol{X}$ 的均值向量 $\boldsymbol{X}_j^{\mathrm{E}}=(x_{j1}^{\mathrm{E}},x_{j2}^{\mathrm{E}},\cdots,x_{jn}^{\mathrm{E}})^{\mathrm{T}},j=1,2,3$,和标准差向量 $\boldsymbol{X}_j^{\mathrm{SD}}=(x_{j1}^{\mathrm{SD}},x_{j2}^{\mathrm{SD}},\cdots,x_{jn}^{\mathrm{SD}})^{\mathrm{T}},j=1,2,3$,这样就得到了 n 个观测对象在解释变量 $\boldsymbol{X}$ 上的取值 $\boldsymbol{X}_j=(N(x_{j1}^{\mathrm{E}},x_{j1}^{\mathrm{SD2}}),N(x_{j2}^{\mathrm{E}},x_{j2}^{\mathrm{SD2}}),\cdots,N(x_{jn}^{\mathrm{E}},x_{jn}^{\mathrm{SD2}}))^{\mathrm{T}},j=1,2,3$. 再生成随机误差向量 $\boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)^{\mathrm{T}}$ 以及 $\beta_0,\beta_j,j=1,2,3$. 被解释变量 $\boldsymbol{Y}$ 的均值向量 $\boldsymbol{Y}^{\mathrm{E}}=(y_1^{\mathrm{E}},y_2^{\mathrm{E}},\cdots,y_n^{\mathrm{E}})^{\mathrm{T}}$

通过 $y_i^{\mathrm{E}}=\beta_0+\beta_1 x_{1i}^{\mathrm{E}}+\varepsilon_i$(一元)和 $y_i^{\mathrm{E}}=\beta_0+\beta_1 x_{1i}^{\mathrm{E}}+\beta_2 x_{2i}^{\mathrm{E}}+\beta_1 x_{3i}^{\mathrm{E}}+\varepsilon_i$(三元),$i=1,2,\cdots,n$ 生成. 另外,随机生成被解释变量 $\boldsymbol{Y}$ 的标准差向量 $\boldsymbol{Y}^{\mathrm{SD}}=(y_1^{\mathrm{SD}},y_2^{\mathrm{SD}},\cdots,y_n^{\mathrm{SD}})^{\mathrm{T}}$,则 n 个观测对象在被解释变量上的取值为 $\boldsymbol{Y}=(N(y_1^{\mathrm{E}},y_1^{\mathrm{SD2}}),N(y_2^{\mathrm{E}},y_2^{\mathrm{SD2}}),\cdots,N(y_n^{\mathrm{E}},y_n^{\mathrm{SD2}}))^{\mathrm{T}}$.

为了研究被解释变量和解释变量的方差以及随机误差对回归建模的影响,本文的蒙特卡洛实验中考虑了 8 种情形. 表 1 给出了 $x_{ji}^{\mathrm{E}},x_{ji}^{\mathrm{SD}},\beta_0,\beta_j,\varepsilon_i,y_i^{\mathrm{SD}}$ 分别服从的分布.

表 1 蒙特卡洛仿真的 8 种情形

情形	$\boldsymbol{X}_j^{\mathrm{E}}$	$\boldsymbol{X}_j^{\mathrm{SD}}$	$\boldsymbol{\varepsilon}$	β_0	β_j	$\boldsymbol{Y}^{\mathrm{SD}}$
C_1	$U[-10,10]$	$U[0.2,0.4]$	$N(0,0.5^2)$	$U[-5,5]$	$U[-10,10]$	$U[0.2,0.4]$
C_2	$U[-10,10]$	$U[0.2,0.4]$	$N(0,1^2)$	$U[-5,5]$	$U[-10,10]$	$U[0.2,0.4]$
C_3	$U[-10,10]$	$U[0.5,1]$	$N(0,0.5^2)$	$U[-5,5]$	$U[-10,10]$	$U[0.5,1]$
C_4	$U[-10,10]$	$U[0.5,1]$	$N(0,1^2)$	$U[-5,5]$	$U[-10,10]$	$U[0.5,1]$
C_5	$U[-10,10]$	$U[1,2]$	$N(0,0.5^2)$	$U[-5,5]$	$U[-10,10]$	$U[1,2]$
C_6	$U[-10,10]$	$U[1,2]$	$N(0,1^2)$	$U[-5,5]$	$U[-10,10]$	$U[1,2]$
C_7	$U[-10,10]$	$U[2,3]$	$N(0,0.5^2)$	$U[-5,5]$	$U[-10,10]$	$U[2,3]$
C_8	$U[-10,10]$	$U[2,3]$	$N(0,1^2)$	$U[-5,5]$	$U[-10,10]$	$U[2,3]$

2) 生成与正态的分布型数据矩阵相对应的单值型数据矩阵. 对于第 i 个观测对象,随机生成 M 个单值型数据点 $(x_{i_l1}^{\mathrm{NUM}},x_{i_l2}^{\mathrm{NUM}},x_{i_l3}^{\mathrm{NUM}},y_{i_l}^{\mathrm{NUM}})$,$l=1,2,\cdots,M$,其中 $x_{i_lj}^{\mathrm{NUM}}\sim N(x_{ij}^{\mathrm{E}},x_{ij}^{\mathrm{SD2}})$,$j=1,2,3$;$y_{i_l}^{\mathrm{NUM}}\sim N(y_i^{\mathrm{E}},y_i^{\mathrm{SD2}})$,从而得到与正态的分布型数据矩阵相对应的 $(n\times M)\times 4$ 维的单值型数据矩阵,该矩阵可视为原始的大规模数据矩阵.

2.2 评价指标

为了评价回归模型的解释能力,本文以采用经典最小二乘法对相应的单值型数据矩阵进行回归建模得到的回归系数为标杆,计算本文提出的方法以及中心法得到的回归系数与标杆之间的差异(coefficients dissimilarity):

$$d_{\mathrm{CD}}=\sum_{j=0}^{k}\left|\frac{\hat{\beta}_j^{\mathrm{NUM}}-\hat{\beta}_j}{\hat{\beta}_j^{\mathrm{NUM}}}\right|\qquad k=1\text{ 或 }3$$

其中,$\hat{\beta}_j$ 表示本文提出的方法或中心法得到的回归系数;$\hat{\beta}_j^{\mathrm{NUM}}$ 表示经典最小二乘法得到的回归系数. d_{CD} 值越小,则说明回归模型的解释能力越强.

对于回归模型预测能力,本文利用后 $n_2\times M$ 个单值型数据点进行评价,定义平均绝对误差(mean absolute error)如下:

$$d_{\mathrm{MAE}}=\frac{1}{n_2\times M}\sum_{i=n_1+1}^{n}\sum_{l=1}^{M}|y_{i_l}^{\mathrm{NUM}}-\hat{y}_{i_l}|$$

其中,$\hat{y}_{i_l}$ 表示本文提出的方法或中心法的预测结果;$y_{i_l}^{\mathrm{NUM}}$ 表示后 $n_2\times M$ 个单值型数据点在被解释变量上的取值. 显然,d_{MAE} 越小,模型的预测效果越好.

2.3 仿真步骤

蒙特卡洛仿真实验可通过以下 4 个步骤完成:

1) 对于表 1 中展示的每种情形,随机生成包含 n 个观测对象的正态的分布型数据矩阵和与之对应的单值型数据矩阵. 为了避免随机因素带来的偏差,实验将重复进行 K 次.

2) 在第 k 次实验中($k=1,2,\cdots,K$),根据前 n_1 个观测对象,采用本文提出的方法以及中心法建立一元和三元线性回归模型,同时采用经典的最小二乘法对相应的单值型数据矩阵建立一元和三元线性回归模型.

3) 记本文提出的方法为方法 A,中心法为方法 B,在第 k 次实验中($k=1,2,\cdots,K$),计算 d_{CD}^A,d_{CD}^B 以及 $d_{\mathrm{MAE}}^A,d_{\mathrm{MAE}}^B$.

4) 计算 K 次实验中 $d_{\mathrm{CD}}^A,d_{\mathrm{CD}}^B$ 以及 $d_{\mathrm{MAE}}^A,d_{\mathrm{MAE}}^B$ 的平均值,分别记作 $d_{\overline{\mathrm{CD}}}^A,d_{\overline{\mathrm{CD}}}^B,d_{\overline{\mathrm{MAE}}}^A,d_{\overline{\mathrm{MAE}}}^B$. 此外,计算 K 次实验中,$d_{\mathrm{CD}}^A\leqslant d_{\mathrm{CD}}^B$,$d_{\mathrm{MAE}}^A\leqslant d_{\mathrm{MAE}}^B$ 出现次数所占的百分比.

在本次实验中,涉及到的相关实验参数分别为:$n=375$,$n_1=250$,$n_2=125$,$M=100$,$K=1\,000$.

2.4 实验结果及分析

表 2 为蒙特卡洛仿真的实验结果. 从 d_{CD} 指标来看,对于实验中的 8 种情形,都有 $d_{\overline{\mathrm{CD}}}^A\leqslant d_{\overline{\mathrm{CD}}}^B$,,并且 1 000 次实验中 $d_{\mathrm{CD}}^A\leqslant d_{\mathrm{CD}}^B$ 出现的比例(R_{CD})最小达到 87.3%,这说明本文提出的方法是有效的,并且能较中心法更好地体现原始大规模数据包含的信息,解释能力更强. 另外,对于每一种情形,在三元线性回归模型中 $d_{\overline{\mathrm{CD}}}^A$ 比 $d_{\overline{\mathrm{CD}}}^B$ 小的更多,并且 R_{CD} 更高,即本文提出的方法的优势更显著.

从实验结果中还可以发现,随着解释变量和被解释变量方差的增大,$d_{\overline{\mathrm{CD}}}^A$ 比 $d_{\overline{\mathrm{CD}}}^B$ 小的更多,并且 1 000 次实验中 $d_{\mathrm{CD}}^A\leqslant d_{\mathrm{CD}}^B$ 出现的次数逐渐增加. 这说明当解释变量和被解释变量的方差越大时,本文提出的方法的优势越显著.

从 d_{MAE} 指标的实验结果来看,对于实验中的

8 种情形,都有 $d^A_{\overline{\mathrm{MAE}}} \leqslant d^B_{\overline{\mathrm{MAE}}}$,也就是说本文提出的方法比中心法对单值型样本点的预测效果更好. 虽然 $d^A_{\overline{\mathrm{MAE}}}$和 $d^B_{\overline{\mathrm{MAE}}}$之间的差距不是很大,但是可以发现,随着从解释变量和被解释变量方差的增大,$d^A_{\overline{\mathrm{MAE}}}$和 $d^B_{\overline{\mathrm{MAE}}}$之间的差距有所增加,而且 1 000 次实验中 $d^A_{\mathrm{MAE}} \leqslant d^B_{\mathrm{MAE}}$出现的比例显著增大,也就是说本文提出的方法较中心法的预测优势更加显著. 另外,对于解释变量和被解释变量的方差取自相同的均匀分布(如 C_1 与 C_2)的情形,当随机误差 ε 较小,即解释变量和被解释变量的线性关系更显著时,R_{MAE}一般较大,本文方法的优势更明显.

将一元回归模型与三元回归模型对比来看,对于每种情形,当是三元回归模型时,1 000 次实验中 $d^A_{\mathrm{MAE}} \leqslant d^B_{\mathrm{MAE}}$出现的次数比一元回归模型时的明显大,说明在建立多元回归模型时本文提出的方法的优势更加明显.

表 2　蒙特卡洛仿真实验结果

情形	p	d^A_{CD}	d^B_{CD}	R_{CD}/%	$d^A_{\overline{\mathrm{MAE}}}$	$d^B_{\overline{\mathrm{MAE}}}$	R_{MAE}/%
C_1	1	0.014 497	0.018 376	87.3	0.010 668	0.010 679	58.6
	3	0.068 307	0.080 796	88.7	0.019 214	0.019 240	64.5
C_2	1	0.010 198	0.013 417	88.2	0.012 435	0.012 446	55.8
	3	0.028 091	0.040 320	90.2	0.020 066	0.020 106	62.3
C_3	1	0.026 809	0.066 232	95.0	0.024 824	0.025 009	70.1
	3	0.100 195	0.192 796	95.5	0.047 384	0.047 768	75.6
C_4	1	0.034 835	0.068 905	94.7	0.025 213	0.025 401	65.5
	3	0.062 165	0.144 673	95.4	0.046 733	0.047 110	72.9
C_5	1	0.062 898	0.190 695	98.5	0.049 185	0.050 657	75.2
	3	0.127 964	0.521 688	98.4	0.090 849	0.093 821	79.1
C_6	1	0.113 276	0.301 109	97.9	0.049 730	0.051 126	72.5
	3	0.117 040	0.541 600	99.0	0.090 396	0.093 436	79.7
C_7	1	0.123 303	0.941 791	99.2	0.077 894	0.084 682	81.2
	3	0.242 365	1.904 576	99.8	0.142 139	0.155 159	81.2
C_8	1	0.059 629	0.463 617	99.2	0.078 156	0.084 732	78.5
	3	0.662 289	4.955 586	99.7	0.142 291	0.154 746	82.1

注:$p=1$ 表示一元回归;$p=3$ 表示三元回归.

3　结 束 语

文献[7]指出“分布是未来的数据”,而正态分布是一个在自然科学和行为科学等领域都非常重要的分布形式,它有着极其广泛的实际背景,提出有效的针对正态的分布型数据的线性回归分析方法是有重要意义的. 本文以经典概率论中连续随机变量的数字特征积分计算理论为基础,提出了正态的分布型变量的数字特征的定义和计算方法,并在此基础上推导了线性回归模型的估计系数. 与已有研究中所使用的局部信息替代的方法不同,仿真实验结果表明本文提出的方法能够体现原始大规模数据的全部信息. 该方法为分析大规模数据库中各变量的线性相关关系提供了新的途径.

参考文献 (References)

[1] Diday E. Introduction à l´approche symbolique en analyse des données[J]. RAIRO Recherche Opérationnelle, 1989, 23(2): 193 – 236

[2] Billard L, Diday E. Regression analysis for interval-valued data [C]//Data Analysis, Classification and Related Methods. Belgium: Springer, 2000: 369 – 374

[3] Billard L, Diday E. Symbolic regression analysis[C]//Classification, Clustering, and Data Analysis. Poland: Springer, 2002: 281 – 288

[4] Lima N E, de Carvalho F. Centre and range method to fitting a linear regression model on symbolic interval data[J]. Computational Statistics and Data Analysis, 2008, 52: 1500 – 1515

[5] Billard L. Dependencies and variation components of symbolic interval-valued data[C]//Selected Contributions in Data Analysis and Classification. Berlin: Springer, 2007: 3 – 12

[6] Verde R, Irpino A. Ordinary least squares for histogram data based on wasserstein distance[C]//Proceedings of COMPSTAT ‘2010. Berlin: Springer-Verlag, 2010: 581 – 588

[7] Schweizer B. Distributions are the numbers of the future[C]// Proceedings of the Mathematics of Fuzzy Systems Meeting. Naples: University of Naples, 1984: 137 – 149

(编　辑:文丽芳)