考虑内部散点的区间型符号数据的回归分析①

郭均鹏, 赵 茹, 李汶华
(天津大学管理与经济学部, 天津 300072)

摘要: 研究通过对样本"点"数据打包形成的区间型符号数据的回归分析. 针对现有区间数回归分析只利用区间数的端点信息这一问题,分析如何充分利用原始的样本"点"数据信息,即区间数的内部散点信息. 首先从理论上推导了当假设原始样本点数据误差项满足回归分析所假定的三条性质时,区间数据回归分析的误差项也满足这三条性质. 然后,在考虑散点的区间型符号数据描述性统计量的基础上,提出了一种新的区间型符号数据回归分析的参数估计方法. 随之给出了区间预测方法. 最后选取常用的CCRM作为对比算法,分别通过随机模拟和实例分析,验证了文中方法的有效性.

关键词: 回归分析; 区间型符号数据; 描述统计量
中图分类号: O212.1 **文献标识码:** A **文章编号:** 1007-9807(2018)04-0114-13

0 引 言

回归分析是一种常用的确定因变量与自变量关系的分析方法,传统的回归分析研究,主要是针对点数据进行的. 实际中,经常需用区间数来描述样本. 区间样本数据主要有两种来源:一是在传统的区间分析研究领域,区间数由观测误差或者专家的不确定判断而得到[1];二是在符号数据分析(symbolic data analysis)方法体系中,区间数是符号数据的一种主要类型[2],称为区间型符号数据. 本文主要研究区间型符号数据的线性回归分析. 符号数据分析是一种研究如何从海量数据中发掘系统知识的理论和方法,区间型符号数据基于符号数据分析的思想,通过对点数据进行"数据打包"而形成[3]. "数据打包"是根据研究问题所需、按照数据的自然类别来进行汇总的一个过程. 例如,为了研究股指变动受宏观经济因素变动的影响,可按年度将每年内的股指及影响因素的点数据(例如工业生产指数增长率)进行打包汇总,形成区间型符号数据,具体的,以年内12个月的最低和最高点数据,作为区间数的下限和上限,构成区间数. 相应的,样本性质随之发生变化:由原来点数据构成的样本,变为区间数描述的样本. 数据打包后的样本个体称为符号对象.

区间型符号数据的回归分析研究是符号数据分析的一个分支,有许多学者进行了卓有成效的研究. Billard和Diday[4]提出了第一个拟合区间型符号数据的模型,该模型分别对区间的左端点和右端点进行了拟合,并且具有相同的系数. 总误差为左端点与右端点产生的误差的平均值,通过最小化总误差来求解模型系数. 随后,Billard和Diday[5]又提出了新的模型,模型同样针对左端点和右端点建立多元线性关系,但与文献[4]不同的是,针对两个端点建立的线性关系具有不同的系数,总误差项由左端点的误差的平方与右端点误差的平方和构成,同样通过最小化总误差得到系数. 实质上这种方法相当于将区间数划分为两个点序列,一个是左端点序列一个是右端点序列,分

① 收稿日期: 2016-12-11; 修订日期: 2017-10-16.
基金项目: 国家自然科学基金资助项目(71671121); 天津市哲学社会科学研究规划项目(TJGL17-011).
作者简介: 郭均鹏(1973—), 男, 博士, 教授, 博士生导师. Email: guojp@tju.edu.cn

别利用最小二乘法求解即可. 上述两种方法都是考虑的区间数的端点,而区间数除了用两个端点表示,还可以用中点半径表示,从这个角度出发,Lima Neto 和 De Carvalho[6] 提出了中点半径法(center and range method,CRM). 这种方法将区间数分成两个点序列,一个是中点点序列,一个是半径点序列. 针对两个序列分别建立多元线性回归,并用最小二乘法进行求解. 由于半径具有非负性,用这种方法可能出现预测半径为负的情况,这就会导致预测区间的不合理:有可能出现区间左端点值大于右端点值的情况. 为了避免这种情况,Lima Neto 和 De Carvalho[7] 对半径点序列的参数进行限制,使得参数不小于零,提出了 CCRM 方法,该方法由于半径具有非负性,参数也大于或等于零,因而预测区间肯定不小于零,所以保证了预测区间的合理性. Giordani[8] 指出了 CCRM 的不足之处,认为其对半径参数非负的约束过强,并借鉴 Tibshirani[9] 的 LASSO(least absolute shrinkage and selection operator)约束应用于区间数回归,提出 LASSO 约束区间数回归方法. Lima Neto 和 De Carvalho[10] 将无约束非线性规划方法应用于区间数回归,Hladík 等[11] 则研究了区间数回归的两个约束非线性规划问题. Hao 和 Guo[12] 提出的约束中点半径组合方法(CCRJM)既保证了区间估计结果的合理性,又对观测区间和预测区间施加交叉约束,提高了预测精度. Zhou 和 Feng 等[13] 将分位数回归方法应用于区间数回归分析. Souza 等[14] 将区间数参数化表示为实数点后进行回归,使用 Box-Cox 变换对估计区间上限小于下限的不合理结果施加修正. 除了上述方法外,还有一些文献研究了进化计算、机器学习方法等求解区间数回归问题[15-22]. 李汶华等[23] 从误差传递的理论出发,借助传递误差估计区间半径,最小二乘法估计区间中点,提出基于误差传递的区间型符号数据的回归分析方法.

如前所述,本文主要研究通过对点数据进行"数据打包"而形成的区间型符号数据的线性回归分析,而上述文献只考虑了区间的端点信息,没有考虑区间内部所包含的点数据信息,即散点信息. 这样处理的优点是方法简单、易操作,其缺点也很明显:导致了原始数据信息的丢失. 鉴于此,本文提出一种充分利用区间内部散点信息,基于区间型符号数据的描述性统计量来估计参数的区间回归模型(descriptive statistics based method,DSM). 换言之,本文的研究出发点是充分考虑区间内部散点,而不仅仅考虑区间的端点数据. 例如,为了研究股票板块的性质,可按板块类别将板块内个股的点数据(例如某日的收盘价数据)进行打包汇总,形成区间型符号数据,本文提出的回归分析方法将充分利用所有板块的所有个股数据. 随之而来的问题是,为何不直接利用所有这些个股数据进行传统的回归分析? 事实上,二者在研究目标上有着本质区别:本文的研究目标是研究板块特性,并不关注个股特性,而传统的点数据回归分析研究目标则是个股特性.

1　区间分析基础

本节对将要用到的相关理论基础进行概述与推导.

1.1　区间数运算基础

$$A = [\underline{a},\bar{a}] = \{x:\underline{a} \leqslant x \leqslant \bar{a}, x \in \Re\}$$

称为一个区间数,$\underline{a}$ 和 $\bar{a}$ 分别称为区间数的下限(或左端点)和上限(或右端点). 区间数的全体记为 $I(\Re)$.

$\forall A、B \in I(\Re)$,$* \in \{+,-,\times,/\}$ 是一个二元运算符,则两个区间数的运算定义为[1]

$$A * B = \{x * y \mid x \in A, y \in B\} \tag{1}$$

(当 $*$ 为除法时,$0 \notin B$).

$\forall \lambda_1 \in \Re, \forall \lambda_2 \in \Re, \forall A = [\underline{a},\bar{a}] \in I(\Re)$,则

$$\lambda_1 A + \lambda_2 A = (\lambda_1 + \lambda_2)A \tag{2}$$

区间数在本质上是一个集合,因此也具有集合的包含关系,当且仅当 $\bar{a} \leqslant \bar{b}$ 及 $\underline{b} \leqslant \underline{a}$ 时

$$A \subseteq B \tag{3}$$

区间四则运算的加法和乘法均满足结合律、交换律,但一般不满足分配律,$\forall A,B,C \in I(\Re)$,即一般情况下,$A(B + C) \neq AB + AC$. 但具有次分配率,即[1]

$$A(B + C) \subseteq AB + AC \tag{4}$$

为了后文的回归系数的参数估计,需研究区间多项式函数的导数. 考虑函数 $f(x) = (a + bx)^2$,其

中 $a,b \in I(\Re)$,$x \in \Re$,根据式(4)以及导数的定义,可推导得出

$$f'(x) \subseteq 2b^2x + 2ab \tag{5}$$

显然有,若 $g'(x) = 0$,则 $f'(x) = 0$.

1.2 考虑散点的区间型符号数据描述性统计量

该节的结论主要来自于作者之前的研究成果[24].

定义 1 若随机变量 X 的观测值在 $[a,b]$ 取值,则称 X 为区间变量.

设 $X_1, \cdots, X_p$ 为 p 个区间变量,$k = 1,\cdots,n$ 表示 n 个符号对象,$[a_{kj},b_{kj}]$ 表示变量 X_j 在第 k 个符号对象的区间观测值.

区间变量 X_j 的经验均值为

$$\overline{X}_j = \frac{1}{n}\sum_{k=1}^{n}\overline{X}_{kj} \tag{6}$$

其中 $\overline{X}_{kj}$ 为"打包"之前的原始样本的点数据的样本均值.

经验方差为

$$S_j^2 = \frac{1}{n}\sum_{k=1}^{n}[S_{kj}^2 + (\overline{X}_j - \overline{X}_{kj})^2] \tag{7}$$

其中 $\overline{X}_{kj}$ 和 S_{kj}^2 为"打包"之前的原始样本的点数据的样本均值和样本方差.

两个区间变量的经验协方差为

$$S_{X_1,X_2} = \frac{1}{n}\sum_{k=1}^{n}S_{X_{k1},X_{k2}} + \frac{1}{n}\sum_{k=1}^{n}(\overline{X}_{k1}\overline{X}_{k2}) - \frac{1}{n^2}\sum_{k=1}^{n}\overline{X}_{k1}\sum_{k=1}^{n}\overline{X}_{k2} \tag{8}$$

其中 $S_{X_{k1},X_{k2}}$ 为第 k 个符号对象"打包"之前的原始样本的点数据的协方差.

不做特别说明,下文中所提到的区间数等相关概念均指本文所研究的区间型符号数据.

2 基于描述统计量的回归参数估计(DSM)

本节中涉及运算符号较多,为便于区分,凡是区间数、区间向量、区间矩阵及区间变量,均加中括号.

2.1 区间数据回归分析误差项的性质研究

设 $E = \{e_1,\cdots,e_n\}$ 为 $p+1$ 个区间变量 $Y, X_1,\cdots,X_p$ 的区间数观测样本,假设 e_i 包含 $m_i(i = 1,\cdots,n)$ 个散点,Y^s、$X_j^s(s = 1,\cdots,m_i, j = 1,\cdots,p)$ 为区间变量的内部散点变量. 对于每个观测 $e_i(i = 1,\cdots,n)$,其内部散点为 $e_i^s = (\boldsymbol{x}_i^s, y_i^s)$,其中 $\boldsymbol{x}_i^s = (x_{i1}^s,\cdots,x_{ip}^s)$.

设 Y^s 为因变量、X_j^s 为自变量,它们之间的线性关系如式(9)所示

$$y_i^s = \beta_0 + \beta_1 x_{i1}^s + \cdots + \beta_p x_{ip}^s + \varepsilon_i^s \tag{9}$$

其中 ε_i^s 为误差项,假定误差项具有三条性质:期望满足 $E(\varepsilon_i^s) = 0$,其方差满足 $D(\varepsilon_i^s) = \sigma^2$,$\sigma$ 为常量;其协方差满足 $Cov(\varepsilon_a^g,\varepsilon_b^h) = 0$,$g \neq h$ 且 $a \neq b$;ε_i^s 服从正态分布.

设 $\boldsymbol{y}_i = (y_i^1,\cdots,y_i^{m_i})^{\mathrm{T}}$,$x_i = (\mathbf{1}_{m_i\times 1},\boldsymbol{x}_{i1},\cdots,\boldsymbol{x}_{ip})$,$\boldsymbol{x}_{ij} = (x_{ij}^1,\cdots,x_{ij}^{m_i})^{\mathrm{T}}$,$\boldsymbol{\varepsilon}_i = (\varepsilon_i^1,\cdots,\varepsilon_i^{m_i})^{\mathrm{T}}$ 则有

$$y_i = \beta_0\mathbf{1}_{m_i\times 1} + \beta_1\boldsymbol{x}_{i1} + \cdots + \beta_p\boldsymbol{x}_{ip} + \boldsymbol{\varepsilon}_i \tag{10}$$

将 $\boldsymbol{y}_i$、$\mathbf{1}_{m_i\times 1}$、x_{ij}、$\boldsymbol{\varepsilon}_i$ 视为区间型符号数据,对应符号数据记为 $[\boldsymbol{y}_i]$、$[\mathbf{1}_{m_i\times 1}]$、$[\boldsymbol{x}_{ij}]$、$[\boldsymbol{\varepsilon}_i]$,则式(10)表示为

$$[y_i] = \beta_0[\mathbf{1}_{m_i\times 1}] + \beta_1[\boldsymbol{x}_{i1}] + \cdots + \beta_p[\boldsymbol{x}_{ip}] + [\boldsymbol{\varepsilon}_i] \tag{11}$$

式(11)即为多元区间型符号数据回归模型. 需要注意的是,式(11)并不是满足区间数四则运算的等式,只是表示区间数内某些样本点满足该线性关系.

容易推导得出,区间型符号数据回归模型的误差项也具有三条性质:$E([\boldsymbol{\varepsilon}_i]) = 0$,$D([\boldsymbol{\varepsilon}_i]) = \sigma^2$,$\sigma$ 为常量;$Cov([\boldsymbol{\varepsilon}_a],[\boldsymbol{\varepsilon}_b]) = 0$,$a \neq b$;$[\boldsymbol{\varepsilon}_i]$ 服从正态分布.

2.2 区间数据回归分析的参数估计

设 $f(\boldsymbol{\beta}) = \sum_{i=1}^{n}([y_i] - \beta_0[1_{m_1\times 1}] - \beta_1[x_{i1}] - \cdots - \beta_p[x_{ip}])^2$,则 $f(\hat{\boldsymbol{\beta}}) = \min f(\boldsymbol{\beta})$,即 $\hat{\boldsymbol{\beta}}$ 使误差达到最小. 根据式(4)和式(5),分别对 $\beta_0,\cdots,\beta_p$ 求一阶偏导数,并令其为 0,可得

$$([\boldsymbol{x}]^{\mathrm{T}}[\boldsymbol{x}])\hat{\boldsymbol{\beta}} = ([\boldsymbol{x}]^{\mathrm{T}}[\boldsymbol{x}]) \tag{12}$$

其中

$$[\boldsymbol{x}] = \begin{bmatrix} 1 & [x_{11}] & \cdots & [x_{1p}] \\ 1 & [x_{21}] & \cdots & [x_{2p}] \\ \vdots & \vdots & \ddots & \vdots \\ 1 & [x_{n1}] & \cdots & [x_{np}] \end{bmatrix},$$

$[\boldsymbol{y}]=\begin{bmatrix}[y_{1p}]\\ [y_{2p}]\\ \vdots\\ [y_{np}]\end{bmatrix}$. 式(12)的解可表示为

$$\hat{\beta}=([\boldsymbol{x}]^{\mathrm{T}}[\boldsymbol{x}])^{-1}([\boldsymbol{x}]^{\mathrm{T}}[\boldsymbol{y}]) \tag{13}$$

由于区间矩阵的逆矩阵运算较为复杂,且区间矩阵求逆及求和的结果仍为区间数,因此难以求出式(13)的确切解,本文考虑计算 $[\boldsymbol{x}]^{\mathrm{T}}[\boldsymbol{x}]$、$[\boldsymbol{x}]^{\mathrm{T}}[\boldsymbol{y}]$ 的近似点矩阵,进而求出系数的近似解.

根据协方差的定义,可以推导(篇幅所限,推导过程从略)

$$[x]^{\mathrm{T}}[x]=\begin{bmatrix} n & \sum_{i=1}^{n}E(x_{i1}) & \cdots & \sum_{i=1}^{n}E(x_{ip})\\ \sum_{i=1}^{n}E(x_{i1}) & \sum_{i=1}^{n}E(x_{i1}x_{i1}) & \cdots & \sum_{i=1}^{n}E(x_{i1}x_{ip})\\ \vdots & \vdots & \ddots & \vdots\\ \sum_{i=1}^{n}E(x_{ip}) & \sum_{i=1}^{n}E(x_{ip}x_{i1}) & \cdots & \sum_{i=1}^{n}E(x_{ip}x_{ip})\end{bmatrix} \tag{14}$$

$$[x]^{\mathrm{T}}[y]=\begin{bmatrix}\sum_{i=1}^{n}E(y_i)\\ \sum_{i=1}^{n}(Cov(x_{i1},y_i)+E(x_{i1})E(y_i))\\ \vdots\\ \sum_{i=1}^{n}(Cov(x_{ip},y_i)+E(x_{ip})E(y_i))\end{bmatrix} \tag{15}$$

最后,将式(6)和式(8)代入上述两式,然后再代入式(13),可以求解估计值 $\hat{\boldsymbol{\beta}}$.

3　因变量的区间值预测

给定自变量的一个区间观测值 $\dot{x}=(\dot{x}_1,\cdots,\dot{x}_p)$,其中 $\dot{x}_j=(\dot{x}_j^1,\cdots,\dot{x}_j^{m_j})$ 由 m_j 个散点构成,考虑对因变量的区间值进行预测. 基本思路是先计算各区间自变量的内部散点的均值,并基于此来预测区间因变量的均值. 然后计算区间自变量的均值到区间左端点和右端点的距离,并基于此来预测区间因变量的均值与其区间左端点和右端点的距离. 最后计算得到因变量的区间值的预测.

具体的,令 $x_j^M=\frac{1}{m_j}\sum_{s=1}^{m_j}x_j^s$ 表示区间自变量 x_j 的内部散点的均值,并记 $x^M=(1,x_1^M,\cdots,x_p^M)$,则由

$$y^M=x^M\hat{\beta} \tag{16}$$

估计得到区间因变量的均值.

为构成区间,首先建立区间均值到区间左端点的距离间、区间均值到区间右端点的距离的线性关系. 设

$$y_i^L=\beta_0^L+\beta_1^Lx_{i1}^L+\cdots+\beta_p^Lx_{ip}^L+\varepsilon_i^L \tag{17}$$

$$y_i^R=\beta_0^R+\beta_1^Rx_{i1}^R+\cdots+\beta_p^Rx_{ip}^R+\varepsilon_i^R \tag{18}$$

其中 y_i^L、y_i^R 分别为区间因变量 Y 的第 i($i=1,\cdots,n$)个观测值的区间均值到区间左端点的距离、区间均值到区间右端点的距离,x_{ij}^L、x_{ij}^R 为区间自变量 X_j($j=1,\cdots,n$)的第 i 个观测值的区间均值到区间左端点的距离、区间均值到区间右端点的距离,用 y_i^l、y_i^r 表示区间因变量 Y 的第 i 个观测的区间左端点、区间右端点,用 x_{ij}^l、y_{ij}^r 表示区间自变量 X_j 的第 i 个观测的区间左端点、区间右端点,即

$$y_i^L=|\,y_i^M-y_i^l\,|=\frac{1}{m}\sum_{s=1}^{m}y_i^s-\min_{s=1,\cdots,m}(y_i^s),$$

$$y_i^R=|\,y_i^M-y_i^r\,|=-\frac{1}{m}\sum_{s=1}^{m}y_i^s+\max_{s=1,\cdots,m}(y_i^s) \tag{19}$$

$$x_{ij}^L=|\,x_{ij}^M-x_{ij}^l\,|=\frac{1}{m}\sum_{s=1}^{m}x_{ij}^s-\min_{s=1,\cdots,m}(x_{ij}^s),$$

$$x_{ij}^R=|\,x_{ij}^M-x_{ij}^r\,|=-\frac{1}{m}\sum_{s=1}^{m}x_{ij}^s+\max_{s=1,\cdots,m}(x_{ij}^s) \tag{20}$$

设式(17)和式(18)的参数估计值分别为 $\hat{\boldsymbol{\beta}}^L=(\hat{\beta}_0^L,\cdots,\hat{\beta}_p^L)^{\mathrm{T}}$、$\hat{\boldsymbol{\beta}}^R=(\hat{\beta}_0^R,\cdots,\hat{\beta}_p^R)$,进而可以分别预测区间均值到区间左端点的距离 y^L、区间均值到区间右端点的距离 y^R,针对观测 $\dot{x}$,设 $\dot{x}_j^L$、$\dot{x}_j^R$ 分别为变量 X_j 的观测值的区间均值到区间左、右端点的距离,则

$$y^L=\hat{\beta}_0^L+\hat{\beta}_1^L\dot{x}_1^L+\cdots+\hat{\beta}_p^L\dot{x}_p^L,$$

$$y^R=\hat{\beta}_0^R+\hat{\beta}_1^R\dot{x}_1^R+\cdots+\hat{\beta}_p^R\dot{x}_p^R \tag{21}$$

最终形成预测区间为 $[y^M-y^L,y^M-y^R]$.

通过预测均值、均值到左右端点的距离,进而

形成区间,可以更充分的利用区间内部散点信息.

4 方法评价

4.1 蒙特卡洛模拟

4.1.1 数据形成

产生区间型符号数据的基本思路是:首先生成大量服从某分布的数据,进而从中随机选择几个数据作为散点数据. 假设 x 的概率密度函数为 $f(x)$,那么,$f(x)$ 越大,表明 x 出现的可能越大,在生成的大量数据中,x 出现的次数会越多,因而在随机选择时,x 被选中的几率会越大,即生成的散点中,x 出现的概率越大,这与概率密度函数值得大小情况是一致的. 生成散点后,直接根据散点生成区间,符合数据打包过程,即先有散点,进而打包成数据,无需对区间进行调整. 具体生成过程如下:

考虑 $p = 2$ 的情况,总共生成 200 组数据,其中 150 组数据作为训练集,用于系数的估计,其余 50 组数据作为测试集,用于对回归方法进行评价. 数据的生成通过以下几个步骤获得:

1) 针对每一个区间自变量,假设区间内散点服从某分布,生成 10 000 个服从该分布的点数据,假设区间内部散点数为 m ,则随机选取 m 个数据形成区间内部散点.

2) 假设区间因变量的内部散点与区间自变量的内部散点具有线性关系,当 $p = 2$ 时,$y^s = \beta_0 + \beta_1 x_1^s + \beta_2 x_2^s + \varepsilon$. β_0 、β_1 、β_2 为参数,是常数,其取值在某均匀分布的区间内随机选取. ε 为误差项,服从标准正态分布.

3) 选取内部散点的最小值、最大值构成区间的左端点、右端点.

4.1.2 模拟实验

为考察散点服从不同分布时各方法的优劣,自变量散点分布考虑三种情况:$x_{ij}^s \sim B(2,5)$,$x_{ij}^s \sim B(3,3)$,$x_{ij}^s \sim B(4,1)$ 三种情况,分别为右偏态分布、对称分布和左偏态分别的情况.

为考虑系数在不同范围内取值时各方法的优劣,系数考虑五类:$\beta_j \sim U(15,30)$ 、$\beta_j \sim U(0,1)$、$\beta_j \sim U(-1,1)$、$\beta_j \sim U(-1,0)$、$\beta_j \sim U(-30,-15)$. 前两种考虑了系数全为正值的情况,后两种考虑了系数全为负的情况,第三种考虑的是既有正值、又有负值的情况. 从另一个角度看,第一种和最后一种情况,考虑的是系数较大的情形,而中间三种考虑的是系数较小的情况,由于误差假定服从的都是正态分布,所以当系数越大时,误差对因变量值得影响会越小,而系数越小时,误差对因变量的影响会越大.

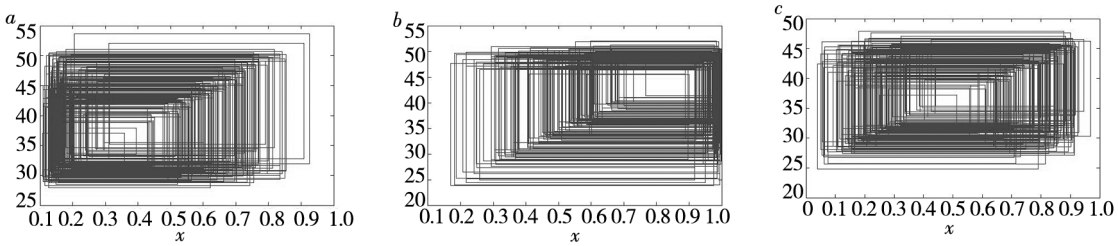
将散点服从的分布、系数取值的不同范围进行组合,共有 15 种情况,如表 1 所示.

表 1 蒙特卡洛模拟的几种情形

Table 1 Configurations of Monte Carlo simulation

系数分布	散点分布		
	$x_{ij}^s \sim B(2,5)$	$x_{ij}^s \sim B(4,1)$	$x_{ij}^s \sim B(3,3)$
$\beta_j \sim U(15,30)$	C1	C6	C11
$\beta_j \sim U(0,1)$	C2	C7	C12
$\beta_j \sim U(-1,1)$	C3	C8	C13
$\beta_j \sim U(-1,0)$	C4	C9	C14
$\beta_j \sim U(-30,-15)$	C5	C10	C15

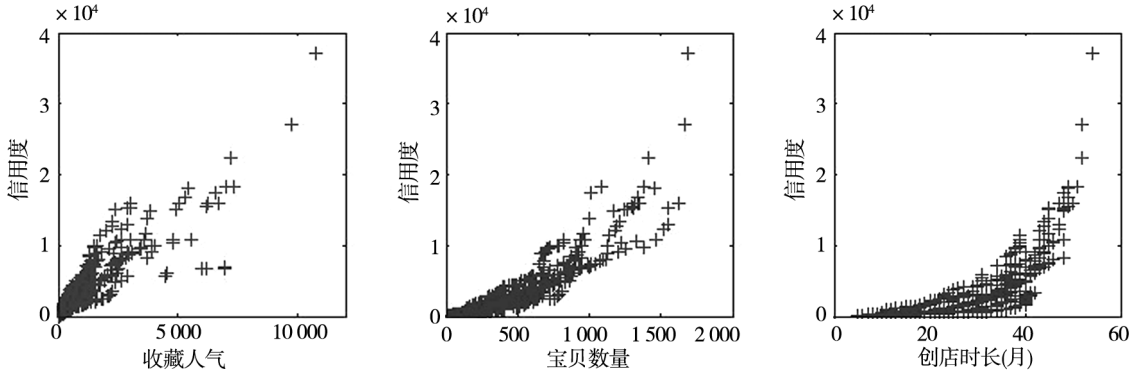
图 1 是 $p = 1$,系数在区间[15,30]内服从均匀分布,散点分别服从 $B(2,5)$ 、$B(4,1)$ 、$B(3,3)$ 分布时随机生成的数据绘制的矩形图. 可以看出,区间自变量和区间因变量之间的线性关系比较明显. 图 2(a)显示出区间内部散点呈右偏态分布,图 2(b)显示出内部散点呈左偏态分布,而图 2(c)中区间内部散点呈对称分布.



(a) 右偏态分布　　(b)左偏态分布　　(c)对称分布

图 1　$p=1$ 情况下随机生成的数据矩形图

Fig. 1 Random generated rectangular data at $p=1$



(a)　　(b)　　(c)

图 2　收藏人气、宝贝数量、创店时长与信用度间散点图

Fig. 2 Collection popularity, quantity, store duration and credit point data

针对表 1 中的几种情形,首先生成一组系数,然后按照上述数据生成步骤,生成 200 组实验数据,任选其中 150 组数据作为训练集,剩余 50 组作为测试集. 上述实验重复 100,以避免偶然情况的出现. 选用 $RMSE_I$ 、AR 、PCO 三个评价指标,具体解释如下:

Chuang[25] 基于 Hausdorff 距离[26] 提出一种均方根误差 $RMSE_I$

$$RMSE_I=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(|y_i^c-\hat{y}_i^c|+|y_i^r-\hat{y}_i^r|)^2} \quad (22)$$

其中 $y_i^c=(y_{Li}+y_{Ui})/2$, $y_i^r=(-y_{Li}+y_{Ui})/2$.

Hu 和 He[27] 定义了准确率的概念,平均准确率 AR 的计算公式如下

$$AR=\frac{1}{n}\sum_{i=1}^{n}\frac{w(Y_i\cap\hat{Y}_i)}{w(Y_i\cup\hat{Y}_i)} \quad (23)$$

Mehran 等[28] 提出可以通过计算预测区间包含的观测区间的比率 OCP 、以及观测区间包含的预测区间的比率 PCO 对预测效果进行评价

$$OCP=\frac{1}{n}\sum_{i=1}^{n}\frac{w(y_i\cap\hat{y}_i)}{w(\hat{y}_i)},PCO=\frac{1}{n}\sum_{i=1}^{n}\frac{w(y_i\cap\hat{y}_i)}{w(y_i)} \quad (24)$$

4.1.3　实验结果

采用 t 检验对五种方法进行比较,显著水平为 1%. $RMSE_I$ 值越小表明方法越有优势,而 AR 和 PCO 两个指标,值越大则方法越好. 若以方法 A 不优于方法 B 为零假设,方法 A 优于方法 B 为备择假设,则对于指标 $RMSE_I$ $H_0:RMSE_I^A\geqslant RMSE_I^B$, $H_1:RMSE_I^A<RMSE_I^B$. 而对于 AR 和 PCO 两个指标, $H_0:AR^A\leqslant AR^B$, $H_1:AR^A>AR^B$, 以及 $H_0:PCO^A\leqslant PCO^B$, $H_1:PCO^A>PCO^B$. 若假设方法 A 和 B 相同,则对于任意指标,零假设为方法 A 与方法 B 的指标值相同,备择假设为不相同.

表 2 为区间内散点服从右偏态分布时,本文所提出的 DSM 方法与现有的常用方法 CCRM[7] 的对比结果. 表中 A 表示本文所提出的 DSM 方法,B 表示 CCRM 方法.

表 2 散点服从右偏态实验结果

Table 2 Test result of right-skewed distribution

参数分布	备择假设 A:DSM;B:CCRM 法	$RMSE_I$(%)	AR(%)	PCO(%)
$\beta_j \sim U(15,30)$	A 优于 B	100.00	100.00	100.00
	A 劣于 B	0.00	0.00	0.00
	A 不等同于 B	100.00	100.00	100.00%
$\beta_j \sim U(0,1)$	A 优于 B	58.00	34.00	46.00%
	A 劣于 B	0.00	0.00	0.00%
	A 不等同于 B	48.00	26.00	40.00%
$\beta_j \sim U(-1,1)$	A 优于 B	44.00	40.00	38.00
	A 劣于 B	0.00	0.00	0.00
	A 不等同于 B	36.00	30.00	28.00
$\beta_j \sim U(-1,0)$	A 优于 B	34.00	28.00	38.00
	A 劣于 B	0.00	0.00	0.00
	A 不等同于 B	26.00	18.00	24.00
$\beta_j \sim U(-30,-15)$	A 优于 B	0.00	0.00	0.00
	A 劣于 B	100.00	100.00	70.00
	A 不等同于 B	100.00	100.00	64.00

类似的,可以得到当区间内散点服从左偏态分布和对称分布时,DSM 与 CCRM 的对比结果. 结果显示,无论散点服从何种分布,只有当系数在[-30,-15]内取值时,CCRM 方法更优,而系数在其他范围内取值时,本文所提出的 DSM 更优.

4.2 实例 1——淘宝卖家信用回归预测研究

淘宝网是 C2C 电子商务的代表,经过十几年的发展,已经成为亚洲最大的零售商圈. 用户在购买时,除关注商品本身的样式、尺寸、价格、评价等因素,往往还会关注店铺信用度,信用度高,则用户更为信任,购买的意愿会更加强烈. 本文对信用度的影响因素进行了研究,数据选取的是包括北京、天津等城市在内的共 10 个城市的淘宝店铺数据,每个店铺数据包括卖家信用、创店时长、宝贝数量、收藏人气五项内容. 图 2 分别为卖家信用与收藏人气、宝贝数量、创店时长的散点图.

图 2(a)和图 2(b)显示出,信用度与收藏人气和宝贝数量具有较好的线性关系,图 2(c)表示信用度与创店时长之间并非线性关系. 因此本文选取收藏人气和宝贝数量作为自变量,其对应的区间变量分别记为$[X_1]$和$[X_2]$,选取信用度为因变量,其对应的区间变量记为$[Y]$. 为验证 DSM 方法的有效性,以城市作为符号对象,进行回归分析. 表 3 为以城市为单位进行“数据打包”形成的区间型符号数据,及符号数据内部散点的均值和偏态系数. 从偏态系数看,全部区间数内部散点均是右偏态.

表 3 数据打包后的区间数及区间内部散点的均值和偏态系数

Table 3 Interval, mean value and skew coefficient of data point within intervals after being packed

城市	$[Y]$	$[X_1]$	$[X_2]$	$\bar{Y}$	$\bar{X}_1$	$\bar{X}_2$	C_Y	C_{X_1}	C_{X_2}
北京	[54,16 038]	[6,3 037]	[23,1 625]	3 220	609	419	2.77	2.46	2.77
长沙	[14,3 410]	[4,1 010]	[8,595]	568	152	195	3.34	1.96	3.45
成都	[30,6 986]	[8,6 960]	[13,945]	1 280	839	276	3.38	2.48	4.08
广州	[59,18 348]	[4,7 338]	[28,1 089]	2 166	753	368	4.34	1.70	4.60
哈尔滨	[27,8 985]	[4,3 900]	[25,930]	1 489	416	251	3.47	2.46	4.19
杭州	[46,18 158]	[2,5 400]	[25,1 455]	3 308	602	415	3.23	2.49	4.24
济南	[12,3 258]	[5,990]	[10,677]	570	148	191	3.39	2.21	3.46
南京	[41,10 547]	[5,2 450]	[29,1 328]	2 248	466	385	2.86	2.12	2.86
上海	[100,37 144]	[20,10 750]	[46,1 686]	4 319	1 368	506	3.69	2.17	3.66
天津	[37,10 814]	[29,4 790]	[17,824]	3 199	1 117	318	2.31	1.64	2.63

表 4 为 DSM 方法和 CCRM 方法的回归分析参数估计值，与直接利用散点进行回归十分相近，这表明 DSM 能够充分利用散点信息，可以获得与直接用散点进行回归近似的结果.

表 4　各方法系数估计结果

Table 4 Estimation results of compared methods

模型		常变量	收藏人气 $[X_1]$	宝贝数量 $[X_2]$
DSM 回归模型	β_{DSM}	−555.30	1.61	5.27
	β_L	0.00	1.31	4.73
	β_R	0.00	1.88	5.66
CCRM 回归模型	β_c	−4 953.46	1.75	13.31
	β_r	0.00	1.77	5.51

根据估计出的系数，可以对原始样本数据进行样本内预测，结果如表 5 所示，$[\hat{Y}_{\mathrm{DSM}}]$ 表示应用 DSM 方法进行预测的结果，$[\hat{Y}_{\mathrm{CCRM}}]$ 为用约束中点半径法回归预测的结果.

表 5　样本内预测结果

Table 5 In-sample estimation results

城市	$[Y]$	$[\hat{Y}_{\mathrm{DSM}}]$	$[\hat{Y}_{\mathrm{CCRM}}]$
北京	[54,16 038]	[−30,14 027]	[1 574,15 782]
长沙	[14,3 410]	[−361,4 594]	[−2 563,2 457]
成都	[30,6 986]	[−83,17 541]	[−1 220,16 254]
广州	[59,18 348]	[7,19 054]	[−529,18 334]
哈尔滨	[27,8 985]	[−172,11 829]	[−1 133,10 768]
杭州	[46,18 158]	[−30,17 507]	[892,18 353]
济南	[12,3 258]	[−354,5 024]	[−2 222,3 202]
南京	[41,10 547]	[−64,11 293]	[478,11 975]
上海	[100,37 144]	[407,28 597]	[1 954,30 034]
天津	[37,10 814]	[70,12 689]	[−1 589,11 307]

表 6 为通过样本预测计算得到的各方法的各项评价指标结果. 从区间整体的误差评价指标 $RMSE_I$ 看，CCRM 略好. 从 AR 、PCO 两个指标看，DSM 方法结果更好.

表 6　样本内预测各评价指标计算结果

Table 6 In-sample estimation evaluation results

方法	$RMSE_I$	AR	PCO
DSM	4 540	0.776	0.960
CCRM	4 391	0.738	0.926

由于样本内预测存在过拟合的情况，会影响结果的评价，因此还需要进行样本外预测以验证方法的有效性. 然而样本数量较少，再划分训练集测试集，训练集和测试集内样本数均较少，考虑到这个问题，本文采取 k − 折交叉验证来进行方法评价，这里 k 取极端值，即 k 的值为样本个数值，此方法又称为留一交叉验证法.

表 7 为采取留一法进行预测的结果，可以看出，通过各方法得出的预测区间，均与实际区间有大范围的重叠区域，由于区间较宽，从预测结果上无法直观的看出各方法的优劣.

表 7 留一交叉验证法预测结果

Table 7 Leave-one-out cross validation estimation results

城市	$[Y]$	$[\hat{Y}_{\mathrm{DSM}}]$	$[\hat{Y}_{\mathrm{CCRM}}]$
北京	[54,16 038]	[11,12 736]	[2 868,15 785]
长沙	[14,3 410]	[-422,4 654]	[-3 511,1 615]
成都	[30,6 986]	[-56,20 256]	[-1 110,19 138]
广州	[59,18 348]	[-1,19 191]	[-693,18 324]
哈尔滨	[27,8 985]	[-189,12 030]	[-1 186,10 909]
杭州	[46,18 158]	[-15,17 358]	[1 077,18 411]
济南	[12,3 258]	[-414,5 169]	[-2 813,2 821]
南京	[41,10 547]	[-92,11 573]	[670,12 549]
上海	[100,37 144]	[414,22 279]	[1 876,23 207]
天津	[37,10 814]	[178,12 784]	[-1 813,11 262]

图 3 为预测区间及实际区间折线图. 其中 Pre_{L}和 Pre_{R}分别表示预测区间数的左、右端点，$Real_{\mathrm{L}}$和 $Real_{\mathrm{R}}$分别表示实际区间数的左、右端点. 从图上看，DSM 模型在左端点上的预测误差较小，而 CCRM 方法在左端点的预测上与实际值有较大偏差；在右端点的预测上，两者差异不明显.

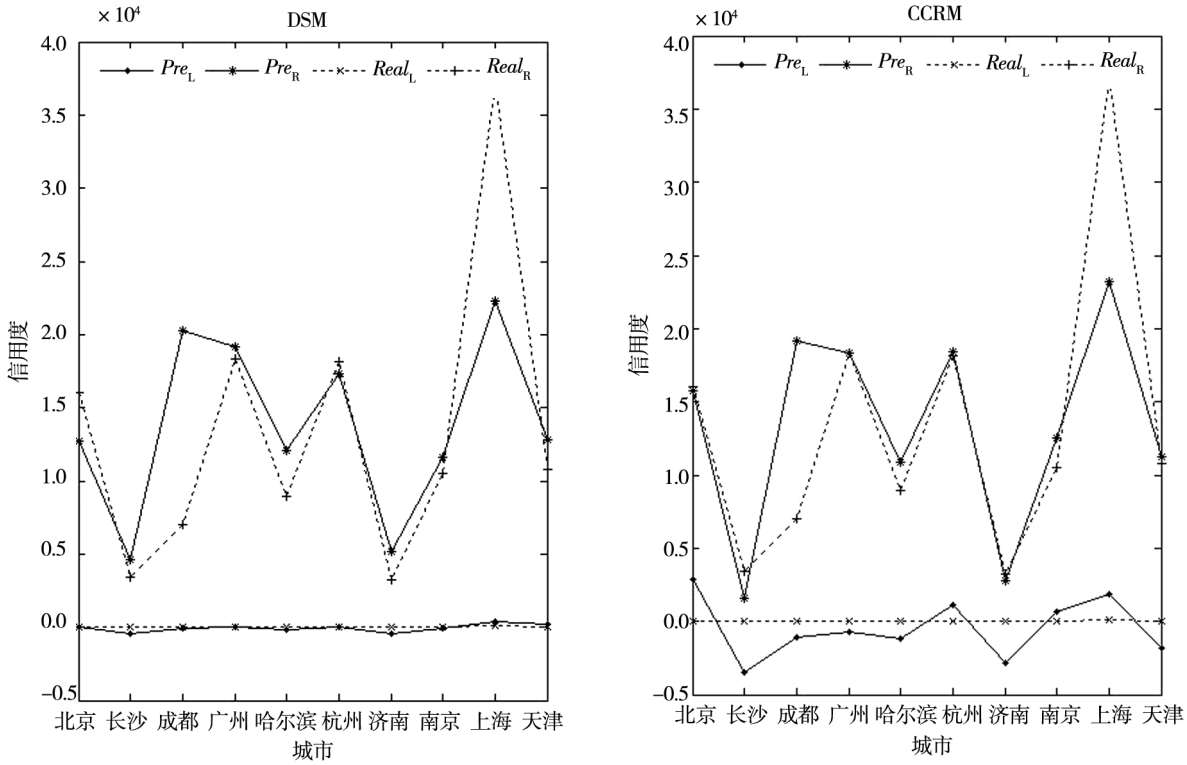


图 3 留一交叉验证法预测结果折线图

Fig. 3 Line graph of leave-one-out cross validation estimation results

表 8 为各项评价指标结果，从区间整体的误差评价指标 $RMSE_I$ 看，CCRM 具有很弱的优势，而从 AR、PCO 两个指标看，DSM 显著优于 CCRM.

表 8　样本外预测各指标评价指标结果

Table 8 Out-of-sample estimation evaluation results

方法	$RMSE_I$	AR	PCO
DSM	6 548	0.735	0.933
CCRM	6 489	0.667	0.860

基于以上分析，由文中提出的区间型符号数据的回归分析方法 DSM，可得到以城市为符号对象样本，卖家信用度与收藏人气和宝贝数量的回归方程为

$$[Y] = -555.30 + 1.61[X_1] + 5.27[X_2]$$

基于该回归方程，在通过数据打包获取样本外其他城市的收藏人气和宝贝数量区间数据后，可以按前文给出的因变量的区间值预测方法，对样本外城市的信用度的区间取值进行预测. 另外，还可看出收藏人气和宝贝数量这两个因素，对卖家的信用度均有正向的影响作用，因而，卖家可以通过提高收藏人气和增加宝贝数量的方式，来提高其信用度，为淘宝商家提供决策支持.

4.3　实例 2——股指变动回归预测研究

该节采用文献[19]中的数据，研究股指变动的回归预测问题. 该文献研究股指的变动(SP)受 5 个宏观经济因素的影响，分别是随季节调整的工业生产指数增长率的变动(IP)、期望通胀变动(UI)、非预期通胀变动(DI)、违约风险溢价(DF)、非预期利率变动(TM). 它们之间在一定时间内具有线性关系，可以表示为

$$SP = a + i * IP + u * UI + d * DI + f * DF + t * TM \tag{26}$$

将一年的数据打包成一个区间数，则每个区间数包含 12 个散点. 一次回归只能得到一组系数，反映股指变动与宏观经济变量的一种关系. 然而现实中，股指变动与宏观经济变量的关系是随时间而变动的，一种关系只能在一段时间内维持，滚动回归则能解决这一问题，每次回归在时间窗口内进行. 文献[19]给出了 1930 年至 2004 年的样本数据. 时间窗口取 10，即每次回归只对连续十年的数据进行分析，回归后将窗口向后移动一个步长，此处取 1 年作为步长，再次进行回归，直到窗口不能再向后移动. 将回归结果用于时间窗口内最后一年预测，称为样本内预测，用于紧随时间窗口下一年预测，称为样本外预测. 考虑到样本内预测可能存在的过拟合问题，此处只进行样本外预测.

图 4 和图 5 分别给出了文中方法 DSM 和对比方法 CCRM 的样本外预测结果. 从图上可以看出，DSM 模型的预测误差较小，其预测效果要优于 CCRM 方法. 特别是 CCRM 方法在几个年份上的预测值与实际值相差较大，如 1946 年的实际区间是[-0.16, 0.06]，而预测区间为[0.02, 0.22]，误差较大.

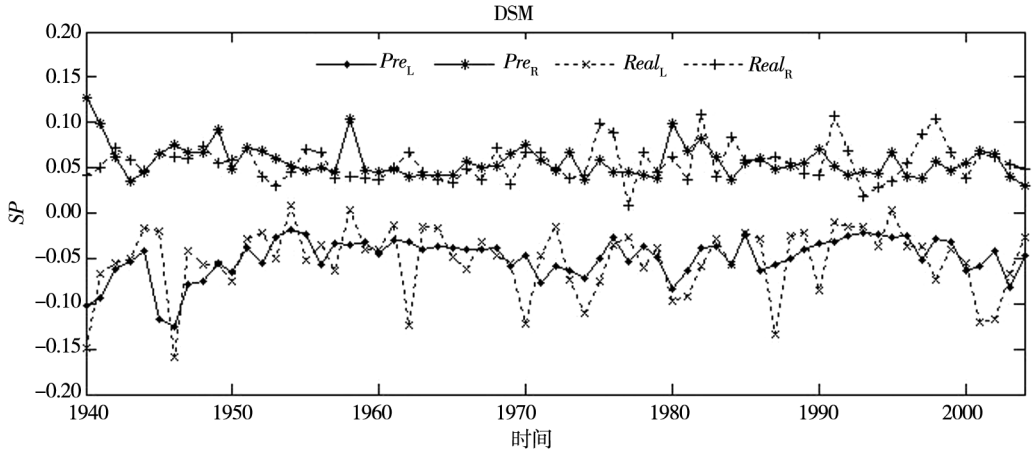


图 4　DSM 方法样本外预测结果

Fig. 4 Out-of-sample estimation results of DSM method

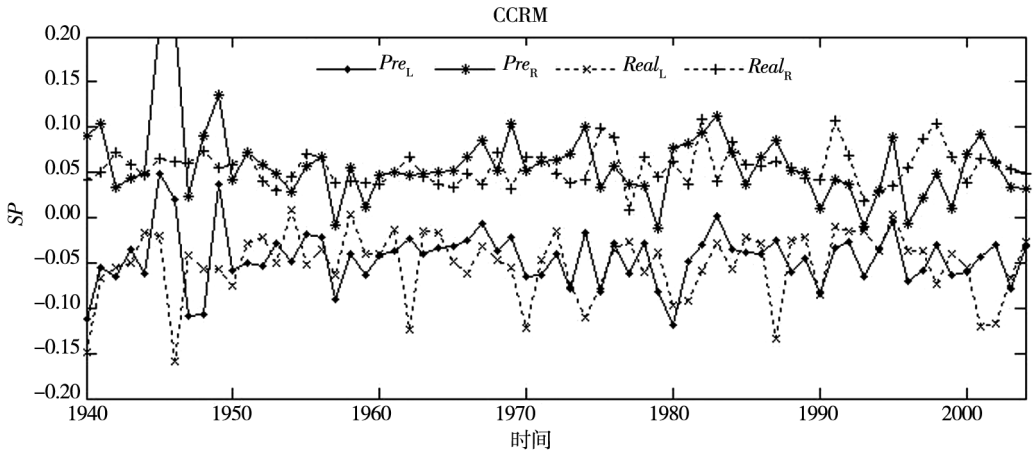


图 5 CCRM 方法样本外预测结果

Fig. 5 Out-of-sample estimation results of CCRM method

表 9 股指变动预测问题——样本外预测各指标评价指标结果

Table 9 Stock index fluctuating forecast problem: Out-of-sample estimation evaluation results

方法	$RMSE_I$	AR	PCO
DSM	0.039 3	0.656 4	0.830 0
CCRM	0.055 9	0.547 9	0.710 8

表 9 结果表明,从各评价指标看,文中方法 DSM 与 CCRM 相比,均具有显著优势. 由此,可以运用 DSM 方法得到该问题的回归方程,进而分析股指变动受工业生产指数增长率等因素的变化而变化的趋势.

5 结束语

本文提出了一种考虑区间内部点数据信息的区间回归模型. 蒙特卡洛模拟实验表明,DSM 方法在多数情况下优于 CCRM 方法,除了系数为负且绝对值较大时. 实例评价结果显示出 DSM 的预测准确率更高.

本文的不足在于,模拟实验的结果,只是说明了何时 DSM 更优,何时 CCRM 更优,因而有必要进一步研究,针对一个具体问题的样本数据,是否存在明确的界限,用以分辨选择 DSM 方法还是 CCRM 方法.

参 考 文 献:

[1] Moore R E. Interval Analysis[M]. New Jersey: Prentice-Hall, Englewood Cliffs, 1966.

[2] Bock H H, Diday E (Eds.). Analysis of Symbolic Data[M]. Berlin: Springer-Verlag, 2000.

[3] 胡 艳, 王惠文. 一种海量数据的分析技术——符号数据分析及应用[J]. 北京航空航天大学学报(社会科学版), 2004, 17(2): 40 – 44.

Hu Yan, Wang Huiwen. A new data mining method based on huge data and its application[J]. Journal of Beijing University of Aeronautics and Astronautics (Social Science Edition), 2004, 17(2): 40 – 44. (in Chinese)

[4] Billard L, Diday E. Regression Analysis for Interval-Valued Data, In Data Analysis, Classification, and Related Methods [M]. Berlin: Springer-Verlag, 2000: 369 – 374.

[5] Billard L, Diday E. Symbolic Regression Analysis. In Classification, Clustering and Data Analysis[M]. Berlin: Springer-Verlag, 2002: 281 – 288.

[6] Lima Neto E A, De Carvalho F D T. Centre and range method for fitting a linear regression model to symbolic interval data

[J]. Computational Statistics & Data Analysis, 2008, 52(3): 1500 – 1515.

[7] Lima Neto E D A, De Carvalho F D A. Constrained linear regression models for symbolic interval-valued variables[J]. Computational Statistics & Data Analysis, 2010, 54(2): 333 – 347.

[8] Giordani P. Lasso-constrained regression analysis for interval-valued data[J]. Advances in Data Analysis and Classification, 2015, 9(1): 5 – 19.

[9] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society Series b-Statistical Methodology, 1996, 58: 267 – 288.

[10] Lima Neto E D A, De Carvalho F D A. Nonlinear regression applied to interval-valued data[J]. Pattern Analysis and Applications, 2017, 20(3): 809 – 824.

[11] Hladík M, Černý M. Two optimization problems in linear regression with interval data[J]. Optimization, 2017, 66(3): 331 – 349.

[12] Hao P, Guo J P. Constrained center and range joint model for interval-valued symbolic data regression[J]. Computational Statistics & Data Analysis, 2017, 116: 106 – 138.

[13] Zhou X Q, Feng Y Q, Du X L. Quantile regression for interval censored data[J]. Communications in Statistics: Theory and Methods, 2017, 46(8): 3848 – 3863.

[14] Souza L C, Souza R M C R, Amaral G J A, et al. A parametrized approach for linear regression of interval data[J]. Knowledge-Based Systems, 2017, 131: 149 – 159.

[15] Maia A L S, De Carvalho F D A. Holt's exponential smoothing and neural network models for forecasting interval-valued time series[J]. Int. J. Forecast, 2011, 27(3): 740 – 759.

[16] Maia A L S, De Carvalho F D A, Ludermir T B. Forecasting models for interval valued time series[J]. Neurocomputing, 2008, 71(16): 3344 – 3352.

[17] Roque S A M, Maté C, Arroyo J, et al. iMLP: Applying multi-layer perceptrons to interval-valued data[J]. Neural Process. Lett, 2007, 25(2): 157 – 169.

[18] Arroyo J, San Roque M A, Maté C, et al. Exponential Smoothing Methods for Interval Time Series[C]. In: Proceedings of the 1st European Symposium on Time Series Prediction, 2007: 231 – 240.

[19] Xiong T, Li C G, Bao Y K, et al. A combination method for interval forecasting of agricultural commodity futures prices [J]. Knowledge-Based Systems, 2015, 77: 92 – 102.

[20] Lim C. Interval-valued data regression using nonparametric additive models[J]. Journal of the Korean Statistical Society, 2016, 45: 358 – 370.

[21] Fagundes R A A, De Souza R M C R, Cysneiros F J A. Robust regression with application to symbolic interval data[J]. Engineering Applications of Artificial Intelligence, 2013, 26(1): 564 – 573.

[22] Xiong T, Bao Y K, Hu Z Y. Multiple-output support vector regression with a firefly algorithm for interval-valued stock price index forecasting[J]. Knowledge-Based Systems, 2014, 55: 87 – 100.

[23] 李汶华, 郭均鹏. 区间型符号数据回归分析及其应用[J]. 管理科学学报, 2010, 13(4): 38 – 43.
Li Wenhua, Guo Junpeng. Methodology and application of regression analysis of interval-type symbolic data[J]. Journal of Management Sciences in China, 2010, 13(4): 38 – 43. (in Chinese)

[24] 郭均鹏, 李汶华, 高　峰. 一般分布区间型符号数据的描述统计与分析[J]. 系统工程理论与实践, 2011, 31(12): 2367 – 2372.
Guo Junpeng, Li Wenhua, Gao Feng. Descriptive statistics and analysis of interval symbolic data with general distribution [J]. System Engineering: Theory & Practice, 2011, 31(12): 2367 – 2372. (in Chinese)

[25] Chuang C C. Extended support vector interval regression networks for interval input-output data[J]. Information Sciences, 2008, 178(3): 871 – 891.

[26] De Carvalho F D A T, De Souza R M C R, Chavent M, et al. Adaptive hausdorff distances and dynamic clustering of sym-

bolic interval data[J]. Pattern Recognition Letters, 2006, 27(3): 167 – 179.

[27] Hu C, He L T. An application of interval methods to stock market forecasting[J]. Reliable Computing, 2007, 13(5): 423 – 434.

[28] Mehran H, Bector C R, Smimou K. A simple method for computation of fuzzy linear regression[J]. European Journal of Operational Research, 2005, 166(1): 172 – 184.

Regression analysis for interval symbolic data considering point data within intervals

GUO Jun-peng, ZHAO Ru, LI Wen-hua

College of Management and Economics, Tianjin University, Tianjin 300072, China

Abstract: The paper studies regression analysis for interval data which are obtained by packaging point data. The existing methods of regression analysis for interval data only consider the endpoints of the intervals. This study focuses on how to use the point data within intervals sufficiently. Firstly, when the error of the regression model for point data satisfies the supposed three characteristics, the error of the regression model for interval data also follows the three characteristics. Then, based on the descriptive statistics of interval symbolic data, a novel approach of estimating the parameters of the regression model is proposed. A method is given to predict the interval values of the dependent variables. Finally, a simulation experiment and an application case are performed following the proposed methodology. The results indicate a better performance of our method than the existing CCRM method.

Key words: regression analysis; interval symbolic data; descriptive statistics