区间型符号数据回归分析及其应用①

李汶华, 郭均鹏
(天津大学管理学院, 天津 300072)

摘要: 介绍了符号数据分析方法的基本理论. 针对一种最常用的符号数据——区间型符号数据, 基于误差传递的理论, 提出了区间回归分析的方法. 方法包括了线性回归分析和可线性化的非线性回归分析两种情形. 讨论了基于 Hausdorff距离的区间数距离, 基于此定义了回归模型的评价指标. 进行了方法的应用研究, 选取沪深 300指数与中信规模风格指数, 从时间维上对其日内数据进行"数据打包", 形成区间型符号数据; 建立了区间线性回归分析模型, 从全局上揭示了两类指数间的相关性. 结论表明, 与针对点数据的传统回归分析相比, 区间型符号数据的回归分析方法不仅实现了样本空间的降维, 而且有利于从整体上把握变量之间的内在关系.

关键词: 符号数据; 回归分析; 区间数; 误差传递; 指数
中图分类号: O212 1 文献标识码: A 文章编号: 1007- 9807(2010) 04- 0038- 06

0 引 言

当样本量巨大时, 传统多元分析方法往往难以把握数据属性的内在关系, 无法获得隐含在数据中的重要知识资源[1], 针对这一类问题, 在 20世纪 90年代的欧洲, 人们提出了一种全新的数据分析思路——符号数据分析 (symbolic data analysis, SDA) [2]. 符号数据分析技术运用"数据打包"的思想, 不仅实现了样本空间的降维, 而且较传统数据分析技术而言, 可以从全局上把握数据对象的内在结构特征.

SDA主要包含两个步骤.

步骤 **1** 称为数据挖掘.

应用数据库查询或分类汇总等方法, 将大样本的原始数据转变为小样本的"符号"数据, 实现了样本空间的降维. 该过程如图 1所示.

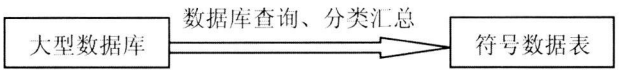


图 **1** 由原始数据库生成符号数据表

Fig 1 From original data base to symbolic data table

符号数据可能是定量数据, 也可能是定性数据; 可有多种表现形式, 常用的符号数据类型有: (1) 区间变量 (Interval-valued Variables). $X = [\underline{x}, \bar{x}] = \{x : \underline{x} \leqslant x \leqslant \bar{x}, x \in R\}$ 称为一个区间变量, $\underline{x}$ 和 $\bar{x}$ 分别称为区间数的下限和上限. (2) 多值变量 (multi-valued variables). (3) 分布式变量 (Distribution Variables). 其中, 区间型符号数据是最常用的一种符号数据类型.

步骤 **2** 称为知识挖掘.

针对符号数据样本, 对其进行数据分析, 常用的分析方法有主成分分析、多元回归分析等多元分析方法. 由于传统的数据分析方法只能处理"点"数据, 因此需对其进行拓展研究, 使其能处理"符号"数据. 该过程如图 2所示.

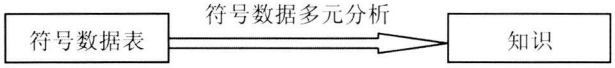


图 **2** 对符号数据表的多元分析

Fig 2 Multivariable analysis of symbolic data table

近些年来, 符号数据分析在理论方法和应用研究中取得了诸多成果[3-7]. Billard和 Diday等

① 收稿日期: 2008- 11- 10; 修订日期: 2009- 12- 02
基金项目: 国家自然科学基金资助项目 (70701026).
作者简介: 李汶华(1974—), 女, 江西南昌人, 博士, 副教授. Email: liwh@ tju. edu. cn

基于对区间数据的协方差的计算, 给出了区间型符号数据的回归方法[8], 计算较为复杂且只适合于线性情形. 事实上, 区间分析的概念最初是从误差分析的角度提出的[9], 区间数 $[a, b]$ 又可表示为 $[c-r, c+r]$, 其中 $c=(a+b)/2$是区间数的中点, $r=(b-a)/2$表示半径, 反映的是该区间数的误差. 这种表示同样适用于符号数据分析理论体系中, 由数据打包所形成的区间数. 本文将从误差传递的理论出发, 研究基于误差传递的区间型符号数据的回归分析方法, 并将其应用于中国的股票市场.

1　误差传递公式

设有 n 个直接测量值 $x_1, \cdots, x_n$, 间接测量值为 y, 它们之间有连续可微的函数关系

$$y=f(x_1, \cdots, x_n) \tag{1}$$

则称 y 的误差与 $x_1, \cdots, x_n$ 各误差之间的关系称为误差传递.

设 $x_1, \cdots, x_n$ 的随机误差分别为 $\delta_1, \cdots, \delta_n$, y 的随机误差分别为 δ, 则误差传递的一般公式为[10-11]

$$\delta^2_{y\lim}=\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2\delta^2_{x_i\lim}+2\sum_{1\leqslant i<j\leqslant n}\frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}\rho_{ij}\delta_{x_i\lim}\delta_{x_j\lim} \tag{2}$$

其中, $\delta_{\lim}$ 表示随机变量 δ的极限误差 (即误差的最大估计), ρ_{ij} 为测量值 x_i 和 x_j 的误差的相关系数. 特别的, 当各测量值 x_i 的随机误差 δ_i $(i=1, \cdots, n)$ 相互独立时, $\rho_{ij}=0$ (2)可化简为

$$\delta^2_{y\lim}=\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2\delta^2_{x_i\lim} \tag{3}$$

由于在实际中各测量值的随机误差之间常是相互独立或近似相互独立, 因此, 公式 (3) 是较为广泛使用的误差传递公式[10]. 公式 (3) 将是本文所提出的基于误差传递公式的区间回归分析方法的主要依据.

2　基于误差传递公式的区间回归分析

2.1　线性回归分析

设有 $p+1$个区间变量 Y和X_j, $y_i=[\underline{y_i}, \overline{y_i}]$表示变量 Y的第 i个区间观测值, $x_{ij}=[\underline{x_{ij}}, \overline{x_{ij}}]$ 表示变量X_j的第 i个区间观测值, $i=1, 2, \cdots, n$, $j=1, \cdots, p$, 则样本矩阵可表示为

$$[Y]=\begin{bmatrix}[\underline{y_1}, \overline{y_1}]\\ [\underline{y_n}, \overline{y_n}]\end{bmatrix},$$

$$[X]=\begin{bmatrix}1 & [\underline{x}_{11}, \overline{x}_{11}] & \cdots & [\underline{x}_{1p}, \overline{x}_{1p}]\\ \vdots & \vdots & & \vdots\\ 1 & [\underline{x}_{n1}, \overline{x}_{n1}] & \cdots & [\underline{x}_{np}, \overline{x}_{np}]\end{bmatrix} \tag{4}$$

希望估计 $[Y]$ 和 $[X]$ 之间的线性回归模型

$$[Y]=[X]\beta+\varepsilon \tag{5}$$

得到区间变量的估计的回归方程

$$[\hat{Y}]=[X]\hat{\beta} \tag{6}$$

令 y_i^c和 y_i^r表示区间因变量 Y的第 i个区间观察值的中点和半径, 即 $y_i^c=\frac{\overline{y}_i+\underline{y}_i}{2}$, $y_i^r=\frac{\overline{y}_i-\underline{y}_i}{2}$, 令 x_{ij}^c和 x_{ij}^r表示第 j个区间自变量X_j 的第 i个区间观察值的中点和半径, 即 $x_{ij}^c=\frac{\overline{x}_{ij}+\underline{x}_{ij}}{2}$, $x_{ij}^r=\frac{\overline{x}_{ij}-\underline{x}_{ij}}{2}$, 其中 $i=1, 2, \cdots, n$, $j=1, \cdots, p$, 则 $y_i=[\underline{y_i}, \overline{y_i}]$可等价表示成 $y_i=[y_i^c-y_i^r, y_i^c+y_i^r]$, $x_{ij}=[\underline{x_{ij}}, \overline{x_{ij}}]$ 可等价表示成 $x_{ij}=[x_{ij}^c-x_{ij}^r, x_{ij}^c+x_{ij}^r]$. 记

$$Y^c=\begin{bmatrix}y_1^c\\ \vdots\\ y_n^c\end{bmatrix},\quad X_j^c=\begin{bmatrix}x_{1j}^c\\ \vdots\\ x_{nj}^c\end{bmatrix} \tag{7}$$

考虑建立 Y^c和X_j^c $(j=1, \cdots, p)$之间的普通线性回归模型

$$Y^c=\beta_0+\beta_1X_1^c+\cdots+\beta_pX_p^c+\varepsilon \tag{8}$$

通过最小二乘法, 可以得到回归方程

$$\hat{Y}^c=\hat{\beta}_0+\hat{\beta}_1X_1^c+\cdots+\hat{\beta}_pX_p^c \tag{9}$$

考虑误差传递公式 (3), 则 $\hat{y_i^c}$的极限误差, 即估计的半径为

$$\hat{y}_i^r=\sqrt{\sum_{j=1}^{p}\left(\frac{\partial\hat{Y}^c}{\partial X_j^c}\right)^2(x_{ij}^r)^2}=\sqrt{\sum_{j=1}^{p}(\hat{\beta}_j)^2(x_{ij}^r)^2} \tag{10}$$

于是可以得到估计的区间因变量 $\hat{y}_i=[\underline{\hat{y}_i}, \overline{\hat{y}_i}]$, 其

中$\underline{\hat{y}}_i = \hat{y}_i^c - \hat{y}_i^r$, $\overline{\hat{y}}_i = \hat{y}_i^c + \hat{y}_i^r$.

2.2 可线性化的非线性回归分析

假设区间变量 Y和$X_j\ (j = 1, \cdots, p)$之间是某种可以线性化的非线性关系, 则可以找到非线性函数$\tilde{X}_j = g_j(X_j)$, 使得 Y和$\tilde{X}_j$之间存在线性关系. 于是可通过变量代换 $\tilde{X}_j^c = g_j(X_j^c)$, 得到 Y^c 和 $\tilde{X}_j^c$ 之间的线性回归模型

$$Y^c = \beta_0 + \beta_1\tilde{X}_1^c + \cdots + \beta_p\tilde{X}_p^c + \varepsilon \qquad (11)$$

进一步, 应用最小二乘方法得到估计的回归方程

$$\hat{Y}^c = \hat{\beta}_0 + \hat{\beta}_1\tilde{X}_1^c + \cdots + \hat{\beta}_p\tilde{X}_p^c \qquad (12)$$

考虑误差传递公式 (3), 则$\hat{y}_i^r$ 的极限误差, 即估计的半径为

$$\hat{y}_i^r = \sqrt{\sum_{j=1}^{p}(\hat{\beta}_j)^2\left(\frac{\partial g_j}{\partial X_j}\right)^2(x_{ij}^r)^2} \qquad (13)$$

于是可以得到估计的区间因变量 $\hat{y}_i = [\underline{\hat{y}}_i, \overline{\hat{y}}_i]$, 其中$\underline{\hat{y}}_i = \hat{y}_i^c - \hat{y}_i^r$, $\overline{\hat{y}}_i = \hat{y}_i^c + \hat{y}_i^r$.

3 模型的评价

回归模型估计得到之后, 常被用于对因变量的预测. 因变量的预测值与实际值之间的误差越小, 就可以在一定程度上说明模型是有效的. 传统的回归分析中, 常用预测值与实际值的均方根误差 (RootMean Square Error)

$RMSE = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(fy_i - y_i)^2}$ 来评价模型预测的优劣, 其中 fy_i 表示因变量的预测值, y_i 表示因变量的实际值, T 表示预测数据个数.

因变量的预测值与实际值的均方根误差定义中 $fy_i - y_i$ 也可以理解成预测值与实际值的距离, 在区间回归分析中, $fy_i = [\underline{fy}_i, \overline{fy}_i]$ 与 $y_i = [\underline{y}_i, \overline{y}_i]$ 均为区间数, 因此, 为研究区间回归模型的评价问题, 需先定义区间数的距离. 下面讨论基于Hausdorff距离的区间数距离.

Hausdorff距离由德国数学家 Felix Hausdorff 于 20世纪初提出, 用来定义 R^P 空间上两个紧集之间的距离[12]. 给定 R^P 上的两点间的一种距离测度 $d(\cdot)$, 设 $x \in R^P$为任意一点, $A \subset R^P$为任一紧集, 定义 x与 A 的距离为

$$d(x, A) = \min_{a\in A} d(x, a) \qquad (14)$$

记 $\mathscr{H}$为 R^P 中所有紧集构成的空间, 定义 $\mathscr{H}$上的 h—距离:

$$h(A, B) = \max_{a\in A} d(a, B),$$

$$h(B, A) = \max_{b\in B} d(b, A),$$

其中 $A \in \mathscr{H}, B \in \mathscr{H}$ (15)

则 $\mathscr{H}$上的 Hausdorff距离定义为

$$H(A, B) = \max\{h(A, B), h(B, A)\} \qquad (16)$$

设 A、B 为任两个区间数, 采用 Minkowski距离作为两点间的一种距离测度 $d(\cdot)$, 则易验证 A、B 间的 Hausdorff距离为:

$$H(A, B) = \max\{|\bar{a} - \bar{b}|, |\underline{a} - \underline{b}|\}$$
$$= |c(A) - c(B)| + |r(A) - r(B)| \qquad (17)$$

其中 $c(X) = \frac{\underline{x} + \bar{x}}{2}$, $r(X) = \frac{\bar{x} - \underline{x}}{2}$分别表示区间数 $X = [\underline{x}, \bar{x}]$ (此处 $X = A$ 或 B) 的中点和半径. 显然, 当两个区间数退化成两个实数时, 公式 (17) 就是这两个实数的绝对值距离.

因此可以定义一个类似于均方根误差的评价模型优劣的指标 —— 基于 Hausdorff距离的均方根误差 $RMSE_H$

$$RMSE_H = \sqrt{\frac{1}{T}\sum_{i=1}^{T}[|fy_i^c - y_i^c| + |fy_i^r - y_i^r|]^2} \qquad (18)$$

其中 fy_i^c、y_i^c 分别表示 fy_i、y_i 的中点, fy_i^r、y_i^r 分别表示 fy_i、y_i 的半径, T 表示预测数据个数.

$RMSE_H$ 度量的是模型预测的绝对误差, 与因变量取值的规模大小有关. 为了消除因变量取值的规模大小的影响, 可以定义度量相对误差的指标:

$$U_H = \frac{\sqrt{\frac{1}{T}\sum_{i=1}^{T}[|fy_i^c - y_i^c| + |fy_i^r - y_i^r|]^2}}{\sqrt{\frac{1}{T}\sum_{i=1}^{T}[|fy_i^c| + |fy_i^r|]^2} + \sqrt{\frac{1}{T}\sum_{i=1}^{T}[|y_i^c| + |y_i^r|]^2}} \qquad (19)$$

U_H 的分子就是 $RMSE_H$, 而分母是 fy_i 和 y_i 分别与 0 的 Hausdorff距离之和, 显然 $0 \leqslant U_H \leqslant 1$ U_H 的取值越大, 说明模型的预测能力越差.

4　应用研究 ——沪深 300指数与中信规模风格指数的回归分析

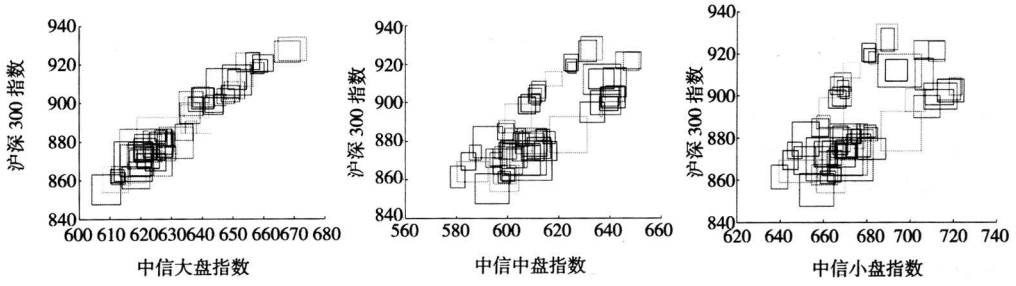
沪深 300指数是沪深证券交易所第一次联合发布的反映 A股市场整体走势的指数, 是由上海和深圳证券市场中选取 300只 A股作为样本编制而成的成份股指数. 中信风格指数以我国证券市场的数据对资产组合定价的多因素模型作了深入的实证研究, 从理论和实证两个方面确定了划分股票风格的指标, 即公司规模 (流通市值) 和净市值比 (净资产除以总市值). 在此基础上, 广泛参考了国际上主流风格指数的编制方法和经验, 先以公司规模将沪深两市股票划分出大、中、小盘, 形成中信大盘、中信中盘及中信小盘三个规模风格指数, 再依净市值比在各规模类中将股票分为价值型或成长型, 形成大盘价值、大盘成长、中盘价值、中盘成长、小盘价值和小盘成长六种风格指数[1 13].

下面采用符号数据分析方法, 研究沪深 300指数与中信大盘、中信中盘及中信小盘三个规模风格指数之间的相关性. 传统方法仅以每日的收盘价为样本, 可能导致大量丢失日内数据信息. 为此, 对日内数据进行打包, 采用最低价和最高价形成区间型符号数据样本, 基于此来研究沪深 300指数与中信规模风格指数之间的相关性.

4 1　区间型符号数据样本的形成

选取 2005年 10月至 12月的沪深 300指数、中信大盘指数、中信中盘指数和中信小盘指数每日的最低价和最高价, 形成区间样本.

图 3给出了沪深 300指数与中信规模风格指数区间样本的矩形图, 其中 (a) 是沪深 300指数与中信大盘指数区间样本的矩形图, (b) 是沪深 300指数与中信中盘指数区间样本的矩形图, (c) 是沪深 300指数与中信小盘指数区间样本的矩形图. 图中显示, 沪深 300指数与三个中信指数之间均为正相关, 且与中信大盘指数的相关程度最为密切, 而与中信小盘指数的相关程度最弱.



(a)　　(b)　　(c)

图 3　沪深 300指数与中信规模风格指数区间样本的矩形图

Fig 3 The rectangle figures of interval samples of CSI300 and CITIC

4 2　回归分析

由于沪深 300指数与中信大盘指数有很强的相关关系, 因此可以考虑建立沪深 300指数与中信大盘指数的区间线性回归模型, 以通过中信大盘指数估计和预测沪深 300指数. 记 $y_i = [\underline{y_i}, \overline{y_i}]$ 为沪深 300指数的区间样本, $x_i = [\underline{x_i}, \overline{x_i}]$ 为中信大盘指数的区间样本, 由 2 1节的方法估计回归系数, 得到估计的区间线性回归方程为

$$[\hat{\underline{y}_i}, \hat{\overline{y}_i}] = 86.463 + 1.265[\underline{x_i}, \overline{x_i}] \qquad (20)$$

$R^2 = 0.961$, 回归方程和各系数显著性检验的 p值均为 0 000, 说明沪深 300指数与中信大盘指数具有显著的线性关系.

在公式 (20) 的基础上, 可由 2 1节的方法, 基于自变量的某一区间值对因变量的区间值进行估计. 以 2005年 10月第一个交易日样本数据为例, 观测值为 $[\underline{y}, \overline{y}] = [906.880, 917.210]$, $[\underline{x}, \overline{x}] = [648.410, 654.450]$. 为估计该交易日的因变量区间值, 将自变量区间中点 $x^c = 651.430$代入公式 (20), 得因变量区间中点的估计值 $\hat{y}^c = 910.522$ 进一步, 由公式 (10) 得因变量区间半径

的估计值 $\hat{y}^{*} = 3.820$ 从而, 因变量估计的区间值为 $[\underline{\hat{y}}, \overline{\hat{y}}] = [\hat{y}^{c} - \hat{y}^{r}, \hat{y}^{c} + \hat{y}^{r}] = [906.702, 914.342]$, 此即为沪深 300指数在该交易日的区间估计值.

4.3 模型评价分析

为了评价模型, 分别建立沪深 300指数与中信大盘指数、中信中盘指数以及中信小盘指数的区间线性回归模型, 并基于样本数据计算评价指标 $RMSE_H$ 和 U_H. 表 1给出了三个模型的评价指标, 中信大盘模型的 $RMSE_H$ 和 U_H 取值均是最小, 说明中信大盘模型的估计、预测能力最好.

表 1 三个区间线性回归模型的评价指标值

Table 1 Evaluation indices of the three regression models

	$RMSE_H$	U_H
中信大盘模型	4.451	0.002
中信中盘模型	14.923	0.017
中信小盘模型	21.061	0.025

5 结 论

基于误差传递公式, 提出了一种针对区间数据的回归分析的新方法, 定义了模型评价指标, 并将该方法应用于我国股票市场的实证研究. 该方法基于区间数据的中点估计回归方程, 并基于区间数据的半径通过误差传递原理来拟合被解释变量, 与基于区间数据的协方差的估计方法等[8] 相比, 具有以下优点:

1) 计算量小, 简单易操作.

2) 估计方法基于实数点, 传统回归分析的显著性检验、拟合优度度量等方法可沿用, 而以前文献中的方法则未涉及此类问题.

3) 不仅适合于线性回归, 而且适合于可线性化的非线性回归分析.

区间回归分析与一般的回归分析相比, 通过样本空间的降维, 更容易从整体上把握样本的属性. 将来可进一步研究区间数据的其它统计方法的求解.

参 考 文 献:

[1]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 Sciences Edition), 2004, 17(2): 40-44. (in Chinese)

[2]Bock H H, Diday E. Analysis of Symbolic Data[M]. New York: Springer-Verlag, 2000.

[3]Billard L, Diday E. From the statistics of data to the statistics of knowledge: Symbolic data analysis[J]. Journal of the American Statistical Association, 2003, 98(462), 470-487.

[4]Billard L, Diday E. Symbolic Data Analysis: Conceptual Statistics and Data Mining[M]. Chichester: John Wiley & Sons, 2006.

[5]Guo J P, Li W H. Integrated Evaluation of Listed Companies by Factor Analysis for Symbolic Data[C]. AMIGE 2008 Proceedings, 2008: 196-199.

[6]Carvalho F D A T, Lechevallier Y. Partitional clustering algorithms for symbolic interval data based on single adaptive distances[J]. Pattern Recognition, 2009, 42(7): 1223-1236.

[7]郭均鹏, 李汶华. 基于经验相关矩阵的区间主成分分析[J]. 管理科学学报, 2008, 11(3), 49-52.
Guo Junpeng, Li Wenhua. Interval PCA based on empirical correlation matrix[J]. Journal of Management Sciences in China, 2008, 11(3): 49-52. (in Chinese)

[8]Billard L, Diday E. Regression analysis for interval-valued data[C] // In Kiers H A L, Rassoon J P, Groenen P J F, etc. Data Analysis, Classification, and Related Methods[C]. Berlin: Springer-Verlag, 2000: 369-374.

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

[10]李金海. 误差理论与测量不确定度评定[M]. 北京: 中国计量出版社, 2003.
Li Jinhai. Error Theory and Measurement of Uncertainty[M]. Beijing: China Metrology Publishing House, 2003. (in Chinese)

[11]郭均鹏, 李汶华. 基于误差理论的区间主成分分析及其应用[J]. 数理统计与管理, 2007, 26(4): 636-640.
Guo Junpeng, Li Wenhua. Principal component analysis based on error theory and its application[J]. Application of Statistics and Management, 2007, 26(4): 636-640. (in Chinese)

[12] Francisco A T, Renata S, Marie C, et al. Adaptive Hausdorff distances and dynamic clustering of symbolic interval data [J]. Pattern Recognition Letters, 2006, 27(3): 167-179.

[13]黄德龙, 杨晓光. 中国证券市场股指收益分布的实证分析[J]. 管理科学学报, 2008, 11(1): 68-77.
Huang Delong, Yang Xiaoguang. Empirical study on distributions of stock index returns in China's securities market[J]. Journal of Management Sciences in China, 2008, 11(1): 68-77. (in Chinese)

Methodology and application of regression analysis of interval-type symbolic data

LI Wen-hua, GUO Jun-peng
School of Management, Tianjin University, Tianjin 300072, China

Abstract The theory of symbolic data analysis(SDA) is introduced first. One type of the most important symbolic data is interval data. The methods of regression analysis of interval-type symbolic data are proposed via error transferring theory. Linear regression analysis and non-linear regression which can be linearized are studied separately. Furthermore, the distance between interval numbers is defined based on Hausdorff distance. And, based upon this, evaluation indices of the regression analysis models are proposed. Finally, application of the correlation between CSI 300 and style indices of Chinese international trust & investment company (CITIC) is carried out. In this application study, the interval-type symbolic data are obtained by data package of daily indices data. The regression analysis models of the interval-type symbolic data are set up, revealing the overall correlation between the CSI 300 and style indices of CITIC. It is concluded that, comparing with the traditional regression analysis of point data, the regression analysis of interval-type symbolic data can not only reduce the dimension of the sample space, but also grasp the inner relations of the variables.

Key words symbolic data; regression analysis; interval numbers; error transferring; index