Tianjin University

Fitting a linear regression model for interval symbolic data considering inner points in the intervals

Junpeng Guo

College of management and economics
Tianjin University, China

Outline

- 1. Introduction
- 2. The Empirical Descriptive Statistics of Interval Symbolic Variable
- 3. Linear Regression Model for Interval Symbolic Data
- 4. Evaluations of the model
- 5. Simulation study
- 6. Conclusions and Future Work

1. Introduction

- **Regarding the regression models for interval symbolic data:**
- **Billard and Diday(2000) developed a methodology, called symbolic method, to fit multiple linear regression equations based on the calculation of covariance and correlation functions for interval symbolic data.**

1. Introduction

- Billard and Diday(2002) extended the work of Billard and Diday(2000) to fit a total of ten possible regression models.

  Of which the first was the symbolic method proposed by Billard and Diday(2000), and the other nine considered taking various specific combinations of the end-points of the interval data and fitting regression models using classical methods.

  It was concluded that the symbolic regression model of Billard and Diday(2000) seems to provide a better fit than the other nine as far as prediction is concerned.

1. Introduction

- Neto and De Carvalho(2008) fitted a linear regression model, called CRM, on the centers and ranges of the interval values.

  Comparison study was made between the CRM method and Center and MinMax method by a Monte Carlo experiment. The results showed that the CRM method performs better than the other two for predicting interval symbolic data.

⋮

1. Introduction

The limitations of the existed methods

- The previous regression methods for interval symbolic data only take into account the special points, such as the endpoints, of the intervals during the calculation, based upon the assumption of uniform distribution of the individuals in the interval.
- However, the assumption is usually not true in many application aspects. Furthermore, in the framework of symbolic data analysis, interval data are usually obtained by summarizing a class of the original sample points (individuals) with classical data. Under this circumstance, the other points within the interval, besides the endpoints of the interval, are also usually important to describe the objects.

1. Introduction

- our methods
- We assume the data has an arbitrary distribution in the interval and all the original individuals will be considered in our analysis.
- For example, let us consider the evaluation of the stocks in a certain stock market. The fact is that we may be interested in the overall behavior of certain blocks such as finance, real estate, oil, instead of the stock individuals.

The sample (object), oil block for instance, might have interval value price, X=[20, 30], which is obtained by the minimum and the maximum price of all the stocks in the oil block.

In our method, we will take into account all the prices of the individual stocks across the interval [20, 30].

Outline

- 1. Introduction
- 2. The Empirical Descriptive Statistics of Interval Symbolic Variable
- 3. Linear Regression Model for Interval Symbolic Data
- 4. Evaluations of the model
- 5. Simulation study
- 6. Conclusions and Future Work

2. The Empirical Descriptive Statistics of Interval Symbolic Variable

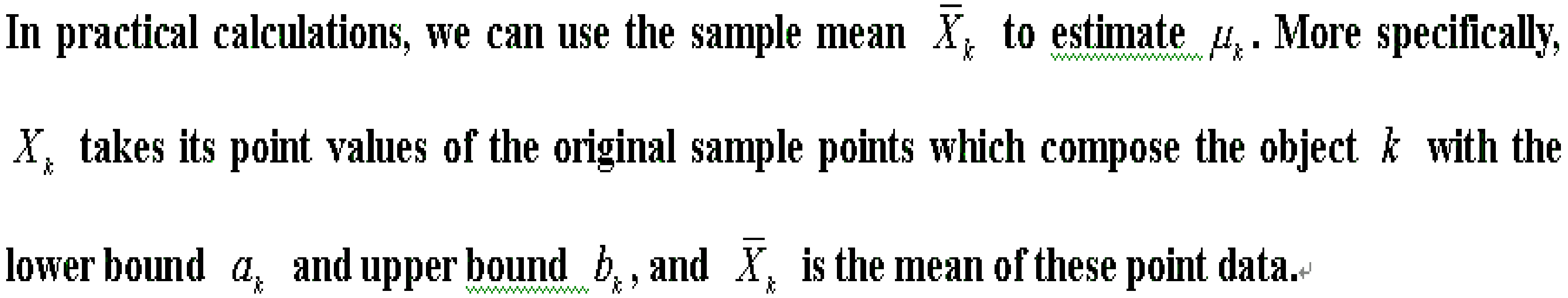
- Let $E = \{1, \cdots, n\}$ be the objects set.
- For the interval symbolic variable *X*, suppose that each object is equally likely to be observed with probability 1/n.
- let random variable $X_k \ (k \in E)$ be a single-valued version of *X*.
- Then, based upon the empirical statistics theory ,we can obtain the empirical descriptive statistics of interval variable.

2. The Empirical Descriptive Statistics of Interval Symbolic Variable

- **the empirical mean of interval variable**

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n}\mu_k$$

where μ_k is the mean of X_k.



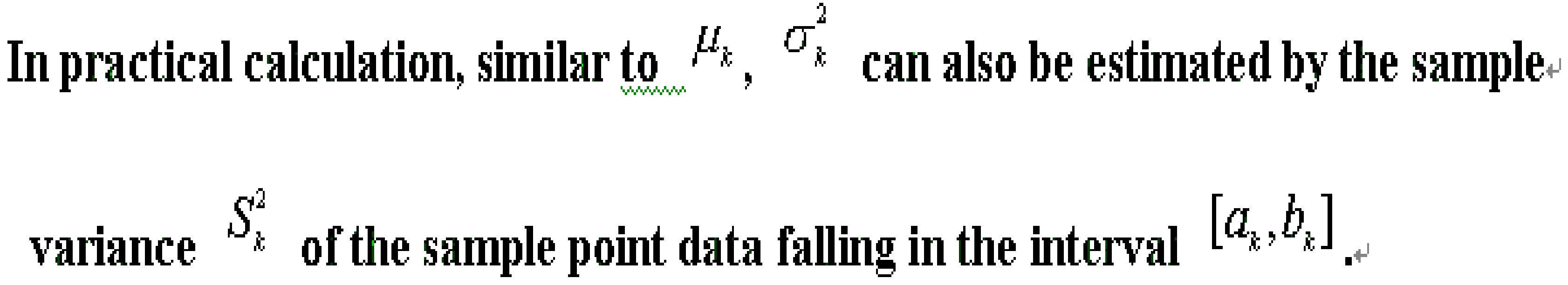
In practical calculations, we can use the sample mean $\bar{X}_k$ to estimate μ_k. More specifically, X_k takes its point values of the original sample points which compose the object k with the lower bound a_k and upper bound b_k, and $\bar{X}_k$ is the mean of these point data.

2. The Empirical Descriptive Statistics of Interval Symbolic Variable

- **the empirical variance of interval variable**

$$S^2=\frac{1}{n}\sum_{k=1}^{n}\left[\sigma_k^2+\left(\bar{X}-\mu_k\right)^2\right]$$

where μ_k is the mean of X_k, and σ_k^2 is the variance of X_k.



In practical calculation, similar to μ_k, σ_k^2 can also be estimated by the sample

variance S_k^2 of the sample point data falling in the interval $[a_k, b_k]$.

2. The Empirical Descriptive Statistics of Interval Symbolic Variable

- **In the same way, we can also obtain the empirical covariance between two interval variables:**

$$S_{12} = \frac{1}{n}\sum_{k=1}^{n}\sigma_{X_{k1},X_{k2}} + \frac{1}{n}\sum_{k=1}^{n}(\mu_{k1}\mu_{k2}) - \frac{1}{n^2}\sum_{k=1}^{n}\mu_{k1}\sum_{k=1}^{n}\mu_{k2}$$

2. The Empirical Descriptive Statistics of Interval Symbolic Variable

Example 1.

k	Object *k*	Classical point data	Interval symbolic data	$\bar{X}_k$	S_k^2
1	C_1	1,2,4,5,6	[1, 6]	3.60	4.30
2	C_2	-2,-1,2,3,6,6,7	[-2,7]	3.00	12.67
3	C_3	-3,-1,0	[-3,0]	-1.33	2.33
				$\bar{X}=1.76$	$S^2=11.26$

$$\bar{X}=\frac{1}{n}\sum_{k=1}^{n}\mu_k \qquad S^2=\frac{1}{n}\sum_{k=1}^{n}\left[\sigma_k^2+\left(\bar{X}-\mu_k\right)^2\right]$$

2. The Empirical Descriptive Statistics of Interval Symbolic Variable

- Note. If X_k is a uniformly distributed variable in the interval $[a_k, b_k]$, the empirical mean and empirical variance of X is simplified as follows:

$$\overline{X} = \frac{1}{2n}\sum_{k=1}^{n}(a_k + b_k)$$

$$S^2 = \frac{1}{3n}\sum_{k=1}^{n}(a_k^2 + a_k b_k + b_k^2) - \frac{1}{4n^2}[\sum_{k=1}^{n}(a_k + b_k)]^2$$

These equations are in accordance with those proposed in some previous work (Bertrand and Goupil, 2000; Billard and Diday, 2000; Billard and Diday, 2003; Billard and Diday, 2006) under the uniform distribution assumption of the interval symbolic data.

Outline

- 1. Introduction
- 2. The Empirical Descriptive Statistics of Interval Symbolic Variable
- 3. Linear Regression Model for Interval Symbolic Data
- 4. Evaluations of the model
- 5. Simulation study
- 6. Conclusions and Future Work

3. Linear Regression Model for Interval Symbolic Data

- For a dependent variable *Y* and *p* independent variables $X_1,\cdots,X_p$, the classical linear regression model is defined by

$$Y = X\beta + \varepsilon$$

where $Y = (Y_1,\cdots,Y_n)^{\mathrm{T}}$ $$X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}$$

The least squares estimators of the parameters are given by

$$\hat{\beta} = (X^{\mathrm{T}} X)^{-1}(X^{\mathrm{T}} Y)$$

3. Linear Regression Model for Interval Symbolic Data

$$\hat{\beta} = (X^{\mathrm{T}} X)^{-1}(X^{\mathrm{T}} Y)$$

(1) To obtain $X^T X$ **, Let** $X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix} = [1 \ \ Z]$

$$X^T X = (1 \quad Z)^T (1 \quad Z) = \begin{pmatrix} n & n\bar{Z} \\ n\bar{Z} & Z^T Z \end{pmatrix}$$

And we can deduce that $Z^T Z = nCov(Z, Z) + n\bar{Z}^T \bar{Z}$

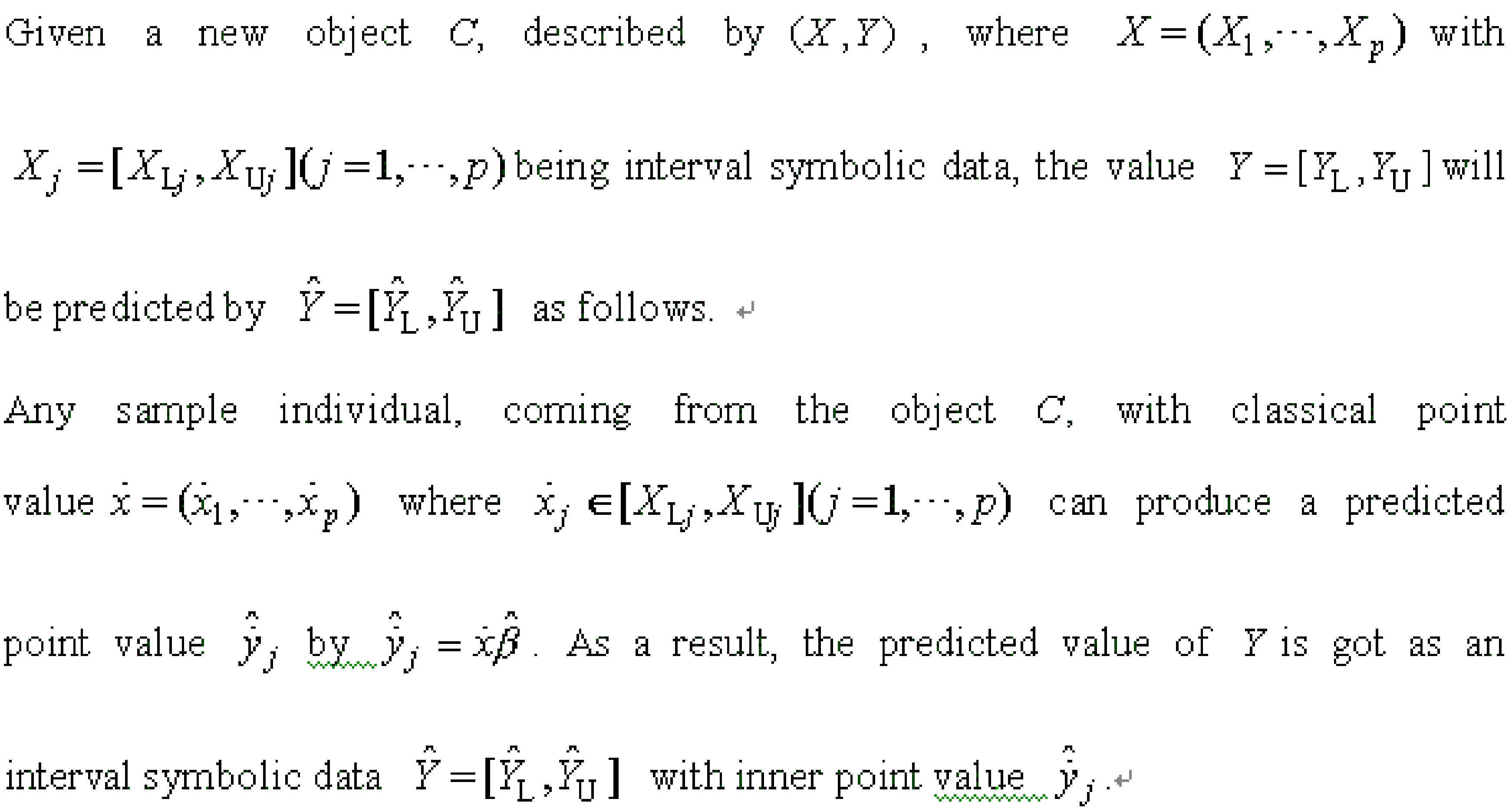
(2) To obtain $X^T Y$

$$X^T X = (1 \quad Z)^T (1 \quad Z) = \begin{pmatrix} n & n\bar{Z} \\ n\bar{Z} & Z^T Z \end{pmatrix}$$

And we can deduce that $Z^T Z = nCov(Z, Z) + n\bar{Z}^T \bar{Z}$

3. Linear Regression Model for Interval Symbolic Data

How to predict using the model.



Given a new object C, described by (X,Y), where $X=(X_1,\cdots,X_p)$ with $X_j=[X_{\mathrm{L}j},X_{\mathrm{U}j}](j=1,\cdots,p)$ being interval symbolic data, the value $Y=[Y_{\mathrm{L}},Y_{\mathrm{U}}]$ will be predicted by $\hat{Y}=[\hat{Y}_{\mathrm{L}},\hat{Y}_{\mathrm{U}}]$ as follows. ↵

Any sample individual, coming from the object C, with classical point value $\dot{x}=(\dot{x}_1,\cdots,\dot{x}_p)$ where $\dot{x}_j\in[X_{\mathrm{L}j},X_{\mathrm{U}j}](j=1,\cdots,p)$ can produce a predicted point value $\hat{\dot{y}}_j$ by $\hat{\dot{y}}_j=\dot{x}\hat{\beta}$. As a result, the predicted value of Y is got as an interval symbolic data $\hat{Y}=[\hat{Y}_{\mathrm{L}},\hat{Y}_{\mathrm{U}}]$ with inner point value $\hat{\dot{y}}_j$. ↵

Outline

- 1. Introduction
- 2. The Empirical Descriptive Statistics of Interval Symbolic Variable
- 3. Linear Regression Model for Interval Symbolic Data
- 4. Evaluations of the model
- 5. Simulation study
- 6. Conclusions and Future Work

4. Evaluations of the model

- **The index for evaluation:**

$$RMSE_S = \sqrt{\frac{1}{T}\sum_{i=1}^{T} D_S(\hat{Y}_i, Y_i)}$$

where $\hat{Y}_i = [\hat{y}_{\mathrm{L}i}, \hat{y}_{\mathrm{U}i}]$ **and** $Y_i = [y_{\mathrm{L}i}, y_{\mathrm{U}i}]$ **are the interval symbolic predicted value and observed value respectively.**

We then defined a new distance measure-----*μσdistance*, for interval symbolic data considering inner points in the intervals, based upon the Hausdorff distance, as follows,

$$D_{\mu\sigma}(A,B) = \left|\overline{A} - \overline{B}\right| + \sqrt{3}\left|S_A - S_B\right|$$

where $\overline{A}$ **and** $\overline{B}$ **are the sample mean and** S_A **and** S_B **the standard deviation of the sample points included in the interval.**

(the Hausdorff distance: $\mathrm{D_H}(A,B) = |\mathrm{c}(A) - \mathrm{c}(B)| + |\mathrm{r}(A) - \mathrm{r}(B)|$ **)**

Outline

- **1. Introduction**
- **2. The Empirical Descriptive Statistics of Interval Symbolic Variable**
- **3. Linear Regression Model for Interval Symbolic Data**
- **4. Evaluations of the model**
- **5. Simulation study**
- **6. Conclusions and Future Work**

5. Simulation study

- We conduct a simulation study to make performance evaluation on our proposed method. Comparisons are made between our method (DSM-Descriptive Statistics based Method) with the other three methods: (a) Center method (CM-fitting the linear regression model to the mid-points of the interval values); (b) MinMax method (MINMAX-fitting the linear regression models to the upper and lower bounds of the interval values); (c) Center and range method (CRM-fitting the linear regression models to the mid-points and ranges of the interval values).

5. Simulation study

- To compare the performances between our method and the previous method which assumes the interval has a uniform distribution, we choose **a uniform and two non-uniform** distributions to generate the point data. (1) One is a symmetric distribution where **normal distribution** is chosen.(2) The other is an asymmetric distribution where χ^2 **distribution** with four degree of freedom is chosen.

- **RESULT:** When the point data follow a non-uniform distribution in the intervals, our method shows the best performance compared to the other methods.

Outline

- 1. Introduction
- 2. The Empirical Descriptive Statistics of Interval Symbolic Variable
- 3. Linear Regression Model for Interval Symbolic Data
- 4. Evaluations of the model
- 5. Simulation study
- 6. Conclusions and Future Work

6. Conclusions and Future Work

- The contribution of our research includes the following points:
- 1) In many practical application cases, individuals are usually non-uniformly distributed in the interval. In our research, the arbitrarily distributed individuals is assumed, which allows a wider application of the proposed method.
- 2) The method presented in this paper does not need the exact form of the distribution function in each interval. Only the original sample data are required.
- 3) The method makes the best of the individual sample information.

6. Conclusions and Future Work

- In our future research, based on the descriptive statistics of interval symbolic data, we can proceed with other data analysis, such as principal component analysis, factor analysis, and so on.

Tianjin University

Thanks!

Fitting a linear regression model for interval symbolic data considering inner points in the intervals

Junpeng Guo