A Bootstrap method for fitting a linear regression model to interval-valued data

by

Muliang Peng

(Under the direction of Jeongyoun Ahn and Cheolwoo Park)

Abstract

We consider interval-valued data that are commonly observed with the advanced technology in the current data collection processes. Interval-valued data are collected as intervals while classical data are formatted as single values. In this thesis, we are particularly interested in regression analysis. This thesis starts with a brief review on the existing methods for the regression analysis of interval-valued data. Then, we propose a new approach to fit a linear regression model to interval-valued data using bootstrap. The proposed method enables one to do statistical inferences concerning regression coefficients where most of the existing methods fail to provide. The proposed and existing methods are applied to the real and simulated data and their performances are compared each other.

Index words: Bootstrap, Interval-valued data, Linear regression, Statistical inference

A Bootstrap method for fitting a linear regression model to interval-valued data

by

Muliang Peng

B.S., Shandong Agriculture University, P.R.China, 1989

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Master of Science

Athens, Georgia

2011

A Bootstrap method for fitting a linear regression model to interval-valued data

by

Muliang Peng

Approved:

Major Professors: Jeongyoun Ahn
Cheolwoo Park

Committee: Lynne Billard
Yehua Li

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2011

Acknowledgments

I would like to express my full appreciation to my major professors, Dr. Jeongyoun Ahn and Dr. Cheolwoo Park, for their outstanding guidance throughout this project. Their enthusiasm, patience, wisdom and countless hours of reading, revising, reflecting and enlightening throughout the overall process inspire me to conduct and complete my research. Without their intelligent direction and steady help the completion of this thesis would not be possible.

I would like to convey my gratitude to Dr. Lynne Billard and Dr. Yehua Li for serving on my thesis committee. I deeply appreciate their comments and directions which have made my thesis quality improved.

I would like to take this unique opportunity to thank all the faculty members and staffs in the Department of Statistics at the University of Georgia for their help throughout my years in the program.

My special thanks go to Dr. Wei Xu and Mr. Kan Bao for their help in R code.

Finally, I would like to thank my family for their love and support to encourage me to finish my program.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the advanced technology, it is common to collect huge sets of data. As a consequence, it becomes important to extract knowledge from large data bases by summarizing these data into new formats. This data process creates a new type of data that might not be properly analyzed by traditional statistical approaches. Symbolic data are one of such examples and they are more complex than the standard ones due to the fact that they contain internal variation and are structured (Diday, 1987).

Xu (2010) described the structure of symbolic data in two-fold. First, a data set is originally structured as symbolic such as intervals, lists and histograms. A common example of this kind of symbolic data is blood pressure measured in an interval range. Because a person's blood pressure varies at different times through a day or from day to day, it is not single-valued but interval-valued. Another example of interval-valued data is weekly household expenditure where expenditure changes from week to week. Second, it is collected as a classical data set at an initial state, but it may become a symbolic data set when it is aggregated. This might be necessary when a huge data set is collected and it is not obvious how to analyze and extract useful information from it. To circumvent the situation, a large data set may need to be reduced into a smaller and more manageable number of groups of interest than its original data, retaining as much interesting information as possible. In a symbolic data set, variables can be quantitative, and also can be categorical. For instance, individual age or height is quantitative variable; cancer variable can be categorical in a data recording different types of cancers. Also, there are different types of symbolic variables including

interval-valued, multi-valued and modal-valued variables. For more details, see Billard and Diday (2007).

In statistics, symbolic data analysis has been introduced as a new approach for extending classical statistical methods to symbolic data (Diday, 1995, Diday et al., 1996, Emilion, 1997, and Diday and Emilion, 1996 and 1998). For more recent developments for different types of symbolic data analysis, see Billard (2011) and Noirhomme-Fraiture and Brito (2011).

In this thesis, we focus on interval-valued data because it is frequently observed among different types of symbolic data and methods developed for them can be extended to other types. Interval-valued data contain ranges, and are not restricted to a single value. Classical data can be considered as special cases of symbolic data, since a single point ($y = a$) in classical data is equivalent to an interval value ($[a, a]$) (Billard and Diday, 2003). Given the important role of interval-valued data in symbolic data analysis, the development of new methods for interval-valued data analysis is demanding.

Among many different statistical analyses, the focus of this thesis lies in regression analysis which refers to any techniques for modeling and analyzing several variables to find the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable change when any one of the independent variables is altered. Regression analysis not only is widely used for prediction and forecasting but also is used to understand which among the independent variables are related to the dependent variable, and explore the forms of these relationships.

There have been several linear regression approaches for interval-valued data and we focus on the following four existing methods in this thesis for comparison purposes. The first work on this problem was proposed by Billard and Diday (2000). Their main idea was to build a classical linear regression model using the center points of intervals, and then the fitted model was used to predict the lower and upper bounds of the interval values of dependent variables. However, they did not take the variations of the observed intervals into

account in the estimation of the parameters of the model (though they did use the ranges for prediction). As an attempt to resolve this problem, Lima Neto et al. (2004) established the center and range method to fit two different linear regression models by using the center points and the ranges of the intervals, and then used the two models to predict the lower and upper bounds of the interval values. To achieve the same goal, Billard and Diday (2007) developed a bivariate center and range method to construct regression models on the center points and ranges of intervals. The main difference from Lima Neto et al.'s approach lies in including both center and range information in a design matrix. However, one issue sometimes encountered from these approaches is that the predicted values of the lower bounds could be bigger than the predicted values of the upper bounds especially when a slope estimate is negative. Xu (2010) proposed a symbolic covariance method to address the two issues - variations within intervals and the switch of lower and upper bounds - based on the exact definition of the symbolic sample covariance suggested by Billard (2007, 2008). We note, however, that statistical inference on regression coefficients has not been studied in the previous works.

In this thesis, we propose a new method for fitting a linear regression model to interval-valued data based on bootstrap resampling. We construct a large number of samples by generating a large number of points within each interval of the observations of the interval-valued data, fit a linear regression model on each sample, calculate the mean regression coefficients, and then use them to produce a final linear regression model. Therefore, the proposed method considers the variability of an interval. Another important characteristic of the proposed method is that it enables one to make inferences concerning the regression coefficients. Using the proposed method, one can conduct the overall model and individual regression coefficients tests, and obtain the standard errors and confidence intervals for the regression coefficients. From these inferences, one can explore the inherent information about population parameters.

Chapter 2 presents literature review of current methods to fit linear regression models on interval-valued data. In Chapter 3, we propose a new bootstrap method to fit a linear regression model to interval-valued data. In Chapter 4, the proposed method is applied to real data and compared with existing methods. A simulation study is done in Chapter 5. Chapter 6 concludes the thesis with summary and future work.

Chapter 2

Review of regression analysis on symbolic interval-valued data

Before proposing a new method for fitting a linear regression model to interval-valued data in the next chapter, we first review a few available and relevant approaches in this chapter. Xu (2010) thoroughly illustrated the four methods, the center method, center and range method, bivariate center and range method, and symbolic covariance method. We summarize them for the sake of completeness of this thesis using the similar settings and notations of Xu (2010).

2.1 The Center Method

The center method (CM) (Billard and Diday, 2000) targets to fit a linear regression model on the mid-points of the interval values.

Let $X_1, \ldots, X_p$ be p independent interval-valued variables, and Y be the dependent interval-valued variable where each observed value $X_{ij} = [a_{ij}, b_{ij}]$ and $Y_i = [c_i, d_i], i = 1, \ldots, n, j = 1, \ldots, p$. Let $X_1^c, \ldots, X_p^c$ be center points of the intervals $X_1, \ldots, X_p$, and Y^c be the center point of the interval Y. From classic linear regression, the model is then

$$Y_i^c = \beta_0 + \beta_1 X_{i1}^c + \beta_2 X_{i2}^c + \cdots + \beta_p X_{ip}^c + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$. This can be written in matrix notation as:

$$\mathbf{Y^c} = \mathbf{X}^c \boldsymbol{\beta}^c + \boldsymbol{\epsilon}^c,$$

where $\mathbf{Y}^c = (Y_1^c, \ldots, Y_n^c)^T, \mathbf{X^c} = (\mathbf{X}_1^c, \ldots, \mathbf{X}_n^c)^T, \mathbf{X}_i^c = (1, X_{i1}^c, \ldots, X_{ip}^c)^T$ for $i = 1, \ldots, n$, $\boldsymbol{\beta}^c = (\beta_0, \beta_1, \ldots, \beta_p)^T, \boldsymbol{\epsilon}^c = (\epsilon_1^c, \ldots, \epsilon_n^c)^T$. Assuming the full rank of $\mathbf{X}^c$, we can show that

the least squares estimate of $\boldsymbol{\beta}^c$ is given by

$$\hat{\boldsymbol{\beta}}^c = ((\mathbf{X}^c)^T\mathbf{X}^c)^{-1}(\mathbf{X}^c)^T\mathbf{Y}^c. \tag{2.1}$$

One can apply equation (2.1) to the lower and upper bounds of the interval values of the independent variables to predict the lower and upper bounds of the dependent variable, respectively. Let a new observation, $\mathbf{X}^0 = (X_1^0, \ldots, X_p^0)$, and $X_j^0 = [a_j, b_j], j = 1, \ldots, p$, $\mathbf{X}_L^0 = (1, a_1, \ldots, a_p)$, and $\mathbf{X}_U^0 = (1, b_1, \ldots, b_p)$. Then, $Y = [Y_L, Y_U]$ can be predicted from $\hat{Y} = [\hat{Y}_L, \hat{Y}_U]$ as follows:

$$\hat{Y}_L = \mathbf{X}_L^0\hat{\boldsymbol{\beta}}^c$$

and

$$\hat{Y}_U = \mathbf{X}_U^0\hat{\boldsymbol{\beta}}^c.$$

However, as Xu (2010) pointed out, this method does not include the information of the ranges of the intervals in prediction. Furthermore, it is not possible to make inferences on the estimated regression coefficients in equation (2.1) because one cannot estimate their variances by this approach.

2.2 Center and Range Method

Lima Neto et al. (2004) improved the CM by considering the ranges of interval-valued in the model. Their approach, called the center and range method (CRM), fits two independent linear regression models on the center points and ranges of the interval-valued data, respectively. The prediction of the intervals of the dependent variable is performed using both the model fitted on the center points of the intervals and the model fitted on the ranges of the intervals. In Section 2.1, the center model based on the center points of the intervals is presented. Thus, we introduce the other model based on the ranges of the intervals only in this section.

Let $X_1^r, \ldots, X_p^r$ be p range of intervals of $X_1, \ldots, X_p$, and Y^r be the range of the interval of Y. Let the observed values of X_j^r be $X_{ij}^r = (b_{ij} - a_{ij})$, and the observed value of Y_i^r be $Y_i^r = (d_i - c_i)$ where $i = 1, \ldots, n, j = 1, \ldots, p$.

The fitted linear regression model on the range of the interval variables is given by

$$\mathbf{Y}^r = \mathbf{X}^r \boldsymbol{\beta}^r + \boldsymbol{\epsilon}^r,$$

where $\mathbf{Y}^r = (Y_1^r, \ldots, Y_n^r)^T$, $\mathbf{X}^r = (\mathbf{X}_1^r, \ldots, \mathbf{X}_n^r)^T$, $\mathbf{X}_i^r = (1, X_{i1}^r, \ldots, X_{ip}^r)^T$, $\boldsymbol{\beta}^r = (\beta_0^r, \beta_1^r, \ldots, \beta_p^r)^T$ and $\boldsymbol{\epsilon}^r = (\epsilon_1^r, \ldots, \epsilon_n^r)^T$. Assuming the full rank of $\mathbf{X}^r$, we can show that the least squares estimate of $\boldsymbol{\beta}^r$ is then

$$\hat{\boldsymbol{\beta}}^r = ((\mathbf{X}^r)^T \mathbf{X}^r)^{-1} (\mathbf{X}^r)^T \mathbf{Y}^r. \tag{2.2}$$

Given a new observation, $\mathbf{X}^0 = (X_1^0, \ldots, X_p^0)$, $\mathbf{X}^{0c} = (X_1^{0c}, \ldots, X_p^{0c})$, $\mathbf{X}^{0r} = (X_1^{0r}, \ldots, X_p^{0r}), X_j^{0c} = (a_j + b_j)/2, X_j^{0r} = (b_j - a_j), j = 1, \ldots, p$, the values of $\hat{Y}^c$ and $\hat{Y}^r$ can be obtained by applying the equations (2.1) and (2.2); therefore, the predicted $\hat{Y} = [\hat{Y}_L, \hat{Y}_U]$ is given by

$$\hat{Y}_L = \hat{Y}^c - \hat{Y}^r/2$$

and

$$\hat{Y}_U = \hat{Y}^c + \hat{Y}^r/2.$$

Xu (2010) pointed out that CRM assumes the independence between center point and range variables, which is not true. Additionally, even though the variations of intervals are explained through the ranges, CRM does not use all the information available in the data. Also, it is not clear how this variation information can be used to account for the variation of the estimated regression coefficients.

2.3 Bivariate Center and Range Method

In the bivariate center and range method (BCRM), introduced by Billard and Diday (2007), the center points and ranges of the interval-valued variables are utilized together to fit two

linear regression models. The two models are expressed as follows:

$$\mathbf{Y}^c = \mathbf{X}^{cr}\boldsymbol{\beta}^c + \boldsymbol{\epsilon}^c$$

and

$$\mathbf{Y}^r = \mathbf{X}^{cr}\boldsymbol{\beta}^r + \boldsymbol{\epsilon}^r,$$

where $\mathbf{Y}^c = (Y_1^c, \ldots, Y_n^c)^T$, $\mathbf{Y}^r = (Y_1^r, \ldots, Y_n^r)^T$; $\mathbf{X}^{cr} = (\mathbf{X}_1^{cr}, \ldots, \mathbf{X}_n^{cr})^T$, $\mathbf{X}_i^{cr} = (1, X_{i1}^c, \ldots, X_{ip}^c, X_{i1}^r, \ldots, X_{ip}^r)^T$, $\boldsymbol{\beta}^c = (\beta_0^c, \beta_1^c, \ldots, \beta_{2p}^c)^T$, $\boldsymbol{\beta}^r = (\beta_0^r, \beta_1^r, \ldots, \beta_{2p}^r)^T$, $\boldsymbol{\epsilon}^c = (\epsilon_1^c, \ldots, \epsilon_n^c)^T$, $\boldsymbol{\epsilon}^r = (\epsilon_1^r, \ldots, \epsilon_n^r)^T$, for $i = 1, \ldots, n$. Note that the number of columns in the design matrix $\mathbf{X}^{cr}$ is $2p+1$ for this model while it was $p+1$ for the CRM. Assuming the full rank of $\mathbf{X}^{cr}$, we can show the least squares estimates of $\boldsymbol{\beta}^c$ and $\boldsymbol{\beta}^r$ are then

$$\hat{\boldsymbol{\beta}}^c = ((\mathbf{X}^{cr})^T\mathbf{X}^{cr})^{-1}(\mathbf{X}^{cr})^T\mathbf{Y}^c \text{ and } \hat{\boldsymbol{\beta}}^r = ((\mathbf{X}^{cr})^T\mathbf{X}^{cr})^{-1}(\mathbf{X}^{cr})^T\mathbf{Y}^r.$$

For a new observation, the value of $Y = [Y_L, Y_U]$ will be predicted from the predicted values of $\hat{Y}^c$ and $\hat{Y}^r$ as follows:

$$\hat{Y}_L = \hat{Y}^c - \hat{Y}^r/2$$

and

$$\hat{Y}_U = \hat{Y}^c + \hat{Y}^r/2.$$

This approach carries some improvements over CRM, but it still does not use all the information available in the data. Also, again it does not provide statistical inference on the estimated regression coefficients.

2.4 Symbolic Covariance Method

Xu (2010) proposed the symbolic covariance method (SCM) to build the least squares estimator of a linear regression model by using the interval-valued variable covariance. All the methods we reviewed above apply classic regression to fit a linear regression model by converting interval-valued data into classical data in which each data point contains only a single

value. Instead, SCM uses the interval-valued data directly to construct a linear regression model by using the exact calculation of covariance of interval-valued variables.

Let us consider the centered model

$$Y - \bar{Y} = \beta_1(X_1 - \bar{X}_1) + \cdots + \beta_p(X_p - \bar{X}_p) + \epsilon,$$

where $\bar{Y}, \bar{X}_j, j = 1, \ldots, p$ are symbolic sample means defined as

$$\bar{X}_j = \frac{1}{2n}\sum_{i=1}^{n}(a_{ij} + b_{ij})$$

for observation $X_{ij} = [a_{ij}, b_{ij}], i = 1, \ldots, n, j = 1, \ldots, p$, and $\bar{Y}$ can be similarly defined (Bertrand and Goupil, 2000).

Billard (2007, 2008) gave the definition of the symbolic covariance between interval-valued variables $X_1([a_{ij}, b_{ij}])$ and $X_2([c_{ij}, d_{ij}])$ as

$$\begin{aligned} Cov(X_1, X_2) &= \tfrac{1}{6n}\textstyle\sum_{i=1}^{n}[2(a_{ij} - \bar{X}_1)(c_{ij} - \bar{X}_2) + (a_{ij} - \bar{X}_1)(d_{ij} - \bar{X}_2) \\ &\quad + (b_{ij} - \bar{X}_1)(c_{ij} - \bar{X}_2) + 2(b_{ij} - \bar{X}_1)(d_{ij} - \bar{X}_2)]. \end{aligned} \tag{2.3}$$

The least squares estimate of $\boldsymbol{\beta}$ in SCM is then

$$\hat{\boldsymbol{\beta}} = ((\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{X} - \bar{\mathbf{X}}))^{-1}(\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{Y} - \bar{\mathbf{Y}}).$$

By means of constructing $(\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{X} - \bar{\mathbf{X}})$ and $(\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{Y} - \bar{\mathbf{Y}})$ in matrix notation, two equations are obtained as follows:

$$(\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{X} - \bar{\mathbf{X}}) = (n \times Cov(X_{j1}, X_{j2}))_{p\times p}$$

and

$$(\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{Y} - \bar{\mathbf{Y}}) = (n \times Cov(X_j, Y))_{p\times 1},$$

where $Cov(X_{j1}, X_{j2})$ is the covariance between X_{j1} and $X_{j2}, Cov(X_j, Y)$ is the covariance between X_j and Y, for $j, j_1, j_2 = 1, \ldots, p$. The estimates of the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ can be finally given by

$$\hat{\boldsymbol{\beta}} = (n \times Cov(X_{j1}, X_{j2}))^{-1}_{p\times p} \times (n \times Cov(X_j, Y))_{p\times 1}. \tag{2.4}$$

Given a new observation $\mathbf{X}^0 = (X_1^0, \ldots, X_p^0)$, according to the prediction approach proposed by other existing methods, prediction of $Y = [Y_L, Y_U]$ is

$$\hat{Y}_L = \mathbf{X}_L^0 \hat{\boldsymbol{\beta}} \text{ and } \hat{Y}_U = \mathbf{X}_U^0 \hat{\boldsymbol{\beta}},$$

where $\mathbf{X}_L^0 = (1, X_{L1}^0, \ldots, X_{Lp}^0), \mathbf{X}_U^0 = (1, X_{U1}^0, \ldots, X_{Up}^0)$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T$. This definition, however, is not complete because when some regression coefficients are negative, the predicted lower bound $\hat{Y}_L$ could be bigger than the predicted upper bound $\hat{Y}_U$. In order to solve this problem, Lima Neto et al. (2005) and Lima Neto and Carvalho (2010) developed constrained methods where the model is fitted under the condition that all the parameters must be positive. However, it is clear that this constraint cannot be met in some circumstances (Xu, 2010). Instead, SCM suggests to make the prediction by using the following equations:

$$\hat{Y}_L = \min(\mathbf{X}_L^0 \hat{\boldsymbol{\beta}}, \quad \mathbf{X}_U^0 \hat{\boldsymbol{\beta}}) \text{ and } \hat{Y}_U = \max(\mathbf{X}_L^0 \hat{\boldsymbol{\beta}}, \quad \mathbf{X}_U^0 \hat{\boldsymbol{\beta}}).$$

We note that the variance of $\hat{\boldsymbol{\beta}}$ in equation (2.4) can be derived, which will lead to statistical inferences on $\boldsymbol{\beta}$, but a relevant study has not been done yet.

Chapter 3

Bootstrap Method for Interval-valued Data

In this chapter, we propose the bootstrap method (BM) to fit a linear regression model on interval-valued data, which is capable of making some statistical inferences on the regression coefficients. We first review some background on bootstrap in Section 3.1, and then introduce the proposed algorithm and explain inference procedures in Section 3.2.

3.1 Bootstrap basics

3.1.1 Introduction

Bootstrap (Efron, 1979) is a modern and computer-intensive method used frequently in applied statistics. It is a general approach to statistical inference based on building a sampling distribution for a statistic by resampling from the data.

The bootstrap is a method to derive properties (standard errors, confidence intervals and critical values) of the sampling distribution of estimators by measuring those properties when sampling from an approximating distribution. However, instead of fully specifying the data generating process, we use information from the sample. One standard choice for an approximating distribution is the empirical distribution of the observed data. In the case where a set of observations can be assumed to be independent and identical from a population distribution, this can be implemented by constructing a number of resamples of the observed data set (and of equal size to that of the observed data set), each of which is obtained by randomly sampling with replacement from the original data set. It may also be used for constructing hypothesis tests. It is often used as an alternative to inference based on parametric assumptions when those assumptions are in doubt, or where parametric inference

is impossible or requires very complicated formulas for the calculation of standard errors. More detailed introduction about the bootstrap method is given in Efron and Tibshirani (1993).

3.1.2 Resampling

There are several different resampling methods which are related to the several forms of the bootstrap, for instance, nonparametric bootstrap, cross-validation, jackknifing, permutation test, and randomizing test. Here we briefly introduce nonparametric bootstrap.

Suppose that there is an independent sample $\mathbf{S} = \{X_1, X_2, \ldots, X_n\}$ drawn from a large population $\mathbf{P} = \{x_1, x_2, \ldots, x_N\}$. Suppose that we are interested in some statistic as an estimate of the corresponding population parameter θ. Let k denote the number of bootstrap replications, that is the number of bootstrap samples we select. The bootstrap procedure can be generalized as follows (Fox, 2002):

1. We proceed to draw a random sample $\mathbf{S}_1^* = \{X_{11}^*, X_{12}^*, \ldots, X_{1n}^*\}$ of size n from the elements in sample $\mathbf{S}$ with replacement. Thus, the observed sample data $\mathbf{S}$ is treated as a "substitute" for the population $\mathbf{P}$; that is, we select each element X_j of $\mathbf{S}$ for the bootstrap sample with the same probability $1/n$, just like the original selection of the sample $\mathbf{S}$ from the population $\mathbf{P}$.

2. We repeat the resampling procedure a large number of times, producing k bootstrap samples. Let $\mathbf{S}_b^* = \{X_{b1}^*, X_{b2}^*, \ldots, X_{bn}^*\}$ denote the bth bootstrap sample, where $b = 1, 2, \ldots, k$.

3. For each bootstrap sample, calculate the estimate $\hat{\theta}_b^*$ of parameter θ of interest.

4. Use the distribution of the $\hat{\theta}_b^*$ to estimate the properties of the sampling distribution of $\hat{\theta}$. For example, we can compute the estimated standard deviation of the k bootstrap $\hat{\theta}_b^*$:

$$SE(\hat{\theta}^*) = \sqrt{\frac{\sum_{b=1}^{k}(\hat{\theta}_b^* - \bar{\theta}^*)^2}{k-1}},$$

where

$$\bar{\theta}^* = \frac{\sum_{b=1}^{k} \hat{\theta}_b^*}{k}.$$

3.1.3 Confidence intervals based on normal theory

There are several approaches to constructing bootstrap confidence intervals, such as normal-theory intervals and percentile intervals. We briefly introduce the normal-theory intervals here. Most statistics including sample mean, are asymptotically normally distributed. In sufficiently large bootstrap samples, we can construct a $100(1-\alpha)\%$ confidence interval by using the bootstrap estimate of the standard error of $\hat{\theta}$, along with the normal distribution. Therefore, the bootstrap interval for θ is given as

$$\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta}^*),$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$th quantile of the standard normal distribution.

3.1.4 Bootstrap regression

By extending the procedure of the previous subsection, we can easily bootstrap a regression model. Fox (2002) introduced two general methods to bootstrap a regression model: the observations of the model variables are resampled randomly, or the model matrix $\mathbf{X}$ is fixed.

Suppose we want to fit a regression model with random predictor variables $X_1, \ldots, X_p$, and response variable Y. The sample has n observations $\mathbf{Z}_i = (Y_i, X_{i1}, \ldots, X_{ip}), i = 1, \ldots, n$. In the resampling, the observations $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are resampled to produce k bootstrap samples $\mathbf{Z}_{b1}^*, \ldots, \mathbf{Z}_{bn}^*, b = 1, \ldots, k$. For each of the bootstrap samples, we calculate the regression estimates, and obtain k sets of bootstrap regression coefficients $\hat{\boldsymbol{\beta}}_b^* = (\hat{\beta}_{b0}^*, \hat{\beta}_{b1}^*, \ldots, \hat{\beta}_{bp}^*)^T, b = 1, \ldots, k$. We can compute the standard error or confidence intervals of the regression estimates by applying the method described in the previous subsections.

In the other method of resampling, the predictor $\mathbf{X} = (X_1, \ldots, X_p)$ is fixed when we generate bootstrap samples. The procedure of the bootstrap resampling is as follows:

1. Construct the regression model for the observed sample, and compute the fitted values of response variables Y and the residual for each observation:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip},\ E_i = Y_i - \hat{Y}_i,$$

   where $i = 1, \ldots, n$.

2. Generate k bootstrap samples of the residuals, $\mathbf{E}_b^* = (E_{b1}^*, \ldots, E_{bn}^*)^T$, attach the bootstrap errors to each $\hat{Y}_i$, and produce the bootstrap $\mathbf{Y}_b^* = (Y_{b1}^*, \ldots, Y_{bn}^*)^T, b = 1, \ldots, k$.

3. Regress the bootstrap $\mathbf{Y}_b^*$ on the fixed $\mathbf{X}$ to obtain the bootstrap estimates of the regression coefficients, $\hat{\boldsymbol{\beta}}_b^* = (\hat{\beta}_{b0}^*, \hat{\beta}_{b1}^*, \ldots, \hat{\beta}_{bp}^*)^T$.

4. Use the bootstrap coefficients $\hat{\boldsymbol{\beta}}_b^*$ to calculate the bootstrap standard errors and confidence intervals for the regression coefficients.

3.2 Bootstrap regression to interval-valued data

In this section, we illustrate how to apply bootstrap resampling and bootstrap regression to interval-valued data.

3.2.1 Bootstrap data sets generated from interval-valued data

Suppose we have an interval-valued data set with p predictor variables $X_1, \ldots, X_p$ where $X_j = [a_{ij}, b_{ij}], i = 1, \ldots, n, j = 1, \ldots, p$, and response variable $Y = [c_i, d_i]$. Then, the symbolic interval-valued independent variables can be expressed in matrix notation as

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2P} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{pmatrix} = \begin{pmatrix} [a_{11}, b_{11}] & [a_{12}, b_{12}] & \ldots & [a_{1p}, b_{1p}] \\ [a_{21}, b_{21}] & [a_{22}, b_{22}] & \ldots & [a_{2p}, b_{2p}] \\ \vdots & \vdots & \vdots & \vdots \\ [a_{n1}, b_{n1}] & [a_{n2}, b_{n2}] & \ldots & [a_{np}, b_{np}] \end{pmatrix},$$

where $a_{ij} \leq b_{ij}, i = 1, 2, \ldots, n, j = 1, 2, \ldots, p$. The response variable $\mathbf{Y}$ can be expressed in matrix notation as

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} [c_1, d_1] \\ [c_2, d_2] \\ \vdots \\ [c_n, d_n] \end{pmatrix},$$

where $c_i \leq d_i, i = 1, 2, \ldots, n$. Similarly, the jth independent variable X_j can be written as

$$\mathbf{X}_j = \begin{pmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{nj} \end{pmatrix} = \begin{pmatrix} [a_{1j}, b_{1j}] \\ [a_{2j}, b_{2j}] \\ \vdots \\ [a_{nj}, b_{nj}] \end{pmatrix}.$$

Assuming the points within the interval $[a_{ij}, b_{ij}]$ are uniformly distributed, we can randomly generate a number of points that are uniformly distributed within each interval $[a_{ij}, b_{ij}]$, and then construct bootstrap data sets. The procedure for producing the bootstrap data sets is as follows:

1. By using the unform distribution, randomly generate k points within each interval $[a_{ij}, b_{ij}]$ of the jth predictor variable X_j, and then get the bootstrap variable X_j^*, which can be written in matrix notation as follows:

$$\mathbf{X}_j^* = \begin{pmatrix} X_{1j}^* \\ X_{2j}^* \\ \vdots \\ X_{nj}^* \end{pmatrix} = \begin{pmatrix} X_{1j}^{*1} & X_{1j}^{*2} & \ldots & X_{1j}^{*k} \\ X_{2j}^{*1} & X_{2j}^{*2} & \ldots & X_{2j}^{*k} \\ \vdots & \vdots & \vdots & \vdots \\ X_{nj}^{*1} & X_{nj}^{*2} & \ldots & X_{nj}^{*k} \end{pmatrix},$$

   where $j = 1, \ldots, p$.

2. Similarly, generate k points within each interval $[c_i, d_i]$ of response variable Y_j, the bootstrap variable $\mathbf{Y}^*$ is obtained as follows:

$$\mathbf{Y}^* = \begin{pmatrix} Y_1^* \\ Y_2^* \\ \vdots \\ Y_n^* \end{pmatrix} = \begin{pmatrix} Y_1^{*1} & Y_1^{*2} & \dots & Y_1^{*k} \\ Y_2^{*1} & Y_2^{*2} & \dots & Y_2^{*k} \\ \vdots & \vdots & \vdots & \vdots \\ Y_n^{*1} & Y_n^{*2} & \dots & Y_n^{*k} \end{pmatrix}.$$

3. From the previous two steps, generate p predictor variables $\mathbf{X}_1^*, \dots, \mathbf{X}_p^*$ and response variable $\mathbf{Y}^*$. The *bth* column from the matrices of variables $\mathbf{Y}^*, \mathbf{X}_1^*, \dots, \mathbf{X}_p^*$, respectively, is selected to construct the *bth* bootstrap data set where $b = 1, \dots, k$. Thus, the *bth* bootstrap data sets can be denoted by

$$D^{*b} = \begin{pmatrix} Y_1^{*b} & X_{11}^{*b} & X_{12}^{*b} & \dots & X_{1p}^{*b} \\ Y_2^{*b} & X_{21}^{*b} & X_{22}^{*b} & \dots & X_{2p}^{*b} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Y_n^{*b} & X_{n1}^{*b} & X_{n2}^{*b} & \dots & X_{np}^{*b} \end{pmatrix}.$$

Therefore, k classical bootstrap data sets with response variable $\mathbf{Y}^*$ and predictor variables $\mathbf{X}_1^*, \dots, \mathbf{X}_p^*$ are generated.

3.2.2 Bootstrap linear regression model

We can fit a linear regression model for each of the bootstrap data sets $D^{*1}, \dots, D^{*k}$. Thus, the bootstrap linear models are:

$$\hat{Y}^{*1} = \hat{\beta}_0^{*1} + \hat{\beta}_1^{*1} X_1^{*1} + \hat{\beta}_2^{*1} X_2^{*1} + \cdots + \hat{\beta}_p^{*1} X_p^{*1},$$

$$\hat{Y}^{*2} = \hat{\beta}_0^{*2} + \hat{\beta}_1^{*2} X_1^{*2} + \hat{\beta}_2^{*2} X_2^{*2} + \cdots + \hat{\beta}_p^{*2} X_p^{*2},$$

$$\vdots$$

$$\hat{Y}^{*k} = \hat{\beta}_0^{*k} + \hat{\beta}_1^{*k} X_1^{*k} + \hat{\beta}_2^{*k} X_2^{*k} + \cdots + \hat{\beta}_p^{*k} X_p^{*k}.$$

For the *bth* bootstrap linear regression model, if it has full rank $p+1 \leq n$, the least squares estimate of $\hat{\boldsymbol{\beta}}^{*b}$ can be calculated as

$$\hat{\boldsymbol{\beta}}^{*b} = ((\mathbf{X}^{*b})'\mathbf{X}^{*b})^{-1}(\mathbf{X}^{*b})'\mathbf{Y}^{*b}, \tag{3.1}$$

where $\mathbf{Y}^{*b} = (Y_1^{*b}, \ldots, Y_n^{*b})^T, \mathbf{X}^{*b} = (\mathbf{X}_1^{*b}, \ldots, \mathbf{X}_n^{*b})^T, \mathbf{X}_i^{*b} = (1, X_{i1}^{*b}, \ldots, X_{ip}^{*b})^T, \hat{\boldsymbol{\beta}}^{*b} = (\hat{\beta}_0^{*b}, \hat{\beta}_1^{*b}, \ldots, \hat{\beta}_p^{*b})^T$, for $i = 1, \ldots, n, b = 1, \ldots, k$.

By using the equation (3.1), k values of $\hat{\beta}_0^*, \hat{\beta}_1^*, \ldots, \hat{\beta}_p^*$ are produced respectively, and the mean of each regression coefficient is obtained. Then the linear regression model on interval-valued data, through the bootstrap method, is fitted as

$$\hat{Y}^* = \bar{\hat{\beta}}_0^* + \bar{\hat{\beta}}_1^* X_1^* + \bar{\hat{\beta}}_2^* X_2^* + \cdots + \bar{\hat{\beta}}_p^* X_p^*,$$

where

$$\bar{\hat{\beta}}_0^* = \frac{\sum_{b=1}^k \hat{\beta}_0^{*b}}{k} \text{ and } \bar{\hat{\beta}}_j^* = \frac{\sum_{b=1}^k \hat{\beta}_j^{*b}}{k},$$

for $j = 1, \ldots, p$.

Now, we can proceed with prediction through the fitted linear regression model to obtain the estimated lower bound $\hat{Y}_L$ and upper bound $\hat{Y}_U$ of an interval-valued observation.

Suppose $\mathbf{X}^0 = (X_1^0, \ldots, X_p^0)$ is a given observation where $X_j^0 = [a_j^0, b_j^0]$ with $j = 1, \ldots, p$. Using the method proposed by Xu (2010), we have that the predicted interval $\hat{Y} = [\hat{Y}_L, \hat{Y}_U]$ of Y is computed as

$$\hat{Y}_L = \min(\mathbf{X}_L^0 \bar{\hat{\boldsymbol{\beta}}}^*,\ \mathbf{X}_U^0 \bar{\hat{\boldsymbol{\beta}}}^*) \ \text{ and } \ \hat{Y}_U = \max(\mathbf{X}_L^0 \bar{\hat{\boldsymbol{\beta}}}^*,\ \mathbf{X}_U^0 \bar{\hat{\boldsymbol{\beta}}}^*),$$

where $\bar{\hat{\boldsymbol{\beta}}}^* = (\bar{\hat{\beta}}_0^*, \bar{\hat{\beta}}_1^*, \ldots, \bar{\hat{\beta}}_p^*)^T, \mathbf{X}_L^0 = (1, a_1^0, \ldots, a_p^0)$ and $\mathbf{X}_U^0 = (1, b_1^0, \ldots, b_p^0)$.

3.2.3 Confidence intervals and hypothesis tests

First, we construct an overall model significance F-test based on the bootstrap method (BM), $H_0 : \beta_1 = \cdots = \beta_p = 0$. The procedure is described as follows:

1. Calculate the F-statistic for the k bootstrap linear regression models in Subsection 3.2.2. The bth F-statistic is given by

$$F^{*b} = \frac{SS^{*b}_{reg}/p}{RSS^{*b}/(n-p-1)},$$

where $RSS^{*b} = \sum_{i=1}^{n}(Y_i^{*b} - \hat{Y}_i^{*b})^2$ and $SS^{*b}_{reg} = \sum_{i=1}^{n}(\hat{Y}_i^{*b} - \bar{Y}^{*b})^2$.

2. Calculate the p-value based on F^{*b} value. The bth p-value is $P(F > F^{*b})$ where F has an F distribution with p degrees of freedom for the numerator and $(n-p-1)$ degrees of freedom for the denominator.

3. Calculate the number N of times that p-values are bigger than $\alpha = 0.05$.

4. The p-value for the overall model significance test is equal to N/k.

Second, we present a method of constructing confidence intervals for the bootstrap coefficient estimates. In Subsection 3.2.2, we have fitted k bootstrap linear regression models based on the k bootstrap samples, and obtained the bootstrap regression coefficient estimates $\bar{\hat{\boldsymbol{\beta}}}^* = (\bar{\hat{\beta}}_1^*, \ldots, \bar{\hat{\beta}}_p^*)$. Note that in Subsection 3.1.3, we introduced the confidence interval for bootstrap estimates.

The standard error of $\bar{\hat{\beta}}_j^*$ is calculated as

$$SE(\bar{\hat{\beta}}_j^*) = \sqrt{\frac{\sum_{b=1}^{k}(\hat{\beta}_j^{*b} - \bar{\hat{\beta}}_j^*)^2}{k-1}},$$

and then the $100(1-\alpha)\%$ confidence interval for the regression coefficients is constructed as

$$\bar{\hat{\beta}}_j^* \pm z_{\alpha/2} SE(\bar{\hat{\beta}}_j^*).$$

This approach will work well if the distribution of the bootstrap coefficient estimates is normally distributed. We can examine the normality of the estimates using histograms and Q-Q plots.

Third, in addition to providing standard errors and confidence intervals, the bootstrap estimates can also be applied to conduct statistical hypothesis tests for the regression coefficients estimate, that is, $H_0 : \beta_j = 0$ for $j = 1, \ldots, p$.

The test statistics can be calculated as

$$Z_j = \frac{\bar{\hat{\beta}}_j^*}{SE(\bar{\hat{\beta}}_j^*)}$$

and Z_j approximately follows a standard normal distribution under the null hypothesis. We can compare the obtained statistic values with the quantile of $z_{\alpha/2}$.

A p-value is a measure of how much evidence we have against the null hypothesis. The smaller the p-value, the more evidence we have against the null hypothesis. For individual regression coefficients test, $H_0 : \beta_j = 0$, we obtain the Z statistic for each of the estimated regression coefficients $\beta_j, j = 1, \ldots, p$, and the p-value is $P(|Z| \geq Z_j)$ where $Z \sim N(0, 1)$ for a two-sided test.

Chapter 4

application

In this chapter, we illustrate the utilization of the bootstrap method (BM) with two symbolic interval-valued data sets. We apply the proposed and other existing methods to the data sets and compare the results. These two data sets were also analyzed by Xu (2010) and we present our results in a similar fashion for direct and easy comparison.

4.1 bats species data set

4.1.1 data

The bats species data set in Table 4.1, taken from Xu (2010), is a naturally occurring interval-valued data which records physical measurements of 21 different species of bats. In the data set, there are four random interval-valued variables, X_1 = head size, X_2 = tail length, X_3 = forearm length and Y = weight.

4.1.2 the linear model and prediction

By using BM described in Section 3.2, the fitted linear regression model for the bats species data set is given by

$$\hat{Y} = -25.941 + 0.490X_1 - 0.147X_2 + 0.448X_3. \tag{4.1}$$

The residuals or the fitting errors for each observation in the data are obtained by using the fitted linear function to compute the distinction between the observed response variable and the estimated response variable. The residuals are calculated as follows:

$$Y_L^{residual} = Y_L - \hat{Y}_L,\ Y_U^{residual} = Y_U - \hat{Y}_U. \tag{4.2}$$

Table 4.1: Bats Species Data

i	Species	Weight	Head	Tail	Forearm
1	PIPC	[3, 8]	[33, 52]	[26, 33]	[27, 32]
2	PHE	[4, 10]	[35, 43]	[24, 30]	[31, 41]
3	MOUS	[4, 7]	[38, 50]	[30, 40]	[32, 37]
4	PIPS	[7, 8]	[43, 48]	[34, 39]	[31, 38]
5	PIPN	[6, 9]	[44, 38]	[34, 44]	[31, 36]
6	MDAUB	[7, 11]	[41, 51]	[30, 39]	[33, 41]
7	MNAT	[5, 10]	[42, 50]	[32, 43]	[36, 42]
8	MDEC	[1, 10]	[40, 45]	[39, 44]	[36, 42]
9	MGP	[8, 12]	[45, 53]	[35, 38]	[39, 44]
10	OCOM	[5, 10]	[41, 51]	[34, 50]	[34, 50]
11	MBEC	[7, 12]	[46, 53]	[34, 44]	[39, 44]
12	SBOR	[8, 13]	[48, 54]	[38, 47]	[37, 42]
13	BARB	[6, 9]	[44, 58]	[41, 54]	[35, 41]
14	OGRIS	[6, 10]	[47, 53]	[43, 53]	[37, 41]
15	SBIC	[12, 14]	[50, 63]	[40, 45]	[40, 47]
16	FCHEV	[13, 34]	[50, 69]	[30, 43]	[51, 61]
17	MSCH	[8, 16]	[52, 60]	[50, 60]	[42, 48]
18	SCOM	[17, 35]	[62, 80]	[46, 57]	[48, 56]
19	NOCT	[15, 40]	[69, 82]	[41, 59]	[45, 55]
20	GMUR	[18, 45]	[65, 80]	[48, 60]	[55, 68]
21	MGES	[20, 50]	[82, 87]	[46, 57]	[58, 63]

We can calculate the estimated weight for each bat species using equation (4.1), and then calculate the residuals of weight of 21 bats species using equation (4.2). The observed and predicted values of bats weight are shown in Table 4.5, and residuals are shown in Table 4.6

4.1.3 INFERENCES ON THE REGRESSION COEFFICIENTS

Using BM, we obtain the statistical inferences concerning the regression coefficients for the bats species example. The results are shown in Table 4.2.

Table 4.2: Inferences Concerning the Regression Coefficients (Bats Species)

		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Z-test	Mean	0.490	-0.147	0.447
	SE	0.2142	0.1526	0.2631
	Z-statistic	2.2875	-0.9633	1.6990
	p-value	0.02218	0.33418	0.08879
	95% Confidence interval	(0.0702, 0.9098)	(-0.4461, 0.1521)	(-0.0687, 0.9627)
F-test	p-value	0.00000	(df1,df2)	(3, 17)

In this table, we can find that the p-value is 0 for the overall F-test, and thus the null hypothesis, $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, is rejected. The z statistics for the regression coefficients of head, tail and forearm, are $2.2875, -0.9633, 1.6990$, respectively. If the critical value for rejection is $z_{0.05/2} = 1.96$, we reject the hypothesis for the head coefficient, but fail to reject the hypotheses for tail and forearm coefficients. If the critical value is $z_{0.1/2} = 1.64$, we reject the hypotheses for head and forearm coefficients, but fail to reject the hypothesis for tail coefficient. We can also find that the head coefficient has the smallest p-value (0.02218), and the tail coefficient has the largest p-value (0.33418). From these inference results, we can conclude that the head has the most significant influence on the weight, and the tail has the least influence on the weight.

In order to investigate the bootstrap distributions, the histograms and Q-Q plots of the estimated regression coefficients of the predictors head, tail and forearm are shown in Figure

4.1. In these figures, we can find that head, tail and forearm coefficients are all reasonably normally distributed.

4.2 BLOOD PRESSURE DATA SET

4.2.1 DATA

We illustrate the application of BM to the second example, the blood pressure data set (Table 4.3) taken from Billard and Diday (2000). The blood pressure data set contains 11 observations with three variables Pulse Rate, Systolic Pressure and Diastolic Pressure. Each value of the data is an interval since pulse rates and blood pressure values fluctuate considerably. Suppose X_1 = Systolic Pressure, X_2 = Diastolic Pressure and Y = Pulse Rate.

Table 4.3: Blood Pressure Data

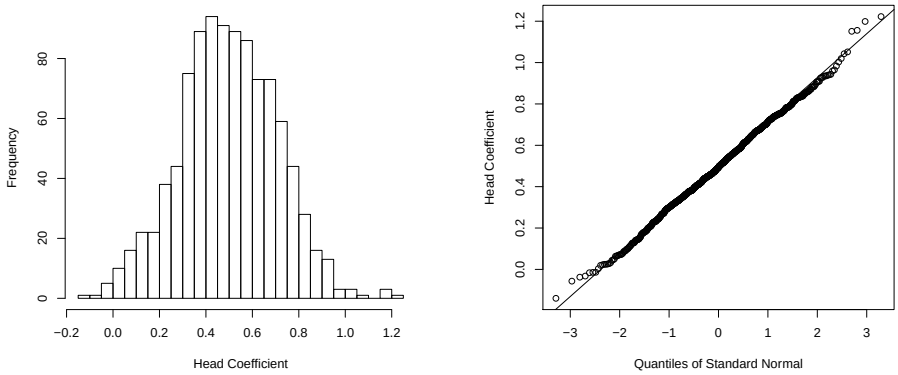
i	Pulse Rate	Systolic Pressure	Diastolic Pressure
1	[44, 68]	[90, 100]	[50, 70]
2	[60, 72]	[90, 130]	[70, 90]
3	[56, 90]	[140, 180]	[90, 100]
4	[70, 112]	[110, 142]	[80. 108]
5	[54, 72]	[90, 100]	[50, 70]
6	[70, 100]	[130, 160]	[80, 110]
7	[72, 100]	[130, 160]	[76, 90]
8	[76, 98]	[110, 190]	[70, 110]
9	[86, 96]	[138, 180]	[90, 110]
10	[86, 100]	[110, 150]	[78, 100]
11	[63, 75]	[60, 100]	[140, 150]

4.2.2 MODEL AND PREDICTION

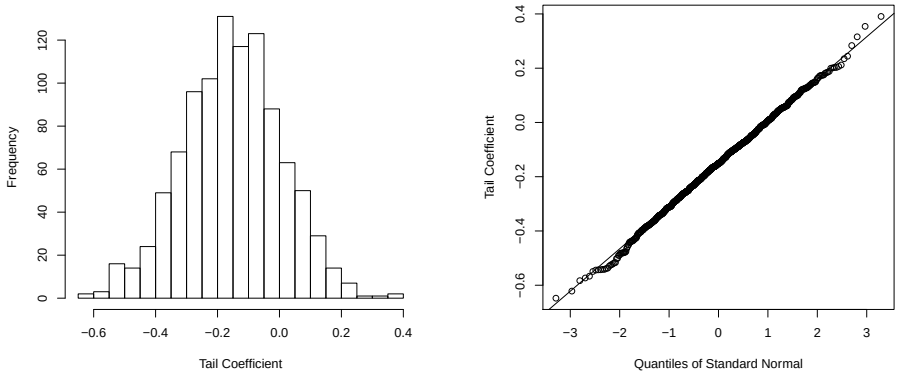
Using BM, the fitted linear regression model for the blood pressure data set is

$$\hat{Y} = 27.301 + 0.291X_1 + 0.155X_2. \tag{4.3}$$

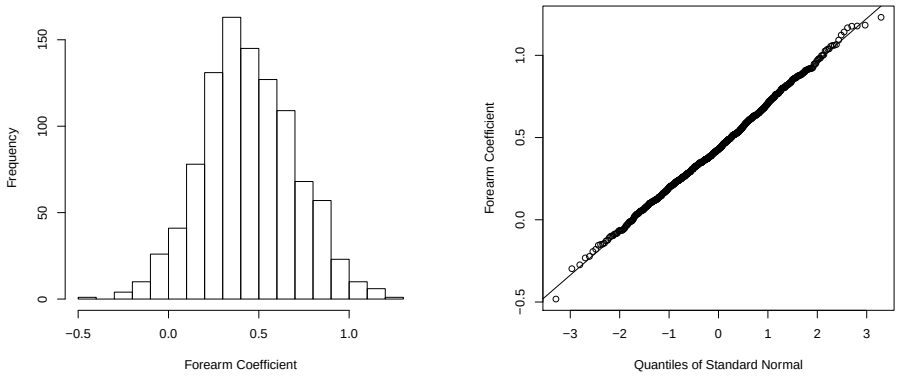
The estimated Pulse Rate for each observation can be calculated using the equation (4.3), and then residual can be calculated using the equation (4.2). The observed and the fitted values of Pulse Rate are shown in Table 4.8, and the residuals are shown in Table 4.9.



(a) Histogram for Head (b) Q-Q plot for Head



(c) Histogram for Tail (d) Q-Q plot for Tail



(e) Histogram for Forearm (f) Q-Q plot for Forearm

Figure 4.1: Histograms and Q-Q plots for the 1000 bootstrap replications of the head, tail and forearm coefficients in the Bats Species data

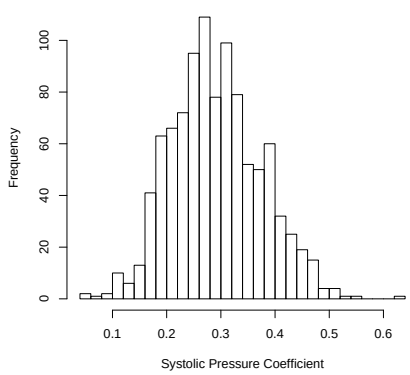
4.2.3 INFERENCES ON THE REGRESSION COEFFICIENTS

Using the methods introduced in the previous chapter, we obtain the confidence intervals and test results which are shown in Table 4.4. In the F-test, we fail to reject the null hypothesis: $\beta_1 = \beta_2 = 0$ when $\alpha = 0.05$. Although the overall test is not rejected, we proceed to test individual coefficients. The p-values for the Z-tests are $0.00069, 0.06047$ for Systolic Pressure and Diastolic Pressure coefficients. If we select $\alpha = 0.05$, we reject the null hypothesis for Systolic Pressure coefficient, but fail to reject the null hypothesis for Diastolic Pressure coefficient. This result looks contradictory to the overall F-test since the p-values for the individual tests are much smaller than that of the overall test. This indicates that there is room for improvement for our testing procedures and we suggest it as our future work.

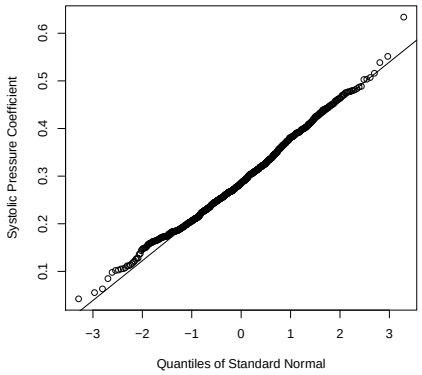
Table 4.4: Inferences Concerning the Regression Coefficients (Blood Pressure)

		$\hat{\beta}_1$	$\hat{\beta}_2$
Z-test	Mean	0.291	0.155
	SE	0.0858	0.0826
	Z-statistic	3.3916	1.8765
	$p-$value	0.00069	0.06047
	95% Confidence interval C.I	(0.1228, 0.4592)	$(-0.0069, 0.3169)$
F-test	p-value= 0.693	(df1,df2)	(2, 8)

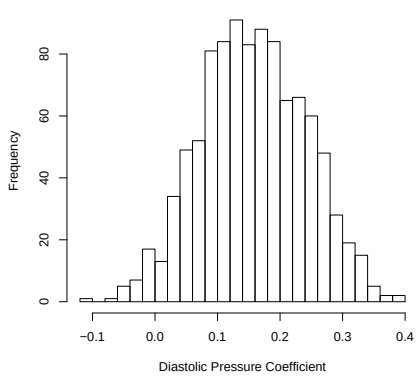
The histograms and Q-Q plots of the estimated coefficients for the independent variables Systolic Pressure and Diastolic Pressure are shown in Figure 4.2. The Systolic Pressure and Diastolic Pressure coefficients are roughly normally distributed, but some outliers can be found. In fact, Xu (2010) excluded the 11th data point from his analysis because it is known that Systolic Pressure should be higher than Diastolic Pressure.



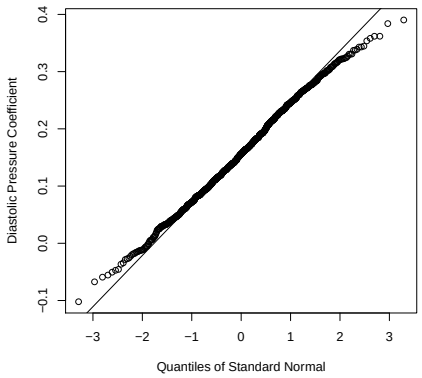
(a) Histogram for Systolic Pressure



(b) Q-Q plot for Systolic Pressure



(c) Histogram for Diastolic Pressure



(d) Q-Q plot for Diastolic Pressure

Figure 4.2: Histograms and Q-Q plots for the 1000 bootstrap replications of the Systolic Pressure and Diastolic Pressure coefficients in the Blood Pressure data

4.3 COMPARISON OF VARIOUS METHODS

In this section, we compare the proposed method, BM, with the four existing methods, CM, CRM, BCRM and SCM, using the bats species and blood pressure data sets.

4.3.1 EVALUATION MEASURES

We use the similar measures considered in Xu (2010) to assess the performance of the five methods. They are the lower bound root mean-square error ($RMSE_L$), the upper bound root mean-square error ($RMSE_U$), and the symbolic correlation coefficient (r). We review these measures using the notations in Xu (2010) before the detailed comparisons.

Lima Neto and de Carvalho (2008) introduced the square root mean-square errors of lower and upper bounds for an interval-valued data to assess the differences between the predicted variable values and the true values. These measures are given by

$$RMSE_L = \sqrt{\frac{\sum_{i=1}^{n}(Y_{L_i} - \hat{Y}_{L_i})^2}{n}} \text{ and } RMSE_U = \sqrt{\frac{\sum_{i=1}^{n}(Y_{U_i} - \hat{Y}_{U_i})^2}{n}}. \tag{4.4}$$

Xu (2010) proposed the correlation coefficient between symbolic variable $\mathbf{Y}$ and the predicted variable $\hat{\mathbf{Y}}$, and the measure is

$$r(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{Cov(\mathbf{Y}, \hat{\mathbf{Y}})}{S_{\mathbf{Y}} S_{\hat{\mathbf{Y}}}}, \tag{4.5}$$

where $\mathbf{Y} = [\mathbf{Y}_L, \mathbf{Y}_U], \hat{\mathbf{Y}} = [\hat{\mathbf{Y}}_L, \hat{\mathbf{Y}}_U], Cov(\mathbf{Y}, \hat{\mathbf{Y}})$ is the symbolic covariance between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, $S_{\mathbf{Y}}$ is the standard deviation of $\mathbf{Y}$, $S_{\hat{\mathbf{Y}}}$ is the standard deviation of $\hat{\mathbf{Y}}$; $Cov(\mathbf{Y}, \hat{\mathbf{Y}})$ can be computed using the equation (2.3), $S_{\mathbf{Y}}$ and $S_{\hat{\mathbf{Y}}}$ can be computed using the following equation: for $Y_i = [c_i, d_i], i = 1, \ldots, n$, the symbolic sample variance (Bertrand and Goupil, 2000) of the random variable $\mathbf{Y}$ is given by

$$S^2_{\mathbf{Y}} = \frac{1}{3n}\sum_{i=1}^{n}(a_{ij}^2 + a_{ij}b_{ij} + b_{ij}^2) - \frac{1}{4n^2}\left[\sum_{i=1}^{n}(a_{ij} + b_{ij})\right]^2.$$

4.3.2 The bats species data set

The four methods are applied to the bats species data set shown in Table 4.1. The observed and predicted lower and upper bounds of the response variable weight are shown in Table 4.5, and the residuals of weight are shown in Table 4.6. The fitted models and the comparison of the performance of these methods based on $RMSE_L, RMSE_U$ and correlation coefficient r are shown in Table 4.7.

Table 4.5: Observed and Predicted Values of Weight by Various Methods (Bats Species)

		Predicted Value				
i	Weight	BM	CM	CRM	BCM	SCM
1	[3, 8]	[-1.61, 8.90]	[-1.30, 10.34]	[0.25, 8.79]	[4.71, 5.07]	[-2.12, 9.35]
2	[4, 10]	[2.83, 9.05]	[3.20, 9.25]	[3.71, 8.75]	[4.12, 8.11]	[2.39, 8.78]
3	[4, 7]	[2.53, 9.22]	[2.66, 9.29]	[1.31, 10.64]	[4.24, 7.45]	[2.10, 9.12]
4	[7, 8]	[3.94, 8.84]	[4.25, 8.74]	[4.88, 8.11]	[5.99, 7.54]	[3.72, 8.62]
5	[6, 9]	[4.42, 7.23]	[4.85, 6.65]	[2.25, 9.26]	[4.38, 6.72]	[4.27, 6.84]
6	[7, 11]	[4.44, 11.66]	[4.87, 11.74]	[3.92, 12.69]	[5.99, 11.04]	[4.20, 11.59]
7	[5, 10]	[6.01, 11.07]	[6.15, 1050]	[3.61, 13.04]	[4.96, 11.06]	[5.69, 10.73]
8	[1, 10]	[4.05, 8.51]	[3.12, 7.22]	[3.71, 6.62]	[4.41, 5.66]	[3.30, 7.76]
9	[8, 12]	[8.41, 14.14]	[8.37, 14.40]	[10.59, 12.17]	[8.12, 14.39]	[8.11, 14.18]
10	[5, 10]	[4.33, 14.21]	[4.23, 12.45]	[-0.61, 17.29]	[5.52, 12.79]	[3.90, 13.47]
11	[7, 12]	[9.03, 13.30]	[9.23, 12.84]	[7.09, 14.98]	[6.48, 14.67]	[8.85, 13.09]
12	[8, 13]	[8.53, 12.45]	[8.61, 11.87]	[6.91, 13.56]	[6.88, 13.07]	[8.37, 12.23]
13	[6, 9]	[5.26, 12.95]	[4.62, 12.07]	[1.82, 14.87]	[5.94, 10.78]	[4.73, 12.75]
14	[6, 10]	[7.34, 10.67]	[6.70, 9.31]	[4.37, 11.65]	[5.56, 9.66]	[6.90, 10.14]
15	[12, 14]	[10.58, 19.37]	[10.48, 19.81]	[12.37, 17.92]	[11.00, 19.93]	[10.42, 19.77]
16	[13, 34]	[16.99, 28.93]	[17.44, 19.50]	[15.58, 31.36]	[12.51, 34.17]	[17.00, 29.54]
17	[8, 16]	[11.06, 16.27]	[9.89, 14.50]	[7.95, 16.44]	[8.32, 15.73]	[10.57,15.79]
18	[17, 35]	[19.20, 30.02]	[19.34, 30.52]	[18.45, 31.41]	[15.90, 34.53]	[19.47, 30.95]
19	[15, 40]	[21.93, 30.26]	[23.68, 30.81]	[17.87, 36.62]	[16.37, 38.72]	[22.98, 31.26]
20	[18, 45]	[23.56, 35.06]	[23.41, 34.50]	[21.64, 36.27]	[17.49, 41.29]	[23.80, 35.59]
21	[20, 50]	[33.45, 36.61]	[35.38, 37.52]	[32.33, 40.57]	[21.10, 50.61]	[34.94, 37.88]

The results in Table 4.7 show that $RMSE_Ls$ for BM, CM, SCM, are $4.096, 4.475, 4.465$ respectively while $RMSE_Us$ for BM, CM, SCM, are $4.932, 4.762, 4.540$. The BM has a smallest $RMSE_L$ among these three methods, but BM has a little bigger $RMES_U$ than do CM and SCM. Therefore, the differences among the $RMSE_Ls$ and $RMSE_Us$ for these methods are not apparent. The correlation coefficients for BM, CM and SCM, are

Table 4.6: Residuals of Weight by Various Methods (Bats Species)

		Residuals				
i	Weight	BM	CM	CRM	BCM	SCM
1	[3, 8]	[4.61, -0.90]	[4.30, -2.34]	[2.75, -0.79]	[-1.71, 2.93]	[5.12, -1.35]
2	[4, 10]	[1.17, 0.95]	[0.80, 0.75]	[0.29, 1.25]	[-0.12, 1.88]	[1.61, 1.22]
3	[4, 7]	[1.47, -2.21]	[1.33, -2.29]	[2.69, -3.64]	[-0.24, -0.45]	[1.90, -2.11]
4	[7, 8]	[3.06, -0.84]	[2.75, -0.74]	[2.12, -0.11]	[1.01, 0.46]	[3.28, -0.62]
5	[6, 9]	[1.58, 1.77]	[1.15, 2.35]	[3.75, -0.26]	[1.62, 2.28]	[1.73, 2.16]
6	[7, 11]	[2.56, -0.66]	[2.13, -0.74]	[3.08, -1.69]	[1.01, -0.04]	[2.80, -0.59]
7	[5, 10]	[-1.01, -1.07]	[-1.15, -0.50]	[1.39, -3.04]	[0.04, -1.06]	[-0.69, -0.73]
8	[1, 10]	[-3.05, 1.49]	[-2.12, 2.78]	[-2.71, 3.38]	[-3.41, 4.34]	[-2.30, 2.24]
9	[8, 12]	[-0.41, -2.14]	[-0.37, -2.40]	[-2.59, -0.17]	[-0.12, -2.39]	[-0.11, -2.18]
10	[5, 10]	[0.67, -4.22]	[0.77, -2.45]	[5.61, -7.29]	[-0.52, -2.79]	[1.10, -3.47]
11	[7, 12]	[-2.03, -1.30]	[-2.23, -0.84]	[-0.09, -2.98]	[0.52, -2.67]	[-1.85, -1.09]
12	[8, 13]	[-0.53, 0.54]	[-0.61, 1.13]	[1.09, -0.56]	[1.12, -0.07]	[-0.37, 0.77]
13	[6, 9]	[0.74, -3.95]	[1.38, -3.07]	[4.18, -5.87]	[0.06, 1.78]	[1.27, -3.75]
14	[6, 10]	[-1.34, -0.67]	[-0.70, 0.69]	[1.63, -1.65]	[0.44, 0.34]	[-0.90, -0.14]
15	[12, 14]	[1.41, -5.37]	[1.52, -5.81]	[-0.37, -3.92]	[1.00, -5.93]	[1.58, -5.77]
16	[13, 34]	[-3.99, 5.07]	[-4.44, 4.50]	[-2.58, 2.64]	[0.49, -0.17]	[-4.00, 4.46]
17	[8, 16]	[-3.06, -0.27]	[-1.89, 1.50]	[0.05, -0.44]	[-0.32, 0.27]	[-2.57, 0.21]
18	[17, 35]	[-2.20, 4.97]	[-2.34, 4.48]	[-1.45, 3.59]	[1.10, 0.47]	[-2.47, 4.05]
19	[15, 40]	[-6.93, 9.74]	[-8.68, 9.19]	[-2.87, 3.38]	[-1.37, 1.28]	[-7.98, 8.74]
20	[18, 45]	[-5.56, 9.94]	[-5.41, 10.50]	[-3.64, 8.73]	[0.51, 3.71]	[-5.80, 9.41]
21	[20, 50]	[-13.45, 13.39]	[-15.38, 12.48]	[-12.33, 9.43]	[-1.10, -0.61]	[-14.94, 12.12]

Table 4.7: Comparison of Methods (Bats Species)

	BM	CM	CRM		BCRM		SCM
	$\hat{\beta}$	$\hat{\beta}^c$	$\hat{\beta}^c$	$\hat{\beta}^r$	$\hat{\beta}^c$	$\hat{\beta}^r$	$\hat{\beta}$
$\hat{\beta}_0$	-25.941	-25.167	-25.167	-5.153	-25.412	-25.550	-27.440
$\hat{\beta}_1$	0.490	0.604	0.604	0.290	0.628	0.411	0.557
$\hat{\beta}_2$	-0.147	-0.260	-0.260	0.941	-0.250	-0.332	-0.183
$\hat{\beta}_3$	0.448	0.396	0.396	0.318	0.344	0.561	0.432
$\hat{\beta}_4$					0.026	-0.025	
$\hat{\beta}_5$					-0.049	0.359	
$\hat{\beta}_6$					0.143	-0.076	
r	0.959	0.958	0.968		0.990		0.959
$RMSE_L$	4.096	4.475	3.742		1.137		4.465
$RMSE_U$	4.932	4.762	4.086		2.320		4.540

$0.959, 0.958, 0.959$, respectively. It can be seen that the correlation coefficients for the three methods are similar. The CRM and BCRM give the smallest $RMSE_Ls$ and $RMSE_Us$, and the biggest correlation coefficients among all the methods.

4.3.3 BLOOD PRESSURE DATA SET

To further assess the performance of the linear regression models from different methods, similarly, the existing methods are applied to the blood pressure data shown in Table 4.3. The observed and predicted values of pulse rate are shown in Table 4.8. The residuals of pulse rate are shown in Table 4.9.

Table 4.8: Observed and Predicted Values of Pulse Rate by Various Methods (Blood Pressure)

		Predicted Value				
i	Weight	BM	CM	CRM	BCM	SCM
1	[44, 68]	[61.2, 67.3]	[59.3, 65.9]	[49.8, 75.5]	[51.6, 73.5]	[58.1, 65.3]
2	[60, 72]	[64.3, 79.1]	[62.7, 79.2]	[60.3, 81.6]	[60.0, 75.8]	[61.9, 79.2]
3	[56, 90]	[82.0, 95.3]	[82.5, 97.4]	[81.0, 98.8]	[70.8, 93.6]	[62.6, 98.1]
4	[70, 112]	[71.7, 85.5]	[70.9, 86.2]	[65.9, 91.2]	[70.8, 97.9]	[70.6, 86.7]
5	[54, 72]	[61.2, 67.3]	[59.3, 65.9]	[49.8, 75.5]	[51.6, 73.5]	[58.1, 65.3]
6	[70, 100]	[77.6, 91.0]	[77.5, 92.5]	[71.9, 98.1]	[76.4, 109.0]	[77.3, 93.2]
7	[72, 100]	[76.9, 87.9]	[76.8, 89.1]	[72.6, 93.3]	[66.2, 90.3]	[76.6, 89.4]
8	[76, 98]	[70.2, 99.7]	[69.2, 102.3]	[74.6, 97.2]	[80.8, 99.9]	[68.7, 103.3]
9	[86, 96]	[81.5, 96.8]	[81.8, 99.1]	[79.9, 101.0]	[76.3, 103.3]	[81.9, 100.0]
10	[86, 100]	[71.4, 86.5]	[70.6, 87.5]	[68.0, 90.0]	[67.8, 89.5]	[70.2, 87.9]
11	[63, 75]	[66.6, 79.8]	[64.7, 79.5]	[63.2, 81.0]	[64.6, 76.6]	[65.0, 80.5]

To evaluate the performance of different methods, the estimated coefficients, $RMSE_L$, $RMSE_U$ and $r(\mathbf{Y}, \hat{\mathbf{Y}})$ are calculated for each model. These measures are shown in Table 4.10.

The $RMSE_Ls$ for BM, CM and SCM, are $11.319, 11.094, 10.988$ respectively while $RMSE_Us$ for BM, CM and SCM, are $10.564, 10.414, 10.393$ respectively, where no apparent differences among the methods can be seen. Also, the correlation coefficients for these methods are almost same. The CRM and BCRM give the smallest $RMSE_Ls$ $(9.810, 8.575)$,

Table 4.9: Residuals of Pulse Rate by Various Methods (Blood Pressure)

		Residuals				
i	Weight	BM	CM	CRM	BCM	SCM
1	[44, 68]	[-17.2, 0.7]	[-15.2, 2.1]	[-5.8, -7.5]	[-7.6, -5.5]	[-14.1, 2.7]
2	[60, 72]	[-4.3, -7.1]	[-2.7, -7.2]	[-0.3, -9.6]	[0.0, -3.8]	[-1.9, -7.2]
3	[56, 90]	[-26.0, -5.3]	[-26.5, -7.4]	[-25.0, -8.8]	[-14.8, -3.6]	[-26.6, -8.1]
4	[70, 112]	[-1.7, 26.5]	[-0.9, 25.8]	[4.1, 20.8]	[-0.8, 14.1]	[-0.6, 25.3]
5	[54, 72]	[-7.2, 4.7]	[-5.2, 6.1]	[4.2, -3.5]	[2.3, -1.5]	[-4.1, 6.7]
6	[70, 100]	[-7.6, 9.0]	[-7.5, 7.5]	[-1.9, 1.9]	[-6.4, -9.0]	[-7.3, 6.8]
7	[72, 100]	[-4.9, 12.1]	[-4.8, 10.9]	[-0.6, 6.7]	[5.8, 9.7]	[-4.6, 10.6]
8	[76, 98]	[5.8, -1.7]	[6.8, -4.3]	[1.4, 1.0]	[-4.7, -1.9]	[7.3, -5.3]
9	[86, 96]	[4.5, -0.8]	[4.2, -3.1]	[6.1, -5.0]	[9.7, -7.3]	[4.1, 4.0]
10	[86, 100]	[14.6, 13.5]	[15.4, 12.5]	[18.0, 10.0]	[18.2, 10.5]	[15.8, 12.1]
11	[63, 75]	[-3.6, -4.8]	[-1.7, -4.5]	[-0.2, -6.0]	[-1.6, -1.6]	[-2.0, -5.5]

Table 4.10: Comparison of Methods (Blood Pressure)

	BM	CM	CRM		BCRM		SCM
	$\hat{\beta}$	$\hat{\beta}^c$	$\hat{\beta}^c$	$\hat{\beta}^r$	$\hat{\beta}^c$	$\hat{\beta}^r$	$\hat{\beta}$
$\hat{\beta}_0$	27.143	21.171	21.171	20.215	1.357	-8.762	18.189
$\hat{\beta}_1$	0.291	0.329	0.329	-0.147	0.332	0.200	0.338
$\hat{\beta}_2$	0.157	0.170	0.170	0.348	0.299	0.104	0.190
$\hat{\beta}_3$					-0.187	-0.371	
$\hat{\beta}_4$					0.680	0.456	
r	0.775	0.775	0.802		0.862		0.775
$RMSE_L$	11.319	11.094	9.810		8.575		10.988
$RMSE_U$	10.564	10.414	8.941		7.418		10.393

and $RMSE_U s$ $(8.941, 7.418)$, and have the biggest correlation coefficients $(0.802, 0.862)$ among all methods.

For both data sets, BM shows the similar performance with CM and SCM, and CRM and BCRM show the better performance than other methods in terms of the measures we consider in this section. However, it does not necessarily mean that CRM or BCRM is a preferable choice for regression analysis to interval-valued data. Xu (2010) addressed this issue by using the following arguments: (i) the purpose of regression analysis is to analyze the relationships among variables, not to analyze the relationships among the ranges of observed intervals; (ii) the fitted models can not answer the question of how much response variable Y changes with changes in each of the explanatory variables X's $(X_1, \ldots, X_p)$; (iii) the range model does not differentiate two different interval observations having the same range point because the centers are neglected when modeling.

Our proposed method randomly generates a large number of points in the ranges of observations through the bootstrap idea to fit a regression model, which sufficiently uses variations in the interval-valued data. Furthermore, we can obtain inferences concerning the regression coefficients such as confidence intervals, standard errors, z-statistics, and p-values.

Chapter 5

simulation

In order to assess further the performance of our new method for fitting a linear regression model to interval-valued data, we implement a simulation study. Section 5.1 describes the data generating procedure. In Section 5.2, we build a linear regression model and make some inferences on regression coefficients using the proposed method.

5.1 Data generation

Let X_1, X_2, X_3 be three independent interval-valued variables, and Y be the dependent interval-valued variable. The procedure for building an interval-valued data set is described as follows:

1. Suppose that X_j^c is the center of jth independent variable, and Y^c is the center of dependent variable. Then, X_1^c, X_2^c, X_3^c and Y^c have a linear relation as follows:

$$Y^c = \beta_0 + \beta_1 X_1^c + \beta_2 X_2^c + \beta_3 X_3^c + \epsilon^c,$$

   where $\beta_0 = 0, \beta_1 = 0.5, \beta_2 = 0, \beta_3 = 0.4$, and $\epsilon^c \sim N(0, 1)$.

2. For the ith observation, generate X_{ij}^c randomly from $\{0, 1, 2, 3, 4\}, j = 1, 2, 3, i = 1, \ldots, n$.

3. Based on the center points X_{ij}^c and ϵ_i^c, the center point Y_i^c is calculated according to the following equation

$$Y_i^c = \beta_0 + \beta_1 X_{i1}^c + \beta_2 X_{i2}^c + \beta_3 X_{i3}^c + \epsilon_i^c.$$

4. For interval-valued variables X_1, X_2 and X_3, generate the lower bound of the ith observation from $[X_{ij}^c - 0.5, X_{ij}^c]$ and the upper bound from $[X_{ij}^c, X_{ij}^c + 0.5]$. Similarly, generate the lower and upper bounds of Y_i.

When $n = 25$, a symbolic interval-valued data set is shown in Table 5.1.

Table 5.1: A Simulated Interval-Valued Dataset

	Independent Variable			Response Variable
i	X_1	X_2	X_3	Y
1	[3.56525, 4.43344]	[1.64020, 2.34241]	[3.79447, 4.25006]	[4.03099, 4.24649]
2	[2.82440, 3.17282]	[1.75447, 2.20790]	[1.72605, 2.36233]	[1.26633, 1.62339]
3	[−0.16056, 0.42416]	[0.65109, 1.15421]	[1.88204, 2.13609]	[0.80583, 1.11676]
4	[0.54336, 1.20667]	[3.64430, 4.18711]	[3.64275, 4.24389]	[0.20497, 0.97314]
5	[2.94911, 3.11854]	[3.56545, 4.28001]	[−0.15328, 0.26757]	[2.28970, 2.77730]
6	[2.68823, 3.02022]	[3.59285, 4.21667]	[3.71474, 4.14360]	[3.25946, 3.75904]
7	[3.59460, 4.33557]	[3.83745, 4.40659]	[2.50934, 3.13610]	[3.07426, 3.79239]
8	[3.71180, 4.35219]	[2.91917, 3.38407]	[0.62890, 1.36490]	[1.17616, 1.96764]
9	[3.54886, 4.24093]	[0.58526, 1.10805]	[2.67077, 3.23707]	[1.52255, 1.91892]
10	[0.81721, 1.04167]	[−0.26536, 0.36028]	[1.82280, 2.28839]	[0.71544, 0.99825]
11	[2.70557, 3.12965]	[−0.43975, 0.30181]	[3.65827, 4.00393]	[2.42591, 2.88045]
12	[3.54491, 4.19186]	[1.98180, 2.30669]	[0.33795, 0.45940]	[1.72763, 1.99572]
13	[3.61078, 4.04897]	[3.54882, 4.38257]	[−0.00218, 0.29647]	[1.58939, 2.28669]
14	[2.98782, 3.15147]	[−0.32492, 0.44369]	[−0.05993, 0.35271]	[2.57739, 3.19166]
15	[1.59326, 2.48921]	[−0.38874, 0.20178]	[0.75487, 1.42600]	[−0.27512, 0.43481]
16	[0.97643, 1.18619]	[0.91333, 1.37959]	[−0.23053, 0.37144]	[0.57675, 1.15448]
17	[2.75798, 3.14488]	[1.76837, 2.27916]	[−0.34833, 0.35859]	[0.70609, 1.16818]
18	[2.62119, 3.49111]	[2.83804, 3.10200]	[2.83346, 3.18564]	[3.76669, 4.29341]
19	[1.86692, 2.30026]	[2.52760, 3.34009]	[−0.13952, 0.38492]	[0.62980, 1.46067]
20	[2.62267, 3.42582]	[1.79942, 2.19594]	[2.69964, 3.18485]	[3.29846, 3.74384]
21	[1.71689, 2.03156]	[2.50947, 3.38296]	[3.63631, 4.46145]	[2.99601, 3.25573]
22	[2.55655, 3.05592]	[−0.40223, 0.47684]	[1.95503, 2.46091]	[0.47979, 0.80821]
23	[0.68049, 1.27623]	[1.70651, 2.04741]	[−0.46688, 0.10526]	[0.92467, 1.33477]
24	[0.75587, 1.13026]	[−0.12268, 0.05730]	[1.59117, 2.13037]	[0.09546, 0.54504]
25	[3.80498, 4.36492]	[2.97990, 3.02244]	[0.69863, 1.37963]	[2.41991, 2.63072]

5.2 SIMULATION RESULTS

5.2.1 Fitted linear models with different methods

The center method (CM), center and range method (CRM) , bivariate center and range method (BCRM), symbolic covariance method (SCM) and bootstrap method (BM) are applied to the simulated data set in Table 5.1 to fit linear regression models.

Using CM, we find that the fitted linear model is

$$\hat{Y}^c = 0.53140X_1^c - 0.02901X_2^c + 0.48310X_3^c.$$

Using CRM, we obtain the fitted linear models as

$$\hat{Y}^c = 0.53140X_1^c - 0.02901X_2^c + 0.48310X_3^c,$$

$$\hat{Y}^r = -0.02345X_1^r + 0.39024X_2^r + 0.64430X_3^r.$$

The center and range models with BCRM are

$$\hat{Y}^c = 0.62669X_1^c - 0.06796X_2^c + 0.54092X_3^c + 0.34951X_1^r - 0.15694X_2^r - 0.66902X_3^r,$$

$$\hat{Y}^r = 0.01882X_1^c + 0.04322X_2^c + 0.02734X_3^c - 0.15720X_1^r + 0.25297X_2^r + 0.55391X_3^r.$$

The SCM produces the model

$$\hat{Y} = 0.57067X_1 - 0.00022X_2 + 0.51470X_3.$$

The proposed BM method is applied to the simulated data set, and the model is given by

$$\hat{Y} = 0.52775X_1 - 0.02313X_2 + 0.47938X_3.$$

From the model obtained using our proposed method, we can see that the coefficients for X_1, X_2 and X_3 are $0.52775, -0.02313$, and 0.47938, respectively. It shows that Y increases 0.52775 units when X_1 increases 1 unit, and Y has a positive relationship with X_1, and X_3 while it has a small negative relationship (The coefficient is close to zero) with X_2. The

coefficients $(0.57067, -0.00022, 0.51470)$ of the SCM model are similar to the results from the model with our proposed method. The CM model shows that the center points of Y has a positive relationship with the center points of X_1 and X_2, and a negative relationship with X_2. Using CRM, we obtain the center and range models. The center model, which is the same as CM model, only reveals the relationship between the center points of explanatory and response variables while the range model only reveals the range relationship between Xs and Y. For example, the range of Y decreases 0.02345 units when the range of X_1 increases 1 unit. Note that it is hard to interpret the BCRM models.

In order to evaluate the performance of the these linear regression models, $RMSE_L$ and $RMSE_U$ defined in equation (4.4), and r defined in equation (4.5) are calculated for each of the models. We run the simulations with 100 repetitions, and calculate the estimates and the measures of $RMSE_L$, $RMSE_U$ and correlation coefficient r for each repetition. Then the means of the estimates and the measures are obtained. From Table 5.2, we can see that BCRM has the largest correlation coefficient (0.724) while BM, CM, CRM and SCM almost have the same correlation coefficients (0.678, 0.678, 0.678, 0.683). After the simulations with 100 repetitions, the means of the regression coefficients for BM are $0.502, 0.007$ and 0.384 which are closer to the true regression coefficients $(0.5, 0, 0.4)$ than a single run of simulation. Like BM, the means of the regression coefficients for CM and SCM are also closer to the true values than single runs. For CRM and BCRM, because they have the center and range models, it is hard to interpret the relationship between their coefficients and the true coefficients.

5.2.2 INFERENCES CONCERNING THE REGRESSION COEFFICIENTS

In this section, we conduct statistical inferences concerning the regression coefficients. Another simulation with 100 replications are implemented, and the means of inferences concerning regression coefficients are obtained. From Table 5.3, we can see that the means of estimated regression coefficients for X_1, X_2 and X_3, respectively, are $0.4910, 0.0033, 0.3985$

Table 5.2: Means of Estimates and Measures with 100 Simulation Repetitions

	BM	CM	CRM		BCRM		SCM
	$\hat{\beta}$	$\hat{\beta}^c$	$\hat{\beta}^c$	$\hat{\beta}^r$	$\hat{\beta}^c$	$\hat{\beta}^r$	$\hat{\beta}$
$\hat{\beta}_1$	0.5023	0.5054	0.5054	0.3191	0.4895	0.0137	0.4924
$\hat{\beta}_2$	0.0071	0.0040	0.0040	0.3116	-0.0047	0.0214	-0.0007
$\hat{\beta}_3$	0.3846	0.3856	0.3856	0.3194	0.3700	0.0183	0.3740
$\hat{\beta}_4$					0.0200	0.2533	
$\hat{\beta}_5$					-0.0513	0.2477	
$\hat{\beta}_6$					0.1882	0.2481	
r	0.6776	0.6777	0.6783		0.7236		0.6834
$RMSE_L$	0.9539	0.9540	0.9511		0.8809		1.0541
$RMSE_U$	0.9498	0.9498	0.9479		0.8796		1.1064

which are very close to the true coefficients $(0.5, 0, 0.4)$. The p-value under the F-test is 0.0000, and thus the null hypothesis: $\beta_1 = \beta_2 = \beta_3 = 0$, is rejected. The p-values obtained from the Z-tests are $4.3170 \times 10^{-13}, 0.1007, 1.1931 \times 10^{-3}$ for the three coefficients respectively. If we select $\alpha = 0.05$, we reject the null hypotheses for X_1 and X_3 coefficients, but fail to reject the hypothesis for X_2 coefficient. These results are consistent with the true values $(\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 0.4)$.

Table 5.3: Means of Inferences for Regression Coefficients with 100 Simulation Repetitions

		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Z-test	Mean	0.4910	0.0033	0.3985
	S.E.	0.0270	0.0276	0.0273
	Z-statistic	18.5616	0.0819	15.1465
	p-value	4.3170×10^{-13}	0.1007	1.1931×10^{-3}
	95% Confidence interval	(0.4380, 0.5439)	$(-0.0507, 0.0573)$	(0.3449, 0.4520)
F-test	p-value	0.0000	(df1,df2)	$(3, 22)$

Chapter 6

summary

In this thesis, we reviewed the linear regression methods currently available to analyze symbolic interval-valued data. We also reviewed the bootstrap basics including resampling, bootstrap standard errors, bootstrap confidence intervals and bootstrap regression. Then, we proposed a new method for fitting a linear regression model to interval-valued data.

The objective of many statistical investigations is to make inferences about population parameters based on sample data. The main contribution of our proposed method is that it enables one to do statistical inferences concerning the regression coefficients. First, by using the bootstrap idea, we conducted the overall F-test on the model significance, and then we made inferences on regression coefficients such as the standard errors, the confidence intervals and the p-values under the null hypothesis $H_0 : \beta_j = 0$ where $j = 1, \ldots, p$.

The proposed method and the existing methods were applied to the real and simulated data sets. The comparisons among these methods demonstrated that the prediction of the proposed method is as accurate as other existing methods. In addition, the proposed method takes into account the variability of an interval as SCM does. Also, it was demonstrated that it provides reasonable and meaningful statistical inferences.

Two future directions of this work can be considered. First, the core idea of this thesis can be extended to other statistical analysis problems such as principal component analysis, classification, and clustering, etc. For example, for binary classification, one can obtain B classification rules $f_1, \ldots, f_B$. When a new interval-valued observation is observed, one can generate B classical valued data points and for each datum point, apply one of the B

classification rules and make the class assignment. The final class assignment for the interval-valued observation is made as the majority votes among the B estimated class memberships. Second, some improvements for the overall F-test procedure in Chapter 4 can be made. The method we used in the thesis does not calculate the exact p-value and the result depends on the significance level used for the k individual F-test. One way to improve is to generate an empirical "null" distribution by generating symbolic data assuming no predictor variable is significant.

Bibliography

[1] Bertrand, P. and Goupil, F. (2000). Descriptive Statistics for Symbolic Data. *Analysis of Symbolic Data,* eds. H.-H. Bock and Diday, Springer-Verlag, Berlin, 103-124.

[2] Billard, L. and Diday, E. (2000). Regression Analysis for Interval-Valued Data. *Data Analysis, Classification, and Related Methods,* eds. H. A. L. Kiers, J.-P. Rassoon, P. J. F. Groenen, and M. Schader. Springer-Verlag, Berlin, 369-374.

[3] Billard, L. and Diday, E. (2003). From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *Journal of American Statistical Association,* 98, 470-487.

[4] Billard, L. and Diday, E. (2007). *Symbolic Data Analysis: Conceptual Statistics and Data Mining,* Wiley, Chichester.

[5] Billard, L. (2007). Dependencies and Variation Components of Symbolic Interval-Valued Data. *Selected Contributions in Data Analysis and Classification*, Springer-Verlag, Berlin, 3-13.

[6] Billard, L. (2008). Sample Covariance Functions for Complex Quantitative Data. *Proceeding in World Conferences Interantional Association of Statistical Computing 2008*, Yokohama, Japan.

[7] Billard, L. (2011). Brief overview of symbolic data and analytic issues. *Statistical Analysis and Data Mining*, 4, 149-156, Wiley.

[8] Diday, E. (1987). à I'Approache Symbolique en Analyse des Dommées. *Premières Journées Symbolique - Numérique,* CEREMADE, Université Paris, 21-56.

[9] Diday, E. (1995). Probabilist, Possibilist and Belief Object for Knowledge Analysis. *Annals of Operations Reseatch,* 55, 227-276.

[10] Diday, E. and Emilion, R. (1996). Lattices and Capacities in Analysis of Probabilist Object. *Studies in Classification*, eds. E. Diday, Y. Lechevallier, and O. Opilz, 13-30.

[11] Diday, E., Emilion, R. and Hillali, Y. (1996). Symbolic Data Analysis of Probabilistic Object by Capacities and Credibilities. *Societea' Italianadi Statistica,* 55-22.

[12] Diday, E. and Emilion, R. (1998). Capacities and Credibilities in Analysis of Probabilistic Objects By Histograms and Lattices. *Data Science, Classification , and Related Methods*, eds. C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. -H. Bock, and Y. Baba, 353-357.

[13] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, 7, 1-26.

[14] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall.

[15] Emilion, R. (1997). Différentiation des Capacitiés. *Comptes Rendus de l'Academie des Sciences - Series I - Mathematics,* 324, 389-392.

[16] Fox, J. (2002). Bootstrapping Regression Medels. Submitted Manuscript.

[17] Lima Neto, E.A., de Carvalho, F.A.T. and Tenorio, C.P. (2004). Univariate and Multivariate Linear Regression Methods to Pridict Interval-Valued Features. *Lecture Notes in Computer Science, AI 2004 Advances in Artificial Intelligence*, Springer-Verlag, Berlin, 526-537.

[18] Lima Neto, E.A., de Carvalho, F.A.T. and Freire, E.S. (2005). Applying Constrained linear regression Models to Predict Interval-Valued Data. *Lecture Notes in Computer*

Science, KI: Advances in Artificial Inteligence (ed. U. Furbach), Springer-Verlag, Berlin, 92-106.

[19] Lima Neto, E.A. and de Carvalho, F.A.T. (2008). Center and Range Method for Fitting a Linear Regression Model to Symbolic Interval Data. *Computational Statistics and Data Analysis*, 52, 1500-1515.

[20] Lima Neto, E.A. and de Carvalho, F.A.T. (2010). Constrained Linear Regression Models for Symbolic Interval-Valued Variables. *Computational Statistics and Data Analysis,* 54, 333-347.

[21] Noirhomme-Fraiture, M. and Brito, P. (2011). Far beyond the classical data models: symbolic data analysis. *Statistical Analysis and Data Mining*, 4, 157-170, Wiley.

[22] Xu, W. (2010). Symbolic Data Analysis: Interval-Valued Data Regression, Ph.D. Dissertation, University of Georgia.

Appendix A

R function for simulating to make statistical inferences by bootstrap method

```
inferences=function(n,b1,b2,b3,sigma)  ##observations and model coefficients
{
cx1=sample(c(0,1,2,3,4),n,replace=TRUE)## center points of x
cx2=sample(c(0,1,2,3,4),n,replace=TRUE)
cx3=sample(c(0,1,2,3,4),n,replace=TRUE)
err=rnorm(n,0,1)
cy=b1*cx1+b2*cx2+b3*cx3+sigma*err      ##center points of y
x11=runif(n,min=cx1-0.5,max=cx1)       ##lower bounds
x12=runif(n,min=cx1,max=cx1+0.5)       ##upper bounds
x21=runif(n,min=cx2-0.5,max=cx2)
x22=runif(n,min=cx2,max=cx2+0.5)
x31=runif(n,min=cx3-0.5,max=cx3)
x32=runif(n,min=cx3,max=cx3+0.5)
y1=runif(n,min=cy-0.5,max=cy)
y2=runif(n,min=cy,max=cy+0.5)
d=data.frame(x11=x11,x12=x12,x21=x21,
  x22=x22,x31=x31,x32=x32,y1=y1,y2=y2)
B=1000
x1=matrix(NA,nrow=n,ncol=B)             ##bootstrap points within intervals
x2=matrix(NA,nrow=n,ncol=B)
x3=matrix(NA,nrow=n,ncol=B)
y=matrix(NA,nrow=n,ncol=B)
for(i in 1:n) {
x1[i,]=runif(B,min=d[i,1],max=d[i,2])
x2[i,]=runif(B,min=d[i,3],max=d[i,4])
x3[i,]=runif(B,min=d[i,5],max=d[i,6])
y[i,]=runif(B,min=d[i,7],max=d[i,8])
}
coe=matrix(NA,nrow=B,ncol=3)             ##regression coefficients
ftest=matrix(NA,nrow=B,ncol=3)           ##f-statistic and degrees of freedom
ftestp=rep(0,times=B)                    ##p-value under f-test
for(j in 1:B) {
       md=lm(y[,j]~x1[,j]+x2[,j]+x3[,j] -1)
       coe[j,1]=md$coef[1]
       coe[j,2]=md$coef[2]
       coe[j,3]=md$coef[3]
       ftest[j,1]=summary(md)$fstatistic[1]
```

```
       ftest[j,2]=summary(md)$fstatistic[2]
       ftest[j,3]=summary(md)$fstatistic[3]
       ftestp[j]=pf(summary(md)$fstatistic[1],
                 summary(md)$fstatistic[2],
                 summary(md)$fstatistic[3],
                 lower.tail=FALSE)
}
beta=c(mean(coe[,1]),mean(coe[,2]),         ##coefficients of fitted model
     mean(coe[,3]))
zstderror=sqrt(c(var(coe[,1]),var(coe[,2]), ##standard error
          var(coe[,3])))
z=beta/(zstderror)                          ##z-statistic
p=2*(1-pnorm(abs(z)))                       ##p-values for coefficient
pf=sum(ftestp>0.05)/B                       ##p-value for model test
infer=rbind(beta,zstderror,z,p,fstat,pf)
infer
}
betamatrix=matrix(NA,nrow=rep,ncol=3)
zstderrormatrix=matrix(NA,nrow=rep,ncol=3)
zmatrix=matrix(NA,nrow=rep,ncol=3)
pmatrix=matrix(NA,nrow=rep,ncol=3)
pfmatrix=matrix(NA,nrow=rep,ncol=3)
for(i in 1:rep)                             ##simulation repetition
{
      infer=inferences(n,b1,b2,b3,sigma)
      betamatrix[i,]=infer[1,]
      zstderrormatrix[i,]=infer[2,]
      zmatrix[i,]=infer[3,]
      pmatrix[i,]=infer[4,]
      fstatmatrix[i,]=infer[5,]
      pfmatrix[i,]=infer[6,]
}
```