Symbolic Data Analysis:
Interval-valued Data Regression

by

Wei Xu

(Under the direction of Lynne Billard)

Abstract

In statistics, usually data are formatted as single values. However, sometimes the data are represented by lists, intervals, histograms or even distributions. To deal with these kinds of data, the concept of symbolic data was introduced by Diday (1987).

Among symbolic data, interval-valued data are the most commonly seen in application. Recently, different approaches have been introduced to analyze interval-valued data, including linear regression, principal component analysis and clustering, etc. This research focuses on interval-valued data regression analyses. The study begins with the concept of symbolic data, definition of symbolic interval-valued data and its descriptive statistics, and existing linear regression approaches. It then proposes new approaches, including the symbolic covariance method and symbolic likelihood method with their algorithms and applications and shows the two methods obtain identical results under certain conditions. The proposed methods are applied to real data and simulated data along with other methods and their performances are discussed.

Index words: Symbolic data, Linear regression, Symbolic covariance, Least squares, Maximum likelihood, Order statistics

Symbolic Data Analysis: Interval-valued Data Regression

by

Wei Xu

B.E., Renmin University of China, China, 2002

M.E., Renmin University of China, China, 2005

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2010

Symbolic Data Analysis:
Interval-valued Data Regression

by

Wei Xu

Approved:

Major Professor: Lynne Billard

Committee: William P. McCormick
T.N. Sriram
Xiangrong Yin
Cheolwoo Park

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2010

Acknowledgments

I would never have been able to finish my dissertation without the guidance of my advisor, support from my committee members, help from my friends, and encouragement from my parents.

I would like to express my deepest appreciation to my advisor, Dr. Lynne Billard, for her excellent guidance that inspires me to conduct research and countless hours of reading, editing, reflecting, encouraging, and most of all patience throughout the entire process. Without her guidance and continuous help this dissertation would not be possible.

I would like to thank my committee members, Dr. William McCormick, Dr. T.N. Sriram, Dr. Xiangrong Yin, and Dr. Cheolwoo Park for the time they spend reading my dissertation and their comments which have led to improvement in the quality of my dissertation.

I also want to take this opportunity to thank all the faculty members, staff and graduate students in this department, who made my life enjoyable and memorable during my five wonderful years at UGA.

Finally, I would like to thank my parents for their unconditional love and endless support to encourage me to pursue my dream.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Typically, data we analyze are classical data, which means each observation is a single point in a p-dimensional space $\mathcal{R}^p$. However, data in the form of lists, intervals, histograms, and the like, are examples of symbolic data introduced by Diday (1987). Unlike classical data, symbolic data can be hypercubes in a p-dimensional space $\mathcal{R}^p$. As a consequence, symbolic data have an internal structure which does not exist in classical data. For this unique internal structure, statistical theory and method for classical data can not be readily applied. Therefore, developing new methods to analyze symbolic data and building the mathematical underpinnings are of great necessity.

Diday (1995), Diday et al. (1996), Emilion (1997) and Diday and Emilion (1996, 1998) established the original mathematical framework of symbolic data in a series of papers. In their papers, it is stated that a realization of a symbolic random variable may take a finite or an infinite set of values in $\mathcal{R}^p$. There are several basic types of symbolic data. One type is multi-valued data, i.e., list data. Another commonly seen type of symbolic data is interval-valued data. A more complex type is modal-valued data in which probabilities, possibilities, credibilities, or other suitable weights are associated with its values. A good example of modal-valued data is a histogram. Classical data are a special case of symbolic data by putting probability 1 on each single value in the internal distribution.

Typically, symbolic data arise in two situations throughout data collecting and processing: some data collected are inherently symbolic; some become symbolic data after processing. First of all, the original data may be collected as lists, intervals, histograms, or the like. An example of naturally collected symbolic data is blood pressure where the measuring device

actually identifies an interval range (even though the value may be recorded as a single value) due to the inherit continual change in a person's blood pressure. In contrast, by recording changes throughout the day and from day to day, the result is not a single value but a range of values, i.e., an interval. Another example of symbolic data is income level. Survey analysts well know that asking a person about his income directly usually does not elicit the correct (if any) answer. Instead, a multiple choice question with different income levels, such as A [20K,30K), B [30K,40K), C [40K,50K), D [50K and up) is usually utilized in a questionnaire.

Another important reason to induce symbolic data is that sometimes datasets are very large. With the advent of the modern computer science, massive size datasets are becoming routine while how to analyze them is not so routine. Even performing a simple exploratory statistical analysis may require a huge mount of computing power. To solve this problem, a lot of work has been done to obtain more efficient algorithms. While these improved algorithms help, they are still limited in usefulness. However, a simple way to solve this problem is to aggregate individual observations into groups of interest, especially when characteristics of groups are of higher interest to an analyst than those of individual observations. Thus the original dataset is summarized into one of a more manageable size while retaining as much interesting knowledge as possible. As an example, suppose a hospital has thousands of patient records. However, usually interest is cast in the characteristics and behaviors of certain groups, not those of a single person. Therefore, we can collapse the original data according to the underlying scientific interest, which results in symbolic data. More examples of natural and aggregated symbolic data can be found in Billard and Diday (2007), and an expanded review is in Chapter 2.

Symbolic data analysis provides a solution not only to the naturally occurring symbolic data, but also to the massive data problems. To analyze symbolic data, usually researchers summarize the original data into single points, such as the center points or the medians of the intervals to conduct their analysis. However, by doing so, either some important information in the original data is lost, such as the range of the interval; or great efforts are taken to

reorganize the data into a manageable format while in the end obtaining a result which is not in the original data format. On the contrary, symbolic data analysis can preserve the information from being lost while conducting analyses directly on the original data without transformation or reorganization. Therefore it is more convenient and efficient to utilize symbolic data analyses. In addition, the analysis result can be more friendly to understand since the data format is not changed.

Regression, as a statistical tool for the investigation of relationships between response variables and explanatory variables, has long been one of the central techniques in the field of statistics. It has also increasingly become an area of active research in economics, psychology, education, bio-related areas, etc. Since Legendre (1805) and Gauss (1809) introduced the least squares method for regression for the first time in order to solve the problem of determining the orbits of objects around the Sun based on astronomical observations, almost every aspect of regression has been dug into in the classical data situation, including the famous maximum likelihood estimation recommended and popularized by Fisher (1922); Aldrich (1997).

Adaption of regression to symbolic data, especially interval-valued data has become an area of active research since Billard and Diday (2000) introduced the first approach to fit a regression model to interval-valued data. Their approach is to fit a linear regression model on the center points of the intervals, then applied the fitted model to the lower and upper bounds of the independent variables to generate predictions respectively. Lima Neto et al. (2004) and de Carvalho et al. (2004) took Billard and Diday's approach one step further by utilizing the ranges of intervals besides the center points to build two independent regression models. A bivariate version center and range method was proposed by Billard and Diday (2007) to build the regression model on center points and ranges of intervals simultaneously. However, all these methods can lead to an unpleasant situation where the predicted lower bound could exceed the predicted upper bound. In order to avoid this situation, Lima Neto et al. (2005, 2010) considered a constrained model that forces all parameters to be positive,

which may be misleading in the case where a negative relationship between variables truly exists.

Apart from the drawbacks of current symbolic regression methods for interval-valued data, the thinking of reducing intervals down to center points and ranges to establish a model can not reflect the internal variation of the data in the best way from a symbolic stand point of view. We are interested in finding a solution which can use the interval-valued data directly. On the foundation of symbolic sample covariance definition by Billard (2007, 2008), we proposed a symbolic covariance method (so called SCM method) to conduct interval-valued data regression. The main idea is to reconstruct the least squares estimator of a regression model in a way that utilizes the symbolic sample covariance. Our proposed method is able to account for the total covariance structure of interval-valued data as well as the dependency among all variables in one comprehensive way. Generalization of the proposed method to other types of symbolic data such as histogram-valued data is considered to be derivable.

Maximum likelihood estimation is, if not any more popular, just as important as least squares techniques for regression. It is well known that in classical data regression the resulting estimator from least squares technique is identical to the estimator from maximum likelihood estimation if only the residuals are assumed to follow a normal distribution with known covariance structure. Starting from the likelihood function proposed by Le-Rademacher and Billard (2010) for a univariate interval-valued random variable, we establish a basic likelihood function for interval-valued data and use it to derive maximum likelihood estimators for linear regression models. We show the least squares procedure once again agrees with maximum likelihood estimation even in the symbolic interval-valued data situation.

This dissertation is organized as follows. Chapter 2 gives a literature review of symbolic data regression: the concept and main types of symbolic data are presented as well as their descriptive statistics; differences between symbolic data and fuzzy data are discussed; current

methods to handle interval-valued data regression are reviewed. In Chapter 3, we propose the symbolic covariance method. Prediction, confidence interval and goodness-of-fit problems in interval-valued data regression are considered. Examples are provided in the end of the chapter to illustrate the method and results. In Chapter 4, we obtain a maximum likelihood estimator for interval-valued data regression. Chapter 5 includes examples with real datasets to compare the current methods and our proposed methods. Measures of performance from Lima Neto and de Carvalho (2010) are also introduced. In Chapter 6, we outline the future work ideas, including the so called order statistic method for interval-valued data regression.

1.1 References

[1] Aldrich, J. (1997). R.A. Fisher and the Making of Maximum Likelihood 1912 - 1922. *Statistical Science* 12(3), 162-176.

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

[3] Billard, L. (2008). Sample Covariance Functions for Complex Quantitative Data. Processing, World Conferences International Association of Statistical Computing 2008, Yokohama, Japan.

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

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

[6] de Carvalho F.A.T., Lima Neto, E.A. and Tenorio, C.P. (2004). A New Method to Fit a Linear Regression Model for Interval-valued Data. *Lecture Notes in Computer Science, KI2004 Advances in Artifical Inteligence*. Springer-Verlag, 295-306.

[7] Diday, E. (1987). Introduction à l'Approache Symbolique en Analyse des Données. *Premières Journées Symbolique - Numérique*. CEREMADE, Université Paris, 21-56.

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

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

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

[11] Diday, E., Emilion, R. and Hillali, Y. (1996). Symbolic Data Analysis of Probabilistic Objects by Capacities and Credibilities. *Societea' Italianadi Statistica*, 5-22.

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

[13] Fisher, R.A. (1922). The Goodness of Fit of Regression Formulae, and the Distribution of Regression Coefficients. *Journal of the Royal Statistical Society* 85, 597612.

[14] Gauss, C.F. (1809). Theoria Motus Corporum Coelestium. *Sectionibus Conicis Solem Ambientum*.

[15] Le-Rademacher, J. and Billard, L. (2010). Likelihood Functions and Some Maximum Likelihood Estimators for Symbolic Data. *Journal of Statistical Planning and Inference*, submitted.

[16] Legendre, A.M. (1805). Sur la Méthode des moindres quarrś. *Nouvelles méthodes pour la détermination des orbites des comètes*.

[17] Lima Neto, E.A. and de Carvalho F.A.T. (2010). Constrained Linear Regression Models for Symbolic Interval-valued Variables. *Computational Statistics & Data Analysis*, 54(2), 333-347.

[18] Lima Neto, E.A., de Carvalho F.A.T. and Freire, E.S. (2005). Applying Constrained Linear Aggression Models to Predict Interval-Valued Data. *Lecture Notes in Computer Science, KI: Advances in Artificial Inteligence* (ed. U. Furbach). Springer-Verlag, Brelin, 92-106.

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

Chapter 2

Literature Review

In order to establish the foundation of our work, a review of the literature is given in this chapter. Therefore, Section 2.1 introduces the concept of symbolic data as well as some notation and terms. Section 2.2 discusses the difference between symbolic data and fuzzy data. Three major types of symbolic data and their descriptive statistics are presented in Section 2.3. Section 2.4 briefly reviews the methodology of classical linear regression. In Section 2.5, currently available symbolic interval-valued data linear regression methods are introduced along with their advantages and disadvantages.

2.1 From Classical Data to Symbolic Data

Statistics is an art and a science of dealing with data. However, the data we usually are able to analyze are classical data in which each data point is a single point in a p-dimensional space $\mathcal{R}^p$. However, with the advent of computers and data collecting devices, very large and complicated data sets have become routine. The data we need to analyze may no longer be a single point. Diday (1987) initiated the concept of symbolic data to handle such phenomenon. Symbolic data can be intervals, lists, histograms, or even distributions, which are all examples of symbolic data.

Let us start with some dataset examples. Suppose we have a dataset comprising the medical records of individuals in a hospital. Each individual represents an observation. For each individual, there may be variables recording geographic location information, including county name (e.g., Athens, Macon,$\cdots$), urban/rural classification (yes or no), and so on. There are also demographic variables, such as gender (male or female), marital status (single,

Table 2.1: Sample Patient Information Dataset

ID	Name	County	Age	Income Level[a]	Pulse Rate	Systolic Pressure	$\cdots$
1	K. Roberts	Duluth	46	[40, 50]	70	[90, 112]	$\cdots$
2	W. Bryant	Athens	30	[90, 100]	62	[110, 140]	$\cdots$
3	R. Adams	Macon	32	[20-30]	75	[80, 100]	$\cdots$
4	L. Duncan	Duluth	45	[50-70]	73	[120, 130]	$\cdots$
5	J. Howard	Athens	[30, 35]	[70, 90]	78	[110, 140]	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

[a] Income in \$1000's

married, or divorced), age, income level, employer, and so on. In addition, there could also be basic health information variables including, say, weight, pulse rate, blood pressure, and the like. Other medical information variables could include incidences of certain ailments and diseases.

The dataset in Table 2.1 is a typical dataset that involves symbolic interval-valued data. For example, it is always a sensitive issue when asking a patient's exact income. In certain circumstances, some patients may be unwilling to answer if such kinds of questions are asked. However, it would be much easier to collect the data if several income ranges are provided in the questionnaire from which patients can choose. Another example is systolic blood pressure, an entity well known to fluctuate constantly making "accurate" measurement difficult. When nurses examine a patient's systolic pressure three times to obtain a more accurate result, usually the mean is recorded as the outcome. However, it would be more informative if the original data can be kept in a format of lists, or transformed into intervals stating the minimum and maximum systolic pressure.

Let us think about this problem on a more general basis. Suppose $\mathbf{X} = (X_{ij})$ is the $n \times p$ matrix of the entire dataset. Let X_j, $j = 1, \ldots, p$, represent the jth variable. Let x_{ij} denote the observed value of the variable X_j for the ith observation. Note n is the number

of observations and p is the number of variables, where n and p can be extremely large. Suppose the domain of X_j is $\mathcal{X}_j$, then $\mathbf{X}=(X_1,\ldots,X_p)$ takes values in $\mathcal{X}=\times_{j=1}^{p}\mathcal{X}_j$.

In a classical data setting, variable X_j can be numerical. For example, pulse rate for K. Roberts is 70, i.e., X_{pulse}=70, X_j in $\mathcal{X}_{pulse\ rate}=\{x \geq 0\}$. Variables can also be categorical, for example, $\mathcal{X}_{gender}$={Male, Female}, or $\mathcal{X}_{city}$={Macon, Athens,$\cdots$}.

However, in the classical setting, either numerical or categorical, all the data share one common characteristic, that is, for each x_{ij} in $\mathbf{X}$, there is precisely only one possible realized value. For example, an individual record of variable X_{name} is W. Bryant, a record of X_{age} is 30, a record of X_{county} is Athens, and a record of $X_{pulse\ rate}$ is 62, and so on. In short, a classical data point is a single point in a p-dimensional space $\mathcal{R}^p$.

In contrast, a symbolic data point can be a hypercube in a p-dimensional space or a Cartesian product of distributions in $\mathcal{R}^p$. Entries in a symbolic dataset are no longer restricted to a single specific value. For example, a record of $X_{income\ level}$ is 40,000-50,000, a record of $X_{systolic\ pressure}$ is 90-112.

There are innumerable examples where symbolic data arise. Table 2.2 gives an example of a symbolic dataset aggregated from a large raw dataset. This is a "mixed" dataset and contains both symbolic multi-valued and interval-valued data along with classical data.

Table 2.2: Sample Census Dataset

ID	Race	Gender	Age	Marital Status	Parents Living	Weight	Income[a]
1	White	Male	[70,76]	{D,M}	{0,1}	[150,230]	[60,80]
2	Black	Female	[20,30]	{S,M}	{1,2}	[158,188]	[30,50]
3	Asian	Male	[30,36]	{S,M}	{1,2}	[120,200]	[50,80]
4	Hispanic	Female	[11,16]	{S}	{0,1,2}	[73,150]	[20,40]
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	Hispanic	Female	[20,26]	{M,S}	{0,1,2}	[120,196]	[30,60]

[a] Income in $1000's

In Table 2.2, marital status can be single, married, and divorced; parents living means number of parents living. In this dataset, race and gender are classical data; marital status

and parents living are symbolic multi-valued data; age, weight, and income are symbolic interval-valued data.

Table 2.3 gives another dataset example containing symbolic histogram-valued data from Billard and Diday (2007). The dataset was extracted from a classical dataset, which records 33 flight performance evaluation variables for 50,680 individual flights into JFK airport, New York, in January 2004. The original dataset can be found at Falduti et al. (2004). Rather than operating performance on any single flight, overall performance by different airlines is much more important. Therefore, the original classical data were aggregated by 16 airlines in terms of 6 performance evaluation variables. The variables are: Y_1 = Flight Time, Y_2 = Time to Taxi In to the gate after landing, Y_3 = Arrival Delay Time, Y_4 = Time to Taxi Out from the gate to the runway, Y_5 = Departure Delay Time, and Y_6 = Weather Delay Time. All times are recorded in minutes and a negative time means the flight left or arrived early.

The data in Table 2.3 are histogram-valued. For example, Y_5 = Departure Delay Time is given by three categories attached with the corresponding relative frequencies. For Airline 1, 44% flights were not delayed, 47% were delayed within 1 hour, and 9% delayed by more than 1 hour; likewise, for other airlines and variables.

Table 2.3: Sample Flights into JFK Airport Dataset

Airline	Y_1=Airtime			Y_2=Taxi In			Y_3=Arrival Delay			Y_4=Taxi Out			Y_5=Departure Delay			Y_6=Weather Delay	
	< 120	[120,220]	> 220	< 4	[4,10]	> 10	< 0	[0,60]	> 60	< 16	[16,30]	> 30	< 0	[0,60]	> 60	No	Yes
1	.15	.62	.23	.12	.64	.24	.42	.46	.12	.38	.47	.15	.44	.47	.09	.92	.08
2	.89	.11	.00	.21	.65	.14	.52	.39	.09	.45	.40	.15	.32	.60	.08	.90	.10
3	.99	.01	.00	.10	.82	.08	.45	.48	.07	.41	.44	.15	.54	.30	.06	.97	.03
4	.22	.60	.18	.27	.67	.06	.47	.50	.03	.55	.37	.08	.37	.60	.03	.99	.01
5	.24	.72	.04	.03	.79	.18	.52	.43	.05	.28	.53	.19	.71	.25	.04	.98	.02
6	.95	.05	.00	.31	.57	.12	.51	.37	.12	.51	.37	.12	.46	.42	.12	.96	.04
7	.56	.35	.09	.09	.72	.19	.47	.49	.04	.32	.52	.16	.46	.51	.03	.99	.01
8	.81	.19	.00	.07	.72	.21	.49	.44	.07	.50	.41	.09	.21	.73	.06	.99	.01
9	.83	.17	.00	.03	.78	.19	.41	.49	.10	.33	.49	.18	.47	.45	.08	1.0	.00
10	.00	.01	(.57,.42)	.08	.67	.25	.42	.52	.06	.37	.53	.10	.47	.48	.05	.99	.01
11	.90	.09	.01	.14	.66	.20	.38	.50	.12	.42	.42	.18	.50	.41	.09	.96	.04
12	.75	.25	.00	.05	.78	.17	.40	.55	.05	.43	.40	.17	.63	.34	.03	.96	.04
13	.85	.14	.01	.08	.72	.20	.49	.40	.11	.36	.46	.18	.58	.34	.08	.99	.01
14	.43	.20	(.21,.16)	.23	.66	.11	.46	.44	.10	.29	.53	.18	.57	.34	.09	.98	.02
15	.98	.02	.00	.24	.70	.06	.63	.35	.02	.58	.33	.09	.69	.30	.01	.99	.01
16	.58	.36	.06	.57	.42	.01	.53	.43	.04	.89	.09	.02	.01	.96	.03	.99	.01

Typically, symbolic data arise in two situations throughout data collecting and processing. The first situation is the original data may be collected in a symbolic data format such as lists, intervals, histograms, or the like. Table 2.1 gives a good example of such kinds of datasets. In order to analyze these kind of datasets, usually we break the original data into single (i.e., classical) values such as taking the center points of the intervals to conduct the analysis. However, by doing so, we either are losing some important information the original data contain, such as the range of the interval, but most especially, the internal variation of the interval; or we have to take extra efforts to reorganize the data into a manageable format, and finally obtain a result which is not consistent with the original data format. On the contrary, symbolic data analyses can preserve all the information in the observation so that no information is lost while conducting analyses directly on the original data without the need to transform or reorganize them. Therefore, it is more convenient and efficient to utilize symbolic data methodology. Another benefit of symbolic data analysis may be that the analysis result is more user friendly since the data type has not been changed.

The second situation where symbolic data arise is that sometimes datasets are so large that they must be summarized into a manageable size while retaining as much knowledge as possible before any analysis can be conducted. As in the example dataset in Table 2.1, typically the hospital has thousands of patient records. Usually, our interest is in the characteristics and behaviors of certain groups, not that of a single person. Therefore, we can aggregate the original data according to our scientific interest, which results in symbolic data. Tables 2.2 and 2.3 are examples of summarized symbolic datasets.

Symbolic data are inherently different from classical data. The biggest difference between classical data and symbolic data is that a classical data value has no interval variation while in contrast a symbolic data value has. This means symbolic data deal with both internal variation within each observation value and external variation between observation values.

Let us take interval-value data as an example to show the difference. In Table 2.4, values from three original observations are as shown in the first column. Classical surrogates and

Table 2.4: Difference Between Classical Data and Symbolic Data

Data	Original Value	Classical Surrogate		Symbolic Surrogate	
		Value	Variance	Value	Variance
Observation 1	35	35	0	35	0
Observation 2	[20,50]	35	0	[20,50]	75
Observation 3	[10,60]	35	0	[10,60]	208.3

symbolic surrogates are also shown. In the absence of knowing how to handle interval-valued data, analysts typically take a classical surrogate such as the center point or median for the original observation. Suppose values follow a uniform distribution within the interval. For a classical surrogate, we only take the center point of the interval to conduct analysis while leaving out the variation information. Thus, for the data 35 and [20, 50], we will obtain the same information using classical data analysis, but actually they are hugely different values since the interval [20, 50] has a bigger range than does the classical point 35. The variance of the interval data in second observation is 75 (see Equation (2.10)), whereas for the classical point in the first observation it is 0. Likewise, the third observation [10, 60] also has a classical surrogate center point value of 35 with variance 0 while actually the interval observation has a variance of 208.3. The classical surrogates for the three different original observations give the same value and variance where the symbolic values reflect the true value and variance of the original observation.

2.2 Fuzzy Data Are Not Symbolic Data

Fuzzy data are another type of data, also called fuzzy sets. Fuzzy sets include fuzzy numbers and fuzzy intervals. Zadeh (1965) initiated fuzzy logic and fuzzy sets. Basically, fuzzy logic is a multivalued logic that allows intermediate values to be defined between conventional evaluations in order to mimic a human-like way of thinking in computer programming and control systems.

A fuzzy set is a pair (A, m) where A is a set and for each $x \in A$, $m(x)$ is the grade of membership of x. It is important to point out the distinction between fuzzy logic and probability. Both operate over the same numeric range, and have similar values: 0 representing False (or non membership), and 1 representing True (or full membership). However, there is a distinction to be made between the two statements: fuzzy set theory uses the concept of fuzzy set membership (i.e., how much a variable is in a set), probability theory uses the concept of subjective probability (i.e., how probable a variable is in a set). While this distinction is mostly philosophical, the fuzzy-logic-derived possibility measure is inherently different from the probability measure; hence they are not directly equivalent. Therefore, Zadeh (1965) argues that fuzzy logic is different in character from probability, and is not a replacement for it.

Fuzzy data are different from symbolic data. First, they are introduced for different reasons. Fuzzy data were initiated to mimic a human-like way of thinking while symbolic data were initiated to build data objects in order to conduct analyses such as regression, principal component analysis, and clustering. Second, fuzzy data only include fuzzy numbers and fuzzy intervals, but symbolic data are concepts of data objects which include points, interval-valued data, multivalued data, histogram-valued data and distribution data. Third, fuzzy data utilizes the concept of grade of membership as introduced above, which is different from the concept of probability used in symbolic data. Last, fuzzy data focus on what the system should do rather than trying to model how it works. It concentrates on solving the problem rather than trying to model the system mathematically. On the contrary, symbolic data are brought out to help in building statistical models and to perform analyses.

The following is an example from Billard and Diday (2007) of fuzzy data and symbolic data. Suppose four individuals Sean, Kevin, Rob and Jack have height (Y_1, in meters), weight (Y_2, in kilograms) and hair color (Y_3) as shown in Table 2.5. Thus, in the classical data setting, e.g., Jack is 1.95 meters in height, weighs 90 kilos and has black hair.

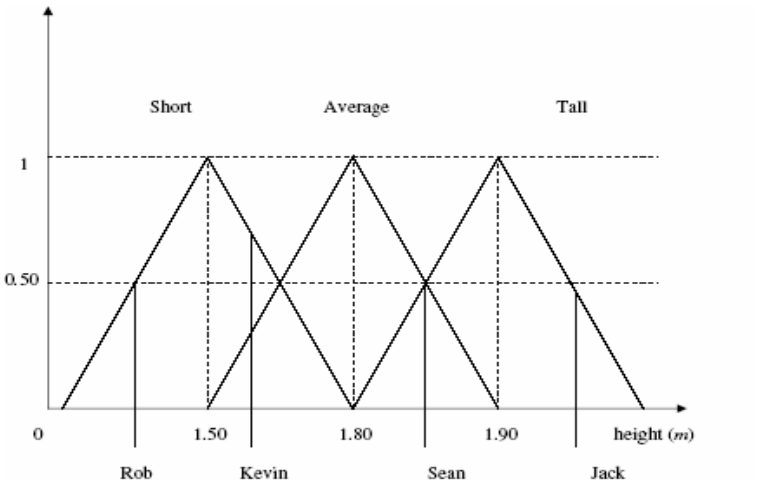

Figure 2.1: Distribution of Height

Assume the random variable height takes three forms: short, average and tall, with each form being triangular distributed centering on 1.50, 1.80 and 1.90 meters, respectively. We can transform the original data into fuzzy data. Suppose the hair color values are kept as classical data. Figure 2.1 (Billard and Diday 2007, Figure 2.6) shows the distribution of height.

From Figure 2.1, we can see that Kevin has a fuzzy measure of 0.30 to be average height and 0.70 to be short. In contrast, Jack has a fuzzy measure of 0.48 to be tall. We can find the

Table 2.5: Individual Appearance Dataset

Individual	Height	Weight	Hair Color
Sean	1.85	80	blonde
Kevin	1.60	45	blonde
Rob	0.65	30	black
Jack	1.95	90	black

Table 2.6: Fuzzy Data on Height

	Height				
Individual	Short	Average	Tall	Weight	Hair Color
Sean	0	0.5	0.5	80	blonde
Kevin	0.7	0.3	0	45	blonde
Rob	0.5	0	0	30	black
Jack	0	0	0.48	90	black

Table 2.7: Symbolic Data Based on Hair Color

	Height			
Hair Color	Short	Average	Tall	Weight
blonde	[0,0.7]	[0.3,0.5]	[0,0.5]	[45,80]
black	[0,0.5]	0	[0,0.48]	[30,90]

fuzzy data about the heights for all four individuals; these are as shown in Table 2.6. Now suppose we want to aggregate the fuzzy data into categories describing hair color. Therefore, we have one category blonde hair color (with its content Sean and Kevin) and the other category dark hair color (with its content Rob and Jack). This aggregation gives us the symbolic data as shown in Table 2.7. See more details about differences between symbolic data and fuzzy data and examples in Billard and Diday (2007).

2.3 Several Types of Symbolic Data

There are several types of symbolic data: interval-valued, multivalued, histogram-valued, and distributions. The most commonly seen methodology to date is concerned with multivalued and interval-valued data.

2.3.1 Notations

Before we go into more details about different types of symbolic data and their descriptive statistics, we want to establish some universal notations. Let $\mathbf{X}$=$(X_1, \ldots, X_p)$ be a random vector with p random variables $X_1, \ldots, X_p$. Furthermore, let the random variable X_{ij} denote the ith observation of jth variable, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let the lower case x_{ij} denote a realization of X_{ij} for a classical value and ξ_{ij} denote a realization of X_{ij} for a symbolic value. The use of ξ is to emphasize the difference between symbolic data and classical data. Note that symbolic data have interval structure.

Before discussing the distribution functions and descriptive statistics of symbolic data, we need to introduce some basic definitions taken from Billard and Diday (2007). See Billard and Diday (2007) for more comprehensive details and information. Many contents in this chapter can also be found in Le-Rademacher (2008).

Let $\mathcal{X}_j$ be the domain of $\mathbf{X}_j$ and $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_p = \times_{j=1}^{p} \mathcal{X}_j$ be the domain of $\mathbf{X} = (X_1, \ldots, X_p)$. We have the following definitions.

Definition 2.3.1. Every point $\mathbf{x} = (x_1, \ldots, x_p) \in \mathcal{X}$ is called a description vector.

Definition 2.3.2. D is called a description set where $D \subseteq \mathcal{X}$ and $D = D_1 \times \ldots \times D_p$, where $D_j \subseteq \mathcal{X}_j$. When $D = D_1 \times \ldots \times D_p$ is a Cartesian product, then D is called a Cartesian description set. When D_j is a singleton, i.e., $D_j = x_j$ for all $j = 1, \ldots, p$, then a description vector $\boldsymbol{D} = (D_1, \ldots, D_p) = (x_1, \ldots, x_p)$ is called a individual description vector.

Definition 2.3.3. Let $A \subseteq D$ and $B \subseteq D$ be two description sets, and $x \in \mathcal{X}$. A logical dependency rule υ is defined as:

$$\upsilon : [x \in A] \Rightarrow [x \in B].$$

The set of all logical dependency rules υ defined on $\mathcal{X}$ is represented as $V_{\mathcal{X}}$.

Definition 2.3.4. The virtual description of a description vector $\boldsymbol{d}$, $vir(\boldsymbol{d})$, is defined as the set of all individual description vectors x that satisfy all the logical dependency rules υ in $\mathcal{X}$, i.e., $vir(\boldsymbol{d}) = \{x \in \boldsymbol{d} | \upsilon(x) = 1, \text{ for all } \upsilon \in V_{\mathcal{X}}\}$.

As an example, consider the baseball dataset of Billard and Diady (2006). Here, Y_1 = number of at-bats and Y_2 = number of hits. Suppose a team has values $Y = (Y_1, Y_2) = \Big([88, 422], \quad [49, 149]\Big)$. Since it is not possible to have more hits than at-bats, it is necessary to include the rule $\upsilon : Y_2 \leq Y_1$. Under this rule the virtual description space for this observation is the hyperrectangle with vertices {(88,88),(149,149,)(149,422),(49,422),(49,88)}. Similar rules hold elsewhere, e.g., diastolic blood pressure < systolic blood pressure. See Billard and Diday (2007) for numerous examples in a variety of settings.

2.3.2 Multi-valued Symbolic Data

Definition 2.3.5. A multi-valued symbolic random variable X is one whose possible value takes one or more values from the list of values in its domain $\mathcal{X}$. The complete list of possible values in $\mathcal{X}$ is finite, and values may be well-defined categorical or quantitative values.

For example, a cancer variable may have a domain $\mathcal{X}$={bone, liver, lung,...} listing all possible cancer locations of the patients in the hospital. For an individual, the particular realization can be ξ_u = {bone, lung, kidney}. In another example, suppose we have records of some criminals in a police station. To investigate a crime, the police may find out a suspect's record showing what kinds of crimes he has previously committed. An observation may be ξ_u = {robbery, fight in street, hit-and-run}.

Description statistics for multi-valued data were given by Bertrand and Goupil (2000) as follows.

Definition 2.3.6. Let $\xi_1, \ldots, \xi_n$ be a a set of observations for the random variable X. Then, we can obtain the observed frequency of ξ by

$$O_X(\xi) = \sum_{i=1}^{n} \pi_X(\xi) = \sum_{i=1}^{n} \frac{|x_i \in vir(d_i)|x_i = \xi|}{|vir(d_i)|} \tag{2.1}$$

where $\pi_X(\xi)$ is the percentage of the individual description vectors in $vir(d_i)$ such that $x_i = \xi$, and for a space A, $|A|$ denotes the number of individual descriptions in the space.

In the summation in Equation (2.1), any dependency rule υ for which $vir(d_i)$ is empty is ignored. It should be noted that this observed frequency $O_X(\xi)$ is a positive real number and not necessarily an integer as for classical data.

Therefore, the empirical distribution function of X is defined by

$$F_X(\xi) = \frac{1}{n'} \sum_{\xi_k \leq \xi} O_X(\xi_k),$$

where $n' = n - n_0$ and n_0 is the number of i for which $|vir(d_i)| = 0$.

Moreover, for quantitative multi-valued data, based on the empirical distribution function, we can derive the symbolic sample mean $\bar{X}$ and sample variance function S^2 as, respectively,

$$\bar{X} = \frac{1}{n'} \sum_{\xi_k \in \mathcal{X}_{(j)}} \xi_k O_X(\xi_k)$$

and

$$S^2 = \frac{1}{n'} \sum_{\xi_k \in \mathcal{X}_{(j)}} (\xi_k - \bar{X})^2 O_X(\xi_k).$$

See Bertrand and Goupil (2000) as well as Billard and Diday (2007) for more details.

2.3.3 Interval-valued Symbolic Data

Among all the formats of symbolic data, interval-valued data plays an important role not only because it is the mostly commonly seen but also because the techniques to analyze it can often be readily generalized to other types of data. Therefore, it is of great importance to develop methodology to analyze interval-valued data. However, since this research focuses on interval-valued data, we review and introduce some basic descriptive statistics for such data more thoroughly.

As an example of interval-valued symbolic data, note that pulse rate can be recorded as an interval, [50,60), [60,70), [70,80),..., income level can be intervals, [20K,30K), [30K,40K), [40K,50K),... Such kinds of data can be encountered when the original recorded data value is in an interval, as the income level case, or the data may correspond to an individual whose precise value is known to be within a range, as the age for Julie Howard in Table 2.1, or

the data value varies over time during the course of the experiment, such as a person's pulse rate which may fluctuate between [60, 70] during a day. Another reason is that sometimes it may be impossible to measure some characteristic so accurately as a single value, but rather gives the result in a $(x \pm \delta)$ format. Confidentiality issues may necessitate transforming an observation to, e.g., $[x - \delta_1, x + \delta_2]$, $\delta_1 \neq \delta_2$.

Bertrand and Goupil (2000) gave the definitions of empirical density function, sample mean and sample variance; many examples can be found in Billard and Diday (2007). Billard (2007, 2008) obtained the sample covariance for interval-valued data. All these definitions are based on the assumption that the internal observation is uniformly distributed within the intervals. This condition may be relaxed in future research. However, for all the content in this dissertation, this condition is assumed to hold.

Definition 2.3.7. An interval-valued symbolic random variable X_j is one that takes values in an interval; i.e., the ith realization of X_j is $X_{ij} = [a_{ij}, b_{ij}] \subset \mathcal{R}$, with $a_{ij} \leq b_{ij}$, $a_{ij}, b_{ij} \in \mathcal{R}$, $i = 1, \ldots, n, j = 1, \ldots, p$. The interval can be closed or open at either end.

Since we assume a point X_{ij} in X_j is uniformly distributed over the interval $[a_{ij}, b_{ij}]$, we have

$$P\{x_{ij} \leq \xi | x_{ij} \in vir(d_i)\} = \begin{cases} 0, & \xi \leq a_{ij}, \\ \frac{\xi - a_{ij}}{b_{ij} - a_{ij}}, & a_{ij} \leq \xi \leq b_{ij}, \\ 1, & b_{ij} \leq \xi, \end{cases}$$

where the virtual space $vir(d)$ was given in Definition 2.3.4.

Definition 2.3.8. For an interval-valued random variable X_j, the empirical density function is

$$f_{X_j}(\xi) = \frac{1}{n} \sum_{i:\xi \in \xi_{ij}} \left(\frac{1}{b_{ij} - a_{ij}}\right) \tag{2.2}$$

or equivalently,

$$f(\xi) = \frac{1}{n} \sum_{i:\xi \in \xi_{ij}} \frac{I_{ij}(\xi)}{||X_{ij}||}, \ \xi \in \Re$$

where $I_{ij}(.)$ is the indicator function that ξ is in the interval X_{ij} and $||X_{ij}||$ is the length of the interval.

Furthermore, Bertrand and Goupil (2000) defines the symbolic mean and variance as follows.

Definition 2.3.9. For an interval-valued random variable X_j, the symbolic sample mean $\bar{X}_j$ is, $j = 1, \ldots, p$,

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

Definition 2.3.10. For an interval-valued random variable X_j, the symbolic sample variance S_j^2 is, $j = 1, \ldots, p$,

$$S_j^2 = \frac{1}{3n} \sum_{i=1}^{n} (a_{ij}^2 + a_{ij} b_{ij} + b_{ij}^2) - \frac{1}{4n^2} [\sum_{i=1}^{n} (a_{ij} + b_{ij})]^2. \tag{2.3}$$

Bertrand and Goupil (2000) and Billard and Diday (2007) showed details of how these two statistics are derived as follows.

For the sample mean $\bar{X}_j$ we have

$$\begin{aligned} \bar{X}_j &= \int_{-\infty}^{\infty} \xi f(\xi) \partial \xi \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{b_{ij} - a_{ij}} \int_{-\infty}^{\infty} \xi \partial \xi \right] \\ &= \frac{1}{2n} \sum_{i=1}^{n} \left[\frac{1}{b_{ij} - a_{ij}} \xi^2 |_{a_{ij}}^{b_{ij}} \right] \\ &= \frac{1}{2n} \sum_{i=1}^{n} (a_{ij} + b_{ij}); \end{aligned} \tag{2.4}$$

and for sample variance S_j^2, we have

$$\begin{aligned} S_j^2 &= \int_{-\infty}^{\infty} (\xi - \bar{X}_j)^2 f(\xi) \partial \xi \\ &= \int_{-\infty}^{\infty} \xi^2 f(\xi) \partial \xi - \bar{X}_j^2, \end{aligned} \tag{2.5}$$

where

$$\begin{aligned}\int_{\infty}^{\infty} \xi^2 f(\xi)\partial\xi &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{b_{ij}-a_{ij}}\int_{-\infty}^{\infty}\xi^2\partial\xi\right] \\ &= \frac{1}{3n}\sum_{i=1}^{n}\left[\frac{1}{b_{ij}-a_{ij}}\xi^3|_{a_{ij}}^{b_{ij}}\right] \\ &= \frac{1}{3n}\sum_{i=1}^{n}\left[\frac{1}{b_{ij}-a_{ij}}(b_{ij}^3-a_{ij}^3)\right] \\ &= \frac{1}{3n}\sum_{i=1}^{n}(a_{ij}^2+a_{ij}b_{ij}+b_{ij}^2). \end{aligned} \tag{2.6}$$

Therefore, the sample variance S_j^2 can be derived by substituting Equations (2.4) and (2.6) into Equation (2.5) and the result in Equation (2.3) follows.

Note, the definition of symbolic data variance is equal to that of the classical data sample variance when $a_{ij} = b_{ij}$.

Billard (2007, 2008) shows that total sum of squares (TSS) of interval-valued symbolic data, i.e., n times the symbolic data variance can be further divided into two parts: within sum of squares (WSS) variation and between sum of squares (BSS) variation. That is, we can write

$$nS_j^2 = TSS = WSS + BSS \tag{2.7}$$

where

$$BSS = \sum_{i=1}^{n}(\bar{X}_{ij} - \bar{X}_j)^2; \tag{2.8}$$

and because

$$b_{ij} - \bar{X}_{ij} = \bar{X}_{ij} - a_{ij} = b_{ij} - a_{ij},$$

we have

$$\begin{aligned} WSS &= \frac{1}{3}\sum_{i=1}^{n}\left[(a_{ij}-\bar{X}_{ij})^2 + (a_{ij}-\bar{X}_{ij})(b_{ij}-\bar{X}_{ij}) + (b_{ij}-\bar{X}_{ij})^2\right] \\ &= \frac{1}{3}\sum_{i=1}^{n}\left[\frac{(b_{ij}-a_{ij})}{2}\right]^2 \\ &= \frac{1}{12}\sum_{i=1}^{n}(b_{ij}-a_{ij})^2. \end{aligned} \tag{2.9}$$

Note, Equation (2.9) is the sample variance of a uniformly distributed random variable obtained by the method of moments; so this result is consistent with the assumption that the values are uniformly distributed in the interval $[a_{ij}, b_{ij}]$. That is, if $X_{ij} \sim uniform(a_{ij}, b_{ij})$, for $i = 1, \ldots, n$, independently, then

$$Var(X_{ij}) = \frac{(b_{ij} - a_{ij})^2}{12}. \tag{2.10}$$

Furthermore, the total within interval variance of the n observations $X_1, \ldots, X_n$ is the sum of the variances in Equation (2.10), which is

$$\frac{1}{12} \sum_{i=1}^{n} (b_{ij} - a_{ij})^2.$$

Let us now consider the bivariate situation. Billard and Diday (2003, 2007) provided the empirical joint density function as

$$f(\xi_1, \xi_2) = \frac{1}{3n} \sum_{i:\xi \in x i_{ij}} \frac{I_{ij}(\xi_1, \xi_2)}{||X_{ij}||}.$$

To obtain the sample covariance, we extend the idea behind Equations (2.7) and (2.9) to the bivariate situation. Billard (2007, 2008) introduced a symbolic interval-valued data covariance function based on the decomposition of the total sum of products. Suppose X_1 and X_2 are two symbolic interval-valued random variables, and suppose $X_{i1} = [a_i, b_i]$, $X_{i2} = [c_i, d_i]$, $i = 1, \ldots, n$.

Definition 2.3.11. For interval-valued random variables X_1 and X_2, Billard (2007, 2008) introduced the definition of the empirical covariance function $Cov(X_1, X_2)$ as

$$\begin{aligned} Cov(X_1, X_2) = \frac{1}{6n} \sum_{i=1}^{n} [&2(a_i - \bar{X}_1)(c_i - \bar{X}_2) + (a_i - \bar{X}_1)(d_i - \bar{X}_2) \\ &+ (b_i - \bar{X}_1)(c_i - \bar{X}_2) + 2(b_i - \bar{X}_1)(d_i - \bar{X}_2)] \end{aligned} \tag{2.11}$$

where $\bar{X}_1 = \frac{1}{n} \sum_{i=1}^{n} (a_i + b_i)/2$, $\bar{X}_2 = \frac{1}{n} \sum_{i=1}^{n} (c_i + d_i)/2$.

Analogously, Billard(2008) showed that for symbolic interval-valued data, we have within sum of products (WSP) and between sum of products (BSP) as, respectively,

$$\text{WSP} = \frac{1}{12}\sum_{i=1}^{n}(a_i - b_i)(c_i - d_i), \tag{2.12}$$

$$\text{BSP} = \sum_{i=1}^{n}(\frac{a_i + b_i}{2} - \bar{X}_1)(\frac{c_i + d_i}{2} - \bar{X}_2). \tag{2.13}$$

Hence, total sum of products (TSP) is given by

$$\text{TSP} = \text{WSP} + \text{BSP}.$$

Therefore, the empirical covariance of X_1 and X_2 can be derived as

$$Cov(X_1, X_2) = \text{TSP}/n,$$

which can be written as in Definition 2.3.11.

The variance and covariance definitions and the calculation method introduced by Billard (2007, 2008) are consistent with results for the classical data case and are easy to compute. A **R** function based on the software package **R** to obtain the sample covariance between interval-valued variables is included in Appendix 2.A. Furthermore, we can divide the total sum of products into within sum of products and between sum of products, which can give us an inside look at how the variables are correlated. Note that the BSS and BSP statistics are based on the interval center points. Thus, any analysis that uses the center points as classical surrogates ignores the within variations (i.e., WSS and WSP).

After we have all the definitions of variance and covariance, it is quite trivial to define the correlation coefficient for interval-valued variables X_1 and X_2.

Definition 2.3.12. Let X_1 and X_2 be two symbolic interval-valued random variables. Then, the sample correlation function $r(X_1, X_2)$ is

$$r(X_1, X_2) = Cov(X_1, X_2)/S_{X_1}S_{X_2}$$

where the covariance $Cov(X_1, X_2)$ is given in Definition 2.3.11 and the standard deviation for X_1 and X_2 are obtained from the sample variances defined in Definition 2.3.10.

2.3.4 Modal-valued Symbolic Data

Symbolic data can be modal-valued. Let X be a random variable taking values on domain $\mathcal{X}$. Then it is called modal-valued if it takes the form

$$X = \{x_k, \pi_k; k = 1, \ldots, m\}$$

where π_k is a nonnegative measure associated with x_k and m is the number of possible x_k values within $\mathcal{X}$.

If the $\{x_k\}$ outcomes take the format of multi-values, then X is called modal multi-valued. If they take the format of subsets of the real-line $\mathcal{R}$, then X is called modal interval-valued, e.g., histogram-valued.

The measures $\{\pi_k\}$ are the support of $\{x_k\}$ in domain $\mathcal{X}$. They can be weights, probabilities, relative frequencies, and the like, corresponding to the respective outcome $\{x_k\}$.

Definition 2.3.13. Let X_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, denote the ith observation of the jth random variable. Let ξ_{ij} be a realization of X_{ij}. Suppose X_j is a quantitative random variable that takes values on a finite number of nonoverlapping intervals $[a_{ij}, b_{ij}]$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, with $a_{ij} \leq b_{ij}$ and relative frequency p_{ij}, that is,

$$\xi_{ij} = \left\{[a_{ij1}, b_{ij1}), p_{ij1};\ [a_{ij2}, b_{ij2}), p_{ij2};\ \ldots;\ [a_{ijs_{ij}}, b_{ijs_{ij}}], p_{ijs_{ij}}\right\}$$

where $([a_{ijk}, b_{ijk}), p_{ijk})$ represents the kth subinterval of ξ_{ij} with relative frequency p_{ijk}. Here, s_{ij} is the number of subintervals in the histogram ξ_{ij}, and $\sum_{k=1}^{s_{ij}} p_{ijk} = 1$. Then X_{ij} is called a histogram-valued random variable.

Based on the assumption that all values within every subinterval $[a_{ijk}, b_{ijk})$, $k = 1, \ldots, s_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, are uniformly distributed, Billard and Diday (2003) extends the empirical distribution function introduced by Bertrand and Goupil (2000) for interval-valued data to histogram-valued data.

The distribution of a point x_u within subinterval $[a_{ijk}, b_{ijk})$ is

$$P(x_u \leq \xi | x_u \in vir(d_i)) = \begin{cases} 0, & \xi < a_{ijk}, \\ \frac{\xi - a_{ijk}}{b_{ijk} - a_{ijk}}, & a_{ijk} \leq \xi < b_{ijk}, \\ 1, & b_{ijk} \leq \xi. \end{cases} \tag{2.14}$$

On the basis of the distribution function (2.14), Billard and Diday (2003) proposed an empirical distribution function for a histogram-valued random variable X_j analogous to that for an interval-valued random variable. The empirical distribution function of X_j is defined by

$$F_{X_j}(\xi) = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k:\xi \in \xi_{ijk}} p_{ijk} (\frac{\xi - a_{ijk}}{b_{ijk} - a_{ijk}}) + \sum_{k:\xi \geq b_{ijk}} p_{ijk} \right) \tag{2.15}$$

where $\xi_{ijk} = [a_{ijk}, b_{ijk})$. Taking the derivative of F_{X_j} in Equation (2.15) with respect to ξ results in the empirical density function of X_j, we obtain

$$f_{X_j}(\xi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k:\xi \in \xi_{ijk}} p_{ijk} (\frac{1}{b_{ijk} - a_{ijk}}).$$

After we have the empirical distribution function and empirical density function, it is relatively easy to obtain the symbolic sample mean $\bar{X}_j$ and sample variance S_j^2 for histogram-valued data, respectively, by

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

and

$$S_j^2 = \frac{1}{3n} \sum_{i=1}^{n} \sum_{k=1}^{s_{ij}} p_{ijk} [a_{ijk}^2 + a_{ijk} b_{ijk} + b_{ijk}^2] - \frac{1}{4n^2} \left[\sum_{i=1}^{n} \sum_{k=1}^{s_{ij}} p_{ijk} (a_{ijk} + b_{ijk}) \right]^2.$$

2.3.5 Summary

More generally, symbolic data can be a probability distribution, or even a model. Schweizer(1984) opined that "Distributions are the numbers of the future".

In summary, classical data on p random variables are single points in a p-dimensional space $\mathcal{R}^p$. For contrast, symbolic data are represented by p-dimensional hypercubes in $\mathcal{R}^p$, or

a Cartesian product of p distributions. Symbolic data can be multi-valued, interval-valued, modal-valued, a distribution, or even a model. Therefore, a classical data value as a single point is a special case of a symbolic data value. In addition, fuzzy data are different from symbolic data. Nowadays, symbolic data are becoming more and more prevalent. Therefore, methods to analyze such kind of data and their mathematical underpinnings are of great necessity.

2.4 Linear Regression in Classical Data Analysis

Regression analysis has long been an important branch in statistics and undergoes constant development. One aim of regression analysis is to construct mathematical models which can describe or explain relationships existing between variables. Usually, interest is centered on one variable, called the response (or dependent) variable, and how it depends on a set of variables called the explanatory (or independent) variables.

If we denote the response variable by Y and the explanatory variables by $X_1, \ldots, X_p$, then a general model can be written as

$$E[Y|X_1 = x_1, \ldots, X_p = x_p] = f(x_1, \ldots, x_p). \tag{2.16}$$

In particular, linear regression models are an important class of regression models in which the response variable is a linear combination of explanatory variables, that is,

$$f(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Suppose $E(Y) = f(x_1, \ldots, x_p) = \eta$, i.e., Y fluctuates about an unknown parameter η, $Y = \eta + \epsilon$, where ϵ is the error. Our focus is to model the linear relationship between Y and $X_j, j = 1, \ldots, p$. Therefore, for observations $(Y_i, X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$, the model of Equation (2.16) can be expressed in the form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

where $i = 1, \ldots, n$, or, equivalently in matrix notation,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.17}$$

where

- $\boldsymbol{X}$ is a non-stochastic $n \times p$ matrix with $p < n$. The matrix $\boldsymbol{X}$ has rank p, that is, $\boldsymbol{X}$ is of full column rank;
- the elements of the $n \times 1$ vector $\boldsymbol{Y}$ are observable random vectors;
- the elements of the $n \times 1$ vector $\boldsymbol{\epsilon}$ are non-observable random variables such that $E(\boldsymbol{\epsilon}) = 0$ and $Cov(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$ with $\sigma^2 > 0$. We will write $\epsilon \sim (0, \sigma^2 \boldsymbol{I}_n)$ for short.

To obtain the estimation of parameters in model (2.17), there are two basic methods: ordinary least squares technique and maximum likelihood estimation.

The principle of the ordinary least squares method is perhaps the best known and most applied method for estimating the regression parameters. If we assume that the Equation (2.17) is solvable with respect to $\boldsymbol{\beta}$, then a solution $\boldsymbol{\beta}^*$ clearly satisfies $||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^*||^2 = 0$. On the other hand, if Equation (2.17) is not solvable, then we can determine a vector $\hat{\boldsymbol{\beta}}$ which satisfies

$$||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2 \leq ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_*||^2, \tag{2.18}$$

for every vector $\boldsymbol{\beta}_* \in \mathcal{R}^p$. Such a vector $\hat{\boldsymbol{\beta}}$ is called the least squares estimator for the parameters $\boldsymbol{\beta}$ in Equation (2.17) because the condition (2.18) can be equivalently expressed as

$$\sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})^2 \leq \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i\boldsymbol{\beta}_*)^2 = \sum_{i=1}^{n} \epsilon_{*i}^2$$

where $\epsilon_{*i} = Y_i - \boldsymbol{X}_i\boldsymbol{\beta}_*$ is the ith residual of the estimate of $\hat{\boldsymbol{\beta}}$. Then the sum of squared residuals is minimized for $\boldsymbol{\beta}_* = \hat{\boldsymbol{\beta}}$, so that $\hat{\boldsymbol{\beta}}$ has the smallest sum of squared residuals.

To obtain $\hat{\boldsymbol{\beta}}$, we can differentiate the function $||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_*||^2$ with respect to $\boldsymbol{\beta}_*$, i.e.,

$$\frac{\partial}{\partial \boldsymbol{\beta}_*} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_*||^2 = 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}_* - 2\boldsymbol{X}'\boldsymbol{Y}. \tag{2.19}$$

Putting Equation (2.19) equal to 0 and solving for $\boldsymbol{\beta}_*$, we can find the least squares estimator of $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.

Definition 2.4.1. Under the linear regression model with assumptions (i) to (iv), the function

$$f(\boldsymbol{\beta}_*) = ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_*||^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_*)'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_*)$$

is minimized for $\boldsymbol{\beta}_* = \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. The vector $\hat{\boldsymbol{\beta}}$ is called the ordinary least squares estimator of $\boldsymbol{\beta}$.

The least squares estimator $\hat{\boldsymbol{\beta}}$ has many favorable properties, including that it is the best linear unbiased estimator of $\boldsymbol{\beta}_*$. More details about linear regression models and related estimation methods can be found in, e.g., Grob (2003).

Another commonly used estimation technique is the maximum likelihood method. Generally speaking, the likelihood of a set of data is the probability of obtaining that particular set of data, given the chosen probability distribution model. This expression contains the unknown model parameters. The values of these parameters that maximize the sample likelihood are known as the Maximum Likelihood Estimators or simply MLE's.

Maximum likelihood estimation is a totally analytic maximization procedure. It has some very desirable large sample properties:

- they become unbiased minimum variance estimators as the sample size increases.
- they have approximate normal distributions and approximate sample variances that can be calculated and used to generate confidence bounds.
- likelihood functions can be used to test hypotheses about models and parameters.

For the linear regression model (2.17), let us assume the error terms ϵ are normally distributed, i.e., $\epsilon \sim N(0, \sigma^2 I)$. Then, given the observations $(Y_i, \boldsymbol{X}_i)$=$(Y_i, X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$, the likelihood function $L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{Y}, \boldsymbol{X})$ for the regression model is the probability

density function of Y,

$$L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{Y}, \boldsymbol{X}) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2\}.$$

The log likelihood, $l(\boldsymbol{\beta}, \sigma^2; \boldsymbol{Y}, \boldsymbol{X})$, is therefore

$$l(\boldsymbol{\beta}, \sigma^2; \boldsymbol{Y}, \boldsymbol{X}) = -\frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2. \tag{2.20}$$

By taking the derivative on both sides of Equation (2.20) and writing $l \equiv l(\boldsymbol{\beta}, \sigma^2; \boldsymbol{Y}, \boldsymbol{X})$,

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma}(-2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}).$$

Setting $\partial l/\partial \boldsymbol{\beta} = 0$, we obtain the maximum likelihood estimator of $\boldsymbol{\beta}$, which exactly equals the least squares estimator in this case. More details about maximum likelihood estimation can be found in, e.g., Seber and Lee (2003).

Since maximum likelihood estimation relies on the likelihood function, it is crucial to find the likelihood function for a symbolic data regression model.

More detailed studies on linear regression models for classical data can be found in, e.g., Scheffe (1959), Draper and Smith (1981), Montgomery and Peck (1982).

2.5 Current Regression Methods For Interval Data

Different approaches have been introduced to conduct linear regression analyses for symbolic interval-valued data since Billard and Diday (2000) presented the first such approach. By fitting a linear regression model to the center point of the intervals, their model then applied the fitted model to the lower and upper bounds of the independent variables to make predictions for the lower and upper bounds, respectively. Lima Neto et al. (2004) and de Carvalho et al. (2004) transformed the original interval variables into center point and range variables; then they conducted a classical regression analysis on each of the center point variables and range variables separately. Later on, Billard and Diday (2007) introduced a bivariate center and range method, which is an improvement over Lima Neta and de Carvalho's approaches

in the way that it builds the regression model on the center point and range variables simultaneously. To try to solve the problem that a lower bound could be bigger than an upper bound in predictions, Lima Neto et al. (2005, 2010) introduced a constrained model. These methods are briefly reviewed in this section. While each advanced the subject at the time, each has limitations. The most serious limitation is that not all the information in the data is utilized. Therefore, in Chapter 3 we introduce a new method which uses all the information.

Probabilistic assumptions are always an important issue in classical data regression. However, for interval-valued data as well as for other symbolic data, probabilistic assumptions that involve the linear regression model theory are still an open research topic. All current available methods use the least squares method, which is an optimization method that minimizes the sum of the squares of errors, and consequently does not require any probabilistic hypothesis on the variables. A proposed likelihood based approach involving probabilistic assumptions will be considered in Chapter 4.

2.5.1 The Center Method

Billard and Diday (2000) introduced the first approach to fit a regression model to interval-valued data. They built a model on the center points of the intervals; then applied this model to the lower and upper bounds of the intervals of the explanatory variables to make predictions for the response variable. Therefore, this approach is called the center method (CM method). They also made a comparison of the goodness-of-fit between the proposed model and two independent linear models over the lower and upper bounds, respectively. In the end they concluded that their approach performs better.

Assume $X_1, \ldots, X_p$ are p independent interval-valued variables, and Y is the dependent interval-valued variable. Let X_i^c and Y_i^c, $i = 1, \ldots, n$, be the center points of the interval-valued data. Let the observed values of X_j be $X_{ij} = [a_{ij}, b_{ij}]$, and observed values of Y_i be $[c_i, d_i]$, $i = 1, \ldots, n$, and $j = 1, \ldots, p$. Hence, we have

$$X_{ij}^c = (a_{ij} + b_{ij})/2,\ Y_i^c = (c_i + d_i)/2,\ i = 1, \ldots, n,\ j = 1, \ldots, p.$$

The fitted univariate linear regression model is then

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

where $\mathbf{Y}^c$=$(Y_1^c, \ldots, Y_n^c)'$, $\mathbf{X}^c$=$(X_1^c, \ldots, X_n^c)'$, $\boldsymbol{\beta}^c$=$(\beta_0, \beta_1, \ldots, \beta_p)$, and $\boldsymbol{X}_i^c = (1, X_{i1}^c, \ldots, X_{ip}^c)'$ for $i = 1, \ldots, n$.

From classical regression, we know that if $\mathbf{X}^c$ has a full rank $(p+1) \leq n$, then the least squares estimator of $\boldsymbol{\beta}^c$ is given by

$$\hat{\boldsymbol{\beta}}^c = ((\mathbf{X}^c)'\mathbf{X}^c)^{-1}(\mathbf{X}^c)'\mathbf{Y}^c. \tag{2.21}$$

For a new observation, $\mathbf{X}^{new} = (X_1^{new}, \ldots, X_p^{new})$, where $X_j^{new} = [X_{jL}^{new}, X_{jU}^{new}], j = 1, \ldots, p$, the predicted $\hat{Y} = [\hat{Y}_L, \hat{Y}_U]$ is given by

$$\hat{\boldsymbol{Y}}_L = (\mathbf{X}_L^{new})\hat{\boldsymbol{\beta}}^c \tag{2.22}$$

and

$$\hat{\boldsymbol{Y}}_U = (\mathbf{X}_U^{new})\hat{\boldsymbol{\beta}}^c. \tag{2.23}$$

For the first time, Diday and Billard (2000) introduced an approach to deal with interval-valued data linear regression. This approach only takes into account the center points of the intervals, while leaving out other important information such as internal variations of the intervals. It was however a prototype of the interval-valued data regression method.

2.5.2 Center and Range Method

Lima Neto et al. (2004) and de Carvalho et al. (2004) introduced a center and range method (CRM method) to estimate the parameter β using both the center points and ranges of the interval-valued data. In the CRM method, the regression model on the center points is the same as that in the CM method obtain in Section 2.5.1.

Therefore, the univariate linear regression model on the center points is given by

$$\boldsymbol{Y}^c = \boldsymbol{X}^c\boldsymbol{\beta}^c + \boldsymbol{\epsilon}^c. \tag{2.24}$$

The parameter estimates in this center point model can be obtained from Equation (2.21).

However, the CRM method builds an additional independent model on ranges of the intervals. Let X_{ij}^r and Y_i^r, $i = 1, \ldots, n$, $j = 1, \ldots, p$, be the ranges of the interval-valued data. Here, the ranges are calculated as

$$X_{ij}^r = (b_{ij} - a_{ij}), \quad Y_i^r = (Y_{Ui} - Y_{Li}) \quad i = 1, \ldots, n, \quad j = 1, \ldots, p.$$

Then, the univariate linear regression model on the range variables is given by

$$\boldsymbol{Y}^r = \boldsymbol{X}^r \boldsymbol{\beta}^r + \boldsymbol{\epsilon}^r \tag{2.25}$$

where $\boldsymbol{Y}^r = (Y_1^r, \ldots, Y_n^r)'$, $\boldsymbol{X}^r = (X_1^{r\prime}, \ldots, X_n^{r\prime})'$, $X_i^r = (1, X_{i1}^r, \ldots, X_{ip}^r)'$, $i = 1, \ldots, n$, and $\boldsymbol{\beta}^r = (\beta_0^r, \beta_1^r, \ldots, \beta_p^r)$.

From classical regression, we know if $\boldsymbol{X}$ has a full rank $(p+1) \leq n$, then the least squares estimator of $\boldsymbol{\beta}^r$ is given by

$$\hat{\boldsymbol{\beta}}^r = ((\boldsymbol{X}^r)'\boldsymbol{X}^r)^{-1}(\boldsymbol{X}^r)'\boldsymbol{Y}^r. \tag{2.26}$$

For a new observation, $\boldsymbol{X}^{new} = (X_1^{new}, \ldots, X_p^{new})'$, the predicted $\hat{Y} = [\hat{Y}_L, \hat{Y}_U]$ is given by

$$\hat{Y}_L = \hat{Y}^c - \hat{Y}^r/2 \text{ and } \hat{Y}_U = \hat{Y}^c + \hat{Y}^r/2 \tag{2.27}$$

where $\hat{Y}^r = \boldsymbol{X}^{new}\hat{\boldsymbol{\beta}}^r$.

2.5.3 Bivariate Center and Range Method

Billard and Diday (2007) introduced the bivariate center and range method (BCRM method) on the basis of the CRM method. In the BCRM method, the center point and range variables are no longer assumed to be independent. Instead, they are used together to build a regression model simultaneously. The model can be represented by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.28}$$

where $\boldsymbol{Y} = (\boldsymbol{Y}^c, \boldsymbol{Y}^r)$, $\boldsymbol{Y}^c = (Y_1^c, \ldots, Y_n^c)'$ and $\boldsymbol{Y}^r = (Y_1^r, \ldots, Y_n^r)'$; where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)'$, $\boldsymbol{X}_i = (1, X_{i1}^c, \ldots, X_{ip}^c, X_{i1}^r, \ldots, X_{ip}^r)$ for $i = 1, \ldots, n$; and where $\boldsymbol{\beta} = (\beta_0, \beta_1^c, \ldots, \beta_p^c, \beta_1^r, \ldots, \beta_p^r)'$.

Therefore, if $\boldsymbol{X}$ has a full rank $(2p+1) \leq n$, then the least squares estimator of $\boldsymbol{\beta}$ becomes

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \tag{2.29}$$

The predicted value for a new observation $\boldsymbol{X}^{new}$ is given by

$$\hat{Y}_L = \hat{Y}^c - \hat{Y}^r/2 \text{ and } \hat{Y}_U = \hat{Y}^c + \hat{Y}^r/2 \tag{2.30}$$

where $(\hat{Y}^c, \hat{Y}^r) = \boldsymbol{X}^{new}\hat{\boldsymbol{\beta}}$.

Further, Billard and Diday (2007) studied the interactions between center points and ranges. By adding the interactions we obtain a new bivariate center and range model with interaction like $X_i = (1, X_{i1}^c, \ldots, X_{ip}^c, X_{i1}^r, \ldots, X_{ip}^r, X_{i1}^c \times X_{i1}^r, \ldots, X_{ip}^c \times X_{ip}^r)'$.

It is clear that the center and range methods, including the CRM method and the BCRM method, have some improvements from the CM method. They take into account both the center points and ranges of the interval-valued data. However, these methods are in fact classical methods. They divide the interval-valued data into center points and ranges, then build classical models. Another drawback of these methods is that the estimates of ranges could be negative which results in a lower bound being bigger than its corresponding upper bound in the prediction.

2.5.4 Constrained Method

There is another totally different method, the constrained method (CONM method). It was proposed by Lima Neto et al. (2005, 2010) to build linear regression models on interval-valued symbolic data in order to solve the problem that a lower bound could be bigger than an upper bound in predictions. In this approach, Lima Neto and de Carvalho built a regression model under the condition that all parameters must be positive. The Constrained Method includes the constrained center method (CONCM method) and the constrained center and range method (CONCRM method).

For the CONCM method, the model is

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

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ with constraints $\beta_j \geq 0$, $j = 0, \ldots, p$.

For the CONCRM method, the model is

$$\boldsymbol{Y}^c = \boldsymbol{X}^c\boldsymbol{\beta}^c + \boldsymbol{\epsilon}^c \text{ and } \boldsymbol{Y}^r = \boldsymbol{X}^r\boldsymbol{\beta}^r + \boldsymbol{\epsilon}^r$$

with constraints $\boldsymbol{\beta}_j^r \geq 0$, $j = 0, \ldots, p$.

In order to guarantee the positiveness of the least squares estimates of the parameters $\boldsymbol{\beta}$, Lima Neto et al. utilized an algorithm that was originally introduced by Lawson and Hanson (1974). The basic idea of the algorithm is to identify the values incompatible with the restrictions and change them to non-negative values through a re-weighting process. The convergence of the algorithm is proved by Lawson and Hanson. However, in the CONM method all parameters are forced to be non-negative, which obviously sometimes can not reflect the true relationship between the explanatory and response variables. See more details about this method in Lima Neto et al. (2005, 2010).

2.6 References

[1] Bertrand, P. and Goupil, F. (2000): Descriptive Statistics for Symbolic Data. *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday). Springer-Verlag, Berlin, 103-124.

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

[3] Billard, L. (2008). Sample Covariance Functions for Complex Quantitative Data. Processing, World Conferences International Association of Statistical Computing 2008, Yokohama, Japan.

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

[5] Billard, L. and Diday, E. (2006). Descriptive Statistics for Interval-valued Observations in the Presence of Rules. *Computational Statistics*, 21, 187-210.

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

[7] Draper and Smith, (1981). *Applied Regression Analysis*. 2nd ed., John Wiley and Sons.

[8] de Carvalho F.A.T., Lima Neto, E.A. and Tenorio, C.P. (2004). A New Method to Fit a Linear Regression Model for Interval-valued Data. *Lecture Notes in Computer Science, KI2004 Advances in Artifical Inteligence*. Springer-Verlag, 295-306.

[9] Diday, E. (1987). Introduction à l'Approache Symbolique en Analyse des Données. *Premières Journées Symbolique - Numérique*. CEREMADE, Université Paris, 21-56.

[10] Falduti, N. and Taibaly, H. (2004). Etude des Retards sur les Vols des Compagnies Aériennes. CEREMADE, Université Paris, Dauphine, 63.

[11] Grob J. (2003). *Linear Regression*. Springer, Berlin, 33-47.

[12] Lawson, C. and Hanson, R. (1974). *Solving Least Squares Problems*. Prentice-Hall.

[13] Le-Rademacher, J.G. (2008). *Principal Component Analysis for Interval-Valued and Histogram-Valued Data and Likelihood Functions and Some Maximum Likelihood Estimators for Symbolic Data*. University of Georgia.

[14] Lima Neto, E.A and de Carvalho F.A.T. (2010). Constrained Linear Regression Models for Symbolic Interval-valued Variables. *Computational Statistics & Data Analysis*, 54(2), 333-347.

[15] Lima Neto, E.A., de Carvalho F.A.T. and Freire, E.S. (2005). Applying Constrained Linear Aggression Models to Predict Interval-Valued Data. *Lecture Notes in Computer Science, KI: Advances in Artificial Inteligence* (ed. U. Furbach). Springer-Verlag, Brelin, 92-106.

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

[17] Montgomery, D.C. and Peck, E.A. (1982). *Introduction to linear regression analysis*. Wiley, New York.

[18] Scheffe, H. (1959). *The Analysis of Variance*. John Wiley and Sons.

[19] Schweizer, B. (1984). Distributions are the Numbers of the Future. *Proceedings of the Mathematics of Fuzzy Systems Meeting* (eds. A. di Nola and A. Ventre). University of Naples, Naples, Italy, 137-149.

[20] Seber, G.A.F. and Lee A.J. (2003). *Linear Regression Analysis*. Wiley Interseience, 35-51.

[21] Zadeh, L.A. (1965): Fuzzy Sets. *Information and Control* 8(3), 338-353.

2.A Appendix

R Function to Calculate Symbolic Covariance

```
intcov=function(...)
{
    #input can be any number of interval-valued variables;
    d=list(...);
    p=length(d); #number of variables
    m=length(d[[1]][,1]); #number of observations
    x=rep(0,m*p*2);
    dim(x)=c(m,2,p);
    cov=matrix(0,p,p); #covariance matrix;
    cov=matrix(0,p,p); #covariance matrix;
    corr=matrix(0,p,p); #correlation coefficient matrix;
    temp=matrix(0,p,2);
    xbar=c(1:p);   #mean for variables:x1, x2, ..., xp;

    #read in data to x;
    for (i in 1:p)
    {
        x[,,i]=matrix(0,m,2);
        x[,,i]=d[[i]];
    }

    #calcualte mean of variables;
    for (i in 1:p)
    {
        for (j in 1:2)
        {
            temp[i,j]=mean(x[,j,i]);
        }
    xbar[i]=mean(temp[i,]);
    }

    #print(xbar);

    #calculate covariance for variables x_k and x_l

    for (k in 1:p)
    {
        for (l in 1:p)
        {
            sum=0;
```

```
            for (i in 1:m)
            {
                sum=sum+2*(x[i,1,k]-xbar[k])*(x[i,1,l]-xbar[l])
                    +(x[i,1,k]-xbar[k])*(x[i,2,l]-xbar[l])
                    +(x[i,2,k]-xbar[k])*(x[i,1,l]-xbar[l])
                    +2*(x[i,2,k]-xbar[k])*(x[i,2,l]-xbar[l]);
            }
            cov[k,l]=sum/6/m;
        }
    }

    #get lower triangle part of corr matrix;
    for (k in 1:p)
    {
        for (l in k:p)
        {
            corr[k,l]=cov[k,l]/sqrt(cov[k,k])/sqrt(cov[l,l]);
        }
    }

    #get upper triangle part of corr matrix;
    for (k in 2:p)
    {
        for (l in 1:(k-1))
        {
            corr[k,l]=corr[l,k]
        }
    }

   list(cov=cov,corr=corr)
}
```

Chapter 3

Symbolic Covariance Method For Interval Data

Section 2.5 described current methods to conduct linear regression analysis on symbolic interval-valued data. However, all these methods have their limitations as mentioned. In this chapter, we propose a Symbolic Covariance Method (SCM method) that overcomes those limitations. We also introduce a simple adjustment to obviate the need for the constraints proposed by Lima Neto et al. (2005, 2010).

In this chapter, Section 3.1 describes some notations and definitions necessary for the contents in this chapter. Section 3.2 gives the methodology of the SCM method, from the model and its algorithm to how to make predictions. Two examples are included in Section 3.3 to illustrate how to build a regression model using SCM method, and a brief comparison between the proposed method and the classical method is provided.

3.1 Preliminaries

Section 2.3 describes several types of commonly seen symbolic data along with their density function, symbolic sample mean, sample variance and covariance. Some notations and results of symbolic interval-valued data necessary for the proposed SCM method are restated in this section without further details. More extensive treatment of interval-valued data can be found in Billard and Diday (2007) and Billard (2007, 2008).

Let $\boldsymbol{X} = (X_1, \ldots, X_p)'$ be a random vector with p random variables $X_1, \ldots, X_p$. Furthermore, suppose we obtain n observations for each random variable, i.e., for $j = 1, \ldots, p$, we

have

$$\boldsymbol{X}_j = \begin{pmatrix} X_{1j} \\ \dots \\ X_{nj} \end{pmatrix}$$

where X_{ij} denotes the ith observation of the jth variable. The lower case x_{ij} denotes the realization of X_{ij}. Note in the classical data situation, X_{ij} is only a point while in the symbolic data situation, X_{ij} can have an internal structure.

For a regression model, suppose Y is the response variable, $X_1, \dots, X_p$, are the explanatory variables, and $\epsilon's$ denote the errors. Therefore, an interval-valued data matrix for explanatory variables has the following form

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} (a_{11}, b_{11}) & \cdots & (a_{1p}, b_{1p}) \\ \vdots & \ddots & \vdots \\ (a_{n1}, b_{n1}) & \cdots & (a_{np}, b_{np}) \end{pmatrix} \tag{3.1}$$

where $a_{ij} \leq b_{ij}$ for all $i = 1, \dots, n$ and $j = 1, \dots, p$. Likewise, the response variable Y has the form

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} (c_1, d_1) \\ \vdots \\ (c_n, d_n) \end{pmatrix}. \tag{3.2}$$

From Equation (2.2), the empirical density function of X_j is

$$f_{X_j(x)} = \frac{1}{n} \sum_{i; x \in \{x_{ij}\}} (\frac{1}{b_{ij} - a_{ij}}).$$

As defined in Definitions 2.3.9 and 2.3.10, for observations $X_{ij} = [a_{ij}, b_{ij}]$, $i = 1, \dots, n$, $j = 1, \dots, p$, the symbolic sample mean and sample variance of the random variable X_j are, respectively,

$$\bar{X_j} = \frac{1}{2n} \sum_{i=1}^{n} (a_{ij} + b_{ij}) \tag{3.3}$$

and

$$S_j^2 = \frac{1}{3n}\sum_{i=1}^{n}(a_{ij}^2 + a_{ij}b_{ij} + b_{ij}^2) - \frac{1}{4n^2}[\sum_{i=1}^{n}(a_{ij} + b_{ij})]^2.$$

In Definition 2.3.11, Billard (2008) shows that the symbolic covariance between variables X_1 ($[a_{ij}, b_{ij}]$) and X_2 ($[c_{ij}, d_{ij}]$) can be calculated as

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

3.2 Methodology

3.2.1 The Model

As shown in Section 2.4, in classical linear regression, if the model is p dimensional, we have the model as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon. \tag{3.4}$$

Let

$$\beta_0 \equiv \bar{Y} - (\beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 + \cdots + \beta_p \bar{X}_p); \tag{3.5}$$

then, we can center the model in Equation (3.4) as

$$Y - \bar{Y} = \beta_1(X_1 - \bar{X}_1) + \cdots + \beta_p(X_p - \bar{X}_p) + \epsilon \tag{3.6}$$

where the means $\bar{Y}$ and $\bar{X}_j$, $j = 1, \ldots, p$, are obtained by applying Equation (3.3). Then, the least squares estimator of parameter $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = ((\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{X} - \bar{\boldsymbol{X}}))^{-1}(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{Y} - \bar{\boldsymbol{Y}}). \tag{3.7}$$

In the classical data situation, estimates of $\boldsymbol{\beta}$ can be obtained by simply putting observation data values into Equation (3.7). However, in the interval-valued data situation, each observation is no longer a point, but an interval, as shown in Equations (3.1) and (3.2). Therefore, we can not use the least squares estimator directly in the symbolic data situation

as in classical data regression. In order to obtain the values of the elements in the matrices $(\boldsymbol{X}-\bar{\boldsymbol{X}})'(\boldsymbol{X}-\bar{\boldsymbol{X}})$ and $(\boldsymbol{X}-\bar{\boldsymbol{X}})'(\boldsymbol{Y}-\bar{\boldsymbol{Y}})$, we need to reconstruct them as

$$
\begin{aligned}
(\boldsymbol{X}-\bar{\boldsymbol{X}})'(\boldsymbol{X}-\bar{\boldsymbol{X}}) &= \begin{pmatrix} \sum_{i=1}^{n}(X_{1i}-\bar{X}_1)^2 & \cdots & \sum_{i=1}^{n}(X_{1i}-\bar{X}_1)(X_{pi}-\bar{X}_p) \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^{n}(X_{pi}-\bar{X}_p)(X_{1i}-\bar{X}_1) & \cdots & \sum_{i=1}^{n}(X_{pi}-\bar{X}_p)^2 \end{pmatrix}_{p\times p} \\
&= \left(\sum_{i=1}^{n}(X_{j_1 i}-\bar{X}_{j_1})(X_{j_2 i}-\bar{X}_{j_2})\right)_{p\times p} \\
&= \left(n\times Cov(X_{j_1},X_{j_2})\right)_{p\times p} \quad (j_1,j_2=1,\cdots,p) \qquad (3.8)
\end{aligned}
$$

and

$$
\begin{aligned}
(\boldsymbol{X}-\bar{\boldsymbol{X}})'(\boldsymbol{Y}-\bar{\boldsymbol{Y}}) &= \left(\sum_{i=1}^{n}(X_{ji}-\bar{X}_j)(Y_i-\bar{Y})\right)_{p\times 1} \\
&= \left(n\times Cov(X_j,Y)\right)_{p\times 1} \quad (i,j=1,\cdots,p). \qquad (3.9)
\end{aligned}
$$

As introduced in Definition 2.3.11, we already have the definition of covariance function $Cov(X_j,Y)$ where X_j and Y are interval-valued variables. Therefore, we can calculate each element of $Cov(X_{j_1},X_{j_2})$ and $Cov(X_j,Y)$ for j, j_1, $j_2=1,\ldots,p$, in Equations (3.8) and (3.9) from Equation (2.11). Then, we can substitute them into the matrices $(\boldsymbol{X}-\bar{\boldsymbol{X}})'(\boldsymbol{X}-\bar{\boldsymbol{X}})$ and $(\boldsymbol{X}-\bar{\boldsymbol{X}})'(\boldsymbol{Y}-\bar{\boldsymbol{Y}})$ to obtain finally the estimates of the parameters $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)$ as

$$
\hat{\boldsymbol{\beta}} = \left(n\times Cov(X_{j_1 i},X_{j_2 i})\right)^{-1}_{p\times p} \times \left(n\times Cov(X_j,Y)\right)_{p\times 1}. \qquad (3.10)
$$

The proposed SCM method gives a solution that is analogous to the classical linear regression method. It combines the definition of covariance function for symbolic interval-valued data with the least squares estimation method. Therefore, unlike other currently available methods on symbolic interval-valued data that divide the interval data into center points and ranges, this new approach is able to build a model directly on the interval-valued data. On behalf of the symbolic variance and covariance functions that take into account both internal variations and external variations of symbolic interval-valued data, the SCM method is inherently a truly "symbolic" approach.

3.2.2 Algorithm

This section contains a brief description of the algorithm to perform the SCM method for interval-valued regression. As described in Section 3.1, suppose Y is a response variable, X_i, $i = 1, \ldots, p$, are the p explanatory variables, β_i, $i = 1, \ldots, p$, are the parameters we want to estimate. There are five steps in estimating $\boldsymbol{\beta}$=$(\beta_1, \ldots, \beta_p)$ as follows:

Step 1. Calculate the symbolic sample mean for Y and X_j, $j = 1, \ldots, p$, based on the definition of interval-valued data symbolic sample mean in Definition 2.3.9.

Step 2. Center the variables $Y, X_1, \ldots, X_p$ around their means $\bar{Y}, \bar{X}_1, \ldots, \bar{X}_p$ and rewrite the model in Equation (3.4) as in Equation (3.6).

Step 3 Based on the definition of interval-valued data symbolic sample covariance in Definition 2.3.11, calculate the covariance $Cov(Y, X_j)$ and $Cov(X_{j_1}, X_{j_2})$ for j, j_1, $j_2 = 1, \ldots, p$.

Step 4. Substitute the covariances calculated in Step 3 into the Equations (3.8) and (3.9) to obtain $(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{X} - \bar{\boldsymbol{X}})$ and $(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{Y} - \bar{\boldsymbol{Y}})$; then substitute these two matrices into Equation (3.10) to obtain the least squares estimates of $\boldsymbol{\beta}$.

Step 5. Substitute the sample means calculated in Step 1 and estimates of $\boldsymbol{\beta}$ calculated in Step 4 into Equation (3.5) to obtain the estimate of β_0.

A complete function to perform this algorithm in software package **R** is given in Appendix 3.A.

3.2.3 Prediction, Confidence Interval and Model Diagnosis

Regression analysis is a big domain containing far more contents than just parameter estimation. In the classical regression situation, a complete set of procedures has already been established including, such as, model fitting, goodness-of-fit tests, confidence interval calculation and prediction. In the symbolic data regression situation, to establish a similar set of procedures is desirable. This section therefore considers some issues relating to prediction, confidence interval and model diagnosis for symbolic interval-valued data regression.

First, let us consider how to calculate the predicted value and confidence interval of a response variable given explanatory variables in symbolic data regression. There is an illustrative application for this topic. For example, a credit card company has the records about customer's age, income, and monthly expenditure on credit card. The company wants to build a model on the aggregated interval-valued age and income level to predict a customer's monthly expenditure. Then, when a new customer with age, e.g., 35 years and income level, e.g., \$60000, enters the system, it would be of great interest for the company to find out the predicted value and confidence interval of this customer's monthly expenditure.

What is suggested by the current available methods in Section 2.5 is lower and upper bounds of a prediction interval $(\hat{Y}_L, \hat{Y}_U)$ can be found from, for $p = 1$, explanatory variable $X = (X_L^{new}, X_U^{new})$ as

$$\hat{Y}_L = \hat{\beta}_0 + \hat{\beta}_1 X_L^{new}, \quad \hat{Y}_U = \hat{\beta}_0 + \hat{\beta}_1 X_U^{new}.$$

However, all currently available methods, except the constrained model, have a common problem in prediction that the predicted lower bound could be bigger than the predicted upper bound as in, e.g., a univariate regression model with negative slope $\hat{\beta} < 0$.

We suggest a solution to solve this problem in the SCM method by letting the lower bounds equal the minimum values and upper bounds equal the maximum values, viz.,

$$\hat{Y}_L = \min_{X \in (X_L^{new}, X_U^{new})} (\boldsymbol{X}^{new} \hat{\boldsymbol{\beta}}), \quad \hat{Y}_U = \max_{X \in (X_L^{new}, X_U^{new})} (\boldsymbol{X}^{new} \hat{\boldsymbol{\beta}}). \tag{3.11}$$

In Equation (3.11), when a new observation $X^{new} = (X_1^{new}, \ldots, X_p^{new})$ arrives, it is substituted into Equation (3.4), or equivalently, Equation (3.5), to obtain the predicted value $\hat{Y}$ which satisfies

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1^{new} + \cdots + \hat{\beta}_p X_p^{new} \equiv \boldsymbol{X}^{new} \hat{\boldsymbol{\beta}}.$$

It is easy to show that the prediction is then equivalent to

$$\hat{Y}_L = \min(X_L^{new} \hat{\beta}, \quad X_U^{new} \hat{\beta}), \quad \hat{Y}_U = \max(X_L^{new} \hat{\beta}, \quad X_U^{new} \hat{\beta}). \tag{3.12}$$

In addition, this formulation holds not just for $p = 1$ but for any p. Note this formula means we are not always use the lower bounds of independent variables to predict the lower bound of the dependent variable, or the upper bounds of independent variables to predict the upper bound of the dependent variable. When the parameter estimate of β_j is negative, then the lower bound X_{jL} is used to predict the upper bound of Y and the upper bound X_{jU} is used to predict the lower bound of Y.

Secondly, confidence intervals of response variables are of high interest in applications. To obtain a confidence interval or prediction interval of the response variable, we have to make some assumptions about the distribution of the errors. We assume the errors ϵ_i, $i = 1, \ldots, n$, follow a Normal distribution; therefore, it is obvious that Y_i is also from a Normal distribution. Note realizations of errors ϵ_i and Y_i are intervals, so they also have internal distributions, typically assumed to be uniform. In Section 4.2 in the likelihood method, we will have more details about the distribution assumptions.

The original model is specified in Equation (3.4). Take the centered model specified in Equation (3.6), then, for a new observation X^{new}, we have

$$\begin{aligned}\hat{Y} - \bar{Y} &= \hat{\beta}_1(X_1^{new} - \bar{X}_1) + \cdots + \hat{\beta}_p(X_p^{new} - \bar{X}_p) + \epsilon \\ &= (\boldsymbol{X}^{new} - \bar{\boldsymbol{X}})\hat{\boldsymbol{\beta}}.\end{aligned}$$

Therefore,

$$Var(\hat{Y} - \bar{Y}) = Var((\boldsymbol{X}^{new} - \bar{\boldsymbol{X}})\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{new} - \bar{\boldsymbol{X}})' Var(\hat{\boldsymbol{\beta}})(\boldsymbol{X}^{new} - \bar{\boldsymbol{X}}) \tag{3.13}$$

where

$$\begin{aligned}Var(\hat{\boldsymbol{\beta}}) &= Var(\boldsymbol{\epsilon}) \left[(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{X} - \bar{\boldsymbol{X}})\right]^{-1} \\ &= Var(\boldsymbol{\epsilon}) \left[nCov(X_i, X_j)\right]^{-1}_{p \times p}\end{aligned} \tag{3.14}$$

and

$$\boldsymbol{\epsilon} = \epsilon_i, \ i = 1, \ldots, n,$$

where

$$\begin{aligned}\epsilon_i &= (\epsilon_{iL}, \epsilon_{iU}) \\ &= (\hat{Y}_{iL} - Y_{iL}, \hat{Y}_{iU} - Y_{iU}).\end{aligned}$$

Based on the covariance Definition 2.3.11, we can calculate $Var(\boldsymbol{\epsilon})$, which can be substituted into Equation (3.14) to obtain $Var(\hat{\boldsymbol{\beta}})$. Further, $Var(\hat{\boldsymbol{\beta}})$ can be substituted back into Equation (3.13) to obtain $Var(\hat{Y} - \bar{Y})$. Therefore, the $100(1-\alpha)\%$ confidence interval of $\hat{Y}$ is given by

$$[\hat{Y} - t_\alpha \times std(\hat{Y} - \bar{Y}) + \bar{Y}, \hat{Y} + t_\alpha \times std(\hat{Y} - \bar{Y}) + \bar{Y}].$$

Finally, it is commonly of interest to evaluate how well a regression model performs after it is established. Common questions may include, for example, what the residuals are and what portion of total variance is explained by the model.

The residuals can be calculated from observed values and predicted values. Let $Y = (Y_L, Y_U)$ be the observed values and $\hat{Y} = (\hat{Y}_L, \hat{Y}_U)$ be the predicted values from Equation (3.11), then residuals are obtained as

$$Y_L^{residual} = \hat{Y}_L - Y_L, \quad Y_U^{residual} = \hat{Y}_U - Y_U. \tag{3.15}$$

In the classical data situation, the quantity R-square is widely used to do an overall goodness-of-fit evaluation. In the symbolic data situation, the counterpart R-square statistics is derived as follows.

Let Y be an interval-valued response variable. Let $\hat{Y}$ be its prediction value, that is,

$$\hat{Y} = \begin{pmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{pmatrix} = \begin{pmatrix} (\hat{Y}_{1L},\ \hat{Y}_{1U}) \\ \vdots \\ (\hat{Y}_{nL},\ \hat{Y}_{nU}) \end{pmatrix}.$$

Then, the explained sum of squares by the model can be written as

$$\begin{aligned}\sum_i (\hat{Y}_i - \bar{Y})^2 &= \sum_i \left[(\hat{Y}_i - \bar{\hat{Y}}) + (\bar{\hat{Y}} - \bar{Y})\right]^2 \\ &= \sum_i (\hat{Y}_i - \bar{\hat{Y}})^2 + \sum_i (\bar{\hat{Y}} - \bar{Y})^2 + 2\sum_i (\hat{Y}_i - \bar{\hat{Y}})(\bar{\hat{Y}} - \bar{Y}) \\ &= \sum_i (\hat{Y}_i - \bar{\hat{Y}})^2 \\ &= n\ Var(\hat{Y}).\end{aligned}$$

Likewise, by replacing $\hat{Y}$ with $\bar{Y}$, the total sum of squares is derived as

$$\sum (Y_i - \bar{Y})^2 = n\ Var(Y).$$

Hence, R-square can be calculated by

$$\begin{aligned}R^2 &= \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} \\ &= \frac{Var(\hat{Y})}{Var(Y)}.\end{aligned} \tag{3.16}$$

Now, by substituting the observed values of Y and predicted values of Y from Equation (3.11) into the interval-valued variable sample variance formula defined in Definition 2.3.10, we can obtain $Var(\hat{Y})$ and $Var(Y)$, which can be substituted into Equation (3.16) to obtain the R-square value.

3.3 Applications

In this section, the proposed SCM method is applied to two datasets to illustrate its usage. The first dataset records some aspects of appearance properties of 21 different European bat species. It is an example of naturally occurring interval-valued dataset. The second dataset contains human blood pressure records. It is an example of data that involve measurements in which there is an inherent imprecision and so are more accurately recorded as intervals.

In Section 5.6, we apply the proposed method to a simulated dataset and compare the results with current methods. The proposed method clearly outperforms current methods.

3.3.1 Bats Example

The Data

First, we apply the proposed SCM method to the Bats dataset. More detailed information about the background of the dataset can be found in Douzal-Chouakria et al. (2009). In the dataset, 21 species of bats living in Europe are recorded in terms of their minimum and maximum head length, tail length, forearm length and weight. Considering the minimum value for each species as the lower bound and maximum value as the upper bound of an interval, we therefore have four interval-valued variables in the dataset, specifically, X_1=head length, X_2=tail length, X_3=forearm length, and Y=weight. Each observation represents a species. It is our intention to find the linear dependency between the weight and other variables. The dataset is given in Table 3.1.

Table 3.1: Bats Dataset

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

Before any modeling, it is always helpful first to compute the symbolic variance-covariance matrix of the data as defined in Definitions 2.3.10 and 2.3.11 in order to understand better the profile of the dataset. The resulting matrix is given by

$$\mathbf{V} = \begin{pmatrix} 94.1 & 112.9 & 51.8 & 78.1 \\ 112.9 & 155.9 & 77.7 & 97.5 \\ 51.8 & 77.7 & 67.9 & 48.7 \\ 78.1 & 97.5 & 48.7 & 77.4 \end{pmatrix}. \tag{3.17}$$

Here, the covariance matrix V takes elements in the order (Y, X_1, X_2, X_3). Matrix $\mathbf{V}$ of Equation (3.17) shows that X_1, Head Length, has the largest variance (155.9) whereas X_2, Tail Length, has the smallest variance (67.9). The variance for X_3, Forearm Length, is 77.4 and the variance for Y, weight, is 94.1.

In addition, the symbolic correlation matrix as defined in Definition 2.3.12 is also presented by

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.932 & 0.648 & 0.916 \\ 0.932 & 1.000 & 0.755 & 0.888 \\ 0.648 & 0.755 & 1.000 & 0.672 \\ 0.916 & 0.888 & 0.672 & 1.000 \end{pmatrix}. \tag{3.18}$$

The elements of the symbolic correlation matrix in Equation (3.18) indicate strong correlations among weight, Head Length and Forearm Length. The correlation coefficients between each two variables are all bigger than 0.9. The coefficient of correlation between weight and Head Length is 0.932. The coefficient of correlation between weight and Forearm Length is 0.916. The coefficient of correlation between Head Length and Forearm Length is 0.9147. Meanwhile, the correlation matrix of Equation (3.18) shows the correlation coefficient between weight and Tail Length is 0.648 and the correlation coefficient between Tail Length and Forearm Length is 0.672, respectively. The variance-covariance matrix of Equation (3.17) and correlation coefficient matrix of Equation (3.18) may help us to choose the appropriate variables to build a model. From the variance-covariance matrix, there could be

a potential multicollinearity; however, as an example only to illustrate how to conduct the SCM method, we utilize all variables.

Analysis Results

We start with building a full model. Let us use weight (Y) as the response variable and Head Length (X_1), Tail Length (X_2) and Forearm Length (X_3) as the explanatory variables.

Following the algorithm in Section 3.2.2, steps to build the full model are as follows.

Step 1. We calculate the sample mean for Y and X_j, $j = 1, 2, 3$, which gives $\bar{Y} = 13.33$, $\bar{X}_1 = 53.50$, $\bar{X}_2 = 41.76$ and $\bar{X}_3 = 42.60$.

Step 2. We center the variables around their means and write the model as

$$Y - 13.33 = \beta_1(X_1 - 53.50) + \beta_2(X_2 - 41.76) + \beta_3(X_3 - 42.60) + \epsilon. \tag{3.19}$$

Step 3. Using the definition of symbolic covariance as specified in Definition 2.3.11, we calculate the covariances between X_i and Y, $i = 1, 2, 3$, respectively. They are readily presented in Equation (3.17).

Step 4. We substitute the variance and covariance in Equation (3.17) into Equations (3.8) and (3.9) to obtain

$$(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{X} - \bar{\boldsymbol{X}}) = \begin{pmatrix} 155.9 & 77.7 & 97.5 \\ 77.7 & 67.9 & 48.7 \\ 97.5 & 48.7 & 77.4 \end{pmatrix} \times n$$

and

$$(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{Y} - \bar{\boldsymbol{Y}}) = \begin{pmatrix} 112.9 \\ 51.8 \\ 78.1 \end{pmatrix} \times n$$

where $n = 10$ in this dataset. Then, substituting these two matrices into Equation (3.7), we can obtain the estimates of $\boldsymbol{\beta}' = (\beta_1, \beta_2, \beta_3)'$ as

$$\begin{aligned}\beta' &= \left[(X - \bar{X})'(X - \bar{X})\right]' (X - \bar{X})'(Y - \bar{Y}) \\ &= (0.513 \quad -0.156 \quad 0.461)'.\end{aligned}$$

Step 5. We substitute the estimates of $\boldsymbol{\beta}'$ into Equation (3.5) to obtain the estimate of β_0. Finally, the full regression model is given by

$$Y = -27.266 + 0.513X_1 - 0.156X_2 + 0.461X_3.$$

Here, X_1=head length, X_2=tail length, X_3=forearm length and Y=weight.

Furthermore, we can calculate predicted values and confidence intervals of the weight. For a new species, say, if we know the information about their head, tail and forearm length as head length is in the interval [46, 64], tail length [30, 40] and forearm length [40, 50]. Then we can predict the weight range using Equation (3.12) as follows.

We first calculate $\boldsymbol{X}_L\hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}_U\hat{\boldsymbol{\beta}}$ as

$$\begin{aligned}\boldsymbol{X}_L\hat{\boldsymbol{\beta}} &= -27.266 + 0.513 \times 46 - 0.156 \times 40 + 0.461 \times 40 \\ &= -27.266 + 23.598 - 6.24 + 18.44 \\ &= 8.532\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{X}_U\hat{\boldsymbol{\beta}} &= -27.266 + 0.513 \times 64 - 0.156 \times 30 + 0.461 \times 50 \\ &= -27.266 + 32.832 - 4.68 + 23.05 \\ &= 23.936.\end{aligned}$$

Therefore, the predicted value will be, from Equation (3.11)

$$\begin{aligned}\hat{Y}_L &= \min(\boldsymbol{X}_L\hat{\boldsymbol{\beta}}, \quad \boldsymbol{X}_U\hat{\boldsymbol{\beta}}) = 8.532, \\ \hat{Y}_U &= \max(\boldsymbol{X}_L\hat{\boldsymbol{\beta}}, \quad \boldsymbol{X}_U\hat{\boldsymbol{\beta}}) = 23.936.\end{aligned}$$

In order to assess the goodness-of-fit of our proposed model, we calculate the predicted weight value for each of the original observations from Equation (3.11), and based on predicted weight we calculate the residuals from Equation (3.15). The predicted weight values and residuals are shown in Table 3.2 along with the original weight values.

Table 3.2: Predicted Values and Residuals - Bats Dataset

	Response Y		Prediction		Residual	
i	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
1	3	8	-1.93	9.04	-4.93	1.04
2	4	10	2.64	9.04	-1.36	-0.96
3	4	7	2.32	9.23	-1.68	2.23
4	7	8	3.81	8.82	-3.19	0.82
5	6	9	4.32	7.12	-1.68	-1.88
6	7	11	4.32	11.74	-2.68	0.74
7	5	10	5.91	11.07	0.91	1.07
8	8	10	3.79	8.35	-4.21	-1.65
9	8	12	8.37	14.31	0.37	2.31
10	5	10	4.16	14.18	-0.84	4.18
11	7	12	9.03	13.38	2.03	1.38
12	8	13	8.52	12.50	0.52	-0.50
13	6	9	5.07	13.00	-0.93	4.00
14	6	10	7.23	10.59	1.23	0.59
15	12	14	10.61	19.74	-1.39	5.74
16	13	34	17.24	29.58	4.24	-4.42
17	8	16	11.01	16.32	3.01	0.32
18	17	35	19.53	30.74	2.53	-4.26
19	15	40	22.52	31.00	7.52	-9.00
20	18	45	23.98	35.81	5.98	-9.19
21	20	50	34.40	37.56	14.40	-12.44

In addition, the R-square value defined in Equation (3.16) is also provided as

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)} = 0.913.$$

We can see $R^2 = 0.913$, which means the model explains around 91.3% of the total variance. Therefore the model fits well in terms of the R-square statistic.

3.3.2 Blood Pressure Example

The Data

In this example we used another dataset on blood pressure to illustrate the steps to conduct interval-valued data regression using the proposed SCM method. The dataset concerns the records of the pulse rate, systolic blood pressure and diastolic blood pressure for each of eleven patients. More details about the dataset can be found in Billard and Diday (2000). Each data point in the dataset was recorded as an interval, reflecting the fluctuation in these variables for each patient. The aim is to build a linear regression model between blood pressure and pulse rate. Let X_1=systolic pressure, X_2=diastolic pressure and Y=pulse rate. The data are shown in Table 3.3.

Table 3.3: Blood Pressure Dataset

ID	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]

Before we build any models, we want first to compute the symbolic sample variance and covariance matrix defined in Definitions 2.3.10 and 2.3.11 for this dataset. Because systolic pressure should be higher than diastolic pressure, observation 11 is deleted from the following analysis. The variance-covariance matrix is shown in Equation (3.20) where the elements in V are in the order of Y, X_1, X_2; i.e.,

$$\mathbf{V} = \begin{pmatrix} 216 & 277 & 185 \\ 277 & 677 & 350 \\ 185 & 350 & 230 \end{pmatrix}. \tag{3.20}$$

Matrix $\mathbf{V}$ of Equation (3.20) shows that X_1, systolic pressure, has the largest variance (677) whereas Y, pulse rate, and X_2, diastolic pressure, have almost the same variance (216 and 230, respectively). Furthermore, the symbolic correlation matrix defined in Definition 2.3.12 is calculated for this dataset as

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.725 & 0.831 \\ 0.725 & 1.000 & 0.889 \\ 0.831 & 0.889 & 1.000 \end{pmatrix}. \tag{3.21}$$

The elements of the symbolic correlation matrix of Equation (3.21) reveals there is a relatively strong correlation between pulse rate and diastolic pressure. The correlation coefficients between pulse rate and diastolic pressure is 0.831. Systolic pressure and diastolic pressure also seem to have a close relationship with coefficient of correlation 0.889, which is as expected. However, the coefficient of correlation between pulse rate and systolic pressure is relatively small, 0.725. The variance-covariance matrix from Equation (3.20) and correlation coefficient matrix from Equation (3.21) may help us look inside the structure of the data. Because systolic pressure and diastolic pressure are considered to be closely related, we only use one in the regression model. From a paper in American Heart Association, in dynamic exercise, a person's heart rate increases in relation to the intensity of the activity. Systolic blood pressure rises progressively, while diastolic blood pressure stays the same or decreases slightly. At the same time, pulse rate rises, and blood flow to the muscles increases. Therefore, we use systolic pressure as the explanatory variable. The analysis results are presented in the following subsection.

Analysis Results

We use pulse rate (Y) as the response variable and systolic pressure (X_1) as the explanatory variable. We calculate the variance of X_1 from Definition 2.3.10. Using the definition of symbolic covariance in Definition 2.3.11, we calculate the covariances between X_1 and Y. Then, we substitute the variance and covariance into Equations (3.8) and (3.9) to obtain the estimate of $\boldsymbol{\beta}$.

Following the algorithm in Section 3.2.2, in step 1, we calculate the sample mean for Y and X_1. We have $\bar{Y} = 79.1$, $\bar{X}_1 = 132$. In step 2, we center the variables around their means and write the model as

$$Y - 79.1 = \beta_1(X_1 - 132) + \epsilon.$$

In step 3, using the Definition of symbolic covariance as specified in Equation (2.3.11), we calculate the covariances between X_1 and Y, which is readily presented in Equation (3.20). In step 4, we substitute the variance and covariance in Equation (3.20) into Equations (3.8) and (3.9) to give

$$(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{X} - \bar{\boldsymbol{X}}) = 677 \times n$$

and

$$(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{Y} - \bar{\boldsymbol{Y}}) = 277 \times n$$

where $n = 21$ in this dataset. Then, substituting these two into Equation (3.7), we can obtain the estimates of β_1 as

$$\begin{aligned}\beta_1 &= \left[(X - \bar{X})'(X - \bar{X})\right]'(X - \bar{X})'(Y - \bar{Y}) \\ &= 0.41.\end{aligned}$$

In step 5, we substitute the estimates of β_1 into Equation (3.5) to obtain the estimate of β_0. Finally, the full regression model is given by

$$Y = 25.228 + 0.410X_1.$$

Here, Y=pulse rate and X_1=systolic pressure.

Furthermore, we can calculate predicted values and confidence intervals of the pulse rate for new individuals, say, if we know the information about their systolic pressure. For example, suppose there is a subject that has systolic pressure measures [100, 116]. Then we can predict the pulse rate range using Equation (3.12) as follows.

We first calculate $\boldsymbol{X}_L\hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}_U\hat{\boldsymbol{\beta}}$ as

$$\begin{aligned}\boldsymbol{X}_L\hat{\boldsymbol{\beta}} &= 25.228 + 0.410 \times 100 \\ &= 66.228\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{X}_U\hat{\boldsymbol{\beta}} &= 25.228 + 0.410 \times 116 \\ &= 72.788.\end{aligned}$$

Therefore, the predicted value will be, from Equation (3.11)

$$\begin{aligned}\hat{Y}_L &= \min(\boldsymbol{X}_L\hat{\boldsymbol{\beta}}, \quad \boldsymbol{X}_U\hat{\boldsymbol{\beta}}) = 66.228, \\ \hat{Y}_U &= \max(\boldsymbol{X}_L\hat{\boldsymbol{\beta}}, \quad \boldsymbol{X}_U\hat{\boldsymbol{\beta}}) = 72.788.\end{aligned}$$

In order to assess the goodness-of-fit of our proposed model, we calculate the predicted pulse rate values from Equation (3.11) and residuals from Equation (3.15). The predicted pulse rate values and residuals are shown in Table 3.4 along with the original observed values. We found that, from Equation (3.33), R^2 =0.526.

3.4 References

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

Table 3.4: Predicted Values and Residuals - Blood Pressure Dataset

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
1	90	100	44	68	62.10	66.20	18.10	-1.80
2	90	130	60	72	62.10	78.49	2.10	6.49
3	140	180	56	90	82.58	98.97	26.58	8.97
4	110	142	70	112	70.29	83.40	0.29	-28.60
5	90	100	54	72	62.10	66.20	8.10	-5.80
6	130	160	70	100	78.49	90.78	8.49	-9.22
7	130	160	72	100	78.49	90.78	6.49	-9.22
8	110	190	76	98	70.29	103.07	-5.71	5.07
9	138	180	86	96	81.76	98.97	-4.24	2.97
10	110	150	86	100	70.29	86.68	-15.71	-13.32

[2] Billard, L. (2008). Sample Covariance Functions for Complex Quantitative Data. Processing, World Conferences International Association of Statistical Computing 2008, Yokohama, Japan.

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

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

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

[6] Douzal-Chouakria, A., Billard, L. and Diday E. (2009). Principal Component Analysis for Interval-valued Observations. Submitted manuscript.

[7] Lima Neto, E.A and de Carvalho F.A.T. (2010). Constrained Linear Regression Models for Symbolic Interval-valued Variables. *Computational Statistics & Data Analysis*, 54(2), 333-347.

[8] Lima Neto, E.A., de Carvalho F.A.T. and Freire, E.S. (2005). Applying Constrained Linear Aggression Models to Predict Interval-Valued Data. *Lecture Notes in Computer Science, KI: Advances in Artificial Inteligence* (ed. U. Furbach). Springer-Verlag, Brelin, 92-106.

3.A Appendix

R Function to Perform the Symbolic Covariance Method

```
intreg=function(...)
    {
    d=list(...);
    p=length(d);            #number of variables including y and xi's.
    m=length(d[[1]][,1]);   #number of observations
    x=rep(0,m*p*2);
    dim(x)=c(m,2,p);
    cov=matrix(0,p,p);      #covariance matrix;
    corr=matrix(0,p,p);     #correlation coefficient matrix;
    temp=matrix(0,p,2);
    xbar=c(1:p);            #mean for variables:x1, x2, ..., xp;

    #read in data to x;
    for (i in 1:p)
    {
        x[,,i]=matrix(0,m,2);
        x[,,i]=d[[i]];
    }

    #calcualte mean of variables;
    for (i in 1:p)
    {
        for (j in 1:2)
        {
            temp[i,j]=mean(x[,j,i]);
        }
    xbar[i]=mean(temp[i,]);
    }

    print(xbar);

    for (k in 1:p)
    {
        for (l in 1:p)
        {
            sum=0;
            for (i in 1:m)
            {
            sum=sum+2*(x[i,1,k]-xbar[k])*(x[i,1,l]-xbar[l])+(x[i,1,k]
            -xbar[k])*(x[i,2,l]-xbar[l])+(x[i,2,k]-xbar[k])*(x[i,1,l]
```

```
            -xbar[l])+2*(x[i,2,k]-xbar[k])*(x[i,2,l]-xbar[l]);
            }
            cov[k,l]=sum/6/m;
        }
    }

    #get lower triangle part of corr matrix;
    for (k in 1:p)
    {
        for (l in k:p)
        {
            corr[k,l]=cov[k,l]/sqrt(cov[k,k])/sqrt(cov[l,l]);
        }
    }

    #get upper triangle part of corr matrix;
    for (k in 2:p)
    {
        for (l in 1:(k-1))
        {
            corr[k,l]=corr[l,k]
        }
    }

    covxy=matrix(cov[-1,1],(p-1),1);
    covxx=matrix(cov[-1,-1],(p-1),(p-1));
    beta=solve(covxx)%*%covxy;
    dbar=c(1:p);

    for (i in 1:p)
    {
        dbar[i]=mean(d[[i]])
    }

    ybar=dbar[1];
    xbar=dbar[2:p];
    beta0=ybar-t(beta)%*%xbar;
    betaall=rbind(beta0,beta);

    list(beta=betaall)
}
```

Chapter 4

Symbolic Likelihood Method for Interval Data

In Chapter 3, we introduced a symbolic regression method for interval-valued data, with the regression parameters estimated by the well-known least squares procedure. This then involves using the moment estimators for underlying covariance terms for interval data. In this chapter, we want to show that those estimators are also maximum likelihood estimators. Therefore, after some preliminary results in Section 4.1, we establish a basic symbolic likelihood function for interval-valued data in Section 4.2. This is then used in Section 4.3 to obtain maximum likelihood estimators for the model parameters.

4.1 Preliminaries

Among many parameter estimation techniques, maximum likelihood estimation stands out as a crucial one in statistics. More details about its general introduction can be found in, e.g., Casella and Berger (2002). In this chapter, we derive the symbolic maximum likelihood estimators for a linear regression model assuming normality of the error terms, without proof or much explanation. The goal is to establish a maximum likelihood estimation method for symbolic data regression.

In a classical conditional model (Casella and Berger, 2002), the observed data are the n pairs, $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$. We consider the values of the explanatory variables, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, to be known, where $X_i = (X_{i1}, \ldots, X_{ip})$, while the values of the response variables, $y_1, \ldots, y_n$, are observed values of random variables, $Y_1, \ldots, Y_n$. For simplicity, let us take the simple linear regression model, i.e., the number of explanatory variables is $p = 1$. We assume the distribution of the error terms and hence of the $Y_i's$ is a normal distribution,

specifically,

$$Y_i \sim N(\alpha + \beta X_i, \sigma^2), \quad i = 1, \ldots, n. \tag{4.1}$$

Therefore, the population regression function is a linear function of X, i.e., $E(Y|X) = \alpha+\beta X$, and we assume all the $Y_i's$ have the same variance σ^2. The model is

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{4.2}$$

where $\epsilon_i = 1, \ldots, n$, are independent and identically distributed $N(0, \sigma^2)$ random variables. Then, the likelihood function, which equals the joint probability density (pdf) function of $Y_1, \ldots, Y_n$, is given as

$$L(\beta, \sigma^2|Y, X) = \prod_{i=1}^{n} f_{Y_i}(Y_i|X_i, \beta, \sigma^2) \tag{4.3}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(Y_i - \alpha - \beta X_i)^2}{2\sigma^2}) \tag{4.4}$$

$$= (\frac{1}{\sqrt{2\pi}\sigma})^2 \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2). \tag{4.5}$$

The maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2 are simply the values of $\boldsymbol{\beta}$ and σ^2 that maximize the likelihood function in Equation (4.5). However, the values that maximize the likelihood function will also maximize the logarithm of the likelihood, viz.,

$$\ln(L(\boldsymbol{\beta}, \sigma^2|Y, X) = -\frac{n}{2} 2\ln(\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2. \tag{4.6}$$

From Equation (4.6), it is clear that maximizing the log likelihood over $\boldsymbol{\beta}$ is the same as minimizing the third item, which, apart from constants, is the residual sum of squares function. Therefore, the maximum likelihood estimators for the normal linear regression model is the same as the ordinary least-square estimators. In particular, the maximum likelihood estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{[\sum_{i=1}^{n}(X_i - \bar{X})^2]^{1/2}} = \frac{\widehat{Cov(X, Y)}}{[\widehat{Var(X)}]^{1/2}} \tag{4.7}$$

where $\widehat{Cov(X,Y)}$ is the sample covariance function between the random variables X and Y, $\widehat{Var(X)}$ is the sample variance of X, and where

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{4.8}$$

Since the numerator in Equation (4.7) equals the covariance function, once the maximum likelihood estimator is established for the bivariate random variables (X, Y), the likelihood basis for $\hat{\boldsymbol{\beta}}$ is established. This is obtained by extending the likelihood function proposed by Le-Rademacher and Billard (2010) for a univariate interval-valued random variable to a bivariate case for interval-valued observations.

4.2 Symbolic Likelihood Function

Suppose we have two random variables (X, Y) with realizations (X_i, Y_i) where $X_i = \xi_{ix} = [c_i, d_i]$ and $Y_i = \xi_{iy} = [a_i, b_i]$ for $i = 1, \ldots, n$. Let (X_i, Y_i) have joint (bivariate) distribution $H_{X,Y}(x, y; \delta) = P\{(X, Y) \leq (x, y)\}$. Hence, the joint probability density function (pdf) is

$$h_{X,Y}(x, y; \delta) = P_\delta\{(X, Y) = (x, y)\}. \tag{4.9}$$

Note that when (X, Y) takes symbolic values, the probability on the right-hand side of Equation (4.9) is non-zero typically since the realizations of $(x, y) = (\xi_x, \xi_y)$ are intervals rather than points in a 2-dimensional space $\mathcal{R}^2$.

Since each variable takes values over an interval, we need to consider the internal distribution of those values. Therefore, let the internal distributions be defined by, respectively,

$$\text{Within } X_i \sim f_i^x(\xi_{ix}; \boldsymbol{\Theta}_i^x), \quad \text{Within } Y_i \sim f_i^y(\xi_{iy}; \boldsymbol{\Theta}_i^y), \quad i = 1, \ldots, n, \tag{4.10}$$

and joint Within (X, Y) distribution $f_i^{xy}(\xi_{ix}, \xi_{iy}; \boldsymbol{\Theta}_i^x, \boldsymbol{\Theta}_i^y)$. These distributions f_i^x, f_i^y and f_i^{xy} along with their parameters $\boldsymbol{\Theta}_i^x$ and $\boldsymbol{\Theta}_i^y$ are internal (or "within" observation) entities, distinct from the overall distributions $h_{X,Y}(x, y; \delta)$.

Also, since (X_i, Y_i), $i = 1, \ldots, n$, are random variables, the parameters $(\boldsymbol{\Theta}_i^x, \boldsymbol{\Theta}_i^y)$ are not fixed, taking variable values as (X_i, Y_i) change with $i = 1, \ldots, n$. That is, $(\boldsymbol{\Theta}_i^x, \boldsymbol{\Theta}_i^y)$ are

themselves random variables. Therefore, let the underlying distribution of $(\mathbf{\Theta}_i^x, \mathbf{\Theta}_i^y)$ be

$$(\mathbf{\Theta}_i^x, \mathbf{\Theta}_i^y) \sim g_{xy}(\mathbf{\Theta}_i^x, \mathbf{\Theta}_i^y; \tau_{xy}),\ i = 1, \ldots, n; \tag{4.11}$$

i.e., we can write

$$g_{xy}(\mathbf{\Theta}_i^x, \mathbf{\Theta}_i^y; \tau_{xy}) = P_{\tau_{xy}}\{(\mathbf{\Theta}_i^x, \mathbf{\Theta}_i^x) = (\theta_i^x, \theta_i^y)\},\ i = 1, \ldots, n. \tag{4.12}$$

Then for these parametric families $g_{xy}(\mathbf{\Theta}_i^x, \mathbf{\Theta}_i^y; \tau_{xy})$, there exist one-to-one correspondences between the (X_i, Y_i) and $(\mathbf{\Theta}_i^x, \mathbf{\Theta}_i^y)$. Hence,

$$P_\delta\{(X_i, Y_i) = (\xi_i^x, \xi_i^y)\} = P_\tau\{(\mathbf{\Theta}_i^x, \mathbf{\Theta}_i^y) = (\theta_i^x, \theta_i^y)\}. \tag{4.13}$$

Therefore, substituting into Equation (4.13) from Equation (4.9) and Equation (4.12), we have

$$h_{X,Y}(x, y; \delta) = g_{xy}(\mathbf{\Theta}_i^x, \mathbf{\Theta}_i^y; \tau_{xy}). \tag{4.14}$$

We can write the likelihood function of the parameter δ given the data $(X_i, Y_i) = (\xi_i^x, \xi_i^y),\ i = 1, \ldots, n$, as

$$L(\delta; \xi_1^x, \ldots, \xi_n^x, \xi_1^y, \ldots, \xi_n^y) = \prod_{i=1}^{n} h_{X,Y}(x_i, y_i; \delta); \tag{4.15}$$

and substituting from Equation (4.14), we have that

$$L(\delta; \xi_1^x, \ldots, \xi_n^x, \xi_1^y, \ldots, \xi_n^y) = \prod_{i=1}^{n} g_{xy}(\theta_i^x, \theta_i^y; \tau_{xy}) = L(\tau_{xy}; \theta_1^x, \ldots, \theta_n^x \theta_i^y, \ldots, \theta_n^y). \tag{4.16}$$

Note that when there is independence, the product of the marginal distributions can replace the joint distribution of Equation (4.12) in Equation (4.14), i.e., we have for $\mathbf{\Theta}_i^x$ and $\mathbf{\Theta}_i^y$, respectively,

$$\mathbf{\Theta}_i^x \sim g^x(\mathbf{\Theta}_i^x; \tau^x), \quad \mathbf{\Theta}_i^y \sim g^y(\mathbf{\Theta}_i^y; \tau^y), \quad i = 1, \ldots, n,$$

where now we can write $g^x(\mathbf{\Theta}_i^x; \tau^x) = P_{\tau^x}(\mathbf{\Theta}_i^x = \theta_i^x)$ and $g^y(\mathbf{\Theta}_i^y; \tau^y) = P_{\tau^y}(\mathbf{\Theta}_i^y = \theta_i^y)$.

Since the data $(\theta_1^x, \ldots, \theta_n^x, \theta_i^y, \ldots, \theta_n^y)$ in Equation (4.16) are now classically valued observations, we can apply well established maximum likelihood methods for classical data to our situation. This is done in the following Section 4.3.

4.3 Maximum Likelihood Estimators

In Section 4.3.1, we obtain the maximum likelihood estimators for the within observation parameters. Then in Section 4.3.2, these are used to obtain the overall (within plus between) variation estimators.

4.3.1 Estimators for Within Observation Parameters

The internal parameters $\mathbf{\Theta}_i^x$ associated with X_i and $\mathbf{\Theta}_i^y$ associated with Y_i, $i = 1, \ldots, n$, introduced in Section 4.2 can be of any dimension, e.g., p. In this section, we take them to be p=2 or 3 dimensional. In particular, we take $\mathbf{\Theta}_{i1} = (\Theta_{i1}^x, \Theta_{i1}^y)$ and $\mathbf{\Theta}_{i2} = (\Theta_{i2}^x, \Theta_{i2}^y, \Theta_{i2}^{xy})$. Thus, the Θ_{i1}^x and Θ_{i1}^y correspond to the internal mean of X_i and Y_i, respectively, for each $i = 1, \ldots, n$; and the Θ_{i2}^x and Θ_{i2}^y correspond to the internal variation of X_i and Y_i, respectively, and Θ_{i2}^{xy} corresponds to the covariance of (X_i, Y_i) for each $i = 1, \ldots, n$. It is not unreasonable to assume the $\mathbf{\Theta}_{i1}$ and $\mathbf{\Theta}_{i2}$ are independent. [The case where these might be dependent will be considered as future work.]

For each $i = 1, \ldots, n$, let us suppose that the joint distribution of the internal means $\mathbf{\Theta}_{i1} = (\Theta_{i1}^x, \Theta_{i1}^y)$ is a bivariate normal distribution $N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$; and let us suppose $\mathbf{\Theta}_{i2} = (\Theta_{i2}^x, \Theta_{i2}^y, \Theta_{i2}^{xy})$ follows the bivariate exponential distribution $f(\theta_{i2}^x, \theta_{i2}^y, \theta_{i2}^{xy}; \gamma_1, \gamma_2, \gamma_3)$ defined by

$$f(x, y; \gamma) = C \frac{(1-\gamma_3)}{\gamma_1 \gamma_2} e^{1/(1-\gamma_3)} \times e^{-x/\gamma_1 - y/\gamma_2 - z/(1-\gamma_3)}, \quad x > 0,\ y > 0,\ z > 0, \tag{4.17}$$

where in Equation (4.17), $x, y, z \equiv xy$ correspond, respectively, to $\theta_{i2}^x, \theta_{i2}^y, \theta_{i2}^{xy}$.

This bivariate exponential distribution is based on an exponential distribution function introduced by Gumbel (1960). Other bivariate exponential distributions have been studied by, e,g., Marshall and Olkin (1967), Freund (1961), Plackett (1965), and Gumbel (1958). Explicit formulations of the pdf $f(\cdot)$ are complicated and among other terms involve the so-called exponential integral,

$$\int_0^\infty \frac{e^{-at}}{(b+t)} dt = e^{ab} E_1(ab),$$

see Abramozwitz and Stegun (1970, Equation(5.1.28)). Abramozwitz and Stegun (1970, Equation(5.1.51)) also give the asymptotic expansion

$$E_1(x) \simeq \frac{e^{-x}}{x}[1 - \frac{1}{x} + \frac{2!}{x^2} - \frac{3!}{x^3} + \dots], \quad |\arg x| < 2\pi/3.$$

For our purposes, it suffices to use the multiplier C of Equation (4.17).

Then, the joint probability density function $g_{xy}(\mathbf{\Theta}_i^x, \mathbf{\Theta}_i^y; \tau_{xy})$ of Equation (4.11) can be written as, with $\tau_{xy} = (\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho, \gamma_1, \gamma_2, \gamma_3)$,

$$g_{xy}(\mathbf{\Theta}_{i1}, \mathbf{\Theta}_{i2}; \tau_{xy}) = g_1(\Theta_{i1}^x, \Theta_{i1}^y; \mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho) \times g_2(\Theta_{i2}^x, \Theta_{i2}^y, \Theta_{i2}^{xy}; \gamma_1, \gamma_2, \gamma_3). \tag{4.18}$$

Suppose the internal distributions within the intervals, i.e., the $f_i^x(\xi_{ix}; \mathbf{\Theta}_i^x)$ and $f_i^y(\xi_{iy}; \mathbf{\Theta}_i^y)$ of Equation (4.10), are uniformly distributed, for each $i = 1, \dots, n$. Hence, for the intervals $Y_i = \xi_i^y = (a_i, b_i)$, realizations of Θ_{i1}^y are $\theta_{i1}^y = (a_i + b_i)/2$; and likewise for the intervals $X_i = \xi_i^x = (c_i, d_i)$ realizations of Θ_{i1}^x are $\theta_{i1}^x = (c_i + d_i)/2$. The variation variables $\Theta_{i2}^x, \Theta_{i2}^y, \Theta_{i2}^{xy}$ have realizations $\theta_{i2}^y = (b_i - a_i)^2/12$, $\theta_{i2}^x = (d_i - c_i)^2/12$ and $\theta_{i2}^{xy} = (b_i - a_i)(d_i - c_i)/12$, respectively.

Let us write the likelihood function based on the intervals, L_I, from Equation (4.16) with $\mathbf{\Theta}_n$ representing the observations, as

$$\begin{aligned} L_I \equiv L_I(\tau; \mathbf{\Theta}_n) \equiv L_I(&\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho, \gamma_1, \gamma_2, \gamma_3; \\ &\theta_{11}^x, \dots, \theta_{n1}^x, \theta_{11}^y, \dots, \theta_{n1}^y, \theta_{12}^x, \dots, \theta_{n2}^x, \theta_{12}^y, \dots, \theta_{n2}^y, \theta_{12}^{xy}, \dots, \theta_{n2}^{xy}). \end{aligned}$$

Then, from Equation (4.18), the likelihood function can be written as

$$L_I = L_{I1} \times L_{I2} \tag{4.19}$$

where

$$\begin{aligned} L_{I1} &= \prod_{i=1}^{n} g_1((c_i + d_i)/2, (a_i + b_i)/2; \mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho) \qquad (4.20) \\ &= \prod_{i=1}^{n} (2\pi\sigma_x\sigma_y(1-\rho^2)^{1/2}) \exp\{\frac{1}{2(1-\rho^2)}[(\frac{\theta_{i1}^x - \mu_x}{\sigma_x})^2 \\ &\quad + (\frac{\theta_{i1}^y - \mu_y}{\sigma_y})^2 - 2\rho\frac{(\theta_{i1}^x - \mu_x)(\theta_{i1}^y - \mu_y)}{\sigma_x\sigma_y}]\} \end{aligned}$$

and

$$L_{I2} = \prod_{i=1}^{n} g_2((d_i - c_i)^2/12, (b_i - a_i)^2/12, (d_i - c_i)(b_i - a_i)/12; \gamma_1, \gamma_2, \gamma_3) \tag{4.21}$$
$$= \prod_{i=1}^{n} C \frac{(1-\gamma_3)}{\gamma_1 \gamma_2} e^{1/(1-\gamma_3) - (d_i-c_i)^2/(12\gamma_1) - (b_i-a_i)^2/(12\gamma_2) - (d_i-c_i)(b_i-a_i)/(12(1-\gamma_3))}$$

Hence, the log likelihood function $\ln L_I$ is

$$\begin{aligned}\ln L_I \propto & -n\ln(\sigma_x^2) - n\ln(\sigma_y^2) - (n/2)\ln(1-\rho^2) - \frac{1}{2(1-\rho^2)}\sum_{i=1}^{n}[(\frac{(c_i+d_i)/2-\mu_x}{\sigma_x})^2 \\ & + (\frac{(a_i+b_i)/2-\mu_y}{\sigma_y})^2 - 2\rho\frac{((c_i+d_i)/2-\mu_x)((a_i+b_i)/2-\mu_y)}{\sigma_x\sigma_y}] \\ & + n\ln(1-\gamma_3) - n\ln(\gamma_1) - n\ln(\gamma_2) + n/(1-\gamma_3) \\ & - \frac{1}{\gamma_1}\sum_{i=1}^{n}\frac{(d_i-c_i)^2}{12} - \frac{1}{\gamma_2}\sum_{i=1}^{n}\frac{(b_i-a_i)^2}{12} - \frac{1}{(1-\gamma_3)}\sum_{i=1}^{n}\frac{(d_i-c_i)(b_i-a_i)}{12}. \end{aligned} \tag{4.22}$$

Then successively differentiating $\ln L_I$ with respect to each of the eight parameters in τ, we obtain

$$\frac{\partial L_I}{\partial \mu_x} = \frac{1}{2(1-\rho^2)\sigma_x^2}\sum_{i=1}^{n}[(c_i+d_i)/2-\mu_x], \tag{4.23}$$

$$\frac{\partial L_I}{\partial \mu_y} = \frac{1}{2(1-\rho^2)\sigma_y^2}\sum_{i=1}^{n}[(a_i+b_i)/2-\mu_y], \tag{4.24}$$

$$\begin{aligned}\frac{\partial L_I}{\partial \sigma_x} = & \frac{-n}{\sigma_x} + \frac{1}{2(1-\rho^2)}[\sum_{i=1}^{n}\frac{[(c_i+d_i)/2-\mu_x]^2}{\sigma_x^3} \\ & + 2\rho\sum_{i=1}^{n}\frac{[(c_i+d_i)/2-\mu_x][(a_i+b_i)/2-\mu_y]}{\sigma_x^2\sigma_y}] \end{aligned} \tag{4.25}$$

$$\begin{aligned}\frac{\partial L_I}{\partial \sigma_y} = & \frac{-n}{\sigma_y} + \frac{1}{2(1-\rho^2)}[\sum_{i=1}^{n}\frac{[(a_i+b_i)/2-\mu_y]^2}{\sigma_y^3} \\ & + 2\rho\sum_{i=1}^{n}\frac{[(c_i+d_i)/2-\mu_x][a_i+b_i)/2-\mu_y]}{\sigma_x^2\sigma_y}], \end{aligned} \tag{4.26}$$

$$\frac{\partial L_I}{\partial \rho} = \frac{n\rho}{(1-\rho^2)} - \frac{\rho}{(1-\rho^2)^2} \times \{\sum_{i=1}^{n}[\frac{(c_i+d_i)/2-\mu_x}{\sigma_x}]^2 + \sum_{i=1}^{n}[\frac{(a_i+b_i)/2-\mu_y}{\sigma_y}]^2 - 2\rho\sum_{i=1}^{n}\frac{[(c_i+d_i)/2-\mu_x][(a_i+b_i)/2-\mu_2]}{\sigma_x\sigma_y}\} - \frac{1}{(1-\rho^2)}\sum_{i=1}^{n}\frac{[(c_i+d_i)/2-\mu_x][(a_i+b_i)/2-\mu_2]}{\sigma_x\sigma_y}, \tag{4.27}$$

$$\frac{\partial L_I}{\partial \gamma_1} = \frac{1}{\gamma_1} - \frac{1}{\gamma_1^2}\sum_{i=1}^{n}(d_i-c_i)^2/12, \tag{4.28}$$

$$\frac{\partial L_I}{\partial \gamma_2} = \frac{1}{\gamma_2} - \frac{1}{\gamma_2^2}\sum_{i=1}^{n}(b_i-a_i)^2/12, \tag{4.29}$$

$$\frac{\partial L_I}{\partial \gamma_3} = \frac{-n\gamma_3}{(1-\gamma_3)^2} + \frac{1}{(1-\gamma_3)^2}\sum_{i=1}^{n}(d_i-c_i)(b_i-a_i)/12. \tag{4.30}$$

Then, substituting the relevant maximum likelihood estimator and setting the derivatives to zero, we obtain the maximum likelihood estimators $\hat{\tau}_{xy} = (\hat{\mu}_x, \hat{\mu}_y, \hat{\sigma}_x^2, \hat{\sigma}_y^2, \hat{\rho}, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3)$ for $\tau_{xy} = (\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho, \gamma_1, \gamma_2, \gamma_3)$ to be given by

$$\hat{\mu}_x = \frac{1}{2n}\sum_{i=1}^{n}(c_i+d_i); \tag{4.31}$$

$$\hat{\mu}_y = \frac{1}{2n}\sum_{i=1}^{n}(a_i+b_i); \tag{4.32}$$

$$\hat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{n}[(c_i+d_i)/2-\hat{\mu}_x]^2; \tag{4.33}$$

$$\hat{\sigma}_y^2 = \frac{1}{n}\sum_{i=1}^{n}[(a_i+b_i)-\hat{\mu}_y]^2; \tag{4.34}$$

$$\hat{\rho} = \frac{\sum_{i=1}^{n}[(a_i+b_i)/2-\hat{\mu}_y][(c_i+d_i)/2-\hat{\mu}_x]}{\{\sum_{i=1}^{n}[(a_i+b_i)/2-\hat{\mu}_y]^2\sum_{i=1}^{n}[(c_i+d_i)/2-\hat{\mu}_x]^2\}^{1/2}}, \tag{4.35}$$

that is, $\hat{\rho} = \hat{\sigma}_{xy}/\hat{\sigma}_x\hat{\sigma}_y$, where $\sigma_{xy} = Cov(X,Y)$, and hence, the estimator for the covariance is

$$\hat{\sigma}_{xy} = \frac{1}{n}\sum_{i=1}^{n}[(a_i+b_i)/2-\hat{\mu}_y][(c_i+d_i)/2-\hat{\mu}_x]; \tag{4.36}$$

$$\hat{\gamma}_1 = \frac{1}{n}\sum_{i=1}^{n}(d_i - c_i)^2/12; \tag{4.37}$$

$$\hat{\gamma}_2 = \frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2/12; \tag{4.38}$$

and

$$\hat{\gamma}_3 = \frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)(d_i - c_i)/12. \tag{4.39}$$

These estimators relate to the within interval means and variations given the observations $Y_i = \xi_{ix} = [a_i, b_i]$ and $X_i = \xi_{ix} = [c_i, d_i]$, $i = 1, \ldots, n$. Thus, for example, the $\hat{\sigma}_y^2$ estimates the variance of the means of the interval values for Y; that is, this is the so-called within observations variance. Likewise, the estimator $\hat{\sigma}_{xy}$ estimates the within observations covariance of (X, Y). We also note that instead of solving the partial derivative in Equation (4.27) for the derivation of the estimator $\hat{\rho}$, we can more easily obtain the result of Equation (4.35) by following, e.g., Casella and Berger (2002, p.358) who suggest using a partially maximized likelihood function.

4.3.2 Overall Parameter Estimators

In this section, we obtain the overall means, variances and covariance of the variables X, Y. It is assumed that the internal within interval observations have the distributions used in Section 4.3. To obtain the overall descriptive statistics, we need conditional expectations.

First, a number of sources provide expressions for a conditional mean and conditional variance; see, e.g., Casella and Berger (2002). Suppose therefore we have two random variables X and U. Then the mean of X, written $E_X(X)$, can be expressed in terms of a conditioning random variable U as

$$E_X(X) = E_U(E_X(X|U)); \tag{4.40}$$

and similarly the variance of X, written $Var_X(X)$, as

$$Var_X(X) = E_U(Var_X(X|U)) + Var_U(E_X(X|U)). \tag{4.41}$$

We can also show that the covariance of two random variables $X,\ Y$ can be expressed in terms of two conditioning random variables U (on X) and V (on Y) as

$$Cov_{XY}(X,Y) = E_{UV}\{Cov_{XY}(X|U,Y|V)\} + Cov_{UV}\{E_X(X|U), E_Y(Y|V)\} \tag{4.42}$$

This can be verified as follows. Let us write the covariance function as

$$\begin{aligned}
&Cov_{XY}(X,Y) \\
&\quad = E_{XY}[(X - E_X(X))(Y - E_Y(Y)) \\
&\quad = E_{XY}\{[X - E_X(X|U) + E_X(X|U) - E_X(X)] \\
&\qquad\qquad \times [Y - E_Y(Y|V) + E_Y(Y|V) - E_Y(Y)]\} \\
&\quad = E_{XY}\{[X - E_X(X|U)][Y - E_Y(Y|V)]\} \\
&\qquad + E_{XY}\{[E_X(X|U) - E_X(X)][E_Y(Y|V) - E_Y(Y)]\} \\
&\qquad + E_{XY}\{[X - E_X(X|U)][E_Y(Y|V) - E_Y(Y)]\} \\
&\qquad + E_{XY}\{[Y - E_Y(Y|V)][E_X(X|U) - E_X(X)]\} \\
&\quad = E_1 + E_2 + E_3 + E_4, \quad say.
\end{aligned} \tag{4.43}$$

The first term E_1 in Equation (4.43) becomes

$$\begin{aligned}
E_1 &= E_{XY}\{[X - E_X(X|U)][Y - E_Y(Y|V)]\} \\
&= E_{UV}(E_{XY}\{[X - E_X(X|U)][Y - E_Y(Y|V)]|UV\}) \\
&= E_{UV}\{Cov_{XY}(X|U,Y|V)\}
\end{aligned} \tag{4.44}$$

The second term E_2 in Equation (4.43) becomes

$$\begin{aligned}
E_2 &= E_{XY}\{[E_X(X|U) - E_X(X)][E_Y(Y|V) - E_Y(Y)]\} \\
&= E_{XY}\{[E_X(X|U) - E_U\{E_X(X|U)\}][E_Y(Y|V) - E_V\{E_Y(Y|V)\}]\} \\
&= E_{UV}(E_{XY}\{[E_X(X|U) - E_U\{E_X(X|U)\}][E_Y(Y|V) - E_V\{E_Y(Y|V)]\}|UV\}) \\
&= E_{UV}\{[E_X(X|U) - E_U\{E_X(X|U)\}][E_Y(Y|V) - E_V\{E_Y(Y|V)\}]\} \\
&= Cov_{UV}\{E_X(X|U), E_Y(Y|V)\}
\end{aligned} \tag{4.45}$$

The third term E_3 in Equation (4.43) becomes

$$\begin{aligned} E_3 &= E_{XY}\{[X - E_X(X|U)][E_Y(Y|V) - E_Y(Y)]\} \\ &= E_{UV}(E_{XY}\{[X - E_X(X|U)][E_Y(Y|V) - E_Y(Y)]\}|UV) \\ &= \{E_Y(Y|V) - E_Y(V)\}\{E_{UV}[E_{XY}\{X - E_X(X|U)\}|UV]\} \\ &= \{E_Y(Y|V) - E_Y(V)\}\{E_{UV}[E_X(X|UV)] - E_{UV}[E_X(X|UV)]\} \\ &= \{E_Y(Y|V) - E_Y(V)\}\{E_X(X) - E_X(X)\} = 0. \end{aligned} \tag{4.46}$$

Similarly, the fourth term $E_4 = 0$. Hence, on substituting back into Equation (4.43), the result of Equation (4.42) follows.

We can use these results to obtain the overall statistics. For clarity, let us denote the overall X variable by W^x to distinguish it from the conditional X_i values. Likewise, let W^y be the overall Y variable. [Here, the $X,\ Y$ variables are the same as that in Section 4.2 with distribution function $H_{X,Y}(x, y; \delta)$.] Then, we have given sets of values of W^x in $X_i = \xi_{ix}$, and given sets of W^y in $Y_i = \xi_{iy}$, with conditional distributions $f_{W^x}(w^x|\xi_{ix})$, with $w^x \in \xi_{ix}$ and $f_{W^y}(w^y|\xi_{iy})$, with $w^y \in \xi_{iy}$, $i = 1, \ldots, n$, and joint distribution $f_{W^x,W^y}(w^x, w^y|\xi_{ix}, \xi_{iy})$ for $(w^x, w^y) \in (\xi_{ix}, \xi_{iy})$.

For the within interval variable X_i, the mean was taken to be Θ^x_{i1}, so that

$$\Theta^x_{i1} = E_{W^x}(W^x|X_i = \xi_{ix}). \tag{4.44}$$

Hence, from Equation (4.40), we have

$$E_{W^x}(W^x) = E_{X_i}[E_{W^x}(W^x|X_i = \xi_{ix}) = E_{X_i}(\Theta^x_{i1}). \tag{4.45}$$

However, the $\mathbf{\Theta}_{i1} = (\Theta^x_{i1}, \Theta^y_{i1})$ were assumed to follow a bivariate normal distribution $N_2(\mu_x, \mu_y, \sigma^2_x, \sigma^2_y, \rho)$ (see Section 4.3.1). Hence, Θ^x_{i1} follows a normal distribution $N(\mu_x, \sigma^2_x)$. Therefore,

$$E_{W^x}(W^x) = \mu_x. \tag{4.46}$$

Similarly, we can show that the overall W^y has mean

$$E_{W^y}(W^y) = \mu_y. \tag{4.47}$$

To calculate the overall variances, we first recognize that the internal within observation variances given the observations $X_i = \xi_{ix}$ and $Y_i = \xi_{iy}$ were set to be Θ^x_{i1} and Θ^y_{i1}, respectively. Therefore, for each $i = 1, \ldots . n$,

$$\Theta^x_{i2} = Var_{W^x}(W^x | X_i = \xi_{ix}). \tag{4.48}$$

From the conditional variance relationship in Equation (4.41), we have

$$\begin{aligned} Var_{W^x}(W^x) &= E_{X_i}[Var_{W^x}(W^x | X_i = \xi_{ix})] + Var_{X_i}[E_{W^x}(W^x | X_i = \xi_{ix})] \qquad (4.7) \\ &= E_{X_i}(\Theta^x_{i2}) + Var_{X_i}(\Theta^x_{i1}). \qquad (4.49) \end{aligned}$$

Now, we have that $\mathbf{\Theta}_{i2} = (\Theta^x_{i2}, \Theta^y_{i2}, \Theta^{xy}_{i2})$ followed the bivariate exponential distribution of equation (4.17) and so we can calculate the first term in Equation (4.49), viz., $E_{X_i}(\Theta^x_{i2})$. We also have Θ^x_{i1} distributed as a normal distribution $N(\mu_x, \sigma^2_x)$. Hence, Equation (4.49) becomes

$$Var_{W^x}(W^x) = \gamma_1 + \sigma^2_x. \tag{4.50}$$

Similarly, we can show that

$$Var_{W^y}(W^y) = \gamma_2 + \sigma^2_y. \tag{4.51}$$

To obtain the corresponding expression for the covariance, we can show, from Equation (4.42), that

$$\begin{aligned} Cov_{W^x,W^y}(W^x, W^y) &= E_{X_i,Y_i}[Cov_{W^x,W^y}(W^x, W^y | X_i = \xi^x_i, Y_i = \xi^y_i,)] \\ &\quad + Cov_{X_i,Y_i}[E_{W^x}(W^x | X_i = \xi^x_i), E_{W^y}(W^y | Y_i = \xi^y_i)] \\ &= E_{X_i,Y_i}(\Theta^{xy}_{i2}) + Cov_{X_i,Y_i}(\Theta^x_{i1}, \Theta^y_{i1}). \qquad (4.52) \end{aligned}$$

The first term in equation (4.52) is the covariance variable in the bivariate exponential distribution of Equation (4.17), from which we can show that the internal expectation $E_{X_i,Y_i}(\Theta^{xy}_{i2}) = \gamma_3$. The second term in Equation (4.52) corresponds to the covariance of the two interval means $\Theta^x_{i1}, \Theta^y_{i1}$. Since these follow the bivariate normal distribution

$N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, it follows that this covariance is $\rho\sigma_x\sigma_y$. Hence, substituting into Equation (4.52) we obtain

$$Cov_{W^x,W^y}(W^x, W^y) = \gamma_3 + \rho\sigma_x\sigma_y. \tag{4.53}$$

All these overall descriptive statistics are expressed in terms of the parameters contained in τ considered in Section 4.3.1. Hence, maximum likelihood estimators are readily found by substituting the relevant values from Equations (4.31)-(4.39). After suitable substitution and rearrangement, we can show that the maximum likelihood estimators are given by, respectively,

$$\widehat{E(W^x)} = \hat{\mu}_x = \frac{1}{2n}\sum_{i=1}^{n}(c_i + d_i); \tag{4.54}$$

$$\widehat{E(W^y)} = \hat{\mu}_y = \frac{1}{2n}\sum_{i=1}^{n}(a_i + b_i); \tag{4.55}$$

$$\widehat{Var(W^x)} = \frac{1}{3n}\sum_{i=1}^{n}[(c_i - \hat{\mu}_x)^2 + (c_i - \hat{\mu}_x)(d_i - \hat{\mu}_x) + (d_i - \hat{\mu}_x)^2]; \tag{4.56}$$

$$\widehat{Var(W^y)} = \frac{1}{3n}\sum_{i=1}^{n}[(a_i - \hat{\mu}_y)^2 + (a_i - \hat{\mu}_y)(b_i - \hat{\mu}_y) + (b_i - \hat{\mu}_y)^2]; \tag{4.57}$$

and

$$\begin{aligned}\widehat{Cov(W^x, W^y)} = \frac{1}{6n}\sum_{i=1}^{n}[&2(a_i - \hat{\mu}_x)(c_i - \hat{\mu}_y) + (a_i - \hat{\mu}_x)(d_i - \hat{\mu}_y) \\ &+ (b_i - \hat{\mu}_x)(c_i - \hat{\mu}_y) + 2(b_i - \hat{\mu}_x)(d_i - \hat{\mu}_y)]\end{aligned} \tag{4.58}$$

In Equations (4.54)-(4.58), it is implicit that these are estimators with respect to the overall variables in that $E_{W^x}(W^x)$, e.g., is written as $E(W^x)$

4.4 Additional Comments

The overall mean μ_x (and similarly for μ_y) is estimated by the average of the interval center points. Thus, regardless of the actual length of an interval, the sample mean is unchanged. This result was first obtained empirically by Bertrand and Goupil (2000) using an empirical distribution approach.

Bertrand and Goupil (2000) also obtained an expression for the sample variance; see Equation (2.3). Later, Billard (2007, 2008) showed that their sample variance was the sum of a so-called within observation variation and a between observation variation; see Equations (2.8) and (2.9), respectively. In the overall variance $Var(W^x)$ of Equation (4.50), the two terms γ_1 and σ_x^2 in fact correspond, respectively, to the within observations variation and the between observations variation. [Clearly, the same applies to $Var(W^y)$ of Equation (4.51).] The same phenomenon applies to the $Cov(X, Y)$ with the first term γ_3 in Equation (4.53) corresponding to the within observation covariation and the second term $\rho\sigma_x\sigma_y = \sigma_{xy}$ corresponding to the between observation covariation; see Equations (2.12) and (2.13), respectively. This result was initially obtained as a moment estimator of the covariance function in Billard (2007, 2008).

The results in this chapter has been derived under an assumption that the within interval observations are uniformly distributed across the given intervals. Other distributions could be used. In those cases, realizations of the interval parameters Θ_{i1}^x, etc. will change. These will give different expressions for the Within Observation quantities; while those for the Between Observation terms are unchanged. The principles are the same however as were followed in the above derivations.

When data are classically-valued, i.e., when $\xi_i^x = [a_i, a_i] = a_i$ and $\xi_i^y = [c_i, c_i] = c_i$, it is easy to show that the within observation variations become zero, while the between observation variations are unchanged since they are based on the interval center points. In this case, all the results in Equations (4.54)-(4.58) reduce to their classical counterparts.

4.5 References

[1] Bertrand, P. and Goupil, F. (2000): Descriptive Statistics for Symbolic Data. *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday). Springer-Verlag, Berlin, 103-124.

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

[3] Billard, L. (2008). Sample Covariance Functions for Complex Quantitative Data. Processing, World Conferences International Association of Statistical Computing 2008, Yokohama, Japan.

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

[5] Casella, G. and Berger, R. L. (2002). *Statistical Inference*. 2nd ed, Duxbury, Pacific Grove CA.

[6] Freund, J. E. (1961). A Bivariate Extension of the Exponential Distribution. *Journal of the American Statistical Association* 56, 971-977.

[7] Gumbel, E. J. (1958). *Statistics of Extremes*. Columbia University Press.

[8] Gumbel, E. J. (1960). Bivariate Dxponential Distributions. *Journal of the American Statistical Association* 55, 698-707.

[9] Le-Rademacher, J. and Billard, L. (2010). Likelihood Functions and Some Maximum Likelihood Estimators for Symbolic Data. *Journal of Statistical Planning and Inference*, Submitted.

[10] Marshall, A. W. and Olkin, I. (1967). A Multivariate Exponential Distribution. *Journal of the American Statistical Association* 62, 30-44.

[11] Plackett, R. L. (1965). A Class of Bivariate Distributions. *Journal of the American Statistical Association* 60, 516-522.

Chapter 5

Comparison of Methods

In this chapter we investigate the performance of the proposed symbolic covariance method (SCM), and compare it with current regression methods for symbolic data. Recall that the center method (CM) introduced by Billard and Diday (2000) uses the center points to carry out the analysis. The center and range method (CRM) proposed by Lima Neto et al. (2004) and de Carvalho et al. (2004) uses the center points and ranges separately. Billard and Diday (2007) introduced the bivariate center and range method without interaction (BCRWO) and bivariate center and range method with interaction (BCRWI). See Section 2.5 for more details about these methods. In this chapter, these different regression methods for symbolic interval-valued data are compared in several examples. We first analyze four datasets. Then, we consider a simulated dataset; it quickly becomes clear that the new symbolic method gives parameter estimates closest to the known parameter values.

The proposed SCM method in Chapter 3 may be able to reveal the true structure of the data by considering both internal variation and external variation as well as the relationship between response and explanatory variables simultaneously. It is also shown to give identical results as the symbolic likelihood method (SLM) proposed in Chapter 4 under certain conditions. Therefore, the relationship between the SCM method and the SLM method are comparable to that between the least squares technique and the maximum likelihood method in classical data regression. Since the SCM method and SLM method will obtain the same results, in this chapter we only use the results from the SCM method to reflect both.

The first dataset is a medical information dataset available in Billard and Diday (2004), and is a simulated dataset obtained by using distribution and parameter values set by consultation with medical personnel and census information. The second dataset used is a mushroom dataset extracted from the Fungi of California Species Index; Billard and Diday (2007) used part of this dataset as an example of interval-valued data and Billard (2007, 2008) used all of it to illustrate symbolic covariance. The bats dataset and blood pressure dataset utilized in Chapter 3 are also analyzed in this chapter to compare the different methods.

This chapter is carried forward in the following structure. In Section 5.1, three performance measures used are described including residuals, root mean square error (RMSE), and correlation coefficient r. Sections 5.2 to 5.5 analyze each dataset separately. Model specification, parameter estimates, as well as performance measures are presented side by side for the current and proposed methods. The simulated dataset is considered in Section 5.6.

5.1 Performance Measures

To assess the performance of methods, the following measures are used: Residuals, the lower bound root mean-square error ($RMSE_L$), the upper bound root mean-square error ($RMSE_U$), the square of the lower bound correlation coefficient (r_L) and the square of the upper bound correlation coefficient (r_U).

Let Y denote an interval-valued variable with observed values $Y_i = [Y_{Li}, Y_{Ui}]$ and corresponding predicted values $\hat{Y}_i = [\hat{Y}_{Li}, \hat{Y}_{Ui}]$, where $i = 1, \ldots, n$. Then residuals for lower bound and upper bound predictions can be obtained from Equation (3.15) as

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

In classical statistics, the root mean squared error (RMSE) of an estimator is one of many measures used to quantify the difference between an estimator and the true value of the quantity being estimated. For an unbiased estimator, the RMSE is the square root of the variance, known as the standard error. Analogously to the thinking underneath RMSE

of a estimator in classical statistics, Lima Neto and de Carvalho (2010) used RMSE of the lower bound and the upper bound, respectively, to evaluate the performance of an interval estimator. These measures are

$$RMSE_L = \sqrt{(\sum_{i=1}^{n}(Y_{iL} - \hat{Y}_{iL})^2)/n}\ ,\ RMSE_U = \sqrt{(\sum_{i=1}^{n}(Y_{iU} - \hat{Y}_{iU})^2)/n} \tag{5.1}$$

where $\boldsymbol{Y}_L = (Y_{1L}, \ldots, Y_{nL})'$, $\boldsymbol{Y}_U = (Y_{1U}, \ldots, Y_{nU})'$, and $\hat{\boldsymbol{Y}}_L = (\hat{Y}_{1L}, \ldots, \hat{Y}_{nL})'$, $\hat{\boldsymbol{Y}}_U = (\hat{Y}_{1U}, \ldots, \hat{Y}_{nU})'$, $\hat{Y}_{iL}$ and $\hat{Y}_{iU}$ are the predicted values of Y_{iL} and Y_{iU}, respectively, $i = 1, \ldots, n$.

Further, in classical statistics, the Pearson product-moment correlation coefficient, or Pearson's correlation, is the most popular measure of dependence between two quantities. It is obtained by dividing the covariance of the two variables by the product of their standard deviations. The Pearson correlation lies between $[-1, 1]$ with -1 indicating perfectly negative linear relationship and 1 indicating perfectly positive linear relationship. Therefore, in addition to RMSE, Lima Neto and De Carvalho (2010) used a correlation coefficient between observed values and predicted values to assess the performance of an estimator. Their interval correlation coefficient (r_L, r_U) is defined as

$$r_L = \frac{Cov(Y_L, \hat{Y}_L)}{S_{Y_L} S_{\hat{Y}_L}}, \quad r_U = \frac{Cov(Y_U, \hat{Y}_U)}{S_{Y_U} S_{\hat{Y}_U}} \tag{5.2}$$

where $Cov(Y_k, \hat{Y}_k)$ is the covariance between Y_k and $\hat{Y}_k$ defined in Definition 2.3.11, and where S_{Y_k} is the standard deviation of Y_k obtained from Definition 2.3.10 and where $S_{\hat{Y}_k}$ is the standard deviation of $\hat{Y}_k$ for $k = L, U$.

Because $\boldsymbol{Y}_L$, $\hat{\boldsymbol{Y}}_L$, $\boldsymbol{Y}_U$ and $\hat{\boldsymbol{Y}}_U$ are no longer symbolic data but written as classical data values, Lima Neto and de Carvalho (2010) used the correlation coefficient function for classical data to calculate the r_L and r_U respectively. However, this approach may lose information about the range of the intervals. Therefore, we propose to calculate the correlation coefficient between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$, that is,

$$r(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = Cov(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) / S_{\boldsymbol{Y}} S_{\hat{\boldsymbol{Y}}} \tag{5.3}$$

where $\boldsymbol{Y} = [\boldsymbol{Y}_L, \boldsymbol{Y}_U]$, $\hat{\boldsymbol{Y}} = [\hat{\boldsymbol{Y}}_L, \hat{\boldsymbol{Y}}_U]$, $Cov(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$ is the symbolic covariance between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ defined in Definition 2.3.11, $S_{\boldsymbol{Y}}$ and $S_{\hat{\boldsymbol{Y}}}$ are the square root of the sample symbolic variance defined in Definition 2.3.10 for $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$, respectively.

An interesting phenomenon to be noticed is that the sum of lower bound residuals equals the negative of the sum of upper bound residuals for the CM method, the SCM method, and the CRM method as should be expected.

The sum of residuals for lower bound and upper bound are, respectively, $\sum_{i=1}^{n}(\hat{Y}_{iL} - Y_{iL})$ and $\sum_{i=1}^{n}(\hat{Y}_{iU} - Y_{iU})$. We want to show that

$$\sum_{i=1}^{n}(\hat{Y}_{iL} - Y_{iL}) = -\sum_{i=1}^{n}(\hat{Y}_{iU} - Y_{iU}). \tag{5.4}$$

Case 1: The CM method

From Equations (2.22) and (2.23), we can obtain $\hat{Y}_{iL}$ and $\hat{Y}_{iU}$. Substituting them into Equation (5.4) gives

$$\sum_{i=1}^{n}(\boldsymbol{X}_{iL}\hat{\boldsymbol{\beta}} - Y_{iL}) = -\sum_{i=1}^{n}(\boldsymbol{X}_{iU}\hat{\boldsymbol{\beta}} - Y_{iU}),$$

which equals

$$\sum_{i=1}^{n}(\boldsymbol{X}_{iL} + \boldsymbol{X}_{iU})\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n}(Y_{iL} + Y_{iU}).$$

Dividing both sides by 2 gives

$$\sum_{i=1}^{n}\boldsymbol{X}_i^c\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n}\frac{(\boldsymbol{X}_{iL} + \boldsymbol{X}_{iU})}{2}\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n}\frac{(Y_{iL} + Y_{iU})}{2} = \sum_{i=1}^{n}Y_i^c.$$

Furthermore, dividing both sides by n gives

$$\bar{\boldsymbol{X}}^c\hat{\boldsymbol{\beta}} = \bar{Y}^c \tag{5.5}$$

where $\bar{\boldsymbol{X}}^c = (\bar{X}_1^c, \ldots, \bar{X}_p^c)$ and $\bar{Y}^c = \sum_{i=1}^{n} Y_i^c/n$. Note, in the CM method $\hat{\boldsymbol{\beta}}$ is estimated by the least squares techniques, i.e., from Equation (2.21); that is,

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

Therefore, Equation (5.5) implies that the fitted line with parameters estimated by least squares computation must pass through the point $(\bar{X}, \bar{Y})$, the center of the data, which is already a proved fact. See proof in, for example, Weisberg (2005).

In summary, Equation (5.4) is proved to hold in the CM method case.

Case 2: The SCM method

Equations (3.12) gives $\hat{Y}_{iL}$ and $\hat{Y}_{iU}$. We substitute them into Equation (5.4) to obtain

$$\sum_{i=1}^{n}(\min(\boldsymbol{X}_{iL}\hat{\beta}, \boldsymbol{X}_{iU}\hat{\beta}) - Y_{iL}) = -\sum_{i=1}^{n}(\max(\boldsymbol{X}_{iL}\hat{\beta}, \boldsymbol{X}_{iU}\hat{\beta}) - Y_{iU}),$$

which equals

$$\sum_{i=1}^{n}(\min(\boldsymbol{X}_{iL}\hat{\beta}, \boldsymbol{X}_{iU}\hat{\beta}) + \max(\boldsymbol{X}_{iL}\hat{\beta}, \boldsymbol{X}_{iU}\hat{\beta})) = \sum_{i=1}^{n}(Y_{iL} + Y_{iU}).$$

Again, dividing both sides by 2 gives

$$\begin{aligned}\sum_{i=1}^{n}\frac{(\min(\boldsymbol{X}_{iL}\hat{\beta}, \boldsymbol{X}_{iU}\hat{\beta}) + \max(\boldsymbol{X}_{iL}\hat{\beta}, \boldsymbol{X}_{iU}\hat{\beta}))}{2} &= \sum_{i=1}^{n}\frac{(\boldsymbol{X}_{iL}\hat{\boldsymbol{\beta}} + \boldsymbol{X}_{iU}\hat{\boldsymbol{\beta}})}{2} \\ &= \sum_{i=1}^{n}\frac{(\boldsymbol{X}_{iL} + \boldsymbol{X}_{iU})}{2}\hat{\boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} Y_i^c.\end{aligned}$$

Furthermore, dividing both sides by n gives

$$\bar{\boldsymbol{X}}^c\hat{\boldsymbol{\beta}} = \bar{Y}^c \tag{5.6}$$

where $\bar{\boldsymbol{X}}^c = (\bar{X}_1^c, \ldots, \bar{X}_p^c)$ and $\bar{Y}^c = \sum_{i=1}^{n} Y_i^c/n$. Note, Equation (5.6) is the same as in Equation (5.5). As stated in the CM method case, the least squares fitted line passes through the center of the data, which means Equation (5.6) holds good as long as the parameters are estimated by the least squares computation.

Case 3: The CRM method

From Equations (2.27), substituting $\hat{Y}_{iL}$ and $\hat{Y}_{iU}$ into Equation (5.4) gives

$$\sum_{i=1}^{n}(\boldsymbol{X}_i^c\hat{\boldsymbol{\beta}}^c - \boldsymbol{X}_i^r\hat{\boldsymbol{\beta}}^r/2 - Y_{iL}) = -\sum_{i=1}^{n}(\boldsymbol{X}_i^c\hat{\boldsymbol{\beta}}^c + \boldsymbol{X}_i^r\hat{\boldsymbol{\beta}}^r/2 - Y_{iU}),$$

which equals

$$2\sum_{i=1}^{n} \boldsymbol{X}_i^c \hat{\boldsymbol{\beta}}^c = \sum_{i=1}^{n} (Y_{iL} + Y_{iU}).$$

Dividing both sides by 2 gives

$$\sum_{i=1}^{n} \boldsymbol{X}_i^c \hat{\boldsymbol{\beta}}^c = \sum_{i=1}^{n} Y_i^c.$$

Furthermore, dividing both sides by n gives

$$\bar{\boldsymbol{X}}^c \hat{\boldsymbol{\beta}}^c = \bar{Y}^c \tag{5.7}$$

where $\hat{\boldsymbol{\beta}}^c$ is estimated from Equation (2.21).

Once again, since $\hat{\boldsymbol{\beta}}^c$ is estimated by least squares computation, so that the fitted line passes through the center of the data, which confirms that Equation (5.4) holds in the CRM method case.

Case 4/5: The BCR methods

Likewise, for the two BCR methods, either with or without interaction, the same relationship holds for the residuals. The details are omitted.

5.2 Medical Dataset

5.2.1 The Data

The medical dataset, from Billard and Diday (2004), is extracted from a large dataset simulated by Billard in 2002, using distribution and parameter values set by consultation with medical personnel and census information (http://www.census.gov). The original dataset has 10,000 classical observations. There are 7 continuous random variables: race, agegroup, diabetes (diabetes type), chol (cholesterol), gluc (glucose), hemo (hemoglobin), hemat (hematocrit), redbld (red blood cell), whdbld (white blood cell, and income. There are 3 discrete random variables: agegroup (1 $\equiv$ 15-24 years old, 2 $\equiv$ 25-34 years old, 3 $\equiv$ 35-44 years old, 4 $\equiv$ 45-54 years old, 5 $\equiv$ 55-64 years old, 6 $\equiv$ 65-74 years old, 7 over 74 years), race (0 $\equiv$

white, 1 ≡ black), diabetes (0 ≡ No, 1 ≡ mild, 2 ≡ yes). The original dataset also has two other structural variables each with numerous other varialbes which are not of interest for the present purpose.

Suppose that the scientific questions of interest relate to agegroup × diabetes × race, rather than to any particular individual. Therefore, we aggregate the entire dataset for each of the 42 categories of agegroup × diabetes × race ($7 \times 3 \times 2 = 42$). The resulting interval-valued dataset is displayed in Table 5.1.

Let X_1 = gluc, X_2 = chol, X_3 = hemo, X_4 = hemat, X_5 = redbld, X_6 = whdbld, and X_7 = income. The covariance matrix and coefficient coefficient matrix for these variables in the medical dataset is shown in Equation (5.8) and (5.9), respectively as follows

$$\mathbf{V} = \begin{pmatrix} 387.98 & 552.3 & 10.717 & 41.75 & 6.830 & 38.63 & 78528 \\ 552.26 & 1819.7 & 22.255 & 93.26 & 14.723 & 97.71 & 226590 \\ 10.72 & 22.3 & 0.891 & 3.10 & 0.400 & 1.94 & 2919 \\ 41.75 & 93.3 & 3.101 & 11.46 & 1.507 & 7.63 & 12614 \\ 6.83 & 14.7 & 0.400 & 1.51 & 0.251 & 1.22 & 2312 \\ 38.63 & 97.7 & 1.937 & 7.63 & 1.221 & 7.33 & 14179 \\ 78528.09 & 226589.7 & 2919.255 & 12614.28 & 2312.450 & 14178.78 & 56633123 \end{pmatrix} \tag{5.8}$$

and

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.657 & 0.576 & 0.626 & 0.693 & 0.724 & 0.530 \\ 0.657 & 1.000 & 0.553 & 0.646 & 0.690 & 0.846 & 0.706 \\ 0.576 & 0.553 & 1.000 & 0.970 & 0.845 & 0.758 & 0.411 \\ 0.626 & 0.646 & 0.970 & 1.000 & 0.889 & 0.832 & 0.495 \\ 0.693 & 0.690 & 0.845 & 0.889 & 1.000 & 0.901 & 0.614 \\ 0.724 & 0.846 & 0.758 & 0.832 & 0.901 & 1.000 & 0.696 \\ 0.530 & 0.706 & 0.411 & 0.495 & 0.614 & 0.696 & 1.000 \end{pmatrix}. \tag{5.9}$$

Table 5.1: Medical Dataset

ID	Race	Agegroup	Diabetes	Gluc	Chol	Hemo	Hemat	Redbld	Whbld	Income
1	0	1	0	[58.5, 138.3]	[74.7, 220.8]	[12.7, 14.9]	[35.8, 46.4]	[3.7, 5.5]	[2.2, 12.9]	[9268, 23369]
2	1	1	0	[64.6, 122.2]	[111.7, 209.3]	[12.9, 14.8]	[37.1, 46]	[3.9, 5.4]	[4.3, 10.9]	[5337, 29897]
3	0	2	0	[58.8, 135.2]	[78, 244]	[10.4, 13.8]	[30.4, 44]	[3.3, 5.6]	[1.1, 13.1]	[11400, 29421]
4	1	2	0	[61.9, 122]	[96.1, 238.8]	[10.3, 13.5]	[31.6, 43.3]	[3.7, 5.4]	[2.8, 11.3]	[6582, 27952]
5	0	3	0	[67.9, 142.7]	[71.2, 243.5]	[10.5, 13.6]	[30.7, 43]	[3.5, 5.6]	[1.8, 13.7]	[12210, 31469]
6	1	3	0	[60.8, 121.6]	[120.7, 257.9]	[11.3, 13.4]	[32.4, 41.6]	[3.8, 5.5]	[3.7, 12.2]	[5849, 32945]
7	0	4	0	[45.9, 143.5]	[106.1, 267.2]	[10.6, 13.9]	[31, 43.2]	[3.4, 5.4]	[1.5, 12.8]	[16263, 41090]
8	1	4	0	[63.9, 123.7]	[116.7, 248.6]	[11, 13.5]	[32.3, 41.7]	[3.4, 5.1]	[3, 12.3]	[8023, 31725]
9	0	5	0	[50.1, 131.6]	[114.5, 280.8]	[10.5, 13.6]	[29.6, 42.5]	[3.4, 5.1]	[2.4, 12.5]	[17808, 44968]
10	1	5	0	[64.8, 123.8]	[112.4, 256.6]	[11.1, 13.2]	[32.1, 39.9]	[3.4, 5]	[4.1, 13.1]	[8342, 31716]
11	0	6	0	[58.2, 124.7]	[127, 268]	[10.9, 13.6]	[31.3, 42.9]	[3.5, 5.1]	[2.3, 12.2]	[12928, 33067]
12	1	6	0	[70.6, 123.4]	[145.6, 276.7]	[10.9, 13.3]	[32.1, 41.3]	[3.6, 4.9]	[3.5, 11.4]	[7751, 32711]
13	0	7	0	[71.5, 149.3]	[124.2, 269.1]	[10.6, 13.7]	[30.6, 43.7]	[3.4, 5]	[1.8, 13]	[10492, 27287]
14	1	7	0	[77.8, 123.5]	[141.2, 251.2]	[11, 13.1]	[32.3, 39.8]	[3.3, 4.7]	[3.5, 9.8]	[6298, 18958]
15	0	1	1	[100.4, 135.5]	[94, 236.9]	[12.8, 14.8]	[36.6, 45.6]	[3.7, 5.4]	[2.9, 13.3]	[10946, 20064]
16	1	1	1	[103.6, 120.9]	[135.7, 207]	[12.8, 14.8]	[37.8, 46.1]	[4.3, 5.4]	[6.2, 9.5]	[8112, 18489]
17	0	2	1	[100, 136.7]	[80.5, 220.3]	[10.7, 13.7]	[31.1, 42.1]	[3.6, 5.5]	[3.4, 11.9]	[11669, 29137]
18	1	2	1	[100.2, 123.8]	[94, 226.3]	[11.3, 13.8]	[32.6, 43.9]	[3.8, 5.2]	[2.6, 12.9]	[6439, 22523]
19	0	3	1	[100, 143.4]	[84.3, 266.3]	[10.9, 13.8]	[31, 41.3]	[3.7, 5.5]	[3.1, 12]	[12436, 31370]
20	1	3	1	[100, 141.6]	[113.6, 211]	[11.3, 13.5]	[31.7, 41.1]	[4, 5.4]	[4, 11.5]	[6836, 26498]
21	0	4	1	[100, 148.9]	[125.9, 267.4]	[10.9, 13.9]	[31.1, 43.3]	[3.5, 5.6]	[0.4, 12.9]	[16186, 40863]
22	1	4	1	[100, 123.8]	[135.8, 256]	[11, 12.9]	[31.5, 41.2]	[3.6, 4.9]	[1.7, 11.3]	[9180, 27951]
23	0	5	1	[100, 137.3]	[118.9, 244.5]	[10.6, 13.6]	[31.9, 41.7]	[3.6, 5.2]	[2.1, 13.2]	[18195, 44771]
24	1	5	1	[101.1, 137.2]	[143.7, 247.4]	[11.2, 12.9]	[33.1, 39.8]	[3.7, 5.1]	[5.4, 11.3]	[10392, 32057]
25	0	6	1	[100.1, 139.1]	[128.1, 274.1]	[10.9, 13.5]	[32.4, 41.3]	[3.5, 4.9]	[1.3, 11.5]	[12898, 32607]
26	1	6	1	[100.5, 130.7]	[156.2, 257.2]	[11.2, 13.2]	[33.2, 40.4]	[3.7, 4.9]	[4.5, 12.7]	[8197, 20126]
27	0	7	1	[100.2, 142.4]	[109.1, 261]	[11, 13.7]	[32, 42.1]	[3.5, 5.1]	[3, 12.6]	[10906, 27454]
28	1	7	1	[101.8, 122.9]	[174, 250]	[11.4, 12.8]	[33.3, 40.3]	[3.7, 4.6]	[3.1, 9.7]	[7215, 21993]
29	0	1	2	[101.1, 143.1]	[108.2, 215.6]	[12.9, 15.2]	[37, 48.1]	[3.8, 5.4]	[2.8, 11.8]	[9913, 23472]
30	1	1	2	[101.5, 144.2]	[106.1, 215.2]	[13, 14.6]	[37.7, 44.4]	[4.2, 5.2]	[4.4, 10.2]	[7327, 18700]
31	0	2	2	[100.1, 151]	[94.9, 211.7]	[11, 13.3]	[32.5, 42.1]	[3.8, 5.3]	[2.4, 12.5]	[11772, 29155]
32	1	2	2	[100.3, 133.6]	[90.1, 208.7]	[11.3, 13.7]	[33, 41.5]	[3.7, 5.3]	[4.8, 11.2]	[5919, 20898]
33	0	3	2	[100, 157.2]	[57.1, 241.1]	[10.7, 13.8]	[31.4, 43.5]	[3.6, 5.7]	[1.8, 12.8]	[12237, 31505]
34	1	3	2	[100, 148.7]	[125.6, 233.6]	[10.8, 13.7]	[31.3, 40.8]	[4, 5.4]	[3.4, 11.2]	[7264, 23741]
35	0	4	2	[100, 154.2]	[111.4, 264.5]	[10.8, 13.4]	[31, 41.7]	[3.4, 5.2]	[3, 14.7]	[16368, 39687]
36	1	4	2	[101.8, 140.6]	[144.1, 239.8]	[10.8, 14]	[31.1, 41.9]	[3.7, 5]	[4.5, 11.6]	[10643, 29799]
37	0	5	2	[100, 150.3]	[128.5, 266.5]	[10.7, 13.9]	[31.1, 41.8]	[3.4, 5]	[2.7, 11.5]	[17991, 44228]
38	1	5	2	[100.3, 135.1]	[142, 271.4]	[11.3, 12.9]	[32.9, 39.2]	[3.7, 5]	[5.7, 10.9]	[10136, 24325]
39	0	6	2	[100, 152.2]	[123.2, 263.4]	[10.9, 13.5]	[32, 43.2]	[3.4, 5]	[2.9, 12.4]	[12845, 32940]
40	1	6	2	[102.2, 143.6]	[161, 239.5]	[11.5, 12.6]	[32.7, 40.1]	[3.7, 4.7]	[2, 10.2]	[9591, 22516]
41	0	7	2	[100.1, 161.3]	[115.9, 257.4]	[10.8, 13.5]	[31.6, 43.1]	[3.3, 5]	[2.1, 13.4]	[10762, 26817]
42	1	7	2	[102.4, 147.7]	[150.9, 248]	[11, 13.4]	[32.9, 40.9]	[3.5, 4.8]	[4.1, 10.5]	[5257, 23259]

5.2.2 Analysis Results

From Equations (5.8) and (5.9) it can be seen that X_5 and X_6 have the biggest correlation coefficient. However, it is of more interest to build a model having real meaning than by only looking at the correlation between variables. Suppose our particular research interest is in the relationship between cholesterol and income level. Let X_2 = chol be the response variable (i.e., Y) and X_7 = income be the explanatory variable (i.e., X). The model to be built is

$$Y = \beta_0 + \beta_1 X_7 + \epsilon. \tag{5.10}$$

Current regression methods and our proposed method are utilized to build the model in Equation (5.10). First, the proposed SCM method is used to build the linear regression model following the algorithm introduced in Section 3.2.2.

We start with calculating the sample mean for Y and X_7, which gives us the results as $\bar{Y} = 181$ and $\bar{X}_7 = 19758$.

Then, we center the variables Y and X_7 around their sample means and write the model in Equation (5.10) as

$$Y - 181 = \beta_1(X_7 - 19758) + \epsilon.$$

From Definition 2.3.10, we calculate the variance of X_7, which is 56633123 from Equation (5.8). From Definition 2.3.11, we calculate the covariance between Y and X_7, which is also presented in Equation (5.8) as 226590.

Now substituting the sample variance and covariance given in Equation (5.8) into Equations (3.8) and (3.9), we obtain

$$(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{X} - \bar{\boldsymbol{X}}) = n \times Var(X_7) = 42 \times 56633123$$

and

$$(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{Y} - \bar{\boldsymbol{Y}}) = n \times Cov(Y, X_7) = 42 \times 226590.$$

Then, substituting into Equation (3.10) gives the estimates of β_1, which is $\hat{\beta}_1$=0.004.

Finally, substituting the estimate of β_1 into Equation (3.5) obtains the estimate of β_0, which is $\hat{\beta}_0$=102.22. Therefore, the model in Equation (5.10) can be written as

$$Y = 102.22 + 0.004X_7 + \epsilon.$$

In addition, the predicted values of Y shown in Table 5.2 can be calculated from Equation (3.11). Also shown are the residuals obtained from Equation (5.1).

The CM method is also applied to the medical dataset. From Billard and Diday (2000), we first calculate the center points of the original intervals-valued data. For each observation, we calculate $Y_i^c = (Y_{iL} + Y_{iU})/2$ and $X_{7i}^c = (X_{7iL} + X_{7iU})/2$, $i = 1, \ldots, n$. Then, we analyze Y^c and X_7^c as though they are classical data. Using the formula in Equation (2.21) gives an estimator of parameters $\boldsymbol{\beta}^c$ as $\hat{\boldsymbol{\beta}}^c = (165.64, 0.00079)'$. The linear regression model in Equation (5.10) can now be written as

$$Y^c = 165.64 + 0.00079X_7^c + \epsilon \tag{5.11}$$

where Y^c = glucose center points and X_7^c = income center points.

In addition, predicted values of Y and residuals can be obtained from Equations (2.22), (2.23) and (5.1); these are shown in Table 5.3.

Another method applied to the medical dataset is the CRM method. In addition to obtaining the center points of the original intervals as shown in CM method, the ranges are also calculated as $Y_i^r = (Y_{iL} - Y_{iU})$ and $X_{7i}^r = (X_{7iL} - X_{7iU})$, $i = 1, \ldots, n$.

Then, the model of center points in Equations (2.24) is identical to that obtained by the CM method, which is the same as shown in Equation (5.11).

The model of ranges in Equation (2.25) can be obtained by using the parameter estimation from Equations (2.26), which gives

$$Y^r = 89.491 + 0.00214X_7^r + \epsilon.$$

Predicted values of Y can be obtained from Equation (2.27) as shown in Table 5.4. Then, the residuals can also be calculated from Equation (5.1) and are also shown in Table 5.4.

Table 5.2: Predicted Values and Residuals with the SCM Method - Medical Dataset

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
1	9268	23369	74.7	220.8	139.3041	195.7224	64.60412	-25.0776
2	5337	29897	111.7	209.3	123.5761	221.841	11.87615	12.54096
3	11400	29421	78	244	147.8343	219.9365	69.83427	-24.0635
4	6582	27952	96.1	238.8	128.5574	214.059	32.4574	-24.741
5	12210	31469	71.2	243.5	151.0751	228.1305	79.87509	-15.3695
6	5849	32945	120.7	257.9	125.6247	234.036	4.924664	-23.864
7	16263	41090	106.1	267.2	167.2912	266.6243	61.19118	-0.57574
8	8023	31725	116.7	248.6	134.3229	229.1548	17.62286	-19.4452
9	17808	44968	114.5	280.8	173.4727	282.1402	58.97274	1.340179
10	8342	31716	112.4	256.6	135.5992	229.1188	23.19918	-27.4812
11	12928	33067	127	268	153.9478	234.5242	26.94781	-33.4758
12	7751	32711	145.6	276.7	133.2346	233.0998	-12.3654	-43.6002
13	10492	27287	124.2	269.1	144.2014	211.3983	20.00135	-57.7017
14	6298	18958	141.2	251.2	127.4211	178.0739	-13.7789	-73.1261
15	10946	20064	94	236.9	146.0178	182.499	52.01781	-54.401
16	8112	18489	135.7	207	134.679	176.1974	-1.02105	-30.8026
17	11669	29137	80.5	220.3	148.9105	218.8002	68.41054	-1.49981
18	6439	22523	94	226.3	127.9853	192.3375	33.98526	-33.9625
19	12436	31370	84.3	266.3	151.9793	227.7344	67.67932	-38.5656
20	6836	26498	113.6	211	129.5737	208.2415	15.97366	-2.75848
21	16186	40863	125.9	267.4	166.9831	265.716	41.08311	-1.68397
22	9180	27951	135.8	256	138.952	214.055	3.152029	-41.945
23	18195	44771	118.9	244.5	175.0211	281.352	56.12114	36.85198
24	10392	32057	143.7	247.4	143.8013	230.4831	0.101253	-16.9169
25	12898	32607	128.1	274.1	153.8278	232.6837	25.72778	-41.4163
26	8197	20126	156.2	257.2	135.019	182.7471	-21.181	-74.4529
27	10906	27454	109.1	261	145.8578	212.0665	36.75777	-48.9335
28	7215	21993	174	250	131.09	190.217	-42.91	-59.783
29	9913	23472	108.2	215.6	141.8848	196.1345	33.68477	-19.4655
30	7327	18700	106.1	215.2	131.5382	177.0416	25.43816	-38.1584
31	11772	29155	94.9	211.7	149.3226	218.8722	54.42265	7.172206
32	5919	20898	90.1	208.7	125.9047	185.8359	35.80474	-22.8641
33	12237	31505	57.1	241.1	151.1831	228.2746	94.08312	-12.8254
34	7264	23741	125.6	233.6	131.2861	197.2107	5.686094	-36.3893
35	16368	39687	111.4	264.5	167.7113	261.0108	56.31129	-3.48916
36	10643	29799	144.1	239.8	144.8055	221.4489	0.705507	-18.3511
37	17991	44228	128.5	266.5	174.2049	279.1794	45.70493	12.67943
38	10136	24325	142	271.4	142.777	199.5473	0.776995	-71.8527
39	12845	32940	123.2	263.4	153.6157	234.016	30.41573	-29.384
40	9591	22516	161	239.5	140.5964	192.3095	-20.4036	-47.1905
41	10762	26817	115.9	257.4	145.2816	209.5178	29.38163	-47.8822
42	5257	23259	150.9	248	123.2561	195.2823	-27.6439	-52.7177

Table 5.3: Predicted Values and Residuals with the CM Method - Medical Dataset

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
1	9268	23369	74.7	220.8	172.9748	184.1334	98.27484	-36.6666
2	5337	29897	111.7	209.3	169.8641	189.2992	58.16411	-20.0008
3	11400	29421	78	244	174.662	188.9225	96.66195	-55.0775
4	6582	27952	96.1	238.8	170.8493	187.7601	74.74932	-51.0399
5	12210	31469	71.2	243.5	175.3029	190.5432	104.1029	-52.9568
6	5849	32945	120.7	257.9	170.2693	191.7112	49.56928	-66.1888
7	16263	41090	106.1	267.2	178.5102	198.1566	72.4102	-69.0434
8	8023	31725	116.7	248.6	171.9896	190.7458	55.28963	-57.8542
9	17808	44968	114.5	280.8	179.7328	201.2254	65.2328	-79.5746
10	8342	31716	112.4	256.6	172.2421	190.7386	59.84206	-65.8614
11	12928	33067	127	268	175.8711	191.8077	48.87111	-76.1923
12	7751	32711	145.6	276.7	171.7744	191.526	26.17439	-85.174
13	10492	27287	124.2	269.1	173.9434	187.2338	49.74343	-81.8662
14	6298	18958	141.2	251.2	170.6246	180.6428	29.42458	-70.5572
15	10946	20064	94	236.9	174.3027	181.518	80.30269	-55.382
16	8112	18489	135.7	207	172.0601	180.2717	36.36006	-26.7283
17	11669	29137	80.5	220.3	174.8748	188.6978	94.37482	-31.6022
18	6439	22523	94	226.3	170.7362	183.4639	76.73616	-42.8361
19	12436	31370	84.3	266.3	175.4818	190.4648	91.18177	-75.8352
20	6836	26498	113.6	211	171.0503	186.6095	57.45032	-24.3905
21	16186	40863	125.9	267.4	178.4493	197.9769	52.54926	-69.4231
22	9180	27951	135.8	256	172.9052	187.7593	37.1052	-68.2407
23	18195	44771	118.9	244.5	180.039	201.0695	61.13905	-43.4305
24	10392	32057	143.7	247.4	173.8643	191.0085	30.16429	-56.3915
25	12898	32607	128.1	274.1	175.8474	191.4437	47.74737	-82.6563
26	8197	20126	156.2	257.2	172.1273	181.5671	15.92732	-75.6329
27	10906	27454	109.1	261	174.271	187.366	65.17104	-73.634
28	7215	21993	174	250	171.3502	183.0445	-2.64977	-66.9555
29	9913	23472	108.2	215.6	173.4852	184.2149	65.28524	-31.3851
30	7327	18700	106.1	215.2	171.4389	180.4387	65.33886	-34.7613
31	11772	29155	94.9	211.7	174.9563	188.712	80.05633	-22.988
32	5919	20898	90.1	208.7	170.3247	182.178	80.22467	-26.522
33	12237	31505	57.1	241.1	175.3243	190.5717	118.2243	-50.5283
34	7264	23741	125.6	233.6	171.389	184.4278	45.78901	-49.1722
35	16368	39687	111.4	264.5	178.5933	197.0463	67.19329	-67.4537
36	10643	29799	144.1	239.8	174.0629	189.2217	29.96292	-50.5783
37	17991	44228	128.5	266.5	179.8776	200.6398	51.37762	-65.8602
38	10136	24325	142	271.4	173.6617	184.8899	31.66171	-86.5101
39	12845	32940	123.2	263.4	175.8054	191.7072	52.60543	-71.6928
40	9591	22516	161	239.5	173.2304	183.4584	12.23044	-56.0416
41	10762	26817	115.9	257.4	174.1571	186.8619	58.25708	-70.5381
42	5257	23259	150.9	248	169.8008	184.0463	18.90081	-63.9537
42	5257	23259	150.9	248	123.2561	195.2823	-27.6439	-52.7177

Table 5.4: Predicted Values and Residuals with the CRM Method - Medical Dataset

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
1	9268	23369	74.7	220.8	118.7396	238.3686	44.03959	17.56864
2	5337	29897	111.7	209.3	108.59	250.5733	-3.10997	41.27329
3	11400	29421	78	244	117.7886	245.7959	39.78858	1.795908
4	6582	27952	96.1	238.8	111.7221	246.8873	15.62209	8.087298
5	12210	31469	71.2	243.5	117.5964	248.2497	46.39639	4.74972
6	5849	32945	120.7	257.9	107.2885	254.692	-13.4115	-3.20803
7	16263	41090	106.1	267.2	117.0564	259.6103	10.95642	-7.58966
8	8023	31725	116.7	248.6	111.293	251.4424	-5.40703	2.842411
9	17808	44968	114.5	280.8	116.7089	264.2492	2.208937	-16.5508
10	8342	31716	112.4	256.6	111.7662	251.2145	-0.63385	-5.38545
11	12928	33067	127	268	117.5723	250.1065	-9.42767	-17.8935
12	7751	32711	145.6	276.7	110.2311	253.0693	-35.3689	-23.6307
13	10492	27287	124.2	269.1	117.8951	243.2821	-6.30486	-25.8179
14	6298	18958	141.2	251.2	117.3591	233.9083	-23.8409	-17.2917
15	10946	20064	94	236.9	123.421	232.3998	29.42097	-4.50024
16	8112	18489	135.7	207	120.331	232.0007	-15.369	25.00072
17	11669	29137	80.5	220.3	118.3736	245.199	37.87361	24.899
18	6439	22523	94	226.3	115.1664	239.0337	21.16637	12.73372
19	12436	31370	84.3	266.3	117.994	247.9527	33.69395	-18.3473
20	6836	26498	113.6	211	113.0726	244.5872	-0.52744	33.58723
21	16186	40863	125.9	267.4	117.0964	259.3298	-8.80356	-8.07024
22	9180	27951	135.8	256	115.5271	245.1374	-20.2729	-10.8626
23	18195	44771	118.9	244.5	117.4082	263.7003	-1.49179	19.2003
24	10392	32057	143.7	247.4	114.5385	250.3342	-29.1615	2.934245
25	12898	32607	128.1	274.1	117.838	249.4531	-10.262	-24.6469
26	8197	20126	156.2	257.2	119.3538	234.3406	-36.8462	-22.8594
27	10906	27454	109.1	261	118.389	243.248	9.288977	-17.752
28	7215	21993	174	250	116.6594	237.7354	-57.3406	-12.2646
29	9913	23472	108.2	215.6	119.6148	238.0854	11.41476	22.48538
30	7327	18700	106.1	215.2	119.0395	232.838	12.93955	17.63799
31	11772	29155	94.9	211.7	118.5123	245.156	23.61232	33.45604
32	5919	20898	90.1	208.7	115.4985	237.0041	25.39854	28.30415
33	12237	31505	57.1	241.1	117.6117	248.2843	60.5117	7.184265
34	7264	23741	125.6	233.6	115.5547	240.262	-10.0453	6.662046
35	16368	39687	111.4	264.5	118.1544	257.4852	6.754388	-7.01477
36	10643	29799	144.1	239.8	116.4257	246.8589	-27.6743	7.058879
37	17991	44228	128.5	266.5	117.4749	263.0425	-11.0251	-3.45754
38	10136	24325	142	271.4	119.3672	239.1844	-22.6328	-32.2156
39	12845	32940	123.2	263.4	117.5363	249.9764	-5.66374	-13.4236
40	9591	22516	161	239.5	119.7866	236.9022	-41.2134	-2.59781
41	10762	26817	115.9	257.4	118.6068	242.4122	2.706811	-14.9878
42	5257	23259	150.9	248	112.9402	240.9069	-37.9598	-7.09306

The BCRMO method is applied to the medical dataset as follows. The Equation (2.28) is now specified as

$$\boldsymbol{Y} = \boldsymbol{X}_7\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{Y} = (\boldsymbol{Y}^c, \boldsymbol{Y}^r)$, $\boldsymbol{Y}^c = (Y_1^c, \ldots, Y_n^c)'$ and $\boldsymbol{Y}^r = (Y_1^r, \ldots, Y_n^r)'$ for $i = 1, \ldots, n$; where $\boldsymbol{X}_7 = (\boldsymbol{X}_{7_1}, \ldots, \boldsymbol{X}_{7_n})'$, $\boldsymbol{X}_{7_i} = (1, X_{7_i}^c, X_{7_i}^r)$ for $i = 1, \ldots, n$; and where

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0^c, \beta_0^r \\ \beta_{X_7^c}^c, \beta_{X_7^c}^r \\ \beta_{X_7^r}^c, \beta_{X_7^r}^r \end{pmatrix}.$$

Using the center point and range values, we obtain the estimate of $\boldsymbol{\beta}$ from Equation (2.29); these are shown in Table 5.6. Hence, the model is

$$(Y) = \begin{pmatrix} Y^c \\ Y^r \end{pmatrix} = \begin{pmatrix} Y^c = 163.325 + 0.00039X_7^c + 0.00054X_7^r + \epsilon^c \\ Y^r = 74.466 + 0.00341X_7^c - 0.00065X_7^r + \epsilon^r \end{pmatrix}.$$

The predicted values are obtained by using Equation (2.30) as

$$\begin{aligned} \hat{Y}_L &= \hat{Y}^c - \hat{Y}^r/2 \\ &= \hat{\beta}_0^c + \hat{\beta}_{X_7^c}^c X_7^c + \hat{\beta}_{X_7^r}^c X_7^r - (\hat{\beta}_0^r + \hat{\beta}_{X_7^c}^r X_7^c + \hat{\beta}_{X_7^r}^r X_7^r)/2 \\ &= 163.325 + 0.00039X_7^c + 0.00054X_7^r - (74.466 + 0.00341X_7^c - 0.00065X_7^r)/2 \end{aligned} \quad (5.12)$$

and

$$\begin{aligned} \hat{Y}_U &= \hat{Y}^c + \hat{Y}^r/2 \\ &= \hat{\beta}_0^c + \hat{\beta}_{X_7^c}^c X_7^c + \hat{\beta}_{X_7^r}^c X_7^r + (\hat{\beta}_0^r + \hat{\beta}_{X_7^c}^r X_7^c + \hat{\beta}_{X_7^r}^r X_7^r)/2 \\ &= 163.325 + 0.00039X_7^c + 0.00054X_7^r + (74.466 + 0.00341X_7^c - 0.00065X_7^r)/2. \end{aligned} \quad (5.13)$$

Predicted values from Equations (5.12) and (5.13) and residuals from Equation (5.1) of Y for the medical dataset are shown in Table 5.5.

The BCRWI method is also applied by specifying $\boldsymbol{X}_i$ as $(1, X_{7_i}^c, X_{7_i}^r, X_{7_i}^c X_{7_i}^r)$ for $i = 1, \ldots, n$.

Once again, based on the center point and range values we obtain the estimate of $\hat{\boldsymbol{\beta}}$ from Equation (2.29); these are presented in Table 5.8. Therefore, the model we obtain is

$$(Y) = \begin{pmatrix} Y^c \\ Y^r \end{pmatrix} = \begin{pmatrix} Y^c = 213.328 - 0.00266X_7^c - 0.00159X_7^r + 0.00000X_7^cX_7^r + \epsilon^c \\ Y^r = -90.452 + 0.01349X_7^c + 0.00641X_7^r + 0.00000X_7^cX_7^r + \epsilon^r \end{pmatrix}.$$

The predicted values can be obtained as

$$\begin{aligned} \hat{Y}_L &= \hat{Y}^c - \hat{Y}^r/2 \\ &= \hat{\beta}_0^c + \hat{\beta}_{X_7^c}^c X_7^c + \hat{\beta}_{X_7^r}^c X_7^r + \hat{\beta}_{X_7^cX_7^r}^c X_7^cX_7^r \\ &\quad - (\hat{\beta}_0^r + \hat{\beta}_{X_7^c}^r X_7^c + \hat{\beta}_{X_7^r}^r X_7^r + \hat{\beta}_{X_7^cX_7^r}^r X_7^cX_7^r)/2 \\ &= 213.328 - 0.00266X_7^c - 0.00159X_7^r + 0.00000X_7^cX_7^r \\ &\quad - (-90.452 + 0.01349X_7^c + 0.00641X_7^r + 0.00000X_7^cX_7^r)/2 \end{aligned} \tag{5.14}$$

and

$$\begin{aligned} \hat{Y}_U &= \hat{Y}^c + \hat{Y}^r/2 \\ &= \hat{\beta}_0^c + \hat{\beta}_{X_7^c}^c X_7^c + \hat{\beta}_{X_7^r}^c X_7^r + \hat{\beta}_{X_7^cX_7^r}^c X_7^cX_7^r \\ &\quad + (\hat{\beta}_0^r + \hat{\beta}_{X_7^c}^r X_7^c + \hat{\beta}_{X_7^r}^r X_7^r + \hat{\beta}_{X_7^cX_7^r}^r X_7^cX_7^r)/2 \\ &= 213.328 - 0.00266X_7^c - 0.00159X_7^r + 0.00000X_7^cX_7^r \\ &\quad + (-90.452 + 0.01349X_7^c + 0.00641X_7^r + 0.00000X_7^cX_7^r)/2 \end{aligned} \tag{5.15}$$

where $\hat{\beta}_{X_7^cX_7^r}^c$ and $\hat{\beta}_{X_7^cX_7^r}^r$ are coefficients for the interaction terms. From the results we can see they are quite small and the interaction term could be actually dropped from the model, which suggests a BCRWO model should be considered. For comparison, calculated predicted values from Equations (5.14) and (5.15) and residuals from Equation (5.1) of Y with the BCRWI method for the medical dataset are shown in Table 5.6.

Table 5.5: Predicted Values and Residuals with the BCRWO Method- Medical Dataset

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
1	9268	23369	74.7	220.8	116.9642	237.8336	42.26423	17.03361
2	5337	29897	111.7	209.3	124.3742	242.8297	12.6742	33.52966
3	11400	29421	78	244	115.0099	247.2642	37.0099	3.264222
4	6582	27952	96.1	238.8	122.0537	241.4024	25.95373	2.602385
5	12210	31469	71.2	243.5	114.2133	250.5291	43.01331	7.029096
6	5849	32945	120.7	257.9	124.2481	247.1126	3.548062	-10.7874
7	16263	41090	106.1	267.2	110.0926	266.0726	3.992575	-1.12745
8	8023	31725	116.7	248.6	120.6645	247.3746	3.964483	-1.22536
9	17808	44968	114.5	280.8	108.5671	272.2641	-5.93295	-8.53591
10	8342	31716	112.4	256.6	120.1752	247.6283	7.77524	-8.97174
11	12928	33067	127	268	113.4604	253.148	-13.5396	-14.852
12	7751	32711	145.6	276.7	121.2922	248.3965	-24.3078	-28.3035
13	10492	27287	124.2	269.1	115.9377	243.8091	-8.26232	-25.2909
14	6298	18958	141.2	251.2	120.5519	229.7836	-20.6481	-21.4164
15	10946	20064	94	236.9	113.6896	235.0448	19.6896	-1.85516
16	8112	18489	135.7	207	117.6799	230.6971	-18.0201	23.69711
17	11669	29137	80.5	220.3	114.5378	247.1283	34.03785	26.82827
18	6439	22523	94	226.3	121.1037	234.4126	27.10374	8.112649
19	12436	31370	84.3	266.3	113.8468	250.5916	29.54676	-15.7084
20	6836	26498	113.6	211	121.3528	239.7732	7.752788	28.77322
21	16186	40863	125.9	267.4	110.1613	265.7213	-15.7387	-1.67871
22	9180	27951	135.8	256	118.0848	243.5595	-17.7152	-12.4405
23	18195	44771	118.9	244.5	107.9335	272.3363	-10.9665	27.83628
24	10392	32057	143.7	247.4	117.117	249.7629	-26.583	2.36292
25	12898	32607	128.1	274.1	113.4072	252.5409	-14.6928	-21.5591
26	8197	20126	156.2	257.2	117.9023	232.8395	-38.2977	-24.3605
27	10906	27454	109.1	261	115.3412	244.3644	6.241193	-16.6356
28	7215	21993	174	250	119.8043	234.3866	-54.1957	-15.6134
29	9913	23472	108.2	215.6	116.0011	238.4998	7.801093	22.89981
30	7327	18700	106.1	215.2	118.9244	230.312	12.82444	15.11199
31	11772	29155	94.9	211.7	114.3844	247.2366	19.48438	35.53662
32	5919	20898	90.1	208.7	121.5484	231.924	31.44836	23.22405
33	12237	31505	57.1	241.1	114.1798	250.5971	57.07981	9.497089
34	7264	23741	125.6	233.6	120.1056	236.6395	-5.49439	3.039537
35	16368	39687	111.4	264.5	109.6302	264.3841	-1.76978	-0.11585
36	10643	29799	144.1	239.8	116.2476	247.1137	-27.8524	7.313722
37	17991	44228	128.5	266.5	108.1282	271.4796	-20.3718	4.97958
38	10136	24325	142	271.4	115.844	239.7646	-26.156	-31.6354
39	12845	32940	123.2	263.4	113.5599	252.9183	-9.64013	-10.4817
40	9591	22516	161	239.5	116.2872	237.0224	-44.7128	-2.47761
41	10762	26817	115.9	257.4	115.4241	243.4386	-0.47593	-13.9614
42	5257	23259	150.9	248	123.0678	234.3622	-27.8322	-13.6378
42	5257	23259	150.9	248	112.9402	240.9069	-37.9598	-7.09306

Table 5.6: Predicted Values and Residuals with the BCRWI Method - Medical Dataset

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
1	9268	23369	74.7	220.8	116.4626	237.9565	41.76264	17.15651
2	5337	29897	111.7	209.3	123.7265	242.9884	12.02649	33.68836
3	11400	29421	78	244	106.5322	249.3415	28.53217	5.341478
4	6582	27952	96.1	238.8	120.4254	241.8014	24.32538	3.001372
5	12210	31469	71.2	243.5	105.311	252.7104	34.11098	9.21039
6	5849	32945	120.7	257.9	126.7773	246.4928	6.077291	-11.4072
7	16263	41090	106.1	267.2	114.5247	264.9866	8.424676	-2.21342
8	8023	31725	116.7	248.6	119.8247	247.5804	3.12466	-1.01959
9	17808	44968	114.5	280.8	126.1868	267.9468	11.68678	-12.8532
10	8342	31716	112.4	256.6	118.9453	247.9296	6.54528	-8.67037
11	12928	33067	127	268	104.8314	255.2623	-22.1686	-12.7377
12	7751	32711	145.6	276.7	122.0958	248.1997	-23.5042	-28.5003
13	10492	27287	124.2	269.1	109.326	245.4291	-14.874	-23.6709
14	6298	18958	141.2	251.2	133.933	226.5049	-7.26702	-24.6951
15	10946	20064	94	236.9	117.5127	234.1081	23.51275	-2.79192
16	8112	18489	135.7	207	131.052	227.4206	-4.64797	20.42059
17	11669	29137	80.5	220.3	105.329	249.3847	24.82899	29.08467
18	6439	22523	94	226.3	125.1905	233.4113	31.19051	7.111287
19	12436	31370	84.3	266.3	104.2454	252.9441	19.94541	-13.3559
20	6836	26498	113.6	211	120.0634	240.0891	6.463431	29.08915
21	16186	40863	125.9	267.4	113.9035	264.8044	-11.9965	-2.59562
22	9180	27951	135.8	256	113.6365	244.6494	-22.1635	-11.3506
23	18195	44771	118.9	244.5	122.6574	268.7285	3.757394	24.22855
24	10392	32057	143.7	247.4	113.0402	250.7618	-30.6598	3.361846
25	12898	32607	128.1	274.1	104.136	254.8126	-23.964	-19.2874
26	8197	20126	156.2	257.2	126.1353	230.8222	-30.0647	-26.3778
27	10906	27454	109.1	261	107.828	246.2053	-1.27199	-14.7947
28	7215	21993	174	250	124.4233	233.2548	-49.5767	-16.7452
29	9913	23472	108.2	215.6	114.2486	238.9292	6.048564	23.32923
30	7327	18700	106.1	215.2	132.4038	227.0092	26.30378	11.80922
31	11772	29155	94.9	211.7	104.9331	249.5524	10.03313	37.85241
32	5919	20898	90.1	208.7	129.5525	229.9628	39.45246	21.26284
33	12237	31505	57.1	241.1	105.248	252.7856	48.14795	11.68562
34	7264	23741	125.6	233.6	121.4439	236.3116	-4.15614	2.711631
35	16368	39687	111.4	264.5	107.7794	264.8376	-3.62061	0.337649
36	10643	29799	144.1	239.8	109.5873	248.7457	-34.5127	8.945669
37	17991	44228	128.5	266.5	120.7527	268.3863	-7.74726	1.886261
38	10136	24325	142	271.4	112.4075	240.6067	-29.5925	-30.7933
39	12845	32940	123.2	263.4	104.9561	255.0265	-18.2439	-8.37351
40	9591	22516	161	239.5	116.821	236.8916	-44.179	-2.60839
41	10762	26817	115.9	257.4	108.4576	245.1455	-7.44237	-12.2545
42	5257	23259	150.9	248	126.2528	233.5817	-24.6472	-14.4183
42	5257	23259	150.9	248	112.9402	240.9069	-37.9598	-7.09306

Performance measures presented in Section 5.1 are calculated for each of these methods. The residuals for the lower bounds and the upper bounds are fairly easy to calculate from Equation (3.15) once the predicted values of Y are obtained. The residuals for these different methods are presented in Tables 5.2 - 5.6 for comparison. Their summations are also provided in Table 5.7 to illustrate the conclusion proved in Section 5.1 that the sum of the lower bound and upper bound residuals is 0.

The other performance measures such as RMSE for the lower bounds and the upper bounds are also easy to calculate from Equation (5.1) given the predicted values of Y. For the correlation coefficient r, instead of calculating it for the lower bounds and the upper bounds separately as in Equation (5.2), we calculate the symbolic correlation coefficient between observed interval response Y and predicted interval response $\hat{Y}$ from Equation (5.3).

Parameter estimates and performance measures are presented in Table 5.7. It can be seen from Table 5.7 that the SCM method outperforms the CM method by giving smaller RMSEs. The CM method gives lower bound RMSE 63.01 and upper bound RMSE 60.30 while the SCM method gives 41.36 and 36.69, respectively.

Table 5.7: Comparison of Methods - Medical Dataset

	Method		Parameter Estimates				Sum of Residuals		RMSE		
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\sum Res_L$	$\sum Res_U$	$RMSE_L$	$RMSE_U$	r
1	SCM method	$\hat{\boldsymbol{\beta}}$	102.223	0.004			1957.05	-1957.05	41.36	36.69	0.706
2	CM method	$\hat{\boldsymbol{\beta}}^c$	165.64	0.00079			2409.18	-2409.18	63.01	60.30	0.706
3	CRM method	$\hat{\boldsymbol{\beta}}^c$	165.64	0.00079			0.00	0.00	26.00	18.14	0.884
		$\hat{\boldsymbol{\beta}}^r$	89.491	0.00214							
4	BCRWO method	$\hat{\boldsymbol{\beta}}^c$	163.325	0.00039	0.00054		0.00	0.00	25.10	17.62	0.888
		$\hat{\boldsymbol{\beta}}^r$	74.466	0.00341	-0.00065						
5	BCRWI method	$\hat{\boldsymbol{\beta}}^c$	213.328	-0.00266	-0.00159	0.000000	0.00	0.00	23.91	17.52	0.893
		$\hat{\boldsymbol{\beta}}^r$	-90.452	0.01349	0.00641	0.000000					

It can be seen from Table 5.7 that the CM method has the biggest RMSEs. The SCM method performs better than the CM method with smaller $RMSE_L$ and $RMSE_U$.

It seems the CRM method gives smaller RMSEs. However, it does not necessarily perform better than other methods in terms of revealing the true relationship between Y and X_7. There are two aspects we need to take into account.

First, note the goal of the modeling is not to analyze the range of observed intervals but to quantify the relationship between the response variable Y and the explanatory variable X_7. The CRM method, by taking into account the center point and range of an interval separately, does not necessarily improve the accuracy of modeling by giving the same parameter estimates for center points. For example, a simple regression was built between Y=cholesterol and X=income. Though the data may be in the format of intervals, our interest is still to find out the relationship between cholesterol and income. If the CM method gives us an estimate of the parameter for X as β, then the CRM method will give exactly the same parameter estimate. We can only say when income increases by 1 unit, cholesterol will increase or decrease by β units. The extra parameter estimates β^r from the CRM method can not help in revealing the relationship between the position of cholesterol and income, though it may reduce the RMSEs when calculating predicted intervals.

Secondly, we know that an analysis of two intervals with the same center points, e.g., $Z_1 = [8, 12]$ and $Z_2 = [5, 15]$ does not distinguish between these two differing observations, in that only the between observation variation is used and that the within observation variation is ignored. There is also a problem with ranges. For example, two intervals $W_1 = [0, 10]$ and $W_2 = [100, 110]$ have the same range but are differently valued observations. See also Douzal-Chouakria et al. (2009) which illustrates how using the center points and ranges as surrogates produces incoherent and inconsistent results for some data. Therefore, while the CRM method is an improvement over the CM method, it is still inadequate. Only the SCM method uses all the variations in the data to fit the model. By accounting for the internal

and external variations in addition to the relationship between variables, the SCM method is believed to be more powerful and less misleading over all types of data.

It is also apparent that of the four methods using classical surrogates, the bivariate center and range method with interaction performs the best, as is to be expected since the center and range variables are considered simultaneously and since the interaction term is included. If no interaction between the center and range variables is present, then this becomes the BCRWO method.

5.3 Mushroom Dataset

5.3.1 The Data

The next dataset used to evaluate the performances of methods is a mushroom dataset. The mushroom dataset contains measurements of three features of 100 species of mushrooms, which are members of the genus Agaricies. All measurements in the dataset are interval-valued, extracted from the Fungi of California Species Index. The original dataset can be extracted from http://www.myknoweb.com/CAF/species_index.html. The three features are represented by three variables X_1 = the pileus cap width, X_2 = the stipe length, and X_3 = the stipe thickness. In all, 274 observations are recorded in the dataset and are provided in Table 5.13 in Appendix 5.A.

It is beneficial to examine the covariance matrix between variables X_1, X_2 and X_3 first. The symbolic covariances calculated from Definition 2.3.11 is

$$\mathbf{V} = \begin{pmatrix} 18.39 & 9.77 & 3.81 \\ 9.77 & 13.12 & 2.00 \\ 3.81 & 2.00 & 1.44 \end{pmatrix}. \tag{5.16}$$

It can be seen from Equation (5.16) that pileus cap width (X_1) has the biggest variance of 18.39, stipe thickness (X_3) has the smallest variance of 1.44, and stipe length (X_2) has a variance of 13.12.

Also, the symbolic correlation coefficients from Definition 2.3.12 is

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.629 & 0.742 \\ 0.629 & 1.000 & 0.462 \\ 0.742 & 0.462 & 1.000 \end{pmatrix}. \tag{5.17}$$

Note from Equation (5.17) that stipe thickness and pileus cap width have the biggest correlation coefficient, which is 0.742, the correlation coefficient between stipe thickness and stipe lengh is 0.462, while that between pileus cap width and stipe length is 0.629.

5.3.2 Analysis Results

As an example, we build a model with pileus cap width X_1 as the response variable (i.e., Y) and stipe length X_2 as the explanatory variable (i.e., X). The model is

$$Y = \beta_0 + \beta_1 X_2 + \epsilon \tag{5.18}$$

where Y=pileus cap width and X_2=stipe length.

Again, the current regression methods and our proposed SCM method are applied to this dataset.

First, the proposed SCM method is used to build the linear regression model following the algorithm introduced in Section 3.2.2. We calculate the sample mean for Y and X_2. The results are $\bar{Y}_1 = 6.10$ and $\bar{X}_2 = 6.32$, respectively.

Then, X_1 and X_2 are centered around their sample means. Then, the model in Equation (5.18) can be written as

$$Y = 6.10 + \beta_1(X_2 - 6.32) + \epsilon.$$

Next, using the equations in Definitions 2.3.10 and 2.3.11, we calculate the variance of X_2, which is 13.12, and the covariance between Y and X_2, which is 12.96. We substitute the sample variance and covariance given in Equation (5.16) into Equations (3.8) and (3.9) to obtain

$$(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{X} - \bar{\boldsymbol{X}}) = n \times Var(X_2) = n \times 23.96$$

and

$$(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{Y} - \bar{\boldsymbol{Y}}) = n \times Cov(Y, X_2) = n \times 12.96$$

where n=274 is the sample size for this dataset.

Therefore, the estimates of β_1 can be obtained by substituting results of $(\boldsymbol{X}-\bar{\boldsymbol{X}})'(\boldsymbol{X}-\bar{\boldsymbol{X}})$ and $(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{Y} - \bar{\boldsymbol{Y}})$ into Equation (3.10). Here, it is $\hat{\beta}_1$=0.745.

Finally, we substitute the estimate of β_1 into Equation (3.5) to obtain the estimate of β_0, which gives $\hat{\beta}_0$=1.21. Therefore, the model in Equation (5.18) is

$$Y = 1.390 + 0.745X_2 + \epsilon.$$

The prediction can then be calculated from Equations (2.22) and (2.23); these are given in Table 5.14 in Appendix 5.A. Also shown in Table 5.14 are the residuals obtained from Equation (5.1).

We also applied the CM method to the mushroom dataset. For each observation, we first calculate $Y_i^c = (Y_{iL} + Y_{iU})/2$ and $X_{2i}^c = (X_{2iL} + X_{2iU})/2$, $i = 1, \ldots, n$, then analyze Y^c and X_2^c as though they are classical data. The estimate of parameters $\boldsymbol{\beta}^c$ as $\hat{\boldsymbol{\beta}}^c$ can be obtained as $(1.83, 0.68)'$ using the formula in Equation (2.21). Therefore, the linear regression model in Equation (5.18) can now be written as

$$Y^c = 1.83 + 0.68X_2^c + \epsilon \tag{5.19}$$

where Y^c=pileus cap width center points and X_2^c=stipe length center points.

Another method applied to the mushroom dataset is the CRM method. In addition to obtaining the center points of the original intervals as shown in the CM method, the ranges of the intervals are also calculated by $Y_i^r = (Y_{iU} - Y_{iL})$ and $X_{2i}^r = (X_{2iU} - X_{2iL})$, $i = 1, \ldots, n$.

The model of center points in Equation (2.24) is identical to that obtained in the CM method, which is presented in Equation (5.19). The model of ranges in Equation (2.25) can be obtained by using the parameter estimation from Equation (2.26) as

$$Y^r = 1.92 + 0.70X_2^r + \epsilon.$$

The BCRMO method is also applied to the mushroom dataset. The Equation (2.28) is structured as

$$\boldsymbol{Y} = \boldsymbol{X}_2\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{Y} = (\boldsymbol{Y}^c, \boldsymbol{Y}^r)$, $\boldsymbol{Y}^c = (Y_1^c, \ldots, Y_n^c)'$ and $\boldsymbol{Y}^r = (Y_1^r, \ldots, Y_n^r)'$ for $i = 1, \ldots, n$; where $\boldsymbol{X}_2 = (\boldsymbol{X}_{2_1}, \ldots, \boldsymbol{X}_{2_n})'$, $\boldsymbol{X}_{2_i} = (1, X_{2_i}^c, X_{2_i}^r)$ for $i = 1, \ldots, n$; and where

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0^c, \beta_0^r \\ \beta_{X_2^c}^c, \beta_{X_2^c}^r \\ \beta_{X_2^r}^c, \beta_{X_2^r}^r \end{pmatrix}.$$

The estimate of $\boldsymbol{\beta}$ are obtained from Equation (2.29) and are shown in Table 5.8. Hence, the model is

$$(Y) = \begin{pmatrix} Y^c \\ Y^r \end{pmatrix} = \begin{pmatrix} Y^c = 1.86 + 0.74X_2^c - 0.09X_2^r + \epsilon^c \\ Y^r = 1.43 + 0.35X_2^c + 0.37X_2^r + \epsilon^r \end{pmatrix}.$$

The predicted values are obtained by using Equation (2.30) as

$$\begin{aligned} \hat{Y}_L &= \hat{Y}^c - \hat{Y}^r/2 \\ &= \hat{\beta}_0^c + \hat{\beta}_{X_2^c}^c X_2^c + \hat{\beta}_{X_2^r}^c X_2^r - (\hat{\beta}_0^r + \hat{\beta}_{X_2^c}^r X_2^c + \hat{\beta}_{X_2^r}^r X_2^r)/2 \\ &= 1.86 + 0.74X_2^c - 0.09X_2^r - (1.43 + 0.35X_2^c + 0.37X_2^r) \end{aligned}$$

and

$$\begin{aligned} \hat{Y}_U &= \hat{Y}^c + \hat{Y}^r/2 \\ &= \hat{\beta}_0^c + \hat{\beta}_{X_2^c}^c X_2^c + \hat{\beta}_{X_2^r}^c X_2^r + (\hat{\beta}_0^r + \hat{\beta}_{X_2^c}^r X_2^c + \hat{\beta}_{X_2^r}^r X_2^r)/2 \\ &= 1.86 + 0.74X_2^c - 0.09X_2^r + (1.43 + 0.35X_2^c + 0.37X_2^r) \end{aligned}$$

Finally, the BCRMI method is also obtained by specifying X_i as $(1, X_{2_i}^c, X_{2_i}^r, X_{2_i}^c X_{2_i}^r)$ for $i = 1, \ldots, n$. From Table 5.8, the model is

$$(Y) = \begin{pmatrix} Y^c \\ Y^r \end{pmatrix} = \begin{pmatrix} Y^c = -0.32 - 1.05X_2^c + 0.30X_2^r - 0.04X_2^c X_2^r + \epsilon^c \\ Y^r = -0.74 + 0.66X_2^c + 0.76X_2^r - 0.04X_7^c X_7^r + \epsilon^r \end{pmatrix}.$$

The predicted values can be obtained as

$$\begin{aligned}\hat{Y}_L &= \hat{Y}^c - \hat{Y}^r/2 \\ &= \hat{\beta}_0^c + \hat{\beta}_{X_2^c}^c X_2^c + \hat{\beta}_{X_2^r}^c X_2^r + \hat{\beta}_{X_2^c X_2^r}^c X_2^c X_2^r \\ &\quad - (\hat{\beta}_0^r + \hat{\beta}_{X_2^c}^r X_2^c + \hat{\beta}_{X_2^r}^r X_2^r + \hat{\beta}_{X_2^c X_2^r}^r X_2^c X_2^r)/2 \\ &= -0.32 - 1.05X_2^c + 0.30X_2^r - 0.04X_2^c X_2^r - (-0.74 + 0.66X_2^c + 0.76X_2^r - 0.04X_7^c X_7^r)\end{aligned}$$

and

$$\begin{aligned}\hat{Y}_U &= \hat{Y}^c + \hat{Y}^r/2 \\ &= \hat{\beta}_0^c + \hat{\beta}_{X_2^c}^c X_2^c + \hat{\beta}_{X_2^r}^c X_2^r + \hat{\beta}_{X_2^c X_2^r}^c X_2^c X_2^r \\ &\quad + (\hat{\beta}_0^r + \hat{\beta}_{X_2^c}^r X_2^c + \hat{\beta}_{X_2^r}^r X_2^r + \hat{\beta}_{X_2^c X_2^r}^r X_2^c X_2^r)/2 \\ &= -0.32 - 1.05X_2^c + 0.30X_2^r - 0.04X_2^c X_2^r + (-0.74 + 0.66X_2^c + 0.76X_2^r - 0.04X_7^c X_7^r)\end{aligned}$$

where $\hat{\beta}_{X_2^c X_2^r}^c$ and $\hat{\beta}_{X_2^c X_2^r}^r$ are coefficients for the interaction terms.

Then, for each method, the fitted models are used to determine prediction and the corresponding residuals. Hence, we can calculate performance measures $\sum Res_L$, $\sum Res_U$, $RMSE_L$, $RMSE_U$ and r for each method. The results are presented in Table 5.8. From these performance measures, it can be seen that the SCM method performs better than the CM method with smaller $RMSE_L$ (1.82 compared to 1.89) and $RMSE_U$ (4.99 compared to 5.02). While the correlation coefficient are identical at r=0.629 as for the mushroom dataset, the BCRMI method performs the best among the four classical surrogate methods.

Table 5.8: Comparison of Methods - Mushroom Dataset

	Method	Parameter Estimates					Sum of Residuals	RMSE		
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\sum Res_L$	$\sum Res_U$	$RMSE_L$	$RMSE_U$	r
1	SCM method	$\hat{\boldsymbol{\beta}}$	1.39	0.75			205.82	-205.82	1.82	4.99	0.629
2	CM method	$\hat{\boldsymbol{\beta}}^c$	1.83	0.68			272.51	-272.51	1.89	5.02	0.629
3	CRM method	$\hat{\boldsymbol{\beta}}^c$	1.83	0.68			0.00	0.00	1.58	4.91	0.642
		$\hat{\boldsymbol{\beta}}^r$	1.92	0.70							
4	BCRWO method	$\hat{\boldsymbol{\beta}}^c$	1.86	0.74	-0.09		0.00	0.00	1.56	4.90	0.643
		$\hat{\boldsymbol{\beta}}^r$	1.43	0.35	0.37						
5	BCRWI method	$\hat{\boldsymbol{\beta}}^c$	-0.32	1.05	0.30	-0.04	0.00	0.00	1.52	4.78	0.667
		$\hat{\boldsymbol{\beta}}^r$	-0.74	0.66	0.76	-0.04					

5.4 Bats Dataset

5.4.1 The Data

The bats dataset is used in Chapter 3 to illustrate how to build a regression on symbolic interval-valued data with the SCM method. The model to be built is

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

where X_1=head length, X_2=tail length, X_3=forearm length and Y=weight.

Similar to what is described in Section 5.2 for the medical dataset, we apply in turn the SCM method, the CM method, the CRM method, the BCRMO method and the BCRMI method to this dataset. The estimated parameters and performance measures are presented in the Table 5.9.

From Table 5.9, the SCM method model is

$$Y = -27.266 + 0.513X_1 - 0.156X_2 + 0.461X_3 + \epsilon.$$

The model from the CM method is

$$Y^c = 25.194 + 0.553X_1^c - 0.227X_2^c + 0.433X_3^c + \epsilon. \tag{5.20}$$

The center points model in CRM method is the same as given by the CM method in Equation (5.20). The model on ranges by CRM method is given by

$$Y^r = -6.836 + 0.358X_1^r + 1.073X_2^r + 0.230X_3^r + \epsilon.$$

The parameter estimates from the BCRWO method are

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} -25.300 & 0.586 & -0.215 & 0.380 & 0.007 & -0.112 & 0.154 \\ -25.774 & 0.496 & -0.403 & 0.489 & 0.012 & 0.484 & -0.098 \end{pmatrix}'. \tag{5.21}$$

Therefore, the BCRWO method gives the model as

$$\begin{aligned} Y^c &= -25.300 + 0.586X_1^c - 0.215X_2^c + 0.380X_3^c \\ &\quad + 0.007X_1^r - 0.112X_2^r + 0.154X_3^r + \epsilon^c \\ Y^r &= -25.774 + 0.496X_1^c - 0.403X_2^c + 0.489X_3^c \\ &\quad + 0.012X_1^r - 0.484X_2^r - 0.098X_3^r + \epsilon^r \end{aligned}$$

For the BCRWI method, the parameter estimates are

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} -10.135 & 0.563 & -0.293 & 0.146 & -0.433 & -0.230 & -1.086 & 0.007 & 0.004 & 0.027 \\ 8.505 & 0.480 & -0.934 & 0.233 & 0.468 & -1.918 & -2.505 & -0.010 & 0.060 & 0.052 \end{pmatrix}' \tag{5.22}$$

and the model is given as

$$\begin{aligned} Y^c &= -10.135 + 0.563X_1^c - 0.293X_2^c + 0.146X_3^c \\ &\quad - 0.433X_1^r - 0.230X_2^r - 1.086X_3^r \\ &\quad + 0.007X_1^cX_1^r + 0.004X_1^cX_1^r + 0.027X_1^cX_1^r + \epsilon^c \\ Y^r &= 8.505 + 0.480X_1^c - 0.934X_2^c + 0.233X_3^c \\ &\quad + 0.468X_1^r - 1.918X_2^r - 2.505X_3^r \\ &\quad - 0.010X_1^cX_1^r + 0.060X_1^cX_1^r + 0.052X_1^cX_1^r + \epsilon^r \end{aligned}$$

We use the fitted models to calculate prediction and the corresponding residuals. Then, we can calculate performance measures $\sum Res_L$, $\sum Res_U$, $RMSE_L$, $RMSE_U$ and r for each method and the results are presented in Table 5.9. The relative merits of the different methods are comparable to those found in the medical dataset.

Table 5.9: Comparison of Methods - Bats Dataset

	Method	Parameter Estimates					Sum of Residuals		RMSE		
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\sum Res_L$	$\sum Res_U$	$RMSE_L$	$RMSE_U$	r
1	SCM method	$\hat{\boldsymbol{\beta}}$	-27.266	0.513	-0.156	0.461	19.86	-19.86	4.42	4.62	0.956
2	CM method	$\hat{\boldsymbol{\beta}}^c$	-25.194	0.553	-0.156	0.461	25.05	-25.05	4.44	4.79	0.955
3	CRM method	$\hat{\boldsymbol{\beta}}^c$	-25.194	0.553	-0.227	0.433	32.58	-32.58	3.55	4.07	0.967
		$\hat{\boldsymbol{\beta}}^r$	-6.836	0.358	1.073	0.230					
4	BCRWO method	$\hat{\boldsymbol{\beta}}$	See Equation (5.21)				0	0	0.882	2.320	0.988
5	BCRWI method	$\hat{\boldsymbol{\beta}}$	See Equation (5.22)				0	0	0.53	1.83	0.993

Table 5.10: Comparison of Methods - Blood Pressure Dataset

	Method	Parameter Estimates					Sum of Residuals		RMSE		
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\sum Res_L$	$\sum Res_U$	$RMSE_L$	$RMSE_U$	r
1	SCM method	$\hat{\boldsymbol{\beta}}$	25.23	0.41			44.49	-44.49	12.31	11.66	0.725
2	CM method	$\hat{\boldsymbol{\beta}}^c$	28.32	0.39			48.65	-48.65	12.43	11.68	0.725
3	CRM method	$\hat{\boldsymbol{\beta}}^c$	28.32	0.39			0	0	10.18	10.29	0.766
		$\hat{\boldsymbol{\beta}}^r$	25.44	-0.06							
4	BCRWO method	$\hat{\boldsymbol{\beta}}^c$	31.79	0.33	0.11		0	0	9.786	10.202	0.776
		$\hat{\boldsymbol{\beta}}^r$	7.87	0.17	-0.19						
5	BCRWI method	$\hat{\boldsymbol{\beta}}^c$	0.65	0.57	1.49	-0.00	0	0	9.121	9.556	0.813
		$\hat{\boldsymbol{\beta}}^r$	7.63	0.17	-0.18	0.00					

5.5 Blood Pressure Dataset

5.5.1 The Data

The blood pressure dataset used in Chapter 3 is also analyzed here to compare different methods. The model we want to build is

$$Y = \beta_0 + \beta_1 X_1.$$

Here, Y=pulse rate and X_1=systolic pressure.

Once again, we apply in turn the proposed SCM method, the CM method, the CRM method, the BCRWO method BCRWI method to this dataset and obtain models as follows. The SCM method gives the model

$$Y = 25.23 + 0.41X_1 + \epsilon.$$

The model from the CM method is

$$Y^c = 28.32 + 0.393X_1^c + \epsilon. \tag{5.23}$$

The center points model in CRM method is the same as given by the CM method in Equation (5.23). The model on ranges by CRM method is given by

$$Y^r = 25.44 - 0.06X_1^r + \epsilon.$$

The parameter estimates from the BCRWO method are

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 31.789 & 0.330 & 0.111 \\ 7.866 & 0.170 & -0.194 \end{pmatrix}'.$$

Therefore, the BCRWO method gives the model as

$$Y^c = 31.789 + 0.330X_1^c + 0.111X_1^r + \epsilon^c$$

$$Y^r = 7.866 + 0.170X_1^c - 0.194X_1^r + \epsilon^r$$

For the BCRWI method, the parameter estimates are

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 0.646 & 0.570 & 1.494 & -0.001 \\ 7.629 & 0.172 & -0.183 & 0.000 \end{pmatrix}'$$

and the model is given as

$$Y^c = 0.646 + 0.570X_1^c + 1.494X_1^r - 0.001X_1^cX_1^r + \epsilon^c$$

$$Y^c = 7.629 + 0.172X_1^c - 0.183X_1^r + 0.000X_1^cX_1^r + \epsilon^r$$

The parameter estimates for interaction term seems to be very small, which suggests the interaction term should be dropped and the BCRWO method should be used.

Prediction and the corresponding residuals can be obtained from the fitted models to calculate. Then, we can calculate performance measures $\sum Res_L$, $\sum Res_U$, $RMSE_L$, $RMSE_U$ and r for each method and the results are presented in in Table 5.10. Once again, the relative merits of the different methods are comparable to those found in the medical dataset.

5.6 Simulation

In this section, we compare all methods by simulated datasets. We first generate X_{iL}, X_{iU} and ϵ_i, where $X_{iL} \sim N(0, \sigma_X^2)$, $X_{iU} \sim N(0, \sigma_X^2)$ and $\epsilon \sim N(0, \sigma_\epsilon^2)$, for $i = 1, \ldots, n$. Then, we generate $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. Data are simulated in two scenarios: a small sample size scenario with n=10 and a large sample size scenario with n=500. For each scenario, the β_0 is set to be 100 and the β_1 is set to be 2. We repeat each scenario 100 times. The results of the parameter estimation include the mean and variance of the 100 parameter estimates are presented in Tables 5.11 and 5.12. We can see the proposed SCM method consistently obtains the closest estimates of β_0 and β_1 to their true values among all methods, and obtains the minimum variance of estimated parameters.

Table 5.11: Comparison of Methods - Simulation With Sample Size 10

Method	β_0		β_1	
	Mean	Var	Mean	Var
SCM Method	100.1331	2.605833	1.984326	0.03025
CM Method	100.1688	2.957451	1.972394	0.071055
CRM Method	100.1688	2.957451	1.972394	0.071055
BCRWO Method	99.97394	7.918947	1.958416	0.079208
BCRWI Method	100.1801	12.18413	2.00487	0.370974

Table 5.12: Comparison of Methods - Simulation With Sample Size 500

Method	β_0		β_1	
	Mean	Var	Mean	Var
SCM Method	100.0033	0.046202	2.000467	0.000426
CM Method	100.0033	0.046267	2.000619	0.000755
CRM Method	100.0033	0.046267	2.000619	0.000755
BCRWO Method	99.98188	0.110911	2.000564	0.000743
BCRWI Method	99.98026	0.11009	2.003105	0.001876

5.7 References

[1] Anderson, E. (1935). The Irises of the Gasp Peninsula. *Bulletin of the American Iris Society*, 59, 25.

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

[3] Billard, L. (2008). Sample Covariance Functions for Complex Quantitative Data. Processing, World Conferences International Association of Statistical Computing 2008, Yokohama, Japan.

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

[5] Billard, L. and Diday, E. (2004): Symbolic Data Analysis: Definitions and Examples. CEREMADE, Université Paris, Dauphine, 61.

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

[7] de Carvalho F.A.T., Lima Neto, E.A. and Tenorio, C.P. (2004). A New Method to Fit a Linear Regression Model for Interval-valued Data. *Lecture Notes in Computer Science, KI2004 Advances in Artifical Inteligence*. Springer-Verlag, 295-306.

[8] Fisher, R.A (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7, 179188.

[9] Lima Neto, E., de Carvalho F.A.T. (2010). Constrained Linear Regression Models for Symbolic Interval-valued Variables. *Computational Statistics & Data Analysis*, 54(2), 333-347.

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

5.A Appendix

Table 5.13: Mushroom Dataset

id	species	Pileus Cap Width	Stipe Length	Stipe Thickness
001	AgaricusArorae	[3.00, 8.00]	[4.00, 9.00]	[0.50, 2.50]
002	arvenis	[6.00, 21.00]	[4.00, 14.00]	[1.00, 3.50]
003	augustus	[6.00, 32.00]	[10.00, 37.00]	.
004	benesi	[4.00, 8.00]	[5.00, 11.00]	[1.00, 2.00]
005	bernardii	[7.00, 16.00]	[4.00, 7.00]	[3.00, 4.50]
006	bisporus	[5.00, 12.00]	[2.00, 5.00]	[1.50, 2.50]
007	bitorquis	[5.00, 15.00]	[4.00, 10.00]	[2.00, 4.00]
008	californinus	[4.00, 11.00]	[3.00, 7.00]	[0.40, 1.00]
009	campestris	[5.00, 10.00]	[3.00, 6.00]	[1.00, 2.00]
010	comtulus	[2.50, 4.00]	[3.00, 5.00]	[0.40, 0.70]
011	cupreoBrunneus	[2.50, 6.00]	[1.50, 3.50]	[1.00, 1.50]
012	diminutives	[1.50, 2.50]	[3.00, 6.00]	[0.25, 0.35]
013	fuseoFibrillosus	[4.00, 15.00]	[4.00, 15.00]	[1.50, 2.50]
014	fuscovelatus	[3.50, 8.00]	[4.00, 10.00]	[1.00, 2.00]
015	hondensis	[7.00, 14.00]	[8.00, 14.00]	[1.50, 2.50]
016	lilaceps	[8.00, 20.00]	[9.00, 19.00]	[3.00, 5.00]
017	micromegathus	[2.50, 4.00]	[2.50, 4.50]	[0.40, 0.70]
018	praeclaresquamosus	[7.00, 19.00]	[8.00, 15.00]	[2.00, 3.50]
019	pattersonae	[5.00, 15.00]	[6.00, 15.00]	[2.50, 3.50]
020	perobscurus	[8.00, 12.00]	[6.00, 12.00]	[1.50, 2.00]
021	semotus	[2.00, 6.00]	[3.00, 7.00]	[0.40, 0.80]
022	silvicola	[6.00, 12.00]	[6.00, 12.00]	[1.50, 2.00]
023	subrutilescens	[6.00, 12.00]	[6.00, 16.00]	[1.00, 2.00]
024	xanthodermus	[5.00, 17.00]	[4.00, 14.00]	[1.00, 3.50]
025	AgrocybePediades	[0.80, 3.00]	[2.00, 4.00]	[0.10, 0.20]

Table 5.13: (continued) Mushroom Dataset

id	species	Pileus Cap Width	Stipe Length	Stipe Thickness
026	praecox	[3.00, 6.00]	[4.00, 9.00]	[0.40, 0.70]
027	Alboleptonia	[1.00, 4.50]	[2.50, 5.50]	[0.40, 0.80]
028	AmanitaLanei	[8.00, 25.00]	[10.00, 20.00]	[1.50, 4.00]
029	constricta	[6.00, 12.00]	[9.00, 17.00]	[1.00, 2.00]
030	franchetti	[4.00, 12.00]	[5.00, 15.00]	[1.00, 2.00]
031	magniverrucata	[7.00, 14.00]	[6.00, 12.00]	.
032	muscanta	[6.00, 39.00]	[6.00, 16.00]	[2.00, 3.00]
033	novinupta	[5.00, 14.00]	[6.00, 12.00]	[1.50, 3.50]
034	ocreata	[5.00, 13.00]	[10.00, 22.00]	[1.50, 3.00]
035	pachycolea	[8.00, 18.00]	[10.00, 25.00]	[1.00, 3.00]
036	pantherina	[4.00, 15.00]	[7.00, 11.00]	[1.00, 2.50]
037	phalloides	[3.50, 15.00]	[4.00, 18.00]	[1.00, 3.00]
038	protecta	[4.00, 14.00]	[5.00, 15.00]	[1.00, 3.00]
039	vaginata	[5.50, 10.00]	[6.00, 13.00]	[1.20, 2.00]
040	velosa	[5.00, 11.00]	[4.00, 11.00]	[1.00, 2.50]
041	AniseClitocycle	[1.00, 2.50]	[1.50, 4.00]	[0.30, 0.50]
042	ArmillariaMellea	[3.00, 13.00]	[5.00, 17.00]	[0.50, 3.00]
043	tricholoma	[5.00, 25.00]	[4.00, 15.00]	[1.00, 6.00]
044	Auriscalpium	[1.00, 3.00]	[2.00, 8.00]	[0.10, 0.30]
045	Battarraer	[4.00, 7.00]	[15.00, 35.00]	[0.50, 1.50]
046	Beefsteak	[5.00, 9.00]	[3.00, 6.00]	[1.00, 3.00]
047	Birchholete	[5.00, 14.00]	[8.00, 14.00]	[2.00, 4.00]
048	BoletusAereus	[7.00, 14.00]	[7.00, 13.00]	[3.00, 4.00]
049	ameygdalinus	[4.00, 10.00]	[4.00, 7.00]	[1.50, 3.00]
050	appendiculatus	[7.00, 14.00]	[5.00, 9.00]	[3.00, 6.00]
051	chrysenteron	[4.00, 9.00]	[5.00, 10.00]	[1.00, 1.50]

Table 5.13: (continued) Mushroom Dataset

id	species	Pileus Cap Width	Stipe Length	Stipe Thickness
052	citriniporus	[4.00, 8.00]	[4.00, 7.00]	[1.00, 3.00]
053	dryophilus	[4.00, 12.00]	[4.00, 8.00]	[1.00, 2.50]
054	edulis	[7.00, 25.00]	[7.00, 20.00]	[3.00, 8.00]
055	flaviporus	[6.00, 11.00]	[6.00, 12.00]	[1.00, 2.00]
056	orovillus	[8.00, 15.00]	[5.00, 9.00]	[2.50, 4.50]
057	puloherrimus	[9.00, 17.00]	[7.00, 14.00]	.
058	regius	[8.00, 15.00]	[5.00, 9.00]	[3.00, 4.00]
059	satanas	[10.00, 22.00]	[7.00, 14.00]	.
060	smithii	[7.00, 15.00]	[7.00, 15.00]	[3.50, 7.00]
061	subtomentosus	[4.00, 12.00]	[4.00, 8.00]	[1.00, 2.00]
062	truncatus	[5.00, 10.00]	[5.00, 10.00]	[1.50, 2.50]
063	zelleri	[4.00, 11.00]	[5.00, 10.00]	[1.00, 3.00]
064	CamPratensis	[3.00, 7.00]	[2.50, 7.00]	[1.00, 2.00]
065	russocoriacus	[1.50, 3.50]	[4.00, 7.50]	[0.40, 0.60]
066	CanthCibarius	[3.00, 11.00]	[2.00, 9.00]	[0.50, 3.50]
067	subalbidus	[5.00, 10.00]	[2.00, 5.00]	[2.00, 3.00]
068	tubaeformis	[2.00, 4.00]	[2.50, 7.00]	[0.50, 1.00]
069	Caulorhizza	[5.00, 15.00]	[4.50, 13.00]	[1.30, 2.50]
070	ChalPiperatoides	[2.50, 6.00]	[3.00, 6.00]	[0.80, 1.50]
071	piperatus	[2.00, 7.00]	[2.00, 4.00]	[0.50, 2.00]
072	Chickenlips	[1.50, 3.00]	[2.50, 6.00]	[0.50, 1.00]
073	Chlorophyllum	[5.00, 15.00]	[2.50, 6.50]	[1.00, 2.00]
074	Chromosera	[1.00, 2.50]	[1.00, 2.50]	[0.10, 0.20]
075	Chroogomphus	[3.00, 9.00]	[4.50, 12.00]	[0.50, 2.50]
076	Chrysomphalina	[1.00, 5.00]	[1.50, 3.50]	[0.20, 0.35]
077	ClitAlbirhiza	[2.00, 9.00]	[2.00, 6.00]	[0.50, 1.20]

Table 5.13: (continued) Mushroom Dataset

id	species	Pileus Cap Width	Stipe Length	Stipe Thickness
078	deceptiva	[1.20, 5.00]	[1.50, 4.00]	[0.30, 0.50]
079	inversa	[2.00, 9.00]	[3.00, 7.00]	[0.40, 0.60]
080	nebularis	[5.00, 25.00]	[5.00, 15.00]	[1.50, 4.00]
081	nuda	[4.00, 14.00]	[3.00, 6.50]	[1.00, 2.50]
082	schlerotoidea	[1.00, 3.00]	[1.00, 4.00]	[0.40, 0.80]
083	tarda	[2.00, 6.00]	[1.50, 5.00]	[0.30, 0.70]
084	ClitopilusPrunulus	[2.50, 9.00]	[2.00, 7.00]	[0.70, 1.50]
085	CollybiaGymoropus	[0.50, 3.00]	[4.00, 10.00]	[0.20, 0.30]
086	butyracea	[2.50, 6.50]	[2.50, 5.50]	[0.50, 1.00]
087	gymnopusdrysophil	[2.00, 5.00]	[2.00, 6.50]	[0.30, 0.60]
088	racemosa	[0.70, 1.50]	[4.00, 6.00]	[0.05, 0.10]
089	Coltricia	[1.00, 1.50]	[1.00, 3.00]	[0.10, 0.30]
090	perennis	[1.00, 7.00]	[0.50, 5.00]	[0.30, 0.70]
091	ConocybeFilaris	[1.00, 2.50]	[1.50, 5.00]	[0.10, 0.30]
092	lactea	[0.50, 2.50]	[1.50, 5.00]	[0.10, 0.20]
093	tenera	[1.00, 2.50]	[3.50, 7.00]	[0.15, 0.40]
094	CopAtramentarius	[3.00, 5.00]	[6.00, 15.00]	[1.00, 2.00]
095	auricomus	[1.50, 4.00]	[4.00, 8.00]	[0.20, 0.40]
096	calyptratus	[4.00, 7.00]	[6.00, 10.00]	[0.50, 0.70]
097	comatus	[5.00, 14.00]	[8.00, 20.00]	[1.00, 1.50]
098	disseminatus	[0.50, 1.50]	[1.50, 3.00]	[0.10, 0.20]
099	epheminatus	[0.40, 0.70]	[2.50, 5.00]	[0.05, 0.10]
100	flocculosus	[2.00, 4.00]	[2.00, 6.00]	[0.20, 0.70]
101	lagopus	[3.00, 6.00]	[5.00, 10.00]	[0.30, 0.50]
102	micaceus	[2.00, 5.00]	[1.50, 5.00]	[0.20, 0.50]
103	plicatilis	[1.00, 3.00]	[2.50, 6.50]	[0.10, 0.20]

Table 5.13: (continued) Mushroom Dataset

id	species	Pileus Cap Width	Stipe Length	Stipe Thickness
104	sterquilinus	[2.00, 2.50]	[4.00, 9.00]	[0.60, 1.00]
105	CortMagnivelatus	[4.00, 8.00]	[3.00, 6.00]	[1.50, 3.00]
106	rubicundulus	[3.50, 8.00]	[5.00, 8.00]	[1.00, 3.00]
107	vanduzerensis	[3.50, 7.00]	[8.00, 16.00]	[1.00, 1.50]
108	verrucisporus	[3.00, 7.00]	[1.00, 3.00]	[1.00, 2.00]
109	CystFallax	[1.50, 4.00]	[2.00, 6.00]	[3.00, 7.00]
110	DeerPluteus	[5.00, 13.00]	[5.00, 12.00]	[0.70, 1.70]
111	HedgehogRepandum	[2.00, 12.00]	[2.00, 7.00]	[1.00, 2.50]
112	DentUmbilicatum	[2.50, 5.00]	[2.50, 6.00]	[0.50, 1.00]
113	Entoloma	[5.00, 13.00]	[5.00, 11.00]	[1.50, 3.00]
114	Fairyring	[1.50, 4.00]	[2.00, 6.00]	[0.20, 0.50]
115	Falsechanterelle	[2.50, 7.00]	[2.00, 7.00]	[0.50, 1.00]
116	FalsemorelGyr	[5.00, 9.00]	[3.00, 6.00]	[1.00, 3.00]
117	Floccularia	[3.00, 15.00]	[3.00, 9.00]	[1.50, 3.50]
118	Friedchicken	[4.00, 14.00]	[5.00, 10.00]	[1.00, 3.00]
119	Funnelchanterelle	[2.00, 4.00]	[2.50, 7.00]	[0.50, 1.00]
120	Galerina	[2.00, 6.00]	[2.00, 10.00]	[0.20, 0.60]
121	Galeroposis	[0.30, 0.70]	[2.00, 6.00]	[0.05, 0.10]
122	GomphiOregonensis	[3.00, 12.00]	[5.00, 11.00]	[1.50, 4.00]
123	subroseus	[3.00, 6.00]	[3.00, 7.00]	[0.70, 1.80]
124	GymnopiLutesfolius	[2.50, 8.00]	[2.00, 8.00]	[0.50, 1.50]
125	sapineus	[2.50, 5.00]	[3.00, 6.00]	[0.50, 0.70]
126	spectabilis	[7.00, 20.00]	[7.00, 21.00]	[1.00, 4.00]
127	GympnopuSubpruinosus	[1.50, 4.00]	[2.00, 5.00]	[0.10, 0.30]
128	villosipe	[1.50, 3.00]	[2.00, 5.00]	[0.10, 0.30]
129	Gyromitracalifornia	[5.00, 12.00]	[4.00, 8.00]	[2.00, 4.00]

Table 5.13: (continued) Mushroom Dataset

id	species	Pileus Cap Width	Stipe Length	Stipe Thickness
130	gyromitragigasMontana	[5.00, 10.00]	[2.00, 8.00]	[3.00, 7.00]
131	infula	[3.00, 8.00]	[2.00, 6.00]	[1.00, 2.50]
132	Haymakers	[1.50, 3.00]	[3.00, 7.00]	[0.20, 0.30]
133	HebCrustuliniforme	[4.00, 9.00]	[4.00, 7.00]	[0.70, 1.40]
134	mesophaeum	[2.50, 5.50]	[3.00, 7.00]	[0.30, 0.80]
135	sinapizans	[4.50, 11.00]	[4.00, 8.00]	[1.50, 2.50]
136	HelvellaCompressa	[2.50, 4.00]	[4.00, 10.00]	[0.50, 1.00]
137	lacunosa	[3.00, 5.00]	[3.00, 15.00]	[1.00, 3.00]
138	Honenbuehelia	[3.00, 6.00]	[1.50, 5.00]	[2.00, 3.00]
139	HygrocybeCoccina	[2.50, 5.00]	[2.50, 5.50]	[0.50, 1.00]
140	conica	[2.00, 9.00]	[5.00, 10.00]	[0.50, 1.00]
141	flavescenis	[2.00, 6.00]	[3.50, 7.00]	[0.70, 1.20]
142	miniata	[1.50, 3.50]	[2.00, 4.00]	[0.30, 0.50]
143	psittacina	[1.50, 4.00]	[4.00, 9.00]	[3.00, 5.00]
144	punicea	[4.00, 12.00]	[3.00, 14.00]	[0.50, 2.00]
145	singeri	[2.00, 5.00]	[4.00, 14.00]	[0.50, 1.00]
146	HygrophorusChrysodon	[3.00, 7.00]	[2.50, 7.00]	[0.50, 1.30]
147	eburneus	[2.50, 7.00]	[4.00, 12.00]	[0.50, 1.50]
148	hypothejus	[2.50, 7.00]	[4.00, 8.00]	[0.50, 1.20]
149	pudorinus	[6.00, 14.00]	[12.00, 15.00]	[1.50, 2.00]
150	purpurascens	[4.00, 11.00]	[2.00, 9.00]	[1.50, 2.50]
151	subalpinus	[5.00, 14.00]	[4.00, 6.00]	[2.00, 4.00]
152	HypholomaAurantiaca	[2.50, 6.00]	[4.00, 6.00]	[0.50, 1.00]
153	capnoides	[2.50, 6.00]	[5.00, 7.00]	[0.40, 1.00]
154	fasciculare	[2.00, 7.00]	[2.00, 9.00]	[0.40, 1.50]
155	InocybeGeophylla	[2.00, 4.00]	[2.50, 5.00]	[0.30, 0.60]

Table 5.13: (continued) Mushroom Dataset

id	species	Pileus Cap Width	Stipe Length	Stipe Thickness
156	iilacina	[1.50, 2.50]	[2.00, 5.00]	[0.20, 0.40]
157	sororia	[2.50, 6.50]	[4.00, 10.00]	[0.30, 0.80]
158	Jackolantern	[6.00, 18.00]	[5.00, 15.00]	[1.00, 4.00]
159	Jellybellies	[1.00, 2.50]	[2.50, 5.00]	[0.50, 1.00]
160	Kuehneromyces	[1.00, 3.50]	[2.50, 5.00]	[0.20, 0.35]
161	LaccariaAmethysteo	[1.00, 6.50]	[2.00, 12.00]	[0.30, 1.20]
162	fraterna	[1.50, 4.00]	[1.50, 5.00]	[0.20, 0.50]
163	pallidifolia	[1.50, 5.00]	[3.00, 6.00]	[0.20, 0.60]
164	LactariusAlnicola	[6.00, 13.00]	[2.00, 5.00]	[1.50, 2.50]
165	argillaceifolius	[9.00, 21.00]	[7.00, 14.00]	[2.00, 5.00]
166	deliciosus	[5.00, 13.00]	[3.00, 6.00]	[1.50, 2.50]
167	pallescens	[5.00, 11.00]	[4.00, 8.00]	[1.20, 2.00]
168	pubescens	[3.00, 7.00]	[2.50, 4.00]	[1.50, 2.00]
169	rubrilacteus	[5.00, 12.00]	[2.00, 5.00]	[1.00, 2.50]
170	xanthogalactus	[4.00, 11.00]	[3.00, 6.00]	[1.00, 2.00]
171	Leccinum	[5.00, 18.00]	[9.00, 17.00]	[2.00, 4.00]
172	LentinusPonderosus	[7.00, 30.00]	[6.00, 16.00]	[2.50, 7.00]
173	torulosus	[3.00, 9.00]	[2.00, 4.00]	[1.50, 2.50]
174	LepiotaAtrodisca	[1.50, 4.00]	[2.00, 8.50]	[0.10, 0.40]
175	cepaestipes	[1.50, 4.50]	[2.00, 6.00]	[0.40, 0.80]
176	magnispora	[3.00, 7.00]	[5.00, 13.00]	[0.30, 0.80]
177	cristata	[1.50, 4.50]	[2.50, 6.00]	[0.20, 0.40]
178	lutea	[2.00, 5.00]	[2.50, 7.00]	[0.20, 0.60]
179	naucina	[4.00, 9.00]	[5.00, 11.00]	[0.70, 1.40]
180	rachodes	[7.00, 20.00]	[6.00, 16.00]	[2.00, 3.00]
181	roseifolia	[2.00, 5.50]	[4.00, 10.00]	[0.40, 0.80]

Table 5.13: (continued) Mushroom Dataset

id	species	Pileus Cap Width	Stipe Length	Stipe Thickness
182	roseilivida	[2.00, 4.00]	[2.00, 6.00]	[0.50, 1.00]
183	rubrotinctus	[2.00, 6.00]	[4.00, 10.00]	[0.40, 0.70]
184	rubamericanus	[5.00, 9.00]	[7.00, 12.00]	.
185	rubleucothites	[4.00, 9.00]	[5.00, 11.00]	[0.70, 1.40]
186	Leucoagaricus	[4.00, 13.00]	[5.00, 10.00]	[1.50, 2.50]
187	LeucopaxillusAlbissimus	[4.00, 20.00]	[3.00, 7.00]	[2.50, 4.00]
188	gentianeus	[5.00, 11.00]	[4.00, 8.00]	[1.00, 2.50]
189	Longula	[4.00, 7.00]	[4.00, 7.00]	[2.00, 3.00]
190	Macrocystidia	[2.00, 5.00]	[3.00, 5.00]	[0.20, 0.60]
191	Manhorseback	[5.00, 13.00]	[4.00, 8.00]	[1.50, 3.00]
192	Matsutake	[5.00, 25.00]	[4.00, 15.00]	[1.00, 6.00]
193	Marasmiellus	[0.60, 4.00]	[7.00, 20.00]	[1.50, 4.00]
194	MarasmiusArmenincus	[4.00, 12.00]	[1.00, 3.00]	.
195	calhouniae	[1.00, 3.00]	[1.50, 4.00]	[0.20, 0.50]
196	copelandii	[0.50, 2.00]	[3.00, 8.00]	[0.10, 0.30]
197	plicatulus	[1.00, 4.00]	[5.00, 11.00]	[0.15, 0.35]
198	quercophilus	[2.00, 5.00]	[1.00, 2.50]	.
199	MegaPlatyphylla	[6.00, 9.00]	[7.00, 12.00]	[1.00, 2.00]
200	MorchDeliciosa	[2.00, 7.00]	[1.00, 4.00]	[0.50, 3.00]
201	elata	[2.00, 8.00]	[2.00, 7.00]	[1.50, 3.00]
202	MycenaAdscendens	[2.00, 4.00]	[0.40, 1.00]	.
203	aurantiomarginata	[1.00, 2.00]	[2.50, 7.00]	[0.10, 0.20]
204	californiensis	[0.70, 2.00]	[2.00, 7.00]	[0.10, 0.20]
205	capillaripes	[1.00, 2.00]	[4.00, 6.00]	[0.10, 0.20]
206	galericulata	[2.00, 5.00]	[3.00, 14.00]	[0.20, 0.50]
207	griseoviridis	[1.50, 3.00]	[2.50, 9.00]	[0.20, 0.30]

Table 5.13: (continued) Mushroom Dataset

id	species	Pileus Cap Width	Stipe Length	Stipe Thickness
208	haematopus	[1.00, 3.00]	[2.50, 7.00]	[0.20, 0.30]
209	masulata	[1.50, 4.00]	[2.00, 9.00]	[0.15, 0.40]
210	oregonensis	[0.20, 0.80]	[1.00, 2.50]	.
211	pura	[1.50, 4.50]	[2.00, 6.00]	[0.20, 0.70]
212	purpureofusca	[0.70, 3.00]	[3.00, 7.00]	[0.10, 0.40]
213	Nivatogastrium	[1.50, 5.00]	[0.50, 2.50]	[0.50, 1.50]
214	Nolanea	[2.00, 6.00]	[3.00, 6.00]	[0.30, 0.60]
215	Omphalina	[0.50, 2.00]	[1.00, 2.50]	[0.10, 0.30]
216	Oyster	[5.00, 25.00]	[0.50, 3.00]	[0.50, 2.00]
217	PanaPapilionaceus	[1.50, 4.00]	[6.00, 12.00]	[0.20, 0.40]
218	semiovatus	[4.00, 6.00]	.	[0.50, 1.00]
219	PaxiInvolutus	[5.00, 15.00]	[3.00, 6.00]	[1.00, 3.00]
220	PholiotaTerrestris	[2.00, 8.00]	[3.50, 9.00]	[0.50, 1.00]
221	Plectania	[0.50, 2.00]	[2.00, 5.00]	[0.10, 0.30]
222	PlutAtromargineus	[6.00, 11.00]	[5.00, 12.00]	[0.70, 1.30]
223	flavofuligineus	[3.00, 6.00]	[4.00, 8.00]	[0.30, 0.60]
224	petasatus	[5.00, 13.50]	[5.00, 9.00]	[1.00, 1.50]
225	Polyporus	[2.00, 5.00]	[0.50, 6.00]	[0.40, 0.80]
226	PsathCandolleana	[1.50, 5.50]	[3.00, 7.00]	[0.30, 0.50]
227	echiniceps	[4.00, 9.00]	[6.00, 13.00]	[0.80, 1.60]
228	gracillis	[1.00, 4.00]	[4.00, 8.00]	[0.15, 0.30]
229	hydrophilia	[2.00, 4.50]	[2.00, 7.00]	[0.30, 0.70]
230	longipes	[2.50, 4.50]	[6.00, 12.00]	[0.30, 0.60]
231	PsiloCoprophilia	[1.00, 2.50]	[1.50, 5.00]	[0.10, 0.30]
232	cyanescens	[2.00, 4.50]	[3.00, 6.00]	[0.30, 0.60]
233	montana	[0.70, 1.50]	[1.00, 3.00]	[0.10, 0.20]

Table 5.13: (continued) Mushroom Dataset

id	species	Pileus Cap Width	Stipe Length	Stipe Thickness
234	subviscida	[1.00, 2.00]	[1.50, 4.00]	[0.10, 0.30]
235	Rickenella	[0.30, 1.10]	[0.70, 1.40]	[0.05, 0.20]
236	RussulaSanguinea	[4.00, 10.00]	[5.00, 10.00]	[1.00, 2.50]
237	amoenolens	[4.00, 11.00]	[3.00, 7.00]	[1.00, 2.50]
238	basifurcata	[4.00, 7.00]	[3.00, 7.00]	[1.00, 3.00]
239	brevipes	[6.00, 12.00]	[4.00, 6.00]	[2.00, 3.00]
240	densifolia	[7.00, 13.00]	[3.00, 7.50]	[2.00, 4.00]
241	eccentrica	[7.00, 12.00]	[2.00, 7.50]	[2.50, 3.50]
242	olivacea	[8.00, 16.00]	[8.00, 13.00]	[2.00, 3.50]
243	silvicola	[4.00, 9.00]	[4.00, 10.00]	[1.00, 3.00]
244	Sarcodon	[5.00, 25.00]	[3.50, 8.00]	[2.00, 3.50]
245	Simocybe	[1.00, 2.50]	[1.50, 3.00]	[0.20, 0.40]
246	Slipperyjack	[5.00, 13.00]	[3.00, 8.00]	[1.50, 2.00]
247	StrobiAlbipilatus	[1.50, 3.00]	[1.50, 6.00]	[0.10, 0.20]
248	trullisatus	[0.40, 1.70]	[1.50, 4.50]	[0.10, 0.20]
249	StrophariaAmbigua	[4.00, 14.00]	[7.00, 17.00]	[1.00, 2.00]
250	coronilla	[2.00, 5.00]	[1.50, 4.50]	[4.00, 7.00]
251	riparia	[2.00, 6.00]	[5.00, 13.00]	[0.30, 0.80]
252	semiglobata	[2.00, 4.00]	[3.00, 8.00]	[0.20, 0.50]
253	SuillusBrevipes	[3.50, 10.00]	[1.50, 6.00]	[1.50, 3.50]
254	caerulescens	[6.00, 13.00]	[2.00, 7.00]	[1.00, 3.50]
255	fuscotomentosus	[4.00, 15.00]	[4.00, 7.00]	[2.00, 3.50]
256	lakei	[4.00, 12.00]	[3.00, 7.00]	[1.50, 2.50]
257	tomentosus	[5.00, 11.00]	[5.00, 9.00]	[1.50, 3.00]
258	Thimblefungus	[2.00, 3.50]	[4.00, 10.00]	[1.00, 2.50]
259	TricholomaAtroviolaceum	[3.50, 9.00]	[4.00, 8.00]	[1.50, 3.00]

Table 5.13: (continued) Mushroom Dataset

id	species	Pileus Cap Width	Stipe Length	Stipe Thickness
260	dryophilum	[5.00, 15.00]	[6.00, 13.00]	[1.00, 4.50]
261	fracticum	[5.00, 10.00]	[3.00, 8.00]	[1.50, 2.50]
262	griseovioloceum	[4.00, 11.00]	[6.00, 14.00]	[1.00, 2.00]
263	imbricatum	[6.00, 15.00]	[5.00, 10.00]	[2.00, 3.50]
264	muricatum	[5.00, 12.00]	[3.00, 6.00]	[1.00, 3.50]
265	myomyces	[1.50, 4.50]	[1.50, 5.00]	[0.50, 0.80]
266	saponaceum	[4.00, 9.00]	[4.50, 8.00]	[1.50, 2.00]
267	vernaticum	[4.00, 14.00]	[5.00, 13.00]	[2.00, 3.50]
268	Tricholomopsis	[3.00, 12.00]	[2.50, 10.00]	[1.00, 2.50]
269	TubariaConfragosa	[2.00, 7.00]	[3.00, 5.00]	[0.20, 0.70]
270	furfuracea	[1.00, 4.00]	[1.00, 5.00]	[0.20, 0.70]
271	Tylopilus	[7.00, 12.00]	[7.00, 15.00]	[1.50, 3.00]
272	Volvariella	[6.00, 14.00]	[9.00, 18.00]	[1.50, 2.00]
273	Weraroa	[0.50, 1.00]	[3.00, 10.00]	[2.00, 4.00]
274	Xeromphalina	[0.50, 2.00]	[1.00, 4.00]	[0.05, 0.25]

Table 5.14: Mushroom Dataset Prediction and Residuals

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
1	3	8	4	9	4.304886	8.168382	1.304886	0.168382
2	6	21	4	14	4.304886	12.03188	-1.69511	-8.96812
3	4	8	5	11	5.077585	9.713781	1.077585	1.713781
4	7	16	4	7	4.304886	6.622984	-2.69511	-9.37702
5	5	12	2	5	2.759487	5.077585	-2.24051	-6.92241
6	5	15	4	10	4.304886	8.941082	-0.69511	-6.05892
7	4	11	3	7	3.532186	6.622984	-0.46781	-4.37702
8	5	10	3	6	3.532186	5.850284	-1.46781	-4.14972
9	2.5	4	3	5	3.532186	5.077585	1.032186	1.077585
10	2.5	6	1.5	3.5	2.373137	3.918536	-0.12686	-2.08146
11	1.5	2.5	3	6	3.532186	5.850284	2.032186	3.350284
12	4	15	4	15	4.304886	12.80458	0.304886	-2.19542
13	3.5	8	4	10	4.304886	8.941082	0.804886	0.941082
14	7	14	8	14	7.395683	12.03188	0.395683	-1.96812
15	8	20	9	19	8.168382	15.89538	0.168382	-4.10462
16	2.5	4	2.5	4.5	3.145837	4.691235	0.645837	0.691235
17	7	19	8	15	7.395683	12.80458	0.395683	-6.19542
18	5	15	6	15	5.850284	12.80458	0.850284	-2.19542
19	8	12	6	12	5.850284	10.48648	-2.14972	-1.51352
20	2	6	3	7	3.532186	6.622984	1.532186	0.622984
21	6	12	6	12	5.850284	10.48648	-0.14972	-1.51352
22	6	12	6	16	5.850284	13.57728	-0.14972	1.577278
23	5	17	4	14	4.304886	12.03188	-0.69511	-4.96812
24	0.8	3	2	4	2.759487	4.304886	1.959487	1.304886
25	3	6	4	9	4.304886	8.168382	1.304886	2.168382

Table 5.14: (continued) Mushroom Dataset Prediction and Residuals

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
26	1	4.5	2.5	5.5	3.145837	5.463935	2.145837	0.963935
27	8	25	10	20	8.941082	16.66808	0.941082	-8.33192
28	6	12	9	17	8.168382	14.34998	2.168382	2.349977
29	4	12	5	15	5.077585	12.80458	1.077585	0.804578
30	6	39	6	16	5.850284	13.57728	-0.14972	-25.4227
31	5	14	6	12	5.850284	10.48648	0.850284	-3.51352
32	5	13	10	22	8.941082	18.21347	3.941082	5.213474
33	8	18	10	25	8.941082	20.53157	0.941082	2.531572
34	4	15	7	11	6.622984	9.713781	2.622984	-5.28622
35	3.5	15	4	18	4.304886	15.12268	0.804886	0.122676
36	4	14	5	15	5.077585	12.80458	1.077585	-1.19542
37	5.5	10	6	13	5.850284	11.25918	0.350284	1.25918
38	5	11	4	11	4.304886	9.713781	-0.69511	-1.28622
39	1	2.5	1.5	4	2.373137	4.304886	1.373137	1.804886
40	3	13	5	17	5.077585	14.34998	2.077585	1.349977
41	5	25	4	15	4.304886	12.80458	-0.69511	-12.1954
42	1	3	2	8	2.759487	7.395683	1.759487	4.395683
43	4	7	15	35	12.80458	28.25857	8.804578	21.25857
44	5	9	3	6	3.532186	5.850284	-1.46781	-3.14972
45	5	14	8	14	7.395683	12.03188	2.395683	-1.96812
46	7	14	7	13	6.622984	11.25918	-0.37702	-2.74082
47	4	10	4	7	4.304886	6.622984	0.304886	-3.37702
48	7	14	5	9	5.077585	8.168382	-1.92241	-5.83162
49	4	9	5	10	5.077585	8.941082	1.077585	-0.05892
50	4	8	4	7	4.304886	6.622984	0.304886	-1.37702

Table 5.14: (continued) Mushroom Dataset Prediction and Residuals

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
51	4	12	4	8	4.304886	7.395683	0.304886	-4.60432
52	7	25	7	20	6.622984	16.66808	-0.37702	-8.33192
53	6	11	6	12	5.850284	10.48648	-0.14972	-0.51352
54	8	15	5	9	5.077585	8.168382	-2.92241	-6.83162
55	8	15	5	9	5.077585	8.168382	-2.92241	-6.83162
56	7	15	7	15	6.622984	12.80458	-0.37702	-2.19542
57	4	12	4	8	4.304886	7.395683	0.304886	-4.60432
58	5	10	5	10	5.077585	8.941082	0.077585	-1.05892
59	4	11	5	10	5.077585	8.941082	1.077585	-2.05892
60	3	7	2.5	7	3.145837	6.622984	0.145837	-0.37702
61	1.5	3.5	4	7.5	4.304886	7.009333	2.804886	3.509333
62	3	11	2	9	2.759487	8.168382	-0.24051	-2.83162
63	5	10	2	5	2.759487	5.077585	-2.24051	-4.92241
64	2	4	2.5	7	3.145837	6.622984	1.145837	2.622984
65	5	15	4.5	13	4.691235	11.25918	-0.30876	-3.74082
66	2.5	6	3	6	3.532186	5.850284	1.032186	-0.14972
67	2	7	2	4	2.759487	4.304886	0.759487	-2.69511
68	1.5	3	2.5	6	3.145837	5.850284	1.645837	2.850284
69	5	15	2.5	6.5	3.145837	6.236634	-1.85416	-8.76337
70	1	2.5	1	2.5	1.986788	3.145837	0.986788	0.645837
71	3	9	4.5	12	4.691235	10.48648	1.691235	1.48648
72	1	5	1.5	3.5	2.373137	3.918536	1.373137	-1.08146
73	2	9	2	6	2.759487	5.850284	0.759487	-3.14972
74	1.2	5	1.5	4	2.373137	4.304886	1.173137	-0.69511
75	2	9	3	7	3.532186	6.622984	1.532186	-2.37702

Table 5.14: (continued) Mushroom Dataset Prediction and Residuals

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
76	5	25	5	15	5.077585	12.80458	0.077585	-12.1954
77	4	14	3	6.5	3.532186	6.236634	-0.46781	-7.76337
78	1	3	1	4	1.986788	4.304886	0.986788	1.304886
79	2	6	1.5	5	2.373137	5.077585	0.373137	-0.92241
80	2.5	9	2	7	2.759487	6.622984	0.259487	-2.37702
81	0.5	3	4	10	4.304886	8.941082	3.804886	5.941082
82	2.5	6.5	2.5	5.5	3.145837	5.463935	0.645837	-1.03607
83	2	5	2	6.5	2.759487	6.236634	0.759487	1.236634
84	0.7	1.5	4	6	4.304886	5.850284	3.604886	4.350284
85	1	1.5	1	3	1.986788	3.532186	0.986788	2.032186
86	1	7	0.5	5	1.600438	5.077585	0.600438	-1.92241
87	1	2.5	1.5	5	2.373137	5.077585	1.373137	2.577585
88	0.5	2.5	1.5	5	2.373137	5.077585	1.873137	2.577585
89	1	2.5	3.5	7	3.918536	6.622984	2.918536	4.122984
90	3	5	6	15	5.850284	12.80458	2.850284	7.804578
91	1.5	4	4	8	4.304886	7.395683	2.804886	3.395683
92	4	7	6	10	5.850284	8.941082	1.850284	1.941082
93	5	14	8	20	7.395683	16.66808	2.395683	2.668075
94	0.5	1.5	1.5	3	2.373137	3.532186	1.873137	2.032186
95	0.4	0.7	2.5	5	3.145837	5.077585	2.745837	4.377585
96	2	4	2	6	2.759487	5.850284	0.759487	1.850284
97	3	6	5	10	5.077585	8.941082	2.077585	2.941082
98	2	5	1.5	5	2.373137	5.077585	0.373137	0.077585
99	1	3	2.5	6.5	3.145837	6.236634	2.145837	3.236634
100	2	2.5	4	9	4.304886	8.168382	2.304886	5.668382

Table 5.14: (continued) Mushroom Dataset Prediction and Residuals

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
101	4	8	3	6	3.532186	5.850284	-0.46781	-2.14972
102	3.5	8	5	8	5.077585	7.395683	1.577585	-0.60432
103	3.5	7	8	16	7.395683	13.57728	3.895683	6.577278
104	3	7	1	3	1.986788	3.532186	-1.01321	-3.46781
105	1.5	4	2	6	2.759487	5.850284	1.259487	1.850284
106	5	13	5	12	5.077585	10.48648	0.077585	-2.51352
107	2	12	2	7	2.759487	6.622984	0.759487	-5.37702
108	2.5	5	2.5	6	3.145837	5.850284	0.645837	0.850284
109	5	13	5	11	5.077585	9.713781	0.077585	-3.28622
110	1.5	4	2	6	2.759487	5.850284	1.259487	1.850284
111	2.5	7	2	7	2.759487	6.622984	0.259487	-0.37702
112	5	9	3	6	3.532186	5.850284	-1.46781	-3.14972
113	3	15	3	9	3.532186	8.168382	0.532186	-6.83162
114	4	14	5	10	5.077585	8.941082	1.077585	-5.05892
115	2	4	2.5	7	3.145837	6.622984	1.145837	2.622984
116	2	6	2	10	2.759487	8.941082	0.759487	2.941082
117	0.3	0.7	2	6	2.759487	5.850284	2.459487	5.150284
118	3	12	5	11	5.077585	9.713781	2.077585	-2.28622
119	3	6	3	7	3.532186	6.622984	0.532186	0.622984
120	2.5	8	2	8	2.759487	7.395683	0.259487	-0.60432
121	2.5	5	3	6	3.532186	5.850284	1.032186	0.850284
122	7	20	7	21	6.622984	17.44077	-0.37702	-2.55923
123	1.5	4	2	5	2.759487	5.077585	1.259487	1.077585
124	1.5	3	2	5	2.759487	5.077585	1.259487	2.077585
125	5	12	4	8	4.304886	7.395683	-0.69511	-4.60432

Table 5.14: (continued) Mushroom Dataset Prediction and Residuals

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
126	5	10	2	8	2.759487	7.395683	-2.24051	-2.60432
127	3	8	2	6	2.759487	5.850284	-0.24051	-2.14972
128	1.5	3	3	7	3.532186	6.622984	2.032186	3.622984
129	4	9	4	7	4.304886	6.622984	0.304886	-2.37702
130	2.5	5.5	3	7	3.532186	6.622984	1.032186	1.122984
131	4.5	11	4	8	4.304886	7.395683	-0.19511	-3.60432
132	2.5	4	4	10	4.304886	8.941082	1.804886	4.941082
133	3	5	3	15	3.532186	12.80458	0.532186	7.804578
134	3	6	1.5	5	2.373137	5.077585	-0.62686	-0.92241
135	2.5	5	2.5	5.5	3.145837	5.463935	0.645837	0.463935
136	2	9	5	10	5.077585	8.941082	3.077585	-0.05892
137	2	6	3.5	7	3.918536	6.622984	1.918536	0.622984
138	1.5	3.5	2	4	2.759487	4.304886	1.259487	0.804886
139	1.5	4	4	9	4.304886	8.168382	2.804886	4.168382
140	4	12	3	14	3.532186	12.03188	-0.46781	0.031879
141	2	5	4	14	4.304886	12.03188	2.304886	7.031879
142	3	7	2.5	7	3.145837	6.622984	0.145837	-0.37702
143	2.5	7	4	12	4.304886	10.48648	1.804886	3.48648
144	2.5	7	4	8	4.304886	7.395683	1.804886	0.395683
145	6	14	12	15	10.48648	12.80458	4.48648	-1.19542
146	4	11	2	9	2.759487	8.168382	-1.24051	-2.83162
147	5	14	4	6	4.304886	5.850284	-0.69511	-8.14972
148	2.5	6	4	6	4.304886	5.850284	1.804886	-0.14972
149	2.5	6	5	7	5.077585	6.622984	2.577585	0.622984
150	2	7	2	9	2.759487	8.168382	0.759487	1.168382

Table 5.14: (continued) Mushroom Dataset Prediction and Residuals

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
151	2	4	2.5	5	3.145837	5.077585	1.145837	1.077585
152	1.5	2.5	2	5	2.759487	5.077585	1.259487	2.577585
153	2.5	6.5	4	10	4.304886	8.941082	1.804886	2.441082
154	6	18	5	15	5.077585	12.80458	-0.92241	-5.19542
155	1	2.5	2.5	5	3.145837	5.077585	2.145837	2.577585
156	1	3.5	2.5	5	3.145837	5.077585	2.145837	1.577585
157	1	6.5	2	12	2.759487	10.48648	1.759487	3.98648
158	1.5	4	1.5	5	2.373137	5.077585	0.873137	1.077585
159	1.5	5	3	6	3.532186	5.850284	2.032186	0.850284
160	6	13	2	5	2.759487	5.077585	-3.24051	-7.92241
161	9	21	7	14	6.622984	12.03188	-2.37702	-8.96812
162	5	13	3	6	3.532186	5.850284	-1.46781	-7.14972
163	5	11	4	8	4.304886	7.395683	-0.69511	-3.60432
164	3	7	2.5	4	3.145837	4.304886	0.145837	-2.69511
165	5	12	2	5	2.759487	5.077585	-2.24051	-6.92241
166	4	11	3	6	3.532186	5.850284	-0.46781	-5.14972
167	5	18	9	17	8.168382	14.34998	3.168382	-3.65002
168	7	30	6	16	5.850284	13.57728	-1.14972	-16.4227
169	3	9	2	4	2.759487	4.304886	-0.24051	-4.69511
170	1.5	4	2	8.5	2.759487	7.782033	1.259487	3.782033
171	1.5	4.5	2	6	2.759487	5.850284	1.259487	1.350284
172	3	7	5	13	5.077585	11.25918	2.077585	4.25918
173	1.5	4.5	2.5	6	3.145837	5.850284	1.645837	1.350284
174	2	5	2.5	7	3.145837	6.622984	1.145837	1.622984
175	4	9	5	11	5.077585	9.713781	1.077585	0.713781

Table 5.14: (continued) Mushroom Dataset Prediction and Residuals

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
176	7	20	6	16	5.850284	13.57728	-1.14972	-6.42272
177	2	5.5	4	10	4.304886	8.941082	2.304886	3.441082
178	2	4	2	6	2.759487	5.850284	0.759487	1.850284
179	2	6	4	10	4.304886	8.941082	2.304886	2.941082
180	4	9	5	11	5.077585	9.713781	1.077585	0.713781
181	4	13	5	10	5.077585	8.941082	1.077585	-4.05892
182	4	20	3	7	3.532186	6.622984	-0.46781	-13.377
183	5	11	4	8	4.304886	7.395683	-0.69511	-3.60432
184	4	7	4	7	4.304886	6.622984	0.304886	-0.37702
185	2	5	3	5	3.532186	5.077585	1.532186	0.077585
186	5	13	4	8	4.304886	7.395683	-0.69511	-5.60432
187	5	25	4	15	4.304886	12.80458	-0.69511	-12.1954
188	0.6	4	7	20	6.622984	16.66808	6.022984	12.66808
189	1	3	1.5	4	2.373137	4.304886	1.373137	1.304886
190	0.5	2	3	8	3.532186	7.395683	3.032186	5.395683
191	1	4	5	11	5.077585	9.713781	4.077585	5.713781
192	6	9	7	12	6.622984	10.48648	0.622984	1.48648
193	2	7	1	4	1.986788	4.304886	-0.01321	-2.69511
194	2	8	2	7	2.759487	6.622984	0.759487	-1.37702
195	1	2	2.5	7	3.145837	6.622984	2.145837	4.622984
196	0.7	2	2	7	2.759487	6.622984	2.059487	4.622984
197	1	2	4	6	4.304886	5.850284	3.304886	3.850284
198	2	5	3	14	3.532186	12.03188	1.532186	7.031879
199	1.5	3	2.5	9	3.145837	8.168382	1.645837	5.168382
200	1	3	2.5	7	3.145837	6.622984	2.145837	3.622984

Table 5.14: (continued) Mushroom Dataset Prediction and Residuals

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
201	1.5	4	2	9	2.759487	8.168382	1.259487	4.168382
202	1.5	4.5	2	6	2.759487	5.850284	1.259487	1.350284
203	0.7	3	3	7	3.532186	6.622984	2.832186	3.622984
204	1.5	5	0.5	2.5	1.600438	3.145837	0.100438	-1.85416
205	2	6	3	6	3.532186	5.850284	1.532186	-0.14972
206	0.5	2	1	2.5	1.986788	3.145837	1.486788	1.145837
207	5	25	0.5	3	1.600438	3.532186	-3.39956	-21.4678
208	1.5	4	6	12	5.850284	10.48648	4.350284	6.48648
209	5	15	3	6	3.532186	5.850284	-1.46781	-9.14972
210	2	8	3.5	9	3.918536	8.168382	1.918536	0.168382
211	0.5	2	2	5	2.759487	5.077585	2.259487	3.077585
212	6	11	5	12	5.077585	10.48648	-0.92241	-0.51352
213	3	6	4	8	4.304886	7.395683	1.304886	1.395683
214	5	13.5	5	9	5.077585	8.168382	0.077585	-5.33162
215	2	5	0.5	6	1.600438	5.850284	-0.39956	0.850284
216	1.5	5.5	3	7	3.532186	6.622984	2.032186	1.122984
217	4	9	6	13	5.850284	11.25918	1.850284	2.25918
218	1	4	4	8	4.304886	7.395683	3.304886	3.395683
219	2	4.5	2	7	2.759487	6.622984	0.759487	2.122984
220	2.5	4.5	6	12	5.850284	10.48648	3.350284	5.98648
221	1	2.5	1.5	5	2.373137	5.077585	1.373137	2.577585
222	2	4.5	3	6	3.532186	5.850284	1.532186	1.350284
223	0.7	1.5	1	3	1.986788	3.532186	1.286788	2.032186
224	1	2	1.5	4	2.373137	4.304886	1.373137	2.304886
225	0.3	1.1	0.7	1.4	1.754978	2.295868	1.454978	1.195868

Table 5.14: (continued) Mushroom Dataset Prediction and Residuals

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
226	4	10	5	10	5.077585	8.941082	1.077585	-1.05892
227	4	11	3	7	3.532186	6.622984	-0.46781	-4.37702
228	4	7	3	7	3.532186	6.622984	-0.46781	-0.37702
229	6	12	4	6	4.304886	5.850284	-1.69511	-6.14972
230	7	13	3	7.5	3.532186	7.009333	-3.46781	-5.99067
231	7	12	2	7.5	2.759487	7.009333	-4.24051	-4.99067
232	8	16	8	13	7.395683	11.25918	-0.60432	-4.74082
233	4	9	4	10	4.304886	8.941082	0.304886	-0.05892
234	5	25	3.5	8	3.918536	7.395683	-1.08146	-17.6043
235	1	2.5	1.5	3	2.373137	3.532186	1.373137	1.032186
236	5	13	3	8	3.532186	7.395683	-1.46781	-5.60432
237	1.5	3	1.5	6	2.373137	5.850284	0.873137	2.850284
238	0.4	1.7	1.5	4.5	2.373137	4.691235	1.973137	2.991235
239	4	14	7	17	6.622984	14.34998	2.622984	0.349977
240	2	5	1.5	4.5	2.373137	4.691235	0.373137	-0.30876
241	2	6	5	13	5.077585	11.25918	3.077585	5.25918
242	2	4	3	8	3.532186	7.395683	1.532186	3.395683
243	3.5	10	1.5	6	2.373137	5.850284	-1.12686	-4.14972
244	6	13	2	7	2.759487	6.622984	-3.24051	-6.37702
245	4	15	4	7	4.304886	6.622984	0.304886	-8.37702
246	4	12	3	7	3.532186	6.622984	-0.46781	-5.37702
247	5	11	5	9	5.077585	8.168382	0.077585	-2.83162
248	2	3.5	4	10	4.304886	8.941082	2.304886	5.441082
249	3.5	9	4	8	4.304886	7.395683	0.804886	-1.60432
250	5	15	6	13	5.850284	11.25918	0.850284	-3.74082

Table 5.14: (continued) Mushroom Dataset Prediction and Residuals

	Predictor X		Response Y		Prediction		Residual	
i	X_{iL}	X_{iU}	Y_{iL}	Y_{iU}	$\hat{Y}_{iL}$	$\hat{Y}_{iU}$	Res_{iL}	Res_{iU}
251	5	10	3	8	3.532186	7.395683	-1.46781	-2.60432
252	4	11	6	14	5.850284	12.03188	1.850284	1.031879
253	6	15	5	10	5.077585	8.941082	-0.92241	-6.05892
254	5	12	3	6	3.532186	5.850284	-1.46781	-6.14972
255	1.5	4.5	1.5	5	2.373137	5.077585	0.873137	0.577585
256	4	9	4.5	8	4.691235	7.395683	0.691235	-1.60432
257	4	14	5	13	5.077585	11.25918	1.077585	-2.74082
258	3	12	2.5	10	3.145837	8.941082	0.145837	-3.05892
259	2	7	3	5	3.532186	5.077585	1.532186	-1.92241
260	1	4	1	5	1.986788	5.077585	0.986788	1.077585
261	7	12	7	15	6.622984	12.80458	-0.37702	0.804578
262	6	14	9	18	8.168382	15.12268	2.168382	1.122676
263	0.5	1	3	10	3.532186	8.941082	3.032186	7.941082
264	0.5	2	1	4	1.986788	4.304886	1.486788	2.304886

Chapter 6

Future Work

6.1 Interval-valued Data Multilevel Modeling

6.1.1 Background

Mixed models, containing both fixed effects and random effects, have long been an active topic in statistical research, particularly in the areas that involve repeated measurements or measurements in clusters, i.e., multilevel data structures. However, no extensions to symbolic data exist at present. As future research, we propose an approach incorporating order statistic ideas with mixed models techniques to conduct linear regression for interval-valued data which can be readily generated to analyze symbolic datasets with multilevel structures as a supplementary material. We call it an order statistic method (OSM).

For symbolic interval-valued data, all currently available methods introduced in Chapters 2-5, including the CM method by Billard and Diday (2000), the CRM method by Lima Neto et al. (2004) and de Carvalho et al. (2004), the BCR methods by Billard and Diday (2007), the Constrained Method by Lima Neto et al. (2005, 2010), and the new symbolic covariance method do not include models to reflect hierarchies in the data.

The necessity to construct multilevel models for symbolic data encourages us to look for a possible solution. As part of this proposal for future research, we include a preliminarily study for a new approach that can perform multilevel modeling on interval-valued data by combining ideas of the symbolic likelihood function with those from order statistics.

The likelihood function is fundamental in statistical analysis including regression analysis. Le-Rademacher and Billard (2010) introduced a likelihood function for symbolic random variables. Their idea was to transform the likelihood function of symbolic random variables into

a likelihood function of classical random variables. Then they built the likelihood function based on distributions of these classical random variables because their distribution function can be obtained under necessary assumptions. Furthermore, this approach was extended to the bivariate case in Chapter 4, so allowing us to have maximum likelihood based estimators for the regression model introduced in Chapter 3.

We wish to explore a different approach, one that combines the symbolic likelihood function approach with order statistics to allow hierarchies in the data. This should then allow us to utilize algorithms used in classical mixed models to estimate the parameters in symbolic data regression models. Section 6.1.2 describes briefly the proposed methodology. In Section 6.1.3,, an application illustrating the steps to utilize this OSM method to build symbolic regression models is outlined.

6.1.2 Proposed Methodology

Likelihood Function for A Symbolic Regression Model

Let Y_i be the dependent variable and X_{ij} be the independent variable, $i = 1, \ldots, n$, and $j = 1, \ldots, p$. Suppose in a symbolic regression model, the response variable Y_i consists of a distribution h with parameter δ. Note that the values within Y_i come from a parametric distribution f_i with parameter vector Θ_i.

The regression model we want to establish is

$$Y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i. \tag{6.1}$$

Here, $\boldsymbol{\beta}$ is the vector of parameters to be estimated, ϵ_i is the residual with mean 0 and variance σ^2, $i = 1, \ldots, n$.

Similar to the approach in Chapter 4, we can show than the likelihood function of $\boldsymbol{\beta}$, σ^2 for the regression model in Equation (6.1) becomes

$$L(\boldsymbol{\beta}, \sigma^2|Y) = \prod_{i=1}^{n} h(y_i; \delta) = \prod_{i=1}^{n} g(\theta_i; \tau).$$

Let ξ_i denote a realization of a symbolic variable. Since Y_i is an interval-value random variable, ξ_i is an interval, that is, Y_i=ξ_i=$[a_i, b_i]$ for $i = 1, \ldots, n$. Similarly, X_{ij}=ξ_{ij}=$[c_i, d_i]$ for $i = 1, \ldots, n$, $j = 1, \ldots, p$. Therefore, the data we have are presented as

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} [a_1 & b_1] \\ \vdots & \vdots \\ [a_n & b_n] \end{pmatrix}$$

and

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} [c_{11}, d_{11}] & \cdots & [c_{1p}, d_{1p}] \\ \vdots & \vdots & \vdots \\ [c_{n1}, d_{n1}] & \cdots & [c_{np}, d_{np}] \end{pmatrix}$$

We want to develop a likelihood function based on the interval endpoints of Y. Suppose f_i is the internal distribution of Y_i with Θ_i as the internal parameters of the distribution. For the interval variable Y_i, suppose f_i is a uniform distribution and $\Theta_i = (Y_{iL}, Y_{iU})$, where Y_{iL} and Y_{iU} are the lower bound and upper bound for the uniform distribution, respectively. I.e., we assume $Y_i \sim uniform(Y_{iL}, Y_{iU})$.

Then, as for classical data, observed values $[a_i, b_i]$ can be used to estimate the parameters $\Theta_i = (Y_{iL}, Y_{iU})$ of the uniform distribution. Note the difference between parameters (Y_{iL}, Y_{iU}) and observations $[a_i, b_i]$. On one hand, Y_{iL} and Y_{iU} are unknown and we want to estimate them from sample data. On the other hand, $[a_i, b_i]$ are known observed values.

An interval can be thought of as two end points filled with infinite points in between. Therefore, for an interval observation $[a_i, b_i]$, we can take it as a sample with infinite sample size m, minimum value a_i and maximum value b_i. That is, there is a sample of Y_i, say, $Y_{i1}, \ldots, Y_{im}$. We order the sample from minimum to maximum as $Y_{i(1)}, \ldots, Y_{i(m)}$; then a_i is the order statistic $Y_{i(1)}$ and b_i is the order statistic $Y_{i(m)}$. Therefore, we have $a_i = Y_{i(1)}$ and $b_i = Y_{i(m)}$. This assumption is especially meaningful when the interval data are aggregated from a large classical data set.

For each Y_i, $i = 1, \ldots, n$, we have Y_{iL} and Y_{iU} as the parameters for its distribution. Note Y_{iL} and Y_{iU} are parameters. Therefore, though Y_i itself is a symbolic variable and has realizations as intervals, Y_{iL} and Y_{iU} are classical. Since it is necessary to assume there is a correlation between the lower bound (Y_{iL}) and the upper bound (Y_{iU}) from the same interval, we assume Y_{iL} and Y_{iU} follow a bivariate normal distribution as

$$(Y_{iL}, Y_{iU}) \sim N(\mu_{iL}, \mu_{iU}, \sigma_L^2, \sigma_U^2, \rho),\ i = 1, \ldots, n.$$

Then the likelihood function can be written as

$$\begin{aligned} L &= \prod_{i=1}^{n} g(Y_{iL}, Y_{iU}; \tau) \\ &= \prod_{i=1}^{n} \frac{1}{2\pi\sigma_L^2\sigma_U^2\sqrt{1-\rho^2}} \\ &\quad \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(Y_{iL}-\mu_{iL})^2}{\sigma_L^2} - \frac{2\rho(Y_{iL}-\mu_{iL})(Y_{iU}-\mu_{iU})}{\sigma_L\sigma_U} + \frac{(Y_{iU}-\mu_{iU})^2}{\sigma_U^2}\right]\right\} \end{aligned} \tag{6.2}$$

where $\tau = (\mu_{iL}, \mu_{iU}, \sigma_L, \sigma_U, \rho)$.

The lower bounds Y_{iL} and upper bounds Y_{iU} in Equation (6.2) are unknown parameters. Therefore, we want to find the relationships between the unknown parameters (Y_{iL}, Y_{iU}) and the observed values $[a_i, b_i]$, $i = 1, \ldots, n$. We need some results from order statistics.

Theorem 6.1.1. Let $X \sim U[0,1]$, then the distribution of the kth order statistic $X_{(k)}$ from a sample with m observations is given by

$$f_{X_{(k)}}(x) = \frac{m!}{(k-1)!(m-k)!}x^{k-1}(1-x)^{m-k}, \quad \leq x \leq 1.$$

Theorem 6.1.2. Let $X_1, \ldots, X_m$ be identically independently distributed uniform random variables on [0,1]. Then, for the kth order statistic $X_{(k)}$, $E[X_{(k)}] = k/(m+1)$.

Proof. The probability distribution function for the kth order statistic is given in 6.1.1. We further show that if $X_i \sim uniform(0,1)$ then

$$f_{X_{(k)}}(x) == \frac{\Gamma(m+1)}{\Gamma(k)\Gamma(m-k+1)}x^{k-1}(1-x)^{(m-k+1)-1}.$$

Therefore, the kth order statistic has a $beta(k, n-k+1)$ distribution. Now it is trivial to show that

$$EX_{(k)} = k/(m+1).$$

Furthermore, it can be proved that if $X \sim uniform(X_L, X_U)$, then,

$$\frac{X - X_L}{X_U - X_L} = Z \sim uniform(0,1). \tag{6.3}$$

Combining Equation (6.3) and Theorem 6.1.2, we have the following theorem.

Theorem 6.1.3. Let $X_1, \ldots, X_m$ be identically independently distributed variables having distribution as $uniform(X_L, X_U)$. For the jth order statistic $X_{(k)}$, we have $E[X_{(k)}] = k(X_U - X_L)/(m+1) + X_L$. $E[X_{(1)}] \to X_L, E[X_{(m)}] \to X_U$ as $m \to \infty$.

Now come back to our symbolic data situation, response variable Y_i is assumed to follow a uniform distribution $uniform(Y_{iL}, Y_{iU})$. The observed interval $[a_i, b_i]$ is thought of as a sample with minimum value $Y_{i(1)}$=a_i, maximum value $Y_{i(m)}$=b_i and sample size m goes to infinity. Therefore, from Theorem 6.1.3, $E[Y_{i(1)}]$ is an unbiased moment estimator for Y_{iL}, and $E[Y_{i(n)}]$ is an unbiased moment estimator for Y_{iU}, that is,

$$E(Y_{i(1)}) = Y_{iL} \text{ and } E(Y_{i(m)}) = Y_{iU}. \tag{6.4}$$

Since $a_i = Y_{i(1)}$ and $b_i = Y_{i(m)}$, from Equation (6.4), we obtain

$$E(a_i) = E(Y_{i(1)}) = Y_{iL} \text{ and } E(b_i) = E(Y_{i(m)}) = Y_{iU}. \tag{6.5}$$

Therefore, a_i and b_i are unbiased estimators for Y_{iL} and Y_{iU}, respectively.

Let μ_{iL} and μ_{iU} be the expected values of Y_{iL} and Y_{iU}, respectively. Then, from Equation (6.5), we have

$$EY_{i(1)} = E(Y_{iL}) = \mu_{iL} \text{ and } EY_{i(m)} = E(Y_{iU}) = \mu_{iU}, \text{ as } m \to \infty. \tag{6.6}$$

The residuals $\epsilon's$ are assumed to have mean 0. Therefore,

$$\begin{aligned} E(Y_{iL}) &= \boldsymbol{c}_i\boldsymbol{\beta} \\ E(Y_{iU}) &= \boldsymbol{d}_i\boldsymbol{\beta} \end{aligned} \tag{6.7}$$

where $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{ip})$ and $\boldsymbol{d}_i = (d_{i1}, \ldots, c_{dp})$ are the observed values for the predictor variables $\boldsymbol{X}$.

Substituting Equation (6.7) into Equation (6.6), we can obtain

$$\begin{aligned} \mu_{iL} &= E(Y_{iL}) = \boldsymbol{c}_i\boldsymbol{\beta} \\ \mu_{iU} &= E(Y_{iU}) = \boldsymbol{d}_i\boldsymbol{\beta}. \end{aligned} \tag{6.8}$$

From Equations (6.5) and (6.8), the likelihood function of $\boldsymbol{\beta}$ in Equation (6.2) can now be written as

$$\begin{aligned} L(\boldsymbol{\beta}; \sigma^2, \rho|Y) = \prod_{i=1}^{n} & \frac{1}{2\pi\sigma_L^2\sigma_U^2\sqrt{1-\rho^2}} \\ & \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(a_i - \boldsymbol{c}_i\boldsymbol{\beta})^2}{\sigma_L^2} - \frac{2\rho(a_i - \boldsymbol{c}_i\boldsymbol{\beta})(b_i - \boldsymbol{d}_i\boldsymbol{\beta})}{\sigma_L\sigma_U} + \frac{(b_i - \boldsymbol{d}_i\boldsymbol{\beta})^2}{\sigma_U^2}\right]\right\}. \end{aligned} \tag{6.9}$$

We have a likelihood function for the interval-valued data regression model (6.1) as shown in Equation (6.9). Based on the likelihood function, classical mixed model theory can be applied to interval-valued data regression to obtain parameter estimators

We can also accommodate a multilevel data structure by including ideas from classical mixed models.

Classical linear mixed models may be expressed in a number of different but equivalent forms. Laird and Ware (1982) used the form of the linear mixed models as

$$Y_{uv} = \beta_0 + \beta_1 X_{1uv} + \cdots + \beta_p X_{puv} + b_{u1} Z_{1uv} + \cdots + b_{uq} Z_{quv} + \epsilon_{uv} \tag{6.10}$$

where $b_{ul} \sim N(0, d_l^2)$, $Cov(b_{ul}, b_{ul'}) = d_{ll'}$, $\epsilon_{uv} \sim N(0, \sigma^2\lambda_{uv})$, $Cov(\epsilon_{uv}, \epsilon_{uv'}) = \sigma^2\lambda_{uvv'}$, and $u = 1, \ldots, N$, $v = 1, \ldots, N_u, i = 1, \ldots, p$, $l = 1, \ldots, q$.

In model (6.10), we have:

- variable Y_{uv} is the value of the response variable for the vth observation in the uth group/subject. Totally, there are N groups/subjects. In the uth group/subject, there are N_u observations.

- variables X_{juv}, $j = 1, \ldots, p$, are the p fixed-effect independent variables for the vth observation in the uth group/subject.

- variables Z_{luv}, $l = 1, \ldots, q$, are the q random-effect independent variables for the vth observation in the uth group/subject.

- parameters β_j, $j = 1, \ldots, p$, are the fixed-effect coefficients, which are identical for all groups/subjects.

- parameters b_{ul}, $l = 1, \ldots, q$, are the random-effect coefficients for group/subject u. Therefore, the random effects vary by group/subject. The b_{ul} are regarded as random variables rather than parameters. Therefore, they are similar in this respect to the errors ϵ_{uv}.

- variables d_l^2 and $d_{ll'}$ are the variances and covariances, respectively, among the random effects. They are assumed to be constant across groups/subjects.

- variableϵ_{uv} is the error for observation v in group/subject u.

- variable $\sigma^2\lambda_{uv}$ is the covariance between errors in group/subject u. The λ_{uv} depends on u only through its dimension N_u, i.e., the set of unknown parameters λ_{uv} in group/subject u will not depend on u. In some cases, this last assumption may be relaxed, as in an example in Lin et al. (1997). In group/subject u, λ_{ijj} are usually assumed to be parameterized into a few basic parameters. For example, when observations are sampled independently within groups/subjects and are assumed to have constant error variances, we have $\lambda_{uv} = \sigma^2$, $\lambda_{uvv'} = 0$ (for $v \neq v'$). Another example is if the observations in a group/subject have autocorrelations; then the structure of the $\lambda's$ will be specified to capture this characteristic.

Then, we can use established procedures to find estimates of the parameters in these mixed models. Several methods to calculate the Maximum Likelihood (ML) or Restricted

Maximum Likelihood (REML) estimates have been developed. Dempster et al. (1977) introduced the EM algorithm for the calculation of MLEs. Larid and Ware (1982) showed the EM algorithm can be applied to find not only MLEs, but also REML estimates. A better algorithm, the Newton-Raphson procedure, is widely used to estimate all parameters in the model nowadays. Details about the Newton-Raphson algorithm can be found in, e.g., Lindstrom and Bates (1988). Several statistical software packages have been developed to build mixed models, such as the **MIXED** procedure or **GLMMIX** procedure in **SAS**, and the **NLME** package in **R**.

Linear mixed models can be used to obtain the parameter estimates of symbolic models from maximizing the likelihood function in Equation (6.9). To link up these two quite different models through their likelihood function, we need first to specify the structure of the mixed model used to estimate symbolic model parameters.

Suppose we have symbolic interval-valued data. We propose to reorganize the original dataset into a new format like the one shown in Table 6.1. Note each symbolic interval-valued variable in the symbolic dataset has been transposed into a new classical random variable in the new dataset.

Table 6.1: Reorganized Interval-valued Dataset

ID	ORDER	Y	X_1	X_2	$\cdots$	X_p
1	1	a_1	c_{11}	c_{12}	$\cdots$	c_{1p}
1	2	b_1	d_{11}	d_{12}	$\cdots$	d_{1p}
2	1	a_2	c_{21}	c_{22}	$\cdots$	c_{2p}
2	2	b_2	d_{21}	d_{22}	$\cdots$	d_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	1	a_n	c_{n1}	c_{n2}	$\cdots$	c_{np}
n	2	b_n	d_{n1}	d_{n2}	$\cdots$	d_{np}

Another important generalization is a multilevel model. This is especially helpful if there is a hierarchy structure in the symbolic data such as symbolic longitudinal data in clinical trials, symbolic panel data in social science, and symbolic space data in geography, etc. As mixed models can analyze classical data with a multilevel structure, the OSM method can

handle symbolic interval-valued data with a multilevel structure, too. We will use an example to explain how we propose to build a multilevel model for symbolic interval-valued data with the proposed OSM method in Section 6.1.3.

6.1.3 Application

We apply our proposed OSM method to the so-called Iris dataset. The Iris data were collected by Edgar Anderson (1935) to determine geographic variation of Iris flowers. Fisher (1936) used the dataset as an application of discriminant analysis. The Iris dataset contains measurements of 50 Iris flowers each from three species: *setosa*, *versicolor*, and *virginica*. For each flower observation, X_1=sepal length, X_2=sepal width, X_3=petal length, and X_4=petal width were measured.

The original Iris dataset has 150 classical observations. Let us suppose each set of five consecutive flowers in the original dataset came from the same location, e.g., a nursery. We aggregate each set of five consecutive classical observations into one interval-valued symbolic observation because we are interested in the characteristics of groups of flowers by nursery instead of features of individual flowers. This results in an interval-valued dataset consisting of 30 observations with 10 observations from each of the three species: *setosa*, *versicolor*, and *virginica*. Table 6.2 shows the interval-valued Iris dataset. More details about the Iris dataset and how to aggregate it into a interval-valued format can be found in Billard and Diday (2007).

Before we conduct any regression analysis, let us first compute the symbolic variance-covariance matrix as defined in Definitions 2.3.10 and 2.3.11 for the data. The resulting matrix is

Table 6.2: Iris Dataset

ID	species	Sepal Length	Sepal Width	Petal Length	Petal Width
1	1	[4.6, 5.1]	[3, 3.6]	[1.3, 1.5]	[0.2, 0.2]
2	1	[4.4, 5.4]	[2.9, 3.9]	[1.4, 1.7]	[0.1, 0.4]
3	1	[4.3, 5.8]	[3, 4]	[1.1, 1.6]	[0.1, 0.2]
4	1	[5.1, 5.7]	[3.5, 4.4]	[1.3, 1.7]	[0.3, 0.4]
5	1	[4.6, 5.4]	[3.3, 3.7]	[1, 1.9]	[0.2, 0.5]
6	1	[4.7, 5.2]	[3, 3.5]	[1.4, 1.6]	[0.2, 0.4]
7	1	[4.8, 5.5]	[3.1, 4.2]	[1.4, 1.6]	[0.1, 0.4]
8	1	[4.4, 5.5]	[3, 3.5]	[1.2, 1.5]	[0.1, 0.2]
9	1	[4.4, 5.1]	[2.3, 3.8]	[1.3, 1.9]	[0.2, 0.6]
10	1	[4.6,5.3]	[3, 3.8]	[1.4, 1.6]	[0.2, 0.3]
11	2	[5.5, 7]	[2.3, 3.2]	[4, 4.9]	[1.3, 1.5]
12	2	[4.9,6.6]	[2.4, 3.3]	[3.3, 4.7]	[1, 1.6]
13	2	[5, 6.1]	[2, 3]	[3.5, 4.7]	[1, 1.5]
14	2	[5.6, 6.7]	[2.2, 3.1]	[3.9, 4.5]	[1, 1.5]
15	2	[5.9, 6.4]	[2.5, 3.2]	[4, 4.9]	[1.2, 1.8]
16	2	[5.7, 6.8]	[2.6, 3]	[3.5, 5]	[1, 1.7]
17	2	[5.4, 6]	[2.4, 3]	[3.7, 5.1]	[1, 1.6]
18	2	[5.5, 6.7]	[2.3, 3.4]	[4, 4.7]	[1.3, 1.6]
19	2	[5, 6.1]	[2.3, 3]	[3.3, 4.6]	[1, 1.4]
20	2	[5.1, 6.2]	[2.5, 3]	[3, 4.3]	[1.1, 1.3]
21	3	[5.8, 7.1]	[2.7, 3.3]	[5.1, 6]	[1.8, 2.5]
22	3	[4.9, 7.6]	[2.5, 3.6]	[4.5, 6.6]	[1.7, 2.5]
23	3	[5.7, 6.8]	[2.5, 3.2]	[5, 5.5]	[1.9, 2.4]
24	3	[6, 7.7]	[2.2, 3.8]	[5, 6.9]	[1.5, 2.3]
25	3	[5.6, 7.7]	[2.7, 3.3]	[4.9, 6.7]	[1.8, 2.3]
26	3	[6.1, 7.2]	[2.8, 3.2]	[4.8, 6]	[1.6, 2.1]
27	3	[6.1, 7.9]	[2.6, 3.8]	[5.1, 6.4]	[1.4, 2.2]
28	3	[6, 7.7]	[3, 3.4]	[4.8, 6.1]	[1.8, 2.4]
29	3	[5.8, 6.9]	[2.7, 3.3]	[5.1, 5.9]	[1.9, 2.5]
30	3	[5, 6.7]	[2.5, 3.4]	[5, 5.4]	[1.8, 2.3]

$$\mathbf{V} = \begin{pmatrix} 0.6007 & -0.1275 & 1.2369 & 0.5271 \\ -0.1275 & 0.1675 & -0.4471 & -0.1854 \\ 1.2369 & -0.4471 & 2.9937 & 1.2711 \\ 0.5271 & -0.1854 & 1.2711 & 0.5592 \end{pmatrix}. \tag{6.11}$$

Matrix **S** of Equation (6.11) shows that the variance for Sepal Length X_1 (0.6007) is similar to the variance for Petal Width X_4 (0.5592). Petal Length, X_3, has the largest variance (2.9937) whereas Sepal Width, X_2, has the smallest variance (0.1675). The symbolic correlation matrix as defined in Equation (2.3.12) for the Iris data is

$$\mathbf{R} = \begin{pmatrix} 1.0000 & -0.4019 & 0.9224 & 0.9095 \\ -0.4019 & 1.0000 & -0.6313 & -0.6056 \\ 0.9224 & -0.6313 & 1.0000 & 0.9825 \\ 0.9095 & -0.6056 & 0.9825 & 1.0000 \end{pmatrix}. \tag{6.12}$$

The elements of the symbolic correlation matrix in Equation (6.12) indicate strong correlation between Sepal Length, Petal Length, and Petal Width. The correlation coefficient between Sepal Length and Petal Length is 0.9224. The coefficient of correlation between Sepal Length and Petal Width is 0.9095. The coefficient of correlation between Petal Length and Petal Width is 0.9825. Meanwhile, the correlation matrix in Equation (6.12) also shows the correlation coefficient between Sepal Width and Petal Length is negative (-0.6313), the correlation coefficient between Sepal Width and Petal Width is also negative (-0.6056).

The variance-covariance matrix of Equation (6.11) and correlation coefficient matrix of Equation (6.12) tell us the inside structure of the data, which helps us to build a model to reflect this structure. Note that Petal Length and Petal Width are almost perfectly related with the correlation coefficient 0.9825. Therefore, we only want to use one from these two variables. Another interesting feature is that Sepal Length and Sepal Width are negatively related. On the contrary, Petal Length and Petal Width are positively related. With this

knowledge of the data structure, one basic multivariate model we can build is using Petal Length (Y) as the response variable, and Sepal Length (X_1) and Sepal Width (X_2) as the two explanatory variables. Another model we build is the a univariate model using Petal Length (Y) as the response variable and Sepal Length (X_1) as the explanatory variable.

We build two models for this dataset. One is a multivariate model; the other is a univariate model. Then we compare the results from the two different models.

Multivariate Model

For the Iris data, after transforming the original dataset into the new dataset, we can build the corresponding mixed model to estimate the parameters $\boldsymbol{\beta}$ in the symbolic interval-valued model. The specific multivariate symbolic model for the Iris dataset is specified in Equation (6.13).

If we assume the lower bound and upper bound of Y_i have the same variance, then the corresponding basic mixed model is

$$\begin{pmatrix} X_{3_{iL}} \\ X_{3_{iU}} \end{pmatrix} = \begin{pmatrix} 1 & X_{1_{iL}} & X_{2_{iL}} \\ 1 & X_{1_{iU}} & X_{2_{iU}} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} b_i + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \tag{6.13}$$

where b_i stands for the random effect, ϵ_i stands for the error, and

$$b_{i1} \sim N(0, d^2), \quad \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N\left(0, \Sigma_i\right).$$

The Iris dataset contains flower information from three different species. The flowers from the same species may carry similar characteristics, while flowers from different species may have distinct features, i.e., there is a hierarchy in the dataset. Obviously, it may not be the best solution if we do not take into account that observations are homogeneous within species while heterogeneous between species. To reflect this structure of the data, we want to build a multilevel model with a hierarchy in the random effects. We can compare the

benefits we obtain from building this more complex model from those from the simple model in Equation (6.13). The corresponding mixed model with a hierarchy, i.e., multilevel model, in the random effects is

$$\begin{pmatrix} X_{3_{iL}} \\ X_{3_{iU}} \end{pmatrix} = \begin{pmatrix} 1 & X_{1_{iL}} & X_{2_{iL}} \\ 1 & X_{1_{iU}} & X_{2_{iU}} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} b_{i1} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} b_{i2} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \tag{6.14}$$

where

$$b_{i1} \sim N(0, d_1^2), \quad b_{i2} \sim N(0, d_2^2), \quad \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N\left(0, \Sigma_i\right).$$

In the Equation (6.14), the random variable b_{i1} stands for the effect of Species, the random variable b_{i2} stands for the effect of ID, i.e., the difference between the lower bound and the upper bound.

The parameter estimates for the basic multilevel model are given in Table 6.3.

Table 6.3: Multivariate Model for Iris Data

Parameters	Basic Model	Multilevel Model
β_0	-2.243	-0.4160
β_1	1.799	0.7596
β_2	-1.461	-0.0874

Univariate Model

Following the steps shown in building multivariate model in the previous section, we have the parameter estimates for the univariate model given in Table 6.4.

Note in the univariate models, parameter estimates for the univariate model and multivariate model are quite different. However, in the multilevel models, they are quite similar.

Table 6.4: Univariate Model for Iris Data

Parameters	Basic Model	Multilevel Model
β_0	-0.9507	-0.3964
β_1	0.8061	0.7106

6.2 Other Future Research Topics

Generalization of the OSM method from interval-valued data to other types of symbolic data such as multi-valued and histogram-valued data is a future research topic. Another possible research topic is to use the connection between symbolic regression models and classical mixed models to conduct statistical inference.

Conditions applied to data result in a very complicated situation in interval-valued regression. In the classical data situation, we can select the observations that satisfy the conditions fairly easily. However, for interval-valued data, there is no straightforward solution. If we map the interval-valued explanatory variable and response variable in a p+1 dimensional space, where p is the number of explanatory variables, we obtain a hypercube in symbolic data situation, no longer a point. Therefore, the conditions applied to the data may cut the hyper cube into different parts, and we need to select those parts that satisfy the conditions. Some preliminary work has been established by Billard on this issue and more formal research will be studied in the future.

6.3 References

[1] Anderson, E. (1935). The Irises of the Gasp Peninsula. *Bulletin of the American Iris Society*, 59, 25.

[2] Billard, L. and Diday, E. (2000). Regression Analysis for Interval-Valued Data. *Data analysis, Classificaiton, 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. (2007). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, Chichester.

[4] de Carvalho F.A.T., Lima Neto, E.A. and Tenorio, C.P. (2004). A New Method to Fit a Linear Regression Model for Interval-valued Data. *Lecture Notes in Computer Science, KI2004 Advances in Artifical Inteligence*. Springer-Verlag, 295-306.

[5] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J.R.Statist. Soc.* B, 39, 1-38.

[6] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 38, 963-974.

[7] Le-Rademacher, J. and Billard, L. (2010). Likelihood Functions and Some Maximum Likelihood Estimators for Symbolic Data. *Journal of Statistical Planning and Inference*, submitted.

[8] Lima Neto, E.A. and de Carvalho F.A.T. (2010). Constrained Linear Regression Models for Symbolic Interval-valued Variables. *Computational Statistics & Data Analysis*, 54(2), 333-347.

[9] Lima Neto, E.A., de Carvalho F.A.T. and Freire, E.S. (2005). Applying Constrained Linear Aggression Models to Predict Interval-Valued Data. *Lecture Notes in Computer Science, KI: Advances in Artificial Inteligence* (ed. U. Furbach). Springer-Verlag, Brelin, 92-106.

[10] Lindstrom, M. and Bates, D. (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. *Journal of the American Statistical Association*, 83,1014 1022.

[11] Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer.