International Journal of Applied Mathematics
& Statistical Sciences (IJAMSS)
ISSN(P): 2319-3972; ISSN(E): 2319-3980
Vol. 7, Issue 2, Feb - Mar 2018; 33 - 44

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations

ANALYSIS OF MULTIPLE LINEAR REGRESSION MODELS USING SYMBOLIC INTERVAL-VALUED VARIABLES

Magda M. M. Haggag

Associate Professor & Head, Department of Statistics, Mathematics and Insurance,

Faculty of Commerce, Damanhour University, Damanhour, Egypt

ABSTRACT

In this paper, a proposed form of dependent regression models are introduced to study symbolic data. The estimation of the proposed linear regression models are based on interval valued data, for which we have lower and upper bounds or center and range values. The least squares method is used to estimate the models. A real example data are used to illustrate the usefulness of the proposed regression models for handling the interval valued data. The estimation results are evaluated using the predicted mean squared errors. The results support the proposed dependent regression models.

***KEYWORDS**: Interval-Valued Data, Least Squares Estimation, Linear Regression Analysis, Symbolic Data*

***Article* History**

***Received:* 23 Dec 2017 | *Revised:* 02 Feb 2018 | *Accepted:* 20 Feb 2018**

1. INTRODUCTION

In multivariate analysis, huge data sets lead to computational difficulties in the standard form of analysis. Therefore, summarizing data into smaller groups than the set of large individuals, has been attracting many statistical researchers. Symbolic data analysis (SDA) is defined as an extension analytical tool for the standard data. SDA has been considered as a domain related to multivariate analysis, pattern recognition, data mining, machine learning, and artificial intelligence. (See Billard and Diday, 2006 [1]; Bock and Diday, 2000 [2]; and Diday and Fraiture-Noirhomme, (2008) [3]).

One of the forms of symbolic data is the interval-valued form, in which the large individuals are summarized into smaller groups or classes. Interval data analysis tools have been used by many authors. (See De Carvalho, 1995 [7]; Ichino et al., 1996 [11]; Cazes et al., 1997 [6]; Bertrand and Goupil, 2000 [1]; Laura and Palumbo, 2000 [12]; Laura Et al., 2000 [13]; Palumbo and Verde, 2000 [16]; Rasson and Lissoir, 2000 [17]; Billard and Diday, 2003 [2]; Gorenen et al., 2006 [10]; Billard et al., 2007 [4]; and Maia et al., 2008 [14]).

Fitting a linear regression analysis using interval-valued data is introduced by Billard and Diday, 2000, their approach is based on fitting a linear regression model to the mid-points of the interval values in the learning set, and applying it to the lower and upper bounds of the interval values of both response and predictor variables. Their work improved after that by Lima Neto and De Carvalho, 2008.

This paper proposes two linear regression models to be fitted using symbolic interval-valued data. The two models are based on center and range of the interval values.

The paper is organized as follows. Section 2 introduces basics of symbolic data. Section 3 considers the linear regression models for classical data. Section 4 describes the proposed linear regression models for symbolic interval data. Section 5 presents application studies. Section 6 considers conclusions and recommendations for future studies.

2. SYMBOLIC DATA

Symbolic data is defined as an approach to summarize large data sets in such a way that the resulting data is in a manageable size. These summarized data contain different types Such as, single quantitative or categorical value, a set of values or categories, and an interval, a set of values with associated probabilities or weights. (See Bock and Diday, 2000 []).

Symbolic data mean rather than a certain value x_i, an observed value may be multi-valued such as {2,7, 12} or {small, medium, large}, it can be interval-valued such as [5, 10), it may be in modal values such as { 1 with probability 0.7, 0 with probability 0.3}, extra. (See Bock and Diday, 2000 []).

Let $E=\{1,2,\ldots,n\}$ denotes the set of units that are described by p symbolic interval-valued variables $X_1,X_2,\ldots,X_p$. For each element $k \in E$, the interval $X(k)$ is denoted by $[\underline{x}_k, \overline{x}_k]$, where $\underline{x}_k$ and $\overline{x}_k$ are the lower and the upper bound of the interval $X(k) \subseteq R$, respectively. It is shown that the variables $X(k)$ are uniformly distributed with mean and variance as in the following definitions. (See Bock and diday, 2000).

Definition (1)

Let X be an interval-valued variable defined on the set $E=\{1,2,\ldots,n\}$. The empirical distribution function of X, denoted by F_x, is the distribution function of n discrete uniform distributions defined on the intervals $X(k)$ for $k \in E$.

The empirical density function of X, which is denoted by f_x, is defined as:

$$f_x = \frac{1}{n}\sum_k \frac{1}{\overline{x}_k - \underline{x}_k} \quad \textbf{(1)}$$

The empirical density function f_x corresponds to the frequency distribution for a multi-valued variable.

Definition (2)

Let X be an interval-valued variable defined on the set $E=\{1,2,\ldots,n\}$, and let f_x is the empirical density function of X defined in (1). Then the empirical mean of X, $\overline{X}$, and the empirical standard deviation, S_x are defined respectively as follows:

$$\overline{x} = \int_{-\infty}^{+\infty} x f_x dx = \frac{1}{n}\sum_{k\in E} \frac{\underline{x}_k + \overline{x}_k}{2}, \quad \textbf{(2)}$$

$$s_x = \sqrt{\int_{-\infty}^{+\infty} (x-\overline{x})^2 f_x dx} = \sqrt{\int_{-\infty}^{+\infty} x^2 f_x dx - \overline{x}^2} \quad . \quad \textbf{(3)}$$

3. LINEAR REGRESSION MODELS FOR CLASSICAL DATA

Linear regression is a widely used method to study the relationship between one response and one or many predictors. The classical regression analysis is used, when observations are specified by numerical data values. (See Draper and Smith, 1981; and Montgomery, 1982).

Consider the following standard multiple linear regression model,

$$Y = X^T \beta + \varepsilon \quad \textbf{(4)}$$

where, Y is a response variable, $X^T = (1, X_1, ..., X_p)$ is a predictor vector, $\beta = (\beta_0, \beta_1, ..., \beta_p)$ is a parameter vector to be estimated, and $\varepsilon \sim (0, \sigma^2)$ is the error term. In this model, each X_j takes specific value $\forall$ j=0,1,2,…,p. The objective of the model (4) is to find the best linear relationship between the two variables Y and X. The estimation problem in model (1) insists in finding a good parameter estimator $\hat{\beta} \in R^p$, such that $(\hat{Y} = X^T \hat{\beta})$, holds, and that $(\hat{Y}_i - X_i^T \hat{\beta})$ exists $\forall$ i={1,2,…,n}.

Recently, different approaches have been considered the analysis of symbolic interval-valued data for regression. (See Tanak and Lee, 1998 []; Billard and Diday, 2000); Lima Neto and De Car; Lima Neto and De Carvalho, 2008, 2010, and 2011; and Sun and Li, 2015).

4. THE PROPOSED LINEAR REGRESSION MODELS FOR SYMBOLIC INTERVAL DATA

4-1. The Methodology

Consider the linear regression model in (4), which is described by existing *(p+1)* symbolic interval-valued variables *Y, X_1,X_2,…,X_p*. $X_i=(x_{i1},x_{i2},\ldots,x_{ip})$, where $x_{ij}=[a_{ij},b_{ij}] \in \{[a,b]: a,b \in R, a \le b\}$ $\forall$ i=1,2,…,n, and j=1,2,…,p, and $Y_i=[y_{Li},y_{Ui}]$ are the observed values of both X_j and Y respectively.

There are two different methods of fitting the linear regression model in (4) that will be considered in this section, the center (CM) and the center and range (CRM) fitting methods.

4-2. Regression Model Using the Center Method (CM)

The center method was proposed by Bilard and Diday (2000), they showed that the two interval-valued variables Y and X are related according to the following relationship:

$$Y^c = X^c \beta + \varepsilon^c, \quad \textbf{(5)}$$

where $Y^c = (Y_1^c, ..., Y_n^c)$, $X^c = \left((x_1^c)^T, ..., (x_n^c)^T\right)^T$, $(x_i^c)^T = (1, x_{i1}^c, ..., x_{ip}^c)$ for i=1,…,n, $\beta = (\beta_0, ..., \beta_p)^T$, $\varepsilon^c = (\varepsilon_1^c, ..., \varepsilon_n^c)$, $x_{ij}^c = (a_{ij} + b_{ij})/2$, and $Y_i^c = (Y_{Li} + Y_{Ui})/2$.

If the matrix X^c has full column rank, then the least squares estimator (LSE) of β will be as follows:

$$\hat{\beta} = \left[(X^c)^T X^c\right]^{-1} (X^c)^T Y^c. \quad \textbf{(6)}$$

The lower and upper bounds of the predicted values of Y, $\hat{Y} = \left[\hat{Y}_L, \hat{Y}_U\right]$, will be defined as follows:

$$\hat{Y}_L = (X_L)^T \hat{\beta} \text{ and } \hat{Y}_U = (X_U)^T \hat{\beta}, \quad \textbf{(7)}$$

where $(X_L)^T = (1, a_1, a_2, ..., a_p)$, and $(X_U)^T = (1, b_1, b_2, ..., b_p)$.

The coefficient of determination of the center method (CM), R^2_{CM}, will be computed as follows:

$$R^2_{CM} = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i^c - \overline{Y}^c\right)}{\sum_{i=1}^{n} \left(Y_i^c - \overline{Y}^c\right)}, \quad \textbf{(8)}$$

where,

$$\hat{Y}_i^c = \left(\hat{Y}_{Li} + \hat{Y}_{Ui}\right)/2, \text{ and } \overline{Y}^c = \left(\overline{Y}_{Li} + \overline{Y}_{Ui}\right)/2. \quad \textbf{(9)}$$

4-3. Regression Models Using the Center and Range Method (CRM)

Lima Neto and de Carvalho (2008), proposed the center and range method, as an approach, for fitting the linear regression model in (4). The ideal of this approach is to consider both the information contained in the centers and ranges of the interval-valued variables. They showed that the two interval-valued variables Y and X are related, over the centers, according to the following relationship:

$$Y^c = X^c \beta^c + \varepsilon^c, \quad \textbf{(10)}$$

where $Y^c = \left(Y_1^c, ..., Y_n^c\right)$, $X^c = \left(\left(x_1^c\right)^T, ..., \left(x_n^c\right)^T\right)^T$, $\left(x_i^c\right)^T = \left(1, x_{i1}^c, ..., x_{ip}^c\right)$ for i=1,…,n, $\beta^c = \left(\beta_\circ^c, ..., \beta_p^c\right)^T$, $\varepsilon^c = \left(\varepsilon_1^c, ..., \varepsilon_n^c\right)$, $x_{ij}^c = \left(a_{ij} + b_{ij}\right)/2$, and $Y_i^c = \left(Y_{Li} + Y_{Ui}\right)/2$.

Also, Lima Neto and de Carvalho (2008) showed that the two interval-valued variables Y and X are related, over the ranges, according to the following relationship:

$$Y^r = X^r \beta^r + \varepsilon^r, \quad \textbf{(11)}$$

where $Y^r = \left(Y_1^r, ..., Y_n^r\right)$, $X^r = \left(\left(x_1^r\right)^T, ..., \left(x_n^r\right)^T\right)^T$, $\left(x_i^r\right)^T = \left(1, x_{i1}^r, ..., x_{ip}^r\right)$ for i=1,…,n, $\beta^r = \left(\beta_\circ^r, ..., \beta_p^r\right)^T$, $\varepsilon^r = \left(\varepsilon_1^r, ..., \varepsilon_n^r\right)$, $x_{ij}^r = \left(b_{ij} - a_{ij}\right)/2$, for j=1,…, p, and $Y_i^r = \left(Y_{Ui} - Y_{Li}\right)/2$ for i=1,…,n.

If the both the two matrices X^c, X^r and have full column ranks, then the least squares estimators (LSEs) of both β^c and β^r in equations (10) and (11), respectively, will be as follows:

$$\hat{\beta}^c = \left[\left(X^c\right)^T X^c\right]^{-1} \left(X^c\right)^T Y^c, \quad \textbf{(12)}$$

and,

$$\hat{\beta}^{r} = \left[\left(X^{r}\right)^{T} X^{r}\right]^{-1}\left(X^{r}\right)^{T} Y^{r}, \quad \textbf{(13)}$$

The lower and upper bounds of the predicted values of Y, $\hat{Y} = \left[\hat{Y}_L, \hat{Y}_U\right]$, will be defined as follows:

$$\hat{Y}_L = \hat{Y}^c - \hat{Y}^r \quad \text{and} \quad \hat{Y}_U = \hat{Y}^c + \hat{Y}^r, \quad \textbf{(14)}$$

where $\hat{Y}^c = \left(\tilde{X}^c\right)^T \hat{\beta}^c$, and $\hat{Y}^r = \left(\tilde{X}^r\right)^T \hat{\beta}^r$, $\left(\tilde{X}^c\right)^T = \left(1, x_1^c, ..., x_p^c\right)$, $\left(\tilde{X}^r\right)^T = \left(1, x_1^r, ..., x_p^r\right)$, $\hat{\beta}^c = \left(\hat{\beta}_\circ^c, \hat{\beta}_1^c, ..., \hat{\beta}_p^c\right)^T$ and $\hat{\beta}^r = \left(\hat{\beta}_\circ^r, \hat{\beta}_1^r, ..., \hat{\beta}_p^r\right)^T$.

Lima Neto and de Carvalho (2010) introduced the constrained center and range method (CCRM) to insure that $\hat{Y}_{Li} \le \hat{Y}_{Ui}$ as follows:

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

$$Y^r = X^r \beta^r + \varepsilon^r \quad \textbf{(15)}$$

With constraints $\beta_j^r \ge 0$, j=0, 1,…,p.

The least squares estimators of both β^c and β^r in (15) will be estimated as in (12) and (13), respectively.

The coefficient of determination of the center and range method, R^2_{CRM}, can be derived as in the case of CM as follows:

$$R^2_{CRM(c)} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i^c - \overline{Y}^c\right)}{\sum_{i=1}^{n}\left(Y_i^c - \overline{Y}^c\right)}, \quad \textbf{(16)}$$

for the center, and

$$R^2_{CRM(r)} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i^r - \overline{Y}^r\right)}{\sum_{i=1}^{n}\left(Y_i^r - \overline{Y}^r\right)}, \quad \textbf{(17)}$$

for the range, where,

$$\hat{Y}_i^c = \left(\hat{Y}_{Li} + \hat{Y}_{Ui}\right)/2, \quad \text{and } \overline{Y}^c = \left(\overline{Y}_{Li} + \overline{Y}_{Ui}\right)/2, \quad \textbf{(18)}$$

$$\hat{Y}_i^r = \left(\hat{Y}_{Ui} - \hat{Y}_{Li}\right)/2, \quad \text{and } \overline{Y}^r = \left(\overline{Y}_{Ui} - \overline{Y}_{Li}\right)/2. \quad \textbf{(19)}$$

Lima Neto and de Carvalho (2010) suggested different methods to measure the goodness of fit for both CRM and CCRM as follows:

$$R_1^2 = \min\left(R_c^2, R_r^2\right), \text{ or}$$

$$R_2^2 = \frac{R_c^2 + R_r^2}{2}, \text{ or} \qquad \textbf{(20)}$$

$$R_3^2 = \max\left(R_c^2, R_r^2\right).$$

They showed that R_1^2 and R_3^2 are a pessimistic and an optimistic version of the goodness of fit measure. It is shown that R_2^2 lies between R_1^2 and R_3^2.

4-4. The Proposed Dependent *Regression Models Using the Center and Range Method* (DCRM)

The center and range method is based on independency of the two models, the center model defined in (10) and range model defined in (11). That is the predictor interval-valued variables for both center and range models are independent as follows:

The center model: $Y^c = X^c\beta^c + \varepsilon^c$,

and

The range model: $Y^r = X^r\beta^r + \varepsilon^r$.

This work proposes the case where the predictors of the center X^c and of the range X^r are dependent according to the following relation:

$X^r = X^c\alpha + \varepsilon^r$, and then the center and range models are defined as follows:

The center model: $Y^c = X^c\beta^c + \varepsilon^c$, **(19)**

and

The range model: $$\begin{aligned} Y^r &= \left(X^c\alpha\right)\beta^r + \varepsilon^r = X^c\alpha\beta^r + \varepsilon^r \\ \therefore Y^r &= X^c\beta^* + \varepsilon^r \end{aligned}, \qquad \textbf{(20)}$$

where $\beta^* = \alpha\beta^r$. The least squares estimator of β^c in (19) is as defined in (12), and the least squares estimator of β^r in (20) is defined as:

$$\hat{\beta}^{*r} = \left[\left(X^c\right)^T X^c\right]^{-1}\left(X^c\right)^T Y^r. \qquad \textbf{(21)}$$

The derivation of (21) will be presented in the Appendix (A).

5. APPLICATIONS

A proposed dependent center and range linear regression models will be estimated and compared with the independent models using a real interval-valued data. The real data are on cardiology and are obtained in terms of center and range values of intervals.

The cardiological interval data contain a record of the pulse rate (Y), systolic blood pressure (X1), and diastolic blood pressure (X2) taken from 11 patients as shown in Table (1).

Table 1: The Cardiological Interval-Valued Data

No	PulseC	SystC	DiastC	PulseR	SystR	DiastR
1	56	95	60	24	10	20
2	66	110	80	12	40	20
3	73	160	95	34	40	10
4	91	126	94	42	32	28
5	63	95	60	18	10	20
6	85	145	95	30	3	30
7	69	80	145	12	40	10
8	86	145	83	28	30	14
9	87	150	90	22	80	40
10	91	159	100	10	42	20
11	93	130	89	14	40	22

The least squares estimators and the estimated standard deviations of the response center and range ($\sigma\left(\hat{Y}_c\right), \sigma\left(\hat{Y}_r\right)$), for the different estimation methods CM, CRM, CCRM, DCRM, and DCCRM are given in Table (2). It is found that the estimated standard deviations of the response center $\sigma\left(\hat{Y}_c\right)$ are equal for all center models, but the estimated standard deviations of the response range $\sigma\left(\hat{Y}_r\right)$ differ for all models. Also, it is found that $\sigma\left(\hat{Y}_r\right)$ for DCRM is less than CRM, and $\sigma\left(\hat{Y}_r\right)$ for DCCRM is also less than CCRM. (See Figure (1)). This means that the linear relation between the center and range interval-valued variables should be considered when handling linear regression models for interval-valued variables.

6. CONCLUSIONS AND RECOMMENDATIONS

The main objective of this paper is fitting the linear regression model using interval-valued variables. The method of least squares is used for fritting the independent and dependent regression models. The estimation results support the dependent regression models. This means that, for the interval-valued variables, the relation between the center and the range of the data is a real relation for this type of data. Therefore, this paper recommends using the dependency relation between the center and the range when handling the interval-valued data. In the future, this dependency relation will be considered and studied for collinear, influential, outlying interval-valued data.

Table 2: Least Squares Estimation Results for the Different Models (CM, CRM, CCRM), and the Proposed Models (DCRM, DCCRM) for the Cardiological Interval-Valued Data

Method	Intercept $\hat{\beta}_0$	Systolic $\hat{\beta}_1$	Diastolic $\hat{\beta}_2$	$\sigma\left(\hat{Y}_c\right)$ or $\sigma\left(\hat{Y}_r\right)$
CM	21.171	0.32889	0.16985	9.517
CRM: C	21.171	0.32889	0.16985	9.517
CRM: R	20.215	-0.1467	0.34801	11.054

Table 2 Contd.,				
Method	**Intercept** $\hat{\beta}_0$	**Systolic** $\hat{\beta}_1$	**Diastolic** $\hat{\beta}_2$	$\sigma(\hat{Y}_c)$ **or** $\sigma(\hat{Y}_r)$
CCRM: C	21.171	0.32889	0.16985	9.517
CCRM: R	17.956	0.00000	0.20722	11.384
DCRM: C	21.171	0.32889	0.16985	9.517
DCRM: R	13.752	0.11740	-0.0697	10.820
DCCRM: C	21.171	0.32889	0.16985	9.517
DCCRM: R	7.4162	0.11787	0.00000	10.963

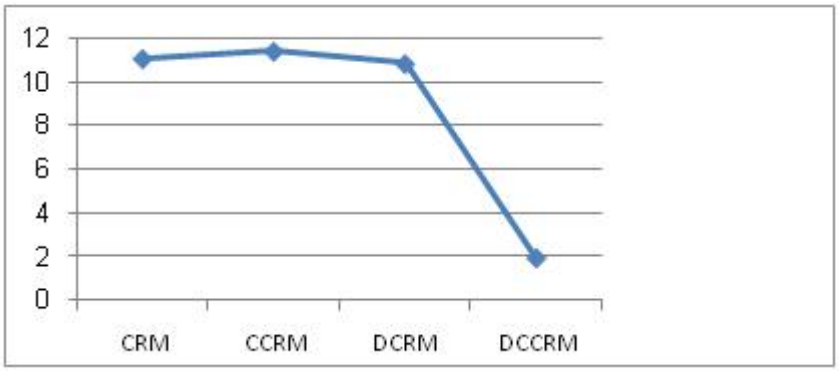

Figure 1: The Estimated Standard Deviations of the Predicted Response for the Models CRM, CCRM, DCRM, and DCCRM

REFERENCES

1. *Bertrand, P., Goupil, F., (2000). Descriptive Statistics for Symbolic Data. In: Bock, H.-H., Diday, E. (Eds.), Analysis of Symbolic Data. Springer, Heidelberg, pp. 106-124.*
2. *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(462), 470-487.*
3. *Billard, L., and Diday, E. (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining, John Wiley, New York.*
4. *Billard, L., Chouakria-Douzal, A., and Diday, E. (2007). Symbolic Principal Components for Interval-Valued Observations. Technical Report. University of Georgia.*
5. *Bock, H.-H., and Diday, E. (2000). Analysis of Symbolic Data, Exploratory Methods of Extracting Statisical information from Complex Data to, Springer-Verlag, Heidelberg.*
6. *Cazes, P., Chouakria, A., Diday, E., Schektman, S. (1997). "Extension de l´analyse en Compsantes Principals des Donnes de Type Intervalle", Rev. Statist. Aplique XLV (3), 5-24.*
7. *De Carvalho, F. A. T. (1995). "Histograms in Symbolic Data Analysis", The Annals of operation Research, 55,229-322.*
8. *Diday, E., and Fraiture-Noirhomme, M. (2008). Symbolic Data Analysis and the SODAS software, Wiley-Interscience, Chichester.*
9. *Draper, N. R., and Smith, H. (1981). Applied Regression Analysis, John Wiley, New York.*

10. *Gorenen, P., Winsberg, S. Rodrigues, O., and Diday E. "Multidimensional Scaling of Interval Dissmilarities, Computational Statistics and Data analysis, 51, pp. 360-378.*
11. *Ichino, M., Yaguchi, H., and Diday, E. (1996). A Fuzzy Symbolic Pattern Classifier. In Diday, E. et al. (Eds.), Ordinal and symbolic Data Analysis. Springer, Berlin, pp. 92-102.*
12. *Laura, N. C., and Palumbo, F. (2000). "Principal Components Analysis of Interval Data: A Symbolic data Analysis Approach", Computational Statistics, 15 (1), 73-87.*
13. *Laura, N. C., Verde, R., and Palumbo, F. (2000). Factorial Discriminant analysis on Symbolic Objects. In: Bock, H.-H., Diday, E. (Eds.), Analysis of Symbolic Data. Springer, Heidelberg, pp. 212-233.*
14. *Lima Neto, E. A., and De Carvalho, F. A. T. (2008). "Center and Range method for fitting a Linear Regression Model on Symbolic Interval Data", Computational Statistics and Data Analysis 52, 1500-1515.*
15. *Duc-Hien Le, Linear Regression and Artificial Neural Networks for Modeling Compressive Strength of Soil-Based CLSMS, International Journal of Civil, Structural, Environmental and Infrastructure Engineering Research and Development (IJCSEIERD), Volume 5, Issue 2, March-April 2015, pp. 25-34*
16. *Lima Neto, E. A., and De Carvalho, F. A. T. (2010). "Constrained Linear Regression Models for Smbolic Interval-Valued Variable", Computational Statistics and Data Analysis 52, 1500-1515.*
17. *Lima Neto, E. A., Gauss M. Cordeiro, and De Carvalho, F. A. T. (2011). " Bivariate Symbolic Regression Models for Interval-valued Variables", Journal of Statistical Computation and Simulation, 81(11), 1727-1744.*
18. *Maia, A. L. S., and De Carvalho, F. A. T., Ludermir, T. B. (2008). "Forcasting Models for Interval-Valued Time Series", Neurocomputing, 71, 3344-3352.*
19. *Montgomery, D. C., and Peck, E. A. (1982). Introduction to Linear Regression Analysis, John Wiley, New York.*
20. *Palumbo, F., Verde, R. (2000). "Non-symmetrical Factorial Discriminant analysis For Symbolic", Applied Stochastic Models Business Industry, 15 (4), 419-427.*
21. *Rasson, J. P., and Lissoir, S. (2000). Symbolic Kernel Discriminant analysis. In: Bock, H.-H., Diday, E. (Eds.), Analysis of Symbolic Data. Springer, Heidelberg, pp. 240-244.*
22. *Sun, Y., and Li, C. (2015). " On Linear Regression for Interval-valued Data in $K_c(R)$".*
23. *Tanaka, H., and Lee, H. (1998). "Interval Regression Analysis by Quadratic programming approach", IEEE Transactions on Fuzzy Systems, 6, 473-481.*

APPENDIX (A)

Derivation of the Least Squares Method for Dependent Regression Models for DCRM and DCCRM:

The center regression model is, $Y^c = X^c\beta^c + \varepsilon^c$, and the proposed range regression model is $Y^r = X^c\beta^* + \varepsilon^r$ as considered in (19), and (20).

Consider the following two center and range regression models:

$$Y_i^c = \beta_\circ^c + \beta_1^c X_{i1}^c + \ldots + \beta_p^c X_{ip}^c + \varepsilon_i^c,$$

$$Y_i^r = \beta_\circ^* + \beta_1^* X_{i1}^c + ... + \beta_p^* X_{ip}^c + \varepsilon_i^r .$$

The sum of squares of the deviations is given by:

$$SS_{DCRM} = \sum_{i=1}^{n} \left(\varepsilon_i^c\right)^2 + \sum_{i=1}^{n} \left(\varepsilon_i^r\right)^2 = \sum_{i=1}^{n} \left(Y_i^c - \beta_\circ^c - \beta_1^c X_{i1}^c ... - \beta_p^c X_{ip}^c\right)^2 + \sum_{i=1}^{n} \left(Y_i^r - \beta_\circ^* - \beta_1^* X_{i1}^c ... - \beta_p^* X_{ip}^c\right)^2. \quad \textbf{(A-1)}$$

Differentiating (A-1) w.r.t the parameters $\beta_\circ^c, \beta_1^c, ..., \beta_p^c$, the following normal equations will be obtained:

$$n\hat{\beta}_\circ^c + \beta_1^c \sum_{i=1}^{n} X_{i1}^c + ... + \hat{\beta}_p^c \sum_{i=1}^{n} X_{ip}^c = \sum_{i=1}^{n} Y_i^c ,$$

$$\hat{\beta}_\circ^c \sum_{i=1}^{n} X_{i1}^c + \beta_1^c \sum_{i=1}^{n} \left(X_{i1}^c\right)^2 + ... + \hat{\beta}_p^c \sum_{i=1}^{n} X_{ip}^c X_{i1}^c = \sum_{i=1}^{n} Y_i^c X_{i1}^c ,$$

$$\vdots$$

$$\hat{\beta}_\circ^c \sum_{i=1}^{n} X_{ip}^c + \beta_1^c \sum_{i=1}^{n} X_{i1}^c X_{ip}^c + ... + \hat{\beta}_p^c \sum_{i=1}^{n} \left(X_{ip}^c\right)^2 = \sum_{i=1}^{n} Y_i^c X_{ip}^c$$

Then the least squares estimators of $\beta_\circ^c, \beta_1^c, ..., \beta_p^c$, which minimize (A-1) are obtained $\beta_\circ^c, \beta_1^c, ..., \beta_p^c$ by solving the above (p+1) normal equations, as follows:

$$\hat{\beta}^c = \left(\hat{\beta}_0^c, \hat{\beta}_1^c, ..., \hat{\beta}_p^c\right)^T = \left[\left(X^c\right)^T \left(X^c\right)\right]^{-1} \left(X^c\right)^T Y^c \quad ,$$

Where, The matrix $\left[\left(X^c\right)^T \left(X^c\right)\right]$ is a (p+1)x(p+1) full column matrix defined as:

$$\left[\left(X^c\right)^T \left(X^c\right)\right] = \begin{bmatrix} n & \sum_i X_{i1}^c ... & \sum_i X_{ip}^c \\ \sum_i X_{i1}^c & \sum_i \left(X_{i1}^c\right)^2 ... & \sum_i X_{ip}^c X_{i1}^c \\ \vdots & \vdots & \vdots \\ \sum_i X_{ip}^c & \sum_i X_{i1}^c X_{ip}^c & \sum_i \left(X_{ip}^c\right)^2 \end{bmatrix}, \quad \textbf{(A-2)}$$

and

$$Y^c = \left(\sum_i Y_i^c, \sum_i Y_i^c X_{i1}^c, ..., \sum_i Y_i^c X_{ip}^c\right)^T .$$

Also, by differentiating (A-1) w.r.t the parameters $\beta_\circ^*, \beta_1^*, ..., \beta_p^*$, the following normal equations will be obtained:

$$n\hat{\beta}_\circ^* + \beta_1^* \sum_{i=1}^{n} X_{i1}^c + ... + \hat{\beta}_p^* \sum_{i=1}^{n} X_{ip}^c = \sum_{i=1}^{n} Y_i^c ,$$

$$\hat{\beta}_{\circ}^{*}\sum_{i=1}^{n} X_{i1}^{c} + \beta_1^{*}\sum_{i=1}^{n}\left(X_{i1}^{c}\right)^2 + ... + \hat{\beta}_{p}^{*}\sum_{i=1}^{n} X_{ip}^{c}X_{i1}^{c} = \sum_{i=1}^{n} Y_i^{c}X_{i1}^{c},$$

.
.
.

$$\hat{\beta}_{\circ}^{*}\sum_{i=1}^{n} X_{ip}^{c} + \beta_1^{*}\sum_{i=1}^{n} X_{i1}^{c}X_{ip}^{c} + ... + \hat{\beta}_{p}^{*}\sum_{i=1}^{n}\left(X_{ip}^{c}\right)^2 = \sum_{i=1}^{n} Y_i^{c}X_{ip}^{c}$$

Then the least squares estimators of $\beta_{\circ}^{*},\beta_1^{*},...,\beta_p^{*}$, which minimize (A-1) are obtained by solving the above (p+1) normal equations, as follows:

$$\hat{\beta}^{*} = \left(\hat{\beta}_0^{*},\hat{\beta}_1^{*},...,\hat{\beta}_p^{*},\right)^{T} = \left[\left(X^{c}\right)^{T}\left(X^{c}\right)\right]^{-1}\left(X^{c}\right)^{T}Y^{r} \quad ,$$

where, the matrix $\left[\left(X^{c}\right)^{T}\left(X^{c}\right)\right]$ is a (p+1)x(p+1) full column matrix defined as in (A-2), and

$$Y^{r} = \left(\sum_i Y_i^{r},\sum_i Y_i^{r}X_{i1}^{c},...,\sum_i Y_i^{r}X_{ip}^{c}\right)^{T}.$$

The estimated center and range responses are defined, respectively as:

$\hat{Y}^{c} = \left(X^{c}\right)^{T}\hat{\beta}^{c}$, and $\hat{Y}^{r} = \left(X^{c}\right)^{T}\hat{\beta}^{r}$.