A Regression Model to Interval-valued Variables based on Copula Approach

Silva, Alisson de Oliveira
Universidade Federal da Paraíba, Departamento de Estatística
Cidade Universitária
João Pessoa, 58051-900, Brazil
E-mail: allysson_jlr@yahoo.com.br

Lima Neto, Eufrásio de Andrade
Universidade Federal da Paraíba, Departamento de Estatística
Cidade Universitária
João Pessoa, 58051-900, Brazil
E-mail: eufrasio@de.ufpb.br

Anjos, Ulisses Umbelino
Universidade Federal da Paraíba, Departamento de Estatística
Cidade Universitária
João Pessoa, 58051-900, Brazil
E-mail: ulisses@de.ufpb.br

1. Introduction

Symbolic Data Analysis (SDA) has been introduced as a domain related to multivariate analysis, pattern recognition and artificial intelligence in order to introduce new methods and to extend classical data analysis techniques and statistical methods to symbolic data (Billard and Diday, 2003). In SDA a variable can assume as a value an interval from a set of real numbers, a set of categories, an ordered list of categories or even a histogram. These new variables could take into account the variability and/or uncertainty presented in the data. The main objective of SDA is to increase the statistical techniques to apply to these types of data.

Moreover, interval-valued data arise in practical situations such as recording monthly interval temperatures in meteorological stations, daily interval stock prices, among others. Another source of interval data is the aggregation of huge data-bases into a reduced number of groups. Therefore, tools for interval-valued data analysis are very much required.

Nowadays, different approaches have been introduced to analyze interval-valued data. In the field of SDA, several suitable tools for managing interval-valued data have been discussed in the literature. Bertrand and Goupil (2000) and Billard and Diday (2003) introduced central tendency and dispersion measures suitable for interval-valued data. De Carvalho (1995) proposed histograms for interval-valued data. Concerning factorial methods, Cazes et al. (1997) and Lauro and Palumbo(2000) proposed principal component analysis methods suitable for interval-valued data. Palumbo and Verde (2000) and Lauro et al. (2000) generalized factorial discriminant analysis (FDA) to interval-valued data. Concerning interval-valued time series, Maia et al. (2008) have introduced autoregressive integrated moving average (ARIMA), artificial neural network (ANN) as well as a hybrid methodology that combines both ARIMA and ANN models in order to forecast interval-valued time series. Other contributions in the SDA field were proposed by Groenen et al. (2006) and Ichino et al. (1996), among others.

In regression analysis of quantitative data, the items are usually represented as a vector of quantitative measurements (Schéffe 1959, Draper and Smith 1981, Montgomery et al. 1982). The

generalized linear models (GLMs) represent a major synthesis of regression models by allowing a wide range of types of response data and explanatory variables to be handled in a single unifying framework. These models are based on the exponential family of distributions and represent a very important regression tool due to their flexibility and applicability in practical situations (McCullagh and Nelder, 1989). In the field of SDA, Billard and Diday (2000) presented the first approach to fit a linear regression model on interval-valued data sets. Their approach consists of fitting a linear regression model on the midpoint of the interval values assumed by the variables in the learning set and to apply this model on the lower and upper boundaries of the interval values of the explanatory variables to predict, respectively, the lower and upper boundaries of the interval values of the dependent variable. Lima Neto and De Carvalho (2008) improved the former approach presenting a new method based on two linear regression models, the first regression model over the midpoints of the intervals and the second one over the ranges, which reconstruct the boundaries of the interval-values of the dependent variable in a more efficient way when compared with the Billard and Diday's method.

At this time, the regression models for interval-valued data attack the problem from an optimization point of view. Recently, Lima Neto et. al. (2011) propose the bivariate symbolic regression models (BSRM) based on the GLM framework. Inference techniques, residual analysis and diagnostic measures are available for this regression approach to interval-valued variables.

Although this approach has had great relevance in the context of regression models for symbolic interval data, since it is possible to perform inference on the model, through hypothesis testing, confidence intervals, and is possible to perform the analysis of residuals and diagnostic, which consists of basic steps to choose accurate models, there is a limitation on the joint distribution of random vector. Based on this fact, in the next section we propose a regression model to symbolic interval-valued data based on copula's theory, which expand the framework of joint distributions for the bivariate random vector $\mathbf{Y} = [Y_1, Y_2]$.

2. Copula Interval Regression Model (CIRM)

Let $Y = \{y_1, \ldots, y_n\}$ be a set of observations that represents a random sample of the interval-valued variable Y. Each observation $y_i = [y_{Li}, y_{Ui}] \in Y$ is defined as an interval $y \in \Im = \{[y_L, y_U] : y_L, y_U \in \Re, y_L \leq y_U\}$ and represents the observed value of the interval variable Y. An interval of real values is an infinity list of values and it is difficult to take into account the whole information inside it. Despite the loss of information, we consider an interval-valued variable Y as a two-dimensional or a bivariate quantitative feature vector $\mathbf{y}_i = (y_{1i}, y_{2i})$, where the variables Y_1 and Y_2 are one-dimensional random variables representing, for example, the lower and upper boundaries or the midpoint and half-range of the intervals or any other pair of interval features possible to be represented.

To model the probability distribution of Y_1 and Y_2 we use copulas approach. Copulas are the part of a multivariate distribution function that completely describes the dependence between the variables of interest and they have become a very popular tool to model dependencies. Copula functions allow for modeling joint multivariate distributions in a simple and extremely flexible way. Copulas are able to yield any kind of dependence structure independently of the marginal distributions. More precisely, in a two dimensional space, the theorem proposed by Sklar (1959) state that: Let H be a joint distribution function with margins F_1 and F_2. Then there exists a copula C such that for all $(y_1, y_2) \in \overline{R}^2$,

$$H(y_1, y_2) = C(F_1(y_1), F_2(y_2)).$$

If F_1 and F_2 are continuous, then C is unique; otherwise, C is uniquely determined on $RanF_1 \times RanF_2$. Conversely, if C is a copula and F_1 and F_2 are distribution functions, then the function H defined above is a joint distribution function with margins F_1 and F_2.

Consider that the joint density probability function of the bivariate quantitative feature vector $\mathbf{y}_i = (y_{1i}, y_{2i})$ is defined by:

$$f(y_1, y_2) = f_1(y_1) f_2(y_2) c(F_1(y_1), F_2(y_2)), \tag{1}$$

where:

$$c(u_1, u_2) = \frac{\partial^2}{\partial u_1 \partial u_2} C(u_1, u_2)$$

is called the copula density, $f_1(y_1)$ e $f_2(y_2)$ are the probability densities and $F_1(y_1)$ and $F_2(y_2)$ are the cumulative distributions of of the one-dimensional variables.

Through the expression (1) we can see the flexibility of the Copula approach, because it separates the choice of dependence from the choice of marginals distributions, on which no restrictions are placed. Among the marginals distributions that could be used we can denote, for example, Normal, Lognormal, Gamma and Inverse Gaussian. For the dependence structure it is possible to consider: Clayton Copula, Gaussian Copula, Frank Copula, among others.

Following the GLM framework, the new approach is defined by two components (a random component and a systematic component) to model interval-valued data. The random component considers the bivariate random vector

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix},$$

having having distribution given by (1). In the systematic component, the explanatory variables X_{1j} and X_{2j} $(j = 1, 2, \ldots, p)$ also represent the lower and upper boundaries or the midpoint and half-range of the explanatory interval-valued variable X_j and are responsible for the variability of Y_1 and Y_2, respectively, and they are defined by

$$\boldsymbol{\eta}_1 = g_1(\boldsymbol{\mu}_1) = \mathbf{X}_1 \boldsymbol{\beta}_1 \text{ and } \boldsymbol{\eta}_2 = g_2(\boldsymbol{\mu}_2) = \mathbf{X}_2 \boldsymbol{\beta}_2, \tag{2}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are known model matrices formed by the observed values of the variables X_{1j} and X_{2j}, respectively, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of parameters to be estimated, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are the linear predictors, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean of the response variables Y_1 and Y_2, respectively, with $\boldsymbol{\eta}_1 = (\eta_{1_1}, \ldots, \eta_{1_n})^T, \boldsymbol{\mu}_1 = (\mu_{1_1}, \ldots, \mu_{1_n})^T$ and $\boldsymbol{\beta}_1 = (\beta_{1_0}, \ldots, \beta_{1_p})^T$. In the same way, we have $\boldsymbol{\eta}_2 = (\eta_{2_1}, \ldots, \eta_{2_n})^T, \boldsymbol{\mu}_2 = (\mu_{2_1}, \ldots, \mu_{2_n})^T$ and $\boldsymbol{\beta}_2 = (\beta_{2_0}, \ldots, \beta_{2_p})^T$. Here, $g_1(\boldsymbol{\mu}_1)$ and $g_2(\boldsymbol{\mu}_2)$ are well-known link functions that connect the mean of the response variables Y_1 and Y_2 with the explanatory variables X_{1j} and X_{2j} $(j = 1, \ldots, p)$, respectively.

3. Determining the Terms of the Joint Distribution and the Parameter Estimation

To fit the regression model based on the copula's theory it is necessary that the terms of the joint density distribution is specified. First we work with the marginals distributions and then choose the dependence structure. This is done using the measures of dependence Spearman and Kendall to select some copulas that are suitable .

The procedure that we use to fit the regression model based on copula, consist of four steps. First the parameters for the marginal models are estimated. In the second step, the copula parameters are estimated with the marginal distribution parameters treated as given. This procedure is know as the inference functions for margins method (IFM). Then we use this estimates as initial values for apply maximum likelihood to jointly estimate the parameters for the marginal models and the copula. Finally, we evaluate the fit of the copula with the estimated parameters by a test of goodness of fit (Genest et al. 2009).

To estimate the parameters of the bivariate model, we used the BFGS method which consists of a nonlinear optimization method, derived from a variation of Newton's method. The log-likelihood function, whose parameters will be optimized can be written as:

$$\begin{aligned} l(y_1, y_2; \beta_1, \beta_2, \rho) &= \sum_{i=1}^{n} [log\left(f(y_{1i}; \beta_1)\right) + log\left(f(y_{2i}; \beta_2)\right) \\ &+ log\left(c(F(y_{1i}; \beta_1), F(y_{2i}; \beta_2); \rho)\right)] \end{aligned} \tag{3}$$

where terms are as defined in section 2.

4. Application to Real Interval-Valued Data Set

In this section, the CIRM will be applied to a real interval-valued data set. This data set gives the records of the weight (Y), height (T_1) and age (T_2) for 531 soccer players grouped in 20 teams of the French Football Professional Championship. We use the new approach to predict the dependent variable Y from the explanatory variables T_j ($j = 1, 2$). Table 1 gives the data which can be free accessed in `http://www.ceremade.dauphine.fr/~touati/foot2.htm`.

Table 1: Soccer interval data set

Team	Y	T_1	T_2	Team	Y	T_1	T_2
A	[58-85]	[164-192]	[21-35]	K	[62-86]	[164-191]	[18-34]
B	[67-84]	[171-190]	[20-30]	L	[62-80]	[168-189]	[19-35]
C	[65-88]	[170-186]	[18-36]	M	[63-85]	[167-190]	[18-31]
D	[60-83]	[162-188]	[19-31]	N	[65-95]	[168-196]	[20-35]
E	[60-84]	[170-189]	[18-34]	O	[63-83]	[170-187]	[18-35]
F	[67-83]	[173-190]	[18-36]	P	[60-87]	[170-197]	[18-37]
G	[69-90]	[176-193]	[19-34]	Q	[67-85]	[168-190]	[18-32]
H	[65-85]	[170-193]	[19-31]	R	[62-83]	[169-192]	[18-35]
I	[63-84]	[168-188]	[18-34]	S	[63-84]	[172-192]	[18-33]
J	[58-88]	[167-197]	[19-35]	T	[63-85]	[169-194]	[20-34]

For this interval-valued data set, we consider to the random component $\mathbf{Y} = [Y_1, Y_2]$ the lower and the upper boundaries of the intervals $[Y_L, Y_U]$ and the midpoint and the range on the intervals $[Y^m, Y^r]$. As possible candidates for the univariate distributions we considered the Gaussian, the Log-Normal and the Gamma distributions; the copulas Gaussian, Clayton, Husler-Reiss and Archimedean (Frank and Clayton) and the link function identity. Based on root mean square error, the best configuration for the CIRM considered the Gaussian distribution for the midpoints of intervals (Y^m), the Log-normal distribution for the ranges of intervals (Y^r and the Gaussian copula for the dependence structure. Moreover, the Kendall (-0.11) and the Spearman (-0.16) coefficients are inside the unit interval $[-1, 1]$.

Table 2: Final estimates to the midpoint explanatory variables

Parameter	Estimates
β_0^m	-18.139
β_1^m	0.567
β_2^m	-0.366

We consider as initial values to estimate the parameters of the copula interval regression model the coefficients of a linear regression models for the midpoints and for the ranges of the intervals. Tables 2 and 3 present the final parameters estimate for the CIRM approach.

Table 3: Final estimates to the range explanatory variables

Parameter	Estimates
β_0^r	1.308
β_1^r	0.064
β_2^r	0.049

The goodness of fit (Genest et al., 2009) suggest that the Gaussian copula presented an satisfactory fitted to the interval-valued data set (p-value = 0.502).

Table 4 suggest that the new approach outperforms the CM and CRM method in both measures ($RMSE_L$ and $RMSE_U$). In the comparison with BSRM, the CIRM approach presented a better performance for the lower bound and a very close performance in the uppper bound. However, a simulated study is strongly recommended for a more consistent results about the CIRM approach.

Table 4: Comparison between the symbolic regression methods

Method	$RMSE_L$	$RMSE_U$
CIRM	1.923	2.651
BSRM1 (Gamma)	2.232	2.570
BSRM1 (Normal)	2.241	2.575
CM	7.540	7.681
CRM	1.946	2.661

Acknowledgments: The authors would like to thanks CNPq (Brazilian Agency) for the financial support.

REFERENCES

1. Bertrand, P. and Goupil, F. *Descriptive statistics for symbolic data.* In: H.-H Bock and E. Diday (Eds.): Analysis of Symbolic Data, Springer, Heidelberg, pp. 106–124, 2000.

2. Billard, L. and Diday, E. *Regression analysis for interval-valued data.* In: Data Analysis, Classification and Related Methods: Proceedings of the Seventh Conference of the International Federation of Classification Societies, Springer-Verlag, Belgium, pp. 369–374, 2000.

3. Billard, L. and Diday, E. *From the statistics of data to the statistics of knowledge: Symbolic Data Analysis*, Journal of American Statistical Association 98, pp. 470–487, 2003.

4. Cazes, P., Chouakria, A., Diday, E. and Schektman, S. *Extension de l'analyse en composantes principales des donnes de type intervalle*, Revue de Statistique Aplique 24, pp. 5–24, 1997.

5. De Carvalho, F. A. T. *Histograms In Symbolic Data Analysis*, Annals of Operations Research 55, pp. 229–322, 1995.

6. Draper, N.R. and Smith, H. *Applied Regression Analysis*, John Wiley, New York, 1981.

7. Genest, C., Rmillard, B. and Beaudoin, D. *Goodness-of-fit tests for copulas: A review and a power study*, Insurance: Mathematics and Economics 44, pp. 199-213, 2009.

8. Groenen, P., Winsberg, S., Rodrigues, O. and Diday, E. *Multidimensional scaling of interval dissimilarities*, Computational Statistics and Data Analysis 51, pp. 360–378, 2006.
9. Ichino, M., Yaguchi, H. and Diday, E. *A fuzzy symbolic pattern classifier,* In: E. Diday et al (Eds.), Ordinal and Symbolic Data Analysis, Springer, Berlin, pp. 92–102, 1996.
10. Lauro, N.C. and Palumbo, F. *Principal component analysis of interval data: A symbolic data analysis approach*, Computational Statistics 15, pp. 73–87, 2000.
11. Lauro, N.C., Verde, R. and Palumbo, F. *Factorial Discriminant Analysis on Symbolic Objects*, In: H.-H Bock and E. Diday (Eds.), Analysis of Symbolic Data, Springer, Heidelberg, pp. 212–233, 2000.
12. Lima Neto, E.A. and De Carvalho, F.A.T. *Centre and range method to fitting a linear regression model on symbolic interval data*, Computational Statistics and Data Analysis 52, pp. 1500–1515, 2008.
13. Lima Neto, E.A., Cordeiro, G.M. and De Carvalho, F.A.T. *Bivariate Symbolic Regression Models for Interval-Valued Variables*, Journal of Statistical Computation and Simulation (on-line), iFrist, pp. 1–18, 2011.
14. Maia, A.L.S., De Carvalho, F.A.T and Ludermir, T.B. *Forecasting models for interval-valued time series*, Neurocomputing 71, pp. 3344–3352, 2008.
15. McCullagh, P. and Nelder, J. *Generalized Linear Models* (2nd Edition), Chapman & Hall, London, 1989.
16. Montgomery, D.C. and Peck, E.A. *Introduction to Linear Regression Analysis*, John Wiley, New York, 1982.
17. Palumbo, F. and Verde, R. *Non-symmetrical factorial discriminant analysis for symbolic objects*, Applied Stochastic Models in Business and Industry 15, pp. 419–427, 2000.
18. Scheffé, H. *The Analysis of Variance*, John Wiley, New York, 1959.
19. Sklar, A. *Fonctions de repartition a n dimensions et leures marges*, Publ. Inst. Statist. Univ. Paris, 8, 229–231, 1959.