Symmetrical Linear Regression models for Symbolic Interval Data

Marco A. O. Domingues[1], Renata M.C.R. de Souza[1],
Francisco José A. Cysneiros[2]

[1] Centro de Informática
[2] Departamento de Estatística,
Universidade Federal de Pernambuco, P.O.Box 7851, Recife (PE), Brazil

Abstract This paper introduces a symmetrical linear regression model as an approach to fit a linear regression for Interval Valued Data.

Keywords: Symmetrical models, symbolic data.

1 Introduction

Symbolic Data Analysis (SDA) could be broadly defined as an extension of standard data analysis to symbolic data. In terms of Regression Analysis, recent works have been proposed to fit the classic linear regression model (CLRM) to symbolic Interval Valued Data (IVD) [1][2]. Those approaches do not consider any probabilistic hypothesis on the response variable and use least squares method to perform parameter estimates whose results are strongly influenced by the presence outliers.

This work introduces a new prediction method for IVD based on the symmetrical linear regression (SLR) analysis. Its main feature is that the response model is less susceptible to the presence of IVD outliers. The model considers the Student-t distribution as an assumption for the errors in the centre of the symbolic interval variables.

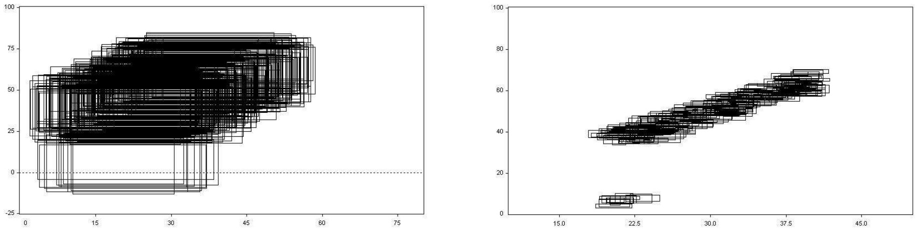
2 Symmetrical linear regression

The SLR model is defined as $Y_i = \mu_i + \epsilon_i, \quad i = 1, \ldots, n$, where $\mu_i = \mathbf{x}_i^t \boldsymbol{\beta}$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ is an unknown parameters vector, $\epsilon_i \sim S(0, \boldsymbol{\phi}, g)$ and $\mathbf{x}_i$ is the vector of explanatory variables. This class of models includes all symmetric continuous distributions, such as normal, Student-t, logistic, among others. The maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ cannot be obtained separately and closed-form expressions for this estimates do not exist. *Scoring Fisher* method can be applied to get $\hat{\boldsymbol{\phi}}$ where the process for $\hat{\boldsymbol{\beta}}$ can be interpreted as a weighted least square. The iterative process for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\phi}}$ takes the form $\boldsymbol{\beta}^{(m+1)} = \{\mathbf{X}\mathbf{D}(\mathbf{v}^{(m)})\mathbf{X}\}^{-1}\mathbf{X}^t\mathbf{D}(\mathbf{v}^{(m)})\boldsymbol{y}$ and $\phi^{(m+1)} = \frac{1}{n}\{\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\}^T\mathbf{D}(\mathbf{v})\{\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\}(m = 0, 1, 2, \ldots)$ with $\mathbf{D}(\mathbf{v}) = \text{diag}\{v_1, \ldots, v_n\}$, $\boldsymbol{y} = (y_1, \ldots, y_n)^t$, $\mathbf{X} = (\mathbf{x}_1^t, \ldots, \mathbf{x}_n^t)^t$ and $v_i = -2\text{W}_g(u_i)$, $\text{W}_g(u) = \frac{g'(u)}{g(u)}$, $g'(u) = \frac{dg(u)}{du}$ and $u_i = (y_i - \mu_i)^2/\phi$.

For the Student-t distribution with ν degrees of freedoms, $g(u) = c(1+u/\nu)^{-(\nu+1)/2}, \nu > 0$ and $u > 0$ so that $\text{W}_g(u_i) = -(\nu+1)/2(\nu+u_i)$ and $v_i = (\nu+1)/(\nu+u_i), \forall i$. In this case the current weight $v_i^{(r)}$ is inversely proportional to the distance between the observed value y_i and its current predicted value $\mathbf{x}_i^t\boldsymbol{\beta}^{(r)}$, so that outlying observations tend to have small weights in the estimation process [3].

3 Experiments

To show the usefulness of SLRM, a small subset of the simulated IVD are clustered and changed into outliers by moving the centre of each observation 1. In order to analyze the proposed method we performed Monte Carlo simulations with 100 iterations considering each data set and their simulated outliers.



(a) *Data set* 1.

(b) *Data set* 2.

Figure 1: Interval-valued data sets containing outlier rectangles.

The performance assessment of the SLR model presented is based on the *pooled root mean-square error* ($PRMSE$) applied for a learning IVD set (n=250) and test IVD set (n=125). $PRMSE^1 = \sqrt{\frac{\sum_{i=1}^{250} \omega(i)[(l_Y(i)-\hat{l}_Y(i))^2+(u_Y(i)-\hat{u}_Y(i))^2]}{250}}$, where $\omega(i)$ is the weight of the residual obtained from SLRM and $PRMSE^2 = \sqrt{\frac{\sum_{i=1}^{125}[(l_Y(i)-\hat{l}_Y(i))^2+(u_Y(i)-\hat{u}_Y(i))^2]}{125}}$

Statistical Student's t-test for paired samples at a significance level of 1% is then applied to compare the proposed SLR model to the linear regression model for IVD. The hypotheses are, respectively: $H_0 : (PRMSE^k)^{Symmetrical} = (PRMSE^k)^{Linear}$ and $H_1 : (PRMSE^k)^{symmetrical} < (PRMSE^k)^{Linear}$.

For all test data sets in this evaluation the rejection ratios of H_0 are equal to 100%.

4 Conclusions

A symmetrical linear prediction model for symbolic IVD is introduced in this paper, and experiments with simulated IVD sets containing IVD outliers are performed. The prediction performance is assessed by a PRMSE applied to learning and test data sets, and results provided by the proposed method are compared to the correspondent results provided by least squares method.The results showed that the symmetrical model is superior to centre-range model in terms of prediction qualities.

References

[1] Billard, L., Diday, E., 2006. Symbolic Data Analysis: Conceptual Statistics and Data Mining, Wiley, West Sussex, England 2006.

[2] Lima Neto, E.A., De Carvalho, F.A.T., 2008. Centre and Range method for fitting a linear regression model to symbolic interval data. In CSDA, v.52 n.3, pp. 1500-1515.

[3] Cysneiros, F.J.A., Paula, G.A., 2005. Restricted methods in symmetrical linear regression models. In CSDA, v. 49, pp 689-708.