A new linear regression model for histogram-valued variables

Dias, Sónia
Instituto Politécnico Viana do Castelo, Escola Superior de Tecnologia e Gestão
Praça General Barbosa
4900-347 Viana do Castelo, Portugal
E-mail: sdias@estg.ipvc.pt

Brito, Paula
Universidade do Porto, Faculdade de Economia and LIAAD-INESC Porto LA
Rua Dr. Roberto Frias
4200-464 Porto, Portugal
E-mail: mpbrito@fep.up.pt

Introduction

In classical data analysis, each individual takes one single "value" on each descriptive variable. Symbolic Data Analysis ([Bock and Diday (2000)], [Billard and Diday (2007)]) generalizes this framework by allowing each individual or class of individuals to take a finite set of values (quantitative multi-valued variables), a finite set of categories (qualitative multi-valued variables), an interval (interval-valued variable) or a distribution on each variable (modal-valued variables). A special case of these latter is when the distribution, for all observations of the modal-valued variable, is given by depicting the probabilities/ frequencies of observations occurring in certain ranges of values - we say then that we are in presence of a histogram-valued variable. Interval-valued variables may be seen as a particular case of the histogram-valued variables if for all observations we have only one interval with probability/frequency one. The variable Y is a random histogram-valued variable if to each observation j, $Y(j)$ corresponds a probability or frequency distribution that can be represented by the histogram ([Bock and Diday (2000)]):

$$H_{Y(j)} = \left\{ \left[\underline{I}_{Y(j)_1}, \overline{I}_{Y(j)_1}\right[, p_{j1}; \left[\underline{I}_{Y(j)_2}, \overline{I}_{Y(j)_2}\right[, p_{j2}; \ldots; \left[\underline{I}_{Y(j)n_j}, \overline{I}_{Y(j)n_j}\right], p_{jn_j} \right\} \tag{1}$$

where p_{ji} is the probability or frequency associated to the sub-interval $\left[\underline{I}_{Y(j)_i}, \overline{I}_{Y(j)_i}\right[$ with $i \in \{1, 2, \ldots, n_j\}$, n_j is the number of sub-intervals for the j^{th} observation, $\sum_{i=1}^{n_j} p_{ji} = 1$, $\underline{I}_{Y(j)_i} \leq \overline{I}_{Y(j)_i}$ and $\underline{I}_{Y(j)_{i+1}} \leq \overline{I}_{Y(j)_i}$.

It is assumed that within each sub-interval $\left[\underline{I}_{Y(j)_i}, \overline{I}_{Y(j)_i}\right[$ the values for the variable Y for the observation j, are uniformly distributed. For different observations, the number of sub-intervals of the histogram-valued variable may be different.

For each observation j, $Y(j)$ can, alternatively, be represented by the inverse cumulative distribution function also called quantile function $\Psi^{-1}_{Y(j)}$ ([Irpino and Verde (2006)]). This function is given by

$$\Psi^{-1}_{Y(j)}(t) = \begin{cases} \underline{I}_{Y(j)_1} + \frac{t}{w_{j1}} a_{Y(j)_1} & if \quad 0 \leq t < w_{j1} \\ \underline{I}_{Y(j)_2} + \frac{t - w_{j1}}{w_{j2} - w_{j1}} a_{Y(j)_2} & if \quad w_{j1} \leq t < w_{j2} \\ \vdots & \\ \underline{I}_{Y(j)_n} + \frac{t - w_{jn_j-1}}{1 - w_{jn_j-1}} a_{Y(j)_{n_j}} & if \quad w_{jn_j-1} \leq t \leq 1 \end{cases} \tag{2}$$

where $w_{jl} = \begin{cases} 0 & if \quad l = 0 \\ \sum_{h=1}^{l} p_{jh} & if \quad l = 1, \ldots, n_j \end{cases}$; $a_{Y(j)_i} = \overline{I}_{Y(j)_i} - \underline{I}_{Y(j)_i}$ with $i = \{1, \ldots, n_j\}$ and n_j is the number of sub-intervals in $Y(j)$.

Every time we use the term "distribution" we are considering a probability or frequency distribution of data of a continuous variable that can be represented by a histogram or a quantile function.

In recent years, some concepts and statistical methods for symbolic variables and in particular for histogram-valued variables were defined ([Billard and Diday (2007)]). Frequently, methods for these variables are a generalization of their counterparts for interval-valued variables.

Linear Regression Model

In 2000, Billard and Diday ([Billard and Diday (2000)]) proposed the first linear regression model for interval-valued variables and later expanded their work to histogram-valued variables ([Billard and Diday (2002)]). For interval-valued variables, several models had already been proposed and compared ([Billard and Diday (2002)], [Neto and Carvalho (2008)], [Neto and Carvalho (2010)]). However, these models present some limitations: firstly, all they are based on differences between real values and do not appropriately quantify the closeness between intervals; then, the elements estimated by the models may fail to build an interval; for this reason, the most recent model imposes non-negativity constraints on the coefficients, therefore forcing a direct linear relationship.

Although interval-valued variables are a particular case of the histogram-valued variables, the limitations of the models proposed for interval-valued variables prevent their generalization to histogram-valued variables; therefore alternative models should be developed ([Irpino and Verde (2010)]). The model that we will next propose for histogram-valued variables is also not a generalization, but the analysis of the limitations present in the models for interval-valued variables has allowed to establish the goals for the model that we define here, as follows: finding an error measure to quantify the difference between the observed and estimated distributions represented by histograms or quantile functions; defining a linear regression model for histogram-valued variables that allows the estimation of histograms or their quantile functions from other histograms or quantile functions, without forcing a direct linear relationship, and measuring the goodness-of-fit of the model.

To quantify the difference between the observed and estimated distributions that can be represented by histograms or quantile functions, we consider the work of Arroyo and Maté ([Arroyo and Maté (2009)]). In their work on forecasting time series, applied to histogram-valued variables, they used the Mallows and Wasserstein distances to measure the error between the observed and forecasted distributions. In using the Wasserstein and Mallows distances, the values that the histogram-valued variables take are represented by their quantile functions and not by their histograms. Given two quantile functions $\Psi^{-1}_{X(j)}$ and $\Psi^{-1}_{Y(j)}$ that represent the distributions that the histogram-valued variables X and Y take for observation j, the Wasserstein distance is defined as

$$D_W(\Psi^{-1}_{X(j)}, \Psi^{-1}_{Y(j)}) = \int_0^1 \left| \Psi^{-1}_{X(j)}(t) - \Psi^{-1}_{Y(j)}(t) \right| dt \tag{3}$$

and the Mallows distance as

$$D_M(\Psi^{-1}_{X(j)}, \Psi^{-1}_{Y(j)}) = \sqrt{\int_0^1 (\Psi^{-1}_{X(j)}(t) - \Psi^{-1}_{Y(j)}(t))^2 dt} \tag{4}$$

For Arroyo and Maté ([Arroyo and Maté (2009)]), these distances are a good measure to analyze the similarity between two distributions. Other works for histogram-valued variables also used these

measures ([Irpino and Verde (2006)]). Therefore, it seems appropriate to choose the Mallows distance to measure the similarity between the observed and the estimated distributions obtained by the linear regression model.

Consider, for each observation j, the quantile function $\Psi^{-1}_{Y(j)}$, that represents the observed distribution of the histogram-valued variable and the quantile function $\Psi^{-1}_{\widehat{Y}(j)}$, that represents the estimated distribution of the histogram-valued variable. For observation j, the error between $Y(j)$ and $\widehat{Y}(j)$ is given by

$$SE(j) = D^2_M(\Psi^{-1}_{Y(j)}, \Psi^{-1}_{\widehat{Y}(j)}) \tag{5}$$

and the total error is

$$SE = \sum_{j=1}^{m} D^2_M(\Psi^{-1}_{Y(j)}, \Psi^{-1}_{\widehat{Y}(j)}) \tag{6}$$

The Mallows distance can be rewritten using the center and half-range of the sub-intervals of the histograms ([Irpino and Verde (2006)]). So, the error between the distributions that the histogram-valued variables Y and $\widehat{Y}$ take, can be defined as follows:

$$SE = \sum_{j=1}^{m}\sum_{i=1}^{n} p_{ji}\left[(c_{Y(j)_i} - c_{\widehat{Y}(j)_i})^2 + \frac{1}{3}(r_{Y(j)_i} - r_{\widehat{Y}(j)_i})^2\right] \tag{7}$$

where $c_{Y(j)_i} = \frac{\overline{I}_{Y(j)_i} + \underline{I}_{Y(j)_i}}{2}$; $c_{\widehat{Y}(j)_i} = \frac{\overline{I}_{\widehat{Y}(j)_i} + \underline{I}_{\widehat{Y}(j)_i}}{2}$; and $r_{Y(j)_i} = \frac{\overline{I}_{Y(j)_i} - \underline{I}_{Y(j)_i}}{2}$; $r_{\widehat{Y}(j)_i} = \frac{\overline{I}_{\widehat{Y}(j)_i} - \underline{I}_{\widehat{Y}(j)_i}}{2}$.

The first option to define the functional linear relation between the observations of the histogram-valued variables was to adapt directly the classical linear regression model:

$$\Psi^{-1}_{\widehat{Y}(j)}(t) = \gamma + \alpha_1\Psi^{-1}_{X_1(j)}(t) + \alpha_2\Psi^{-1}_{X_2(j)}(t) + \ldots + \alpha_p\Psi^{-1}_{X_p(j)}(t). \tag{8}$$

The distributions taken by the histogram-valued variables are represented by their quantile functions, because this is the representation used by the error measure. However, in this model it is also necessary to impose non-negativity constraints on the parameters since quantile functions are necessarily non-decreasing functions. Although we did not generalize the model for interval-valued variables to histogram-valued variables, in defining a model that allows to estimate a quantile function from other quantile functions, we obtain a model with the same limitations as observed before.

To resolve this limitation, we introduce in the above model both the quantile functions that represent the distributions that the independent histogram-valued variables take, $\Psi^{-1}_{X_1(j)}, \Psi^{-1}_{X_1(j)}, \ldots, \Psi^{-1}_{X_p(j)}$ and the quantile functions that represent the distributions that the respective symmetric histogram-valued variables take, $\Psi^{-1}_{\widetilde{X}_1(j)}, \Psi^{-1}_{\widetilde{X}_1(j)}, \ldots, \Psi^{-1}_{\widetilde{X}_p(j)}$. The estimated quantile function $\Psi^{-1}_{\widehat{Y}(j)}$, is then given by:

$$\Psi^{-1}_{\widehat{Y}(j)}(t) = \gamma + \alpha_1\Psi^{-1}_{X_1(j)}(t) + \beta_1\Psi^{-1}_{\widetilde{X}_1(j)}(t) + \alpha_2\Psi^{-1}_{X_2(j)}(t) + \beta_2\Psi^{-1}_{\widetilde{X}_2(j)}(t) + \ldots + \alpha_p\Psi^{-1}_{X_p(j)}(t) + \beta_p\Psi^{-1}_{\widetilde{X}_p(j)}(t). \tag{9}$$

with $\alpha_k, \beta_k \geq 0$, $k = \{1, 2, \ldots, p\}$ and $\gamma \in I\!R$.

In this model, restrictions on the parameters are imposed, but in this case it does not imply a direct linear relationship because the model includes both the quantile functions that represent the distributions taken by the histogram-valued variables and the quantile functions that represent

the distributions taken by the respective symmetric histogram-valued variables. Estimation of the model requires solving a quadratic optimization problem, subject to non-negativity constraints on the unknowns and use the Mallows distance. The parameters of the model are an optimal solution of the minimization problem:

$$(10) \qquad Minimize \qquad SE = \sum_{j=1}^{m} D_M^2(\Psi_{Y(j)}^{-1}, \Psi_{\widehat{Y}(j)}^{-1}) \\ \text{with } \alpha_k, \beta_k \geq 0,\ k = \{1, 2, \ldots, p\} \text{ and } \gamma \in I\!R.$$

Similarly to the classical model, the Kuhn Tucker optimality conditions ([Winston (1994)]) allow defining a measure to evaluate the goodness-of-fit of the model. This is given by

$$(11) \qquad \Omega = \frac{\sum_{j=1}^{m} D_M^2\left(\Psi_{\widehat{Y}(j)}^{-1}(t), \overline{Y}\right)}{\sum_{j=1}^{m} D_M^2\left(\Psi_{Y(j)}^{-1}(t), \overline{Y}\right)},$$

where $\overline{Y} = \frac{1}{m}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}\left(\frac{\underline{I}_{Y(j)_i} + \overline{I}_{Y(j)_i}}{2}\right)p_{ji}\right)$.

It is straigtforward to prove that the goodness-of-fit measure Ω also ranges from 0 to 1.

Application

To illustrate the proposed linear regression model, we choose the example presented by Billard and Diday ([Billard and Diday (2007)]) in their linear regression model for histogram-valued variables. In this case, the hematocrit values and hemoglobin values are represented as histogram-valued data for each of 10 observations.

Observations	*Hematocrit*(Y)	*Hemoglobin*(X)
1	$\{[33.29; 37.52[\,, 0.6; [37.52; 39.61]\,, 0.4\}$	$\{[11.54; 12.19[\,, 0.4; [12.19; 12.8]\,, 0.6\}$
2	$\{[36.69; 39.11[\,, 0.3; [39.11; 45.12]\,, 0.7\}$	$\{[12.07; 13.32[\,, 0.5; [13.32; 14.17]\,, 0.5\}$
3	$\{[36.69; 42.64[\,, 0.5; [42.64; 48.68]\,, 0.5\}$	$\{[12.38; 14.2[\,, 0.3; [14.2; 16.16]\,, 0.7\}$
4	$\{[36.38; 40.87[\,, 0.4; [40.87; 47.41]\,, 0.6\}$	$\{[12.38; 14.26[\,, 0.5; [14.26; 15.29]\,, 0.5\}$
5	$\{[39.19; 50.86]\,, 1\}$	$\{[13.58; 14.28[\,, 0.3; [14.28; 16.24]\,, 0.7\}$
6	$\{[39.7; 44.32[\,, 0.4; [44.32; 47.24]\,, 0.6\}$	$\{[13.81; 14.5[\,, 0.4; [14.5; 15.2]\,, 0.6\}$
7	$\{[41.56; 46.65[\,, 0.6; [46.65; 48.81]\,, 0.4\}$	$\{[14.34; 14.81[\,, 0.5; [14.81; 15.55]\,, 0.5\}$
8	$\{[38.4; 42.93[\,, 0.7; [42.93; 45.22]\,, 0.3\}$	$\{[13.27; 14.0[\,, 0.6; [14.0; 14.6]\,, 0.4\}$
9	$\{[28.83; 35.55[\,, 0.5; [35.55; 41.98]\,, 0.5\}$	$\{[9.92; 11.98[\,, 0.4; [11.98; 13.8]\,, 0.6\}$
10	$\{[44.48; 52.53]\,, 1\}$	$\{[15.37; 15.78[\,, 0.3; [15.78; 16.75]\,, 0.7\}$

Table 1: Example of dataset with two histogram-valued variables

We estimated the quantile function representing the distribution taken by the histogram-valued variable Y from the model, and obtained:

$$\Psi^{-1}_{\widehat{Y}(j)}(t) = -1,953 + 3,5598\Psi^{-1}_{X(j)}(t) + 0,4128\Psi^{-1}_{\widetilde{X}(j)}(t)$$

The goodness-of-fit measure is, for this case, $\Omega = 0,96$.

In *Figure* 1 we can see that the estimated and observed quantile functions for each observation are very similar, in agreement with the value of the coefficient of determination, Ω.

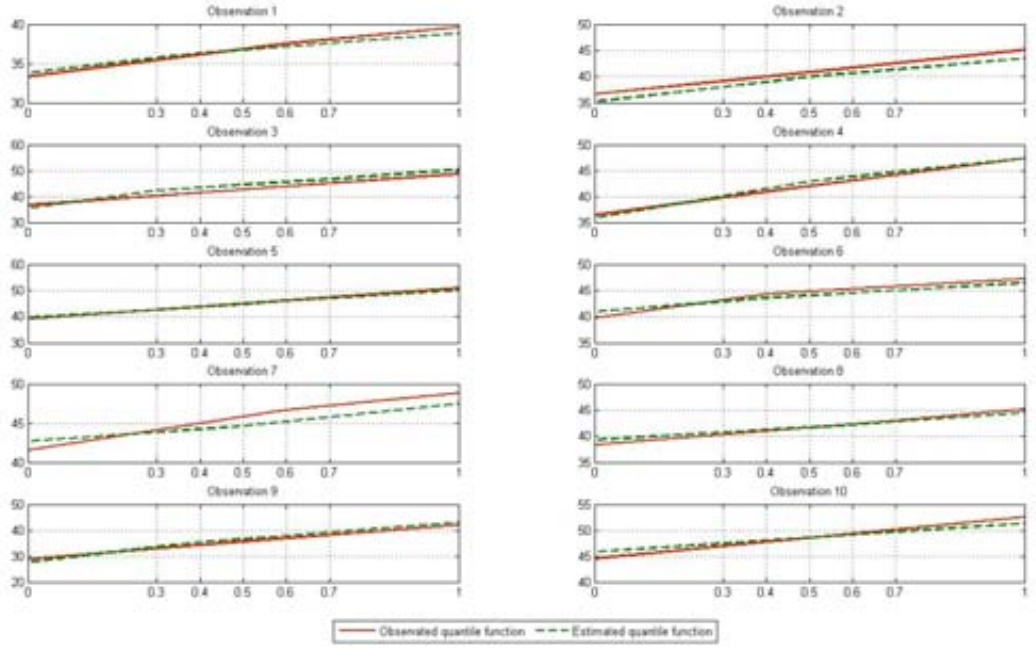

Figure 1: Observed and estimated quantile functions of each observation

Conclusion

With the linear regression model proposed it is possible to estimate the quantile functions that represent the distributions taken by the dependent histogram-valued variable from the quantile functions that represent the distributions taken by the independent histogram-valued variables. Moreover, it is possible to deduce a goodness-of-fit measure from the model. This measure appears to have a good behavior: when we compare the representation of the estimated and observed quantile functions for each observation we have good estimates when the coefficient of determination is close to one whereas the estimated and observed quantile functions are more discrepant when the coefficient of determination is lower. As interval-valued variables are a particular case of the histogram-valued variables it is possible to particularize this model for interval-valued variables. For both types of variables it will be necessary to further study the proposed model, to explore its interpretation and behavior.

REFERENCES

Arroyo, J., Maté, C. (2009). Forecasting Histogram Time Series with K-Nearest Neighbours Methods. International Journal of Forecasting, 25, pp. 182-207.

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

Billard, L., Diday, E. (2002). Symbolic Regression Analysis. In: Classification, Clustering and Data Analysis. Proceedings of the Eighth Conference of the International Federation of Classification Societies (IFCS02). Springer, pp 281-288.

Billard, L., Diday, E. (2000). 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 (IFCS00). Springer, pp 369-374.

Bock, H.-H., Diday, E. (2000). Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Springer-Verlag, Berlin-Heidelberg.

Irpino, A. Verde, R. (2010). Ordinary Least Squares for Histogram Data Based on Wasserstein Distance. In: COMPSTAT'2010, 19th Conference of IASC-ERS (Physica Verlag), pp. 581-589.

Irpino, A., Verde, R. (2006). A New Wasserstein Based Distance for the Hierarchical Clustering of Histogram Symbolic Data. In: Classification, Data Science and Classification, Proceedings of the Conference of the International Federation of Classification Societies (IFCS06). Springer, Berlin, pp. 185-192.

Lima Neto, E.A., de Carvalho, F.A.T. (2010). Constrained linear regression models for symbolic interval-valued. Computational Statistics and Data Analysis, 54, pp. 333-347.

Lima Neto, E.A., de Carvalho, F.A.T. (2008). Linear regression models for symbolic interval-valued. Computational Statistics and Data Analysis, 54, pp. 333-347.

Winston, W. (1994). Operations Research. Applications and Algorithms - 3rd Edition. Duxbury Press. California.