EKONOMETRIA ECONOMETRICS 2(52) • 2016

ISSN 1507-3866
e-ISSN 2449-9994

Marcin Pełka, Andrzej Dudek
Wrocław Universityt of Economics
e-mails: marcin.pelka@ue.wroc.pl; andrzej.dudek@ue.wroc.pl

REGRESSION ANALYSIS FOR INTERVAL-VALUED SYMBOLIC DATA VERSUS NOISY VARIABLES AND OUTLIERS

REGRESJA LINIOWA DANYCH SYMBOLICZNYCH A ZMIENNE ZAKŁÓCAJĄCE I OBSERWACJE ODSTAJĄCE

DOI: 10.15611/ekt.2016.2.03
JEL Classification: C51, C59, C63, C87

Summary: Regression analysis is perhaps the best known and most widely used method used for the analysis of dependence; that is, for examining the relationship between a set of independent variables (*X*'s) and a single dependent variable (*Y*). In general regression, the model is a linear combination of independent variables that corresponds as closely as possible to the dependent variable [Lattin, Carroll, Green 2003, p. 38]. The aim of the article is to present two suitable adaptations for a regression analysis of symbolic interval-valued data (centre method and centre and range method) and to compare their usefulness when dealing with noisy variables and/or outliers. The empirical part of the paper presents the results of simulation studies based on artificial and real data, without noisy variables and/or outliers and with noisy variable and outliers. The results are compared according to the values of two coefficients of determination R_L^2 and R_U^2. The results show that usually the centre and range method obtains better results even when the data set contains noisy variables and outliers, but in some cases the centre method obtains better results than the centre and range method.

Keywords: regression analysis, interval-valued symbolic data, noisy variables, outliers.

Streszczenie: Analiza regresji jest z pewnością jedną z powszechniej stosowanych technik w analize zależności zmiennych. Pozwala ona na zbadanie zależności pomiędzy zbiorem zmiennych objaśniających a zmienną niezależną. Ogólnie w modelu regresji ujęta jest liniowa zależność między zmiennymi zależnymi. Celem artykułu jest zaprezentowanie odpowiednich metod analizy regresji dla zmiennych symbolicznych interwałowych oraz ich porównanie w sytuacji, gdy mamy do czynienia ze zmiennymi zakłócającymi lub obserwacjami odstającymi. W części empirycznej przedstawiono wyniki badań symulacyjnych z zastosowaniem danych rzeczywistych oraz sztucznych ze zmiennymi zakłócającymi lub obserwacjami odstajacymi. Wyniki porównano, wykorzystując współczynnik R^2.

Słowa kluczowe: analiza regresji, interwałowe zmienne symboliczne, obserwacje odstające, zmienne zakłócające.

1. Introduction

Regression analysis is used to determine the statistical relationship between a set of independent variables (*X*'s) and a single dependent variable (*Y*). Of course this relationship is not captured perfectly, there are usually many other unobserved or unmeasured factors that also influence the value of the dependent variable (*Y*) (see for example: [Lattin, Carroll, Green 2003, p. 39; Hair et. al. 2006, pp. 169–170]).

The basic formulation for regression analysis is (see [Hair et. al. 2006, p. 169; Lattin, Carroll, Green 2003, p. 44]):

$$y = b_1 X_1 + b_2 X_2 + \ldots + b_n X_n + \varepsilon \tag{1}$$

or, in matrix terms:

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon}. \tag{2}$$

Usually the ordinary least squares method is used to estimate a linear regression model. This method has many well-known properties, limitations and disadvantages that are widely described in the literature (see for example: [Hair et al. 2006, pp. 204–208; Walesiak, Gatnar (eds.) 2004, pp. 83–84; Lattin, Carroll, Green 2003, pp. 43–47, 56-65; Welfe 2003, pp. 29–32]).

To estimate model coefficients we use the following formula:

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}. \tag{3}$$

The aim of this article is to present suitable adaptations for the regression analysis of symbolic interval-valued data and to compare their usefulness when dealing with noisy variables and/or outliers. The empirical part of the paper presents the results of simulation studies – based on artificial data, without noisy variables and/or outliers and with noisy variable and outliers. The results are compared according to two coefficients of determination: one for lower bounds R_L^2 and one for upper bounds R_U^2.

2. Symbolic data

Symbolic objects, unlike classical objects, can be described by many different symbolic variable types. They can be described by the following variables [Bock, Diday (eds.) 2000, pp. 2–3; Billard, Diday 2006, pp. 7-30; Dudek 2013, pp. 35–36] – see Table 1 for examples of symbolic variables:

1) Quantitative (numerical) variables:

- numerical single-valued variables,
- numerical multi-valued variables,
- interval-valued variables,
- histogram variables.

2) Qualitative (categorical) variables:

- categorical single-valued variables,
- categorical multi-valued variables,
- categorical modal variables.

Regardless of their type, symbolic variables also can be [Bock, Diday (eds.) 2000, p. 2]:

1. Taxonomic – which present a prior known structure.

2. Hierarchically dependent – rules which decide if a variable is applicable or has not been defined.

3. Logically dependent – logical rules which affect a variable's values have been defined.

Table 1. Examples of symbolic variables

Symbolic variable	Realizations	Variable type
Preferred price of a new car (in PLN)	<25000; 36000>, <28000; 37000>, <30000; 50000>, <33000; 58000>, <65000; 80000>, <66000; 90000>	interval-valued (non-disjoint)
Engine capacity	<1000; 1200>, (1200; 1400>, (1400; 1600>, (1600; 1800>, (1800; 2000>, (2000; 2200>	interval-valued (disjoint)
Colour	{green, black, yellow, red, purple, blue}	multivalued

Preferred brand of a car	{60% Honda, 35% Toyota, 5% Audi}
	{40% Honda, 20% Skoda, 20% Toyota, 20% Audi}
	{80% Audi, 15% Opel, 5% Toyota}
		multivalued with weights

Time spent travelling to work	{60% <10, 20>, 40% <30, 40>}
	{20% <10, 20>, 80% <30, 40>}
		histogram

Gender of the respondent	{M}, {F}	nominal

Source: own research.

There are two main symbolic objects' types:

1. First order objects (simple objects, individuals) – single respondent, product, company, etc., described by symbolic variable types. These objects are individuals that are symbolic by their nature.

2. Second order objects (aggregate objects, super individuals) – more or less homogeneous classes, groups of individuals described by symbolic variables.

3. Regression analysis for interval-valued data

As in the case of symbolic interval-valued data, we deal not with single real numbers, but with intervals $x_i = \left[a_i, b_i\right]$ – where a_i (b_i) is the lower (upper) bound of the *i*-th symbolic interval-valued variable – the standard equation $\hat{\mathbf{b}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ cannot be directly applied. There are two main proposals in the literature as to how to apply this equation in the case of symbolic interval-valued data [Lima-Neto, de

Carvalho 2008, pp. 1500-1515; 2010, pp. 333-347; Billard, Diday 2006, pp. 198–201; Diday, Noirhomme-Fraiture 2008, pp. 360–361]:

a) *centre method* – where the elements of the matrices **X** and **y** are substituted by the centres $\left(\frac{a_i+b_i}{2}\right)$ of intervals for each variable, so in this case we get:

$$\hat{\mathbf{b}} = \left(\left(\mathbf{X}^c\right)^T\left(\mathbf{X}^c\right)\right)^{-1}\left(\mathbf{X}^c\right)^T\mathbf{y}^c, \tag{4}$$

where: $\mathbf{X}^c$ – is the matrix of centres for X's, $\mathbf{y}^c$ – is the matrix of centres for y.

The theoretical values for the lower (denoted by L) and upper (denoted by U) bounds of a symbolic interval-valued variable are obtained by applying the following equations:

$$\hat{\mathbf{y}}_L = \left(\mathbf{X}_L\right)^T\hat{\mathbf{b}} \text{ and } \hat{\mathbf{y}}_U = \left(\mathbf{X}_U\right)^T\hat{\mathbf{b}}; \tag{5}$$

b) *centre and range method* – where the elements of the elements of the matrices **X** and **y** are substituted by the centres $\left(\frac{a_i+b_i}{2}\right)$ and ranges $\left(\frac{b_i-a_i}{2}\right)$ of intervals for each variable, so in this case we get:

$$\begin{aligned}\hat{\mathbf{b}}^C &= \left(\left(\mathbf{X}^c\right)^T\left(\mathbf{X}^c\right)\right)^{-1}\left(\mathbf{X}^c\right)^T\mathbf{y}^c\\ \hat{\mathbf{b}}^r &= \left(\left(\mathbf{X}^r\right)^T\left(\mathbf{X}^r\right)\right)^{-1}\left(\mathbf{X}^r\right)^T\mathbf{y}^r\end{aligned} \tag{6}$$

where: $\hat{\mathbf{b}}^r$ are the estimates for ranges, and $\hat{\mathbf{b}}^c$ are the estimates for centres.

This means we estimate separately the coefficients for centres and ranges for the whole regression model.

The theoretical values for the lower (denoted by L) and upper (denoted by U) bounds of the symbolic interval-valued variable are obtained by applying the following equations:

$$\hat{\mathbf{y}}_L = \hat{\mathbf{y}}^C - \hat{\mathbf{y}}^r \text{ and } \hat{\mathbf{y}}_U = \hat{\mathbf{y}}^C + \hat{\mathbf{y}}^r, \tag{7}$$

where: $\hat{\mathbf{y}}^C = \left(\mathbf{X}^C\right)^T\hat{\mathbf{b}}^C$ – estimated values for centers, and $\hat{\mathbf{y}}^r = \left(\mathbf{X}^r\right)^T\hat{\mathbf{b}}^r$ – estimated values for ranges.

The papers by Lima-Neto and de Carvalho 2008 and 2010 (see: [Lima-Neto, de Carvalho 2010, pp. 333-347; 2008, pp. 1500-1515]), also propose linear regression models with penalty functions: *ridge* regression, *lasso regression* and the *elastic net model*. In the empirical part only the centre method and centre and range method, without any penalty functions, will be used.

As mentioned before, in the case where we are dealing with interval $x_i = \left[a_i, b_i\right]$ and not the single real numbers, there are two R^2 values for this type of data – one for the lower bounds $\left(R_L^2\right)$ and one for the upper bounds $\left(R_U^2\right)$.

For other types of symbolic data there are other solutions that allow us to use the ordinary least squares and linear regression model (see for example: [Billard, Diday 2006, pp. 192–226; Diday, Noirhomme-Fraiture 2008, pp. 361–372]).

However in the case of symbolic interval-valued data, and other types of symbolic data, no methodology for detecting problems with the model (e.g. multicollinearity or heteroscedasticity) and significance testing is proposed. It is hard to tell if centres, ranges (or both) or upper and lower bounds should be used in significance testing or detection of problems.

3. Results of simulation studies

In order to compare the centre method and centre and range method (with the application of R_L^2 and R_U^2), the following artificial data sets were prepared:

1) **Data set I** – 14 symbolic objects, one interval-valued variable *X*, interval-valued variable *y*. Figure 1 presents a correlation plot for the centres of this data set (without noisy variables and/or outliers). Figure 1 presents the correlation plot for the centres of *X* and *y* for this data set.

2) **Data set II** – 50 symbolic objects, three interval-valued *X*'s, interval-valued variable *y*. Figure 2 presents a scatterplot for centres of this data set (without noisy variables and/or outliers).

Besides artificial data sets with noisy variables and/or outliers and without them, two real data sets were used in the study:

1) Wheat harvests depending on fertilizers used (in kg of pure NPK) – data was collected from 2002 until 2010 across Polish regions. Then a time-based aggregation was applied to obtain first-order symbolic objects.

2) Human development index (HDI) value depending on three variables: mean years of schooling, life expectancy at birth and gross national income (GNI) per capita (2011 PPP in USD).

As we can see (Figures 1 and 2), in both cases data sets without noisy variables and/or outliers can be assumed to be linear.

For each artificial data set four simulation paths were prepared:

1. The original data set without noisy variables and/or outliers.
2. Data sets with 10% and 20% of outliers added. The outliers were generated independently for each variable for the whole data set from uniform distribution (with range [1, 10]). The generated values are randomly added to the upper bound of the *j*-th variable or subtracted from the lower bound of the *j*-th variable (see [Walesiak, Dudek 2014]).
3. Data set with 1 or 2 noisy variables. The noisy variables are simulated independently from the uniform distribution. The variations of noisy variables in the generated data are similar to non-noisy variables (see: [Milligan, Cooper 1985; Qiu, Joe 2006, p. 322; Walesiak, Dudek 2014]).
4. Data set with 2 noisy variables and 20% of outliers added.

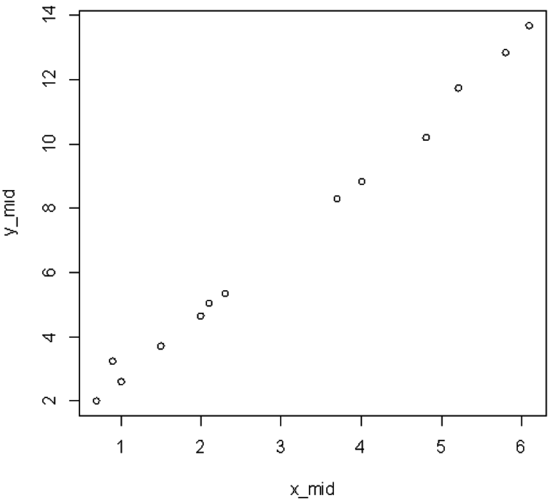

Figure 1. Correlation plot for centers of intervals of *X* and *y* for the first data set

Source: own computation.

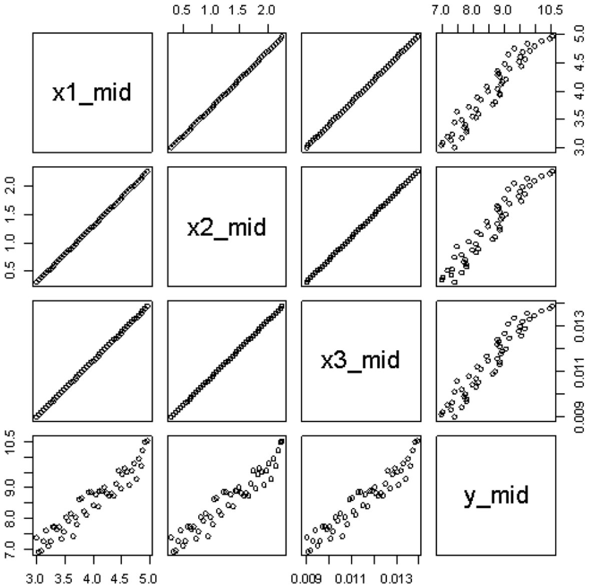

Figure 2. Scatterplot for centres of *X*'s and *y* for the second data set

Source: own computation.

The results for the first data set are shown in Table 2.

Table 2. Results of simulation studies for the first data set

Original data set	Outliers		Noisy variables		2 noisy variables and 20% of outliers
	10%	20%	1	2	
Centre method					
$R_L^2 = 94\%$ $R_U^2 = 96\%$	$R_L^2 = 94.76\%$ $R_U^2 = 95.01\%$	$R_L^2 = 86.88\%$ $R_U^2 = 87.01\%$	$R_L^2 = 98.62\%$ $R_U^2 = 97.77\%$	$R_L^2 = 98.77\%$ $R_U^2 = 97.01\%$	$R_L^2 = 98.57\%$ $R_U^2 = 97.43\%$
Centre and range method					
$R_L^2 = 100\%$ $R_U^2 = 100\%$	$R_L^2 = 94.63\%$ $R_U^2 = 95.12\%$	$R_L^2 = 86.80\%$ $R_U^2 = 86.90\%$	$R_L^2 = 99.62\%$ $R_U^2 = 99.78\%$	$R_L^2 = 99.80\%$ $R_U^2 = 99.17\%$	$R_L^2 = 67.54\%$ $R_U^2 = 68.15\%$

Source: own research with the application of R software.

The centre and range method. in the case of the first. quite simple data set, performs as well as the centre method (when comparing them according to R^2 values). The centre and range method performs worse when dealing with noisy variables and outliers. The results for the first data set are shown in Table 3.

Table 3. Results of simulation studies for the second data set

Original data set	Outliers		Noisy variables		2 noisy variables and 20% of outliers
	10%	20%	1	2	
Centre method					
$R_L^2 = 82.97\%$ $R_U^2 = 82.12\%$	$R_L^2 = 66.04\%$ $R_U^2 = 67.01\%$	$R_L^2 = 60.92\%$ $R_U^2 = 61.17\%$	$R_L^2 = 80.87\%$ $R_U^2 = 81.01\%$	$R_L^2 = 78.01\%$ $R_U^2 = 79.19\%$	$R_L^2 = 66.01\%$ $R_U^2 = 64.44\%$
Centre and range method					
$R_L^2 = 100\%$ $R_U^2 = 100\%$	$R_L^2 = 82.09\%$ $R_U^2 = 83.00\%$	$R_L^2 = 75.67\%$ $R_U^2 = 77.15\%$	$R_L^2 = 89.90\%$ $R_U^2 = 88.54\%$	$R_L^2 = 77.89\%$ $R_U^2 = 78.01\%$	$R_L^2 = 70.01\%$ $R_U^2 = 69.69\%$

Source: own research with application of R software.

In the case of the wheat data and in the case of the HDI data, the better results were obtained by the centre and range method than in the case of the centre method.

4. Summary

There are two main methods proposed for ordinary least squares estimation for regression analysis for interval-valued data: the centre method, and centre and range method. Apart from these, regression methods with penalty functions are proposed.

In most cases the centre and range method obtains better results (in terms of R^2 values) than the simple centre method. This method obtains better results as, besides the centres, the ranges of the intervals are also taken into consideration – so this method should be used for more complex data sets. However, the centre and range method is more sensitive when larger changes of ranges occur.

A problem for future studies is the verification of ordinary least squares assumptions when dealing with symbolic interval-valued data.

Bibliography

Billard L., Diday E., 2006, *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, John Wiley & Sons, Chichester.

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

Diday E., Noirhomme-Fraiture M., 2008, *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, Wiley&Sons, Chichester.

Dudek A., 2013, *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wyd. UE we Wrocławiu, Wrocław.

Hair J.F., Black W.C., Babim B.J., Anderson R.E., Tatham R.L., 2006, *Multivariate Data Analysis*, Prentice Hall, New Jersey.

Lattin J., Carroll J.D., Green P.E., 2003, *Analyzing Multivariate Data*, Thomson Learning, Toronto.

Lima-Neto E.A., de Carvalho F.A.T., 2008, *Centre and range method to fitting a linear regression model on symbolic interval data*, Computational Statistics and Data Analysis, vol. 52, pp. 1500–1515.

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

Milligan G.W., Cooper M.C., *An examination of procedures for determining the number of clusters in a data set*, Psychometrika, vol. 50, no. 2, pp. 159–179.

Qiu W., Joe H., 2006, *Generation of Random Clusters with Specified Degree of Separation*. Journal of Classification, vol. 23, pp. 315-334.

Walesiak M., Dudek A., 2014, *The clusterSim package* [URL:] www.r-project.org.

Walesiak M., Gatnar E. (eds.), 2004, *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wyd. Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.

Welfe A., 2013, *Ekonometria*, PWN, Warszawa.