Bulletin of the *Transilvania* University of Braşov
Series V: Economic Sciences • Vol. 10 (59) No. 1 - 2017

Using the Regression Model in multivariate data analysis

Cristinel CONSTANTIN[1]

Abstract: *This paper is about an instrumental research regarding the using of Linear Regression Model for data analysis. The research uses a model based on real data and stress the necessity of a correct utilisation of such models in order to obtain accurate information for the decision makers. The main scope is to help practitioners and researchers in their efforts to build prediction models based on linear regressions. The conclusion reveals the necessity to use quantitative data for a correct model specification and to validate the model according to the assumptions of the least squares method.*

Key-words: *Regression model, multivariate analysis, model validation, predictions, GDP*

1. Introduction

The present paper is a part of a series of instrumental researches meant to review the main multivariate data analysis models. The research is based on the exemplification of using the Multiple Linear Regression Model starting from the model specification and continuing with the validation of this one. The main issues related to this model are underlined in order to stress the importance of a correct utilisation in the process of data analysis.

2. The Multiple Linear Regression Model

According to Kutner, et al. (2005), "regression analysis has three major purposes: (1) description, (2) control, and (3) prediction". Thus the regression model could be used to describe the relationship between different variables, to control and predict the evolution of a dependent variable according to the evolution of one or more variables used as predictors.

One of the most popular models is the linear one, which starts from the assumption of a linear relationship between the analysed variables. If we take into consideration a dependent variable (Y) and an independent one (X), it would be

[1] Transilvania University of Braşov, cristinel.constantin@unitbv.ro

supposed that the mean of the dependent variable is placed on a straight line determined by the variation of the independent variable. In this respect, two parameters, β_0 and β_1, which determine a straight line, can be calculated based on the observed data. The observations of the dependent variable Y_i is supposed to deviate from the mean with a random error denoted by ε_i (Rawlings, Pantula and Dickey, 1998). Thus, the statistical model of simple regression is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1)$$

In practice, the variation of dependent variable is determined by more than one predictor, so that a multiple regression model is used by adding more independent variables to the above equation.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_k X_{ki} + \varepsilon_i \qquad (2)$$

The estimation of model's parameters (β …β_k) is made by using the least squares method, which can be applied only when the number of observed values for the analysed variables (n) is higher than the number of independent variables (k). Other assumptions should be also considered: the errors (ε_i) have a null mean and a constant variance, the errors are not auto-correlated and the independent variables (X) are not correlated each other (Montgomery, Peck and Vining, 2006).

The variables used in the regression model, both dependent and independent ones, have to be quantifiable (Saunders, Lewis and Thornhill, 2007). In this respect, a ratio scale is used for measurement, which allows the calculation of means and variation indicators. Moreover, in order to make predictions the selected variables have to be in a cause-effect relationship- i.e. the independent variables determine the evolution of the dependent variable. Thus, a strong statistical relationship between variables is not enough and the causal relationship should be interpreted with caution, using supplementary analyses and references to theories (Kutner, et al. 2005).

Conducting multiple regression analysis involves several steps, starting with the estimation of the model's parameters, by using the least squares method, which continues with the test for parameters' significance and other tests conducted for verifying the model's assumptions mentioned above (Malhotra, 2004). Finally, the coefficient of determination (R^2) can be calculated in order to measure the strength of association. The predicted values of the dependent variable ($\hat{Y}$) can also be obtained based on the values of independent variables.

2. Using the Regression Model in data analysis

In the followings we are going to exemplify the applying of the Multiple Linear Regression Model using the IBM SPSS system (Statistical Package for Social Sciences). The dependent variable (Y) is the GDP recorded in 2014 by every county of Romania. The main economic problem related to this indicator is the huge difference that exists among Romanian counties (see Fig. 1).

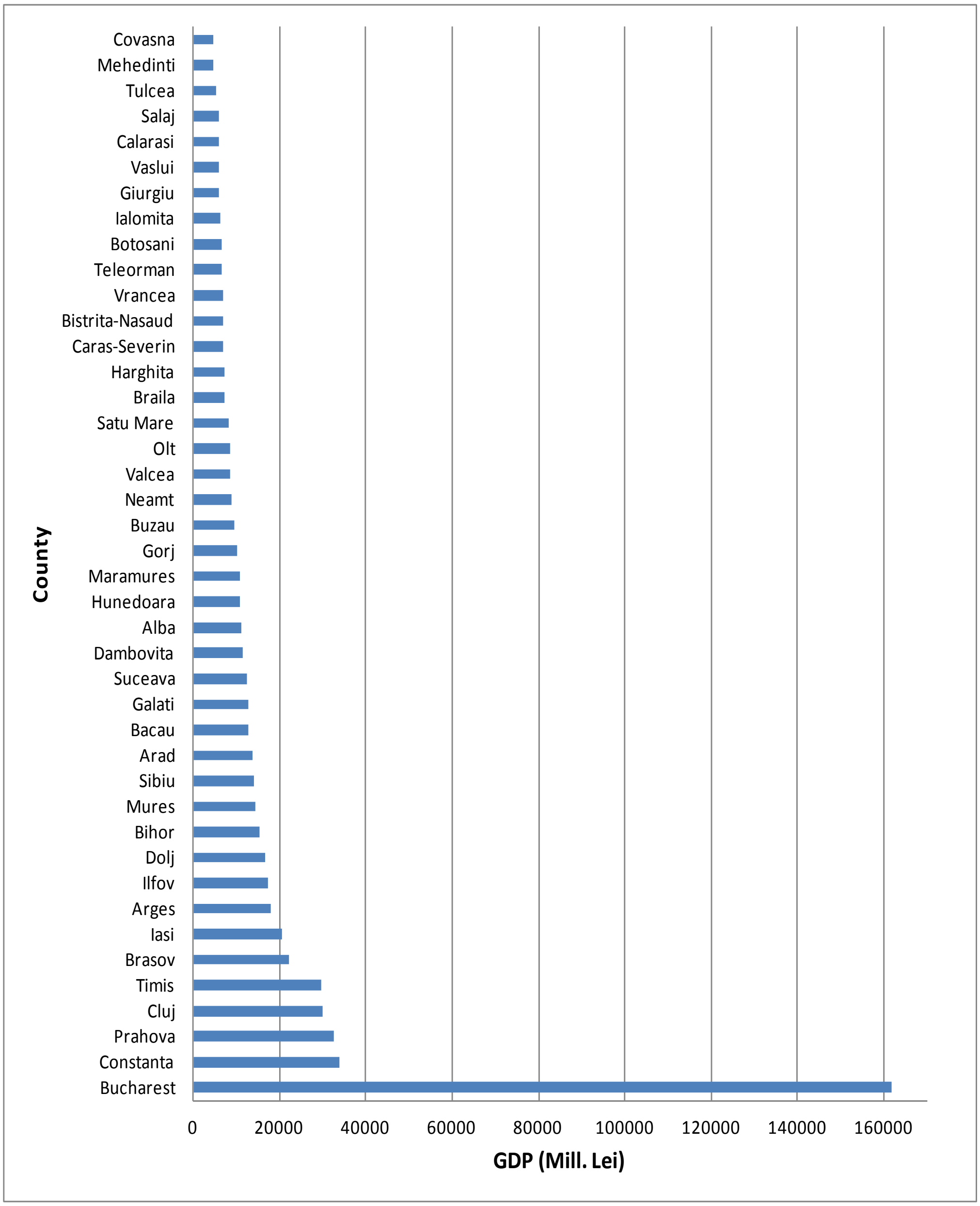

Fig. 1. *The levels of GDP recorded in 2014 by Romanian counties (Bucharest included)*

We can notice that the capital city recorded a very high value of GDP, which is higher than the sum of the first 5 counties but there are also significant discrepancies between counties. Starting from this problem we tried to identify the influence factors that contribute to a higher or a lower level of GDP. In this respect, a Multiple Linear Regression Model has been used having as predictors the following independent variables (see Fig. 2): the population size **(Population)**, the labour resources **(Labour_res)**, the number of active companies **(Company_ number)** and the employed population **(Employed_pop).** These predictors have been chosen starting from the supposition that the population is the main determinant of the total consumption and contribute to a large extent to the production process through the labour resources and the employed population. Another predictor, the number of active companies, can have a direct influence on the production and consequently on the GDP value. As Bucharest recorded an extreme value, it has been excluded from the analysis.

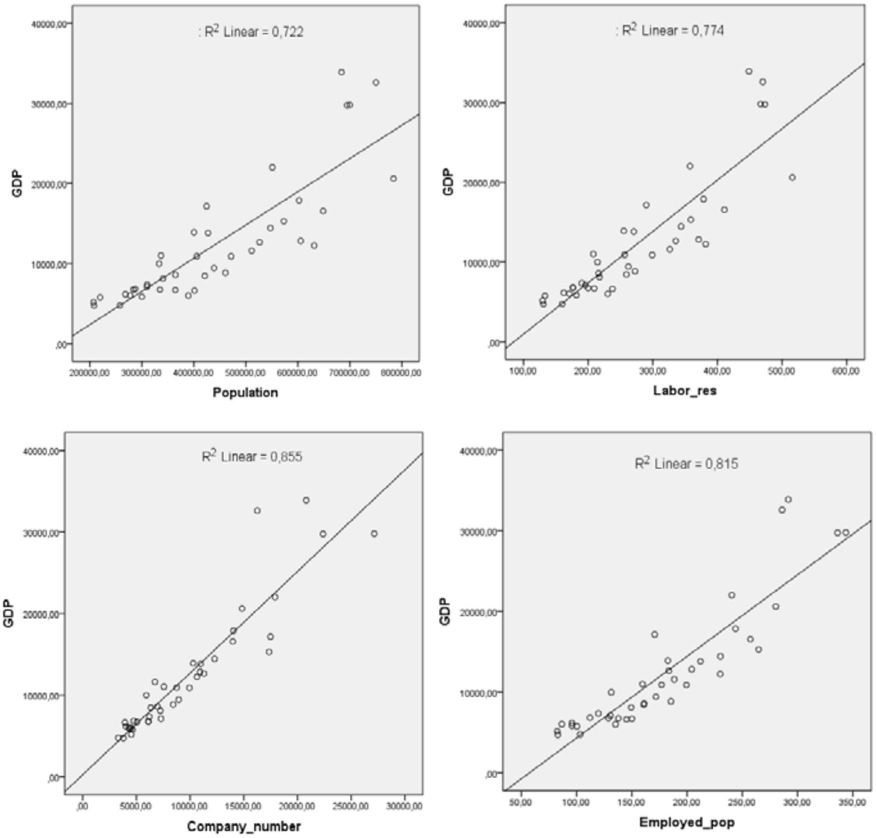

Fig. 2. *The linear dependences between the GDP and every predictor (year 2014)*

In Fig. 2 we can observe the linear dependence between the GDP and every predictor, with determination coefficients (R^2) higher than 0.7. Therefore, we can consider that every variable by itself explains more than 70% of the GDP variation. The highest influence is given by the number of companies, which explains 85.5% of the GDP variation. But the relevance of the statistical relationship is not enough for a scientific explanation of the causal relationship.

In practice, an economic phenomenon is influenced by a mix of factors and the use of Multiple Regression Model is more suitable. Thus we have applied such a model on the above variables, all predictors being included together in the analysis (see Table. 1).

Model	Unstandardized Coefficients		Std. Coeff.	t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
(Constant)	-2065,7	1491,5		-1,38	,175					
Population	-,047	,036	-,95	-1,29	,205	,85	-,21	-,07	,006	172,8
Labour_res	95,069	56,85	1,29	1,67	,103	,88	,26	,09	,005	191,6
Company_number	,668	,25	,49	2,60	,013	,92	,39	,14	,087	11,4
Employed_pop	11,660	30,51	,10	,38	,705	,90	,06	,02	,042	23,5

Table 1. *The regression coefficients and multicolinearity*

The results show that even if every variable by itself is strongly correlated with the dependent variable, when they are put together in the model, the majority of regression coefficients become insignificant. According to the results of t-Student test, the significance level (Sig.) is smaller than 0.05 only for the Company_number. For the rest of variables, Sig.> 0.05 and the coefficients cannot be considered significantly different from zero. This anomaly appears due to a high multicollinearity between variables. This phenomenon is presented in the column “Partial correlation”, where the partial correlation coefficients have quite small values. These coefficients represent the correlation between two variables, which remains after removing their mutual correlation with other variables included in model. The multicolinearity is also revealed in the last two columns of the table. Small values of “Tolerance” and big values of “Variance Inflation Factor (VIF)” underline a low contribution of the variables to the model. It means that the model has computational problem and the predictors have to be reconsidered. In order to avoid such problems it is recommended to use a selection method of predictors. One of the best methods is “Stepwise selection”. This one includes the predictors in model step by step, starting with the variable that has the highest influence on the dependent variable. The variables with small contribution to the variance explanation are excluded from model.

Model	Unstandardized Coefficients		Std. Coeff.	t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
(Constant)	231,501	923,835		,2	,803					
Company_number	1,246	,082	,925	15,1	,000	,925	,92	,925	1,000	1,000
(Constant)	-2631,5	1302,41		-2,0	,050					
Company_number	,87	,15	,649	5,8	,000	,925	,68	,328	,255	3,915
Labour_res	23,411	8,09	,320	2,8	,006	,880	,42	,162	,255	3,915

Table 2. *The results of regression model with stepwise selection*

The results of stepwise selection are presented in Table 2. The variable Company_number has been selected at the first step and Labour_res at the second step. The rest of variables have been rejected. We can observe that by including the second variable the tolerance decreased and the VIF increased due to a certain multicolinearity but the regression coefficients are both significant as Sig.< 0.05.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,925^{a}	,855	,851	2982,58941	
2	,939^{b}	,881	,875	2735,12359	1,881

a. Predictors: (Constant), Company_number

b. Predictors: (Constant), Company_number, Labour_res

c. Dependent Variable: GDP

Table 3. *The coefficient of determination and error autocorrelation testing*

The coefficient of determination (R^2) for the model with two independent variables is 0.855, which means that these variables explain 85.5% of the GDP variation (see Table 3). The result of Durbin-Watson test (DW) is also provided. This is a tool used to test the error autocorrelation, as one of the regression assumption is that the errors are independent (there is no correlation between errors). The interpretation of this test's results is made by comparing the calculated value (d) with two critical values from DW table (d_L and d_U), which lies between 0 and 4. The hypothesis of autocorrelation is rejected if $d_U < d < 4\text{-}d_U$. In our case for a significance level α = 0.05, 2 predictors and 41 observations, the critical values are: d_L=1.391 and d_U=1.600. As the calculated value presented in Table 3, d=1.881 is higher than d_U and lower than 4-d_U, we can reject the hypothesis of error autocorrelation.

Another assumption of the model is the absence of heteroscedasticity, which means that the errors have a constant variance. There is no direct method of identifying heteroscedasticity but some visual methods and empirical tests could be used. One of the visual methods is to plot the standardized residuals (ZRESID) on the standardized predicted values (ZPRED) in a scatterplot diagram (see Fig. 3). This diagram can be made easily with the SPSS just when we perform the regression model by pressing the “Plots” button.

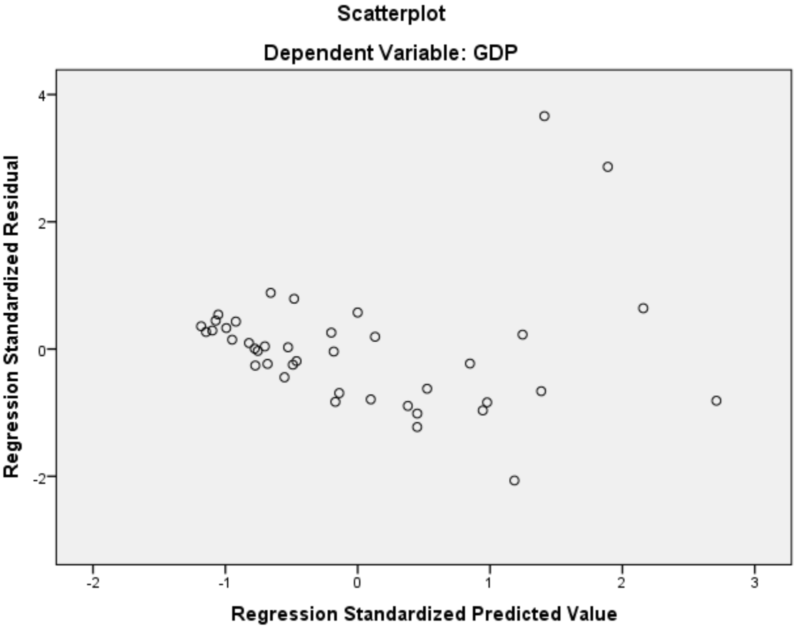

Fig. 3. *The relationship between standardized predicted values and residuals*

As the plot fans out, having some outlier values, we can conclude that there is a sign of heteroscedasticity. In such cases the estimators could be biased or inconsistent so that the results should be interpreted with prudence or new independent variables should be found.

3. Discussions and conclusions

As we presented in the above analysis, the Linear Regression Model is a powerful method used for the description of relationships between variables and for predictions but the results have to be interpreted with caution. First of all, it is very important to have a proper specification of the model taking into account the nature of the variables and of the relationships between these ones. In this respect, both the independent and dependent variables have to be measured with metric scales. Sometimes binary variable, named dummy variables, could be used as independent variables. As regards the relationships between variables, these ones have to be linear but a pure statistical dependence is not enough. The dependence relationship must be based on a theory if we want to use the model for predictions. After the model specification, some validation procedures are necessary in order to accept the model's hypotheses. If certain hypotheses are rejected, the model should be reconsidered or the results should be considered with maximum precaution.

In the above example, we can find that increasing the number of companies and labour resources will lead to higher levels of GDP. Thus new investments are necessary for the regional economic development. Based on different values of the independent variables, predicted values of the GDP could be calculated but we have to take care that the model is susceptible of heteroscedasticity. The results should be also cross validated using other samples from different years.

The results of this instrumental research could be useful both for practitioners and academic researchers in their efforts to build prediction models using the Linear Regression. The overall conclusion is that a superficial using of this model, without a proper validation, could lead to wrong conclusions and predictions. Consequently bad decisions could be made starting from inaccurate information.

6. References

Kutner, M., Nachtsheim, C., Neter, J. and Li, W., 2005. *Applied linear statistical models*, 5th ed. New York: McGraw-Hill Irwin.

Malhotra, N., 2004. *Marketing research. An applied orientation*, 4th ed. New Jersey: Pearson Education.

Montgomery, G., Peck, E. and Vining, G., 2006. *Introduction to Linear Regression Analysis, 4*th ed. New Jersey: John Wiley & Sons.

Rawlings, J., Pentula, S. and Dickey, D., 1998. *Applied regression analysis: a research tool.* 2nd ed. New York: Springer-Verlag.

Saunders, M., Lewis, Ph. and Thornhill, A., 2007. *Research methods for business students, 4*th ed. Harlow: Pearson Education.