Multiple linear regression used to analyse the corelation between GDP and some variables

Prof. Univ. Dr. Constantin Anghelache,
Drd. Cristina Sacala

Abstract

Cercetatorii in domeniul statisticii folosesc adesea o relație liniară pentru a prezice valoarea medie numerică Y pentru o valoare dată a lui X utilizând o linie dreaptă (numită linia de regresie). In cazul in care se cunoaste panta și valoarea interceptpentru acea linie de regresie, atunci avem posibilitatea să corelam o valoare pentru variabila independenta X și sa prezicem valoarea medie pentru variabila dependenta Y.

Dacă am stabilit cel puțin o corelație moderată între X și Y, atat printr-un coeficient de corelație, cat si prin dispersie, atunci consideram că exista un anumit tip de relație liniară.

Abstract

Statistical researchers often use a linear relationship to predict the (average) numerical value of Y for a given value of X using a straight line (called the regression line). If you know the slope and the y-intercept of that regression line, then you can plug in a value for X and predict the average value for Y. In other words, you predict (the average) Y from X. If you establish at least a moderate correlation between X and Y through both a correlation coefficient and a scatterplot, then you know they have some type of linear relationship.

Keywords: *regression, correlation, intercept, variables*

Introduction

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the criterion variable and is referred to as Y. The variable we are basing our predictions on is called the predictor variable and is referred to as X. When there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line

Literature review

Regression analysis is a statistical technique used to relate variables (Bowerman et al., 1990). Its basic aim is to build a mathematical model to relate dependent variables to independent variables. In general, a regression model will be defined as a single algebraic equation of the form (Draper and Smith, 1981).

There are three kinds of regression models. They are namely, a) the Variable-based Degree-Day Model (VBDD), b) the Linear Regression Model, and c) the Change-Point models. They all use generalized least squares regression to determine the model coefficients (Kissock, et al., 2003).

Methodology and data

The multiple linear regressions are used for modelling the relation between two or more explicative variables and the response variable by identifying a linear equation between the observed dates. For each value of the independent variable x it is associated a value of dependent variable y. The individual values of the registered explanatory variables within the linear regression $x_1, x_2, ..., x_p$ is defined as:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

This linear equation describes how μ_y changes whiles the explanatory variables are changing. The values observed for y alters depending of the changes of μ_y, and it supposes having the same standard deviation σ. The values associated to $b_0, b_1, ..., bp$ estimates the parameters $\beta_0, \beta_1, ..., \beta_p$ of the linear regression function.

Having in mind that the values observed for y oscillates around their average value μ_y, the linear regression model is expressed in the form:

DATE = SUITABLE VALUES + REZIDUAL VALUES

where "SUITABLE VALUES" means expression $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$ and the term "REZIDUAL VALUES" represents the deviation of the observed y values from their mean μ_y, which are normally distributed with mean zero, and variance σ. The deviation of the model is noted with ε.

The model of multiple linear regressions with n observation is formally defined by:

$$Y_i = \beta_0 + \beta_1 xi_1 + \beta_2 xi_2 + ... + \beta_p xi_p + \varepsilon_i \text{ , for } i = 1,2, ... n.$$

Within the least square model, the best fitting line for the adjustment of observed data is calculated by minimizing the sum of squares of vertical

deviations of each data point from to the line (in case of a data point is placed on the adjustment line, the deviation is considered zero). Due to the fact that the deviations are first squared, than summed, result that there are no annulation between positive and negative values. The least squares of linear regression functions' estimators b_0, b_1, ... b_p are usually calculated by using specific statistical software.

Within the performed analysis, we took into consideration a multiple linear regression model, where the dependent variable is the GDP, and the independent variables are those for we performed previously a simple regression modelling (inflation rate, yearly average gross income and yearly average net income), in our goal to identify an optimal model to show the influence of the most important variables on the GDP.

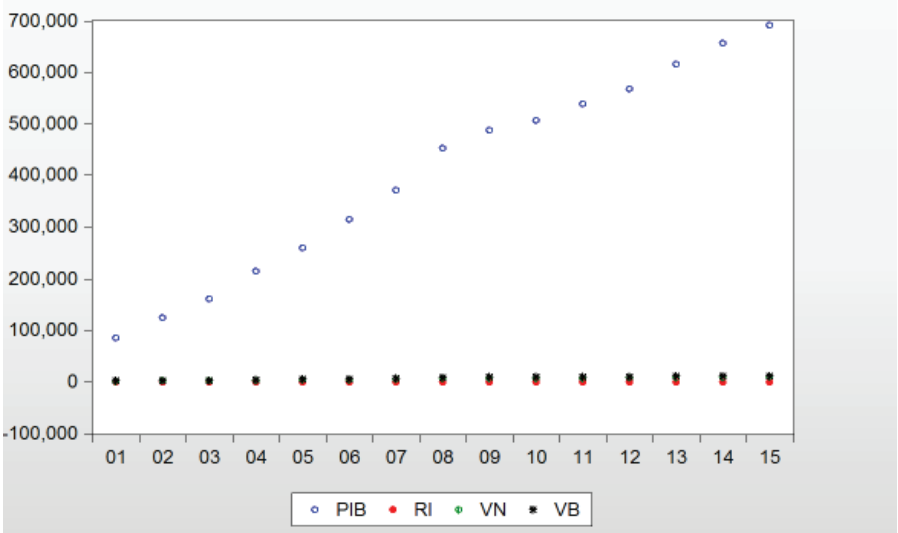
The GDP, inflation rate (IR), yearly average gross income (GI) and yearly average net income (NI) in Romania between 2001 – 2015.

Years	GDP	IR	GI	NI
2001	85820,2	34.5	2170,175	1584,8
2002	124461,5	22.5	2745,295	2018,8
2003	161061,7	15.3	3406,871	2483,2
2004	215374,8	11.9	4416,364	3158,7
2005	259125,1	9.1	4950,6	3613,93
2006	313889,3	6.6	5679,48	4491,6
2007	370821,8	4.9	6932,52	5368,8
2008	453638,3	7.9	8778,36	6738
2009	487331	5.6	9567,12	7518
2010	506446,5	6.1	9543,72	7540,8
2011	539520,4	5.8	10074,36	7609,2
2012	568719,4	3.4	10333,8	7983,6
2013	616393,3	3.2	10750,2	8166
2014	**656318,6**	1.4	11251,8	8707,2
2015	692617,4	-0.4	12128,04	9475,2

Source: National Institute of Statistics of Romania.

Based on the of the data series of the analysed indicators between 2001 – 2015 and the correlogram, we can notice an increase of the GDP, a significant decrease of the inflation rate, and moderate increase of the yearly average gross and net incomes.

Correlogram of GDP and the inflation rate, the yearly average gross income and yearly average net income



The multiple regression function identified within, consider the variation of the GDP as depending from the independent variables (inflation rate, yearly average gross income, and yearly average net income) is:

GDP = -26796.56 – 729.54 IR + 72.78 NI – 19.10 GI

Estimating the regression parameters

Dependent Variable: PIB
Method: Least Squares (Gauss-Newton / Marquardt steps)
Date: 09/13/16 Time: 15:56
Sample: 2001 2015
Included observations: 15
PIB=C(1)+C(2)*RI+C(3)*VB+C(4)*VN

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-26796.56	33103.50	-0.809478	0.4354
C(2)	-729.5425	1100.787	-0.662747	0.5211
C(3)	72.78884	37.09935	1.961998	0.0756
C(4)	-19.10033	46.61009	-0.409790	0.6898

R-squared	0.992040	Mean dependent var	403436.0
Adjusted R-squared	0.989869	S.D. dependent var	200296.3
S.E. of regression	20160.80	Akaike info criterion	22.88405
Sum squared resid	4.47E+09	Schwarz criterion	23.07286
Log likelihood	-167.6304	Hannan-Quinn criter.	22.88204
F-statistic	456.9471	Durbin-Watson stat	0.688971
Prob(F-statistic)	0.000000		

From the result obtained, we can notice a direct relation between the GDP and the yearly average gross income, meaning that increase of the VB determine the increase of the GDP, whilst in relation with the other two variables and the GDP there is a reverse relation.

The value of R-squared is a high one 99.20%, this means that the three independent variables causes in a high extend the variation of the GDP, with other words, the model is well determined; also the model is signisicant, having the Prob(F-statistic) very nearby to zero , the value of F test is 456.94.

Conclusions

One of the features is that the standard errors of the coefficients affected tend to be large. In this case, test the hypothesis that the coefficient is zero can lead to a failure to reject a false null hypothesis, a Type II error. Another problem is that small changes in inputs can lead to large changes in model, even as a result of changes in the sign parameter estimation. Such data redundancy main danger is that of overfitting in regression analysis models. The best regression models predictor variables are highly correlated with each dependent variable (result), but the least correlated with each other. Such a model is often called „noise" and will be statistically robust. In future research we will focus on testing the multicollinearity of the proposed model between variables.

Bibliography:

1. Ang, A., Piazzesi, M., Wei, M. (2006). *What does the yield curve tell us about GDP growth?*, Journal of Econometrics, Volume 131, Issues 1–2, March–April 2006, Pages 359–403
2. Anghelache, C., Anghel, M.G. (2015). *GDP Analysis Methods through the Use of Statistical – Econometric Models*, „ECONOMICA" Scientific and didactic journal, nr. 1 (91), Chişinău, Republica Moldova, pp. 124-130, ISSN: 1810-9136, (print) / ISSN 1844-0029 (online),
3. Anghelache, C., Manole, A., Anghel, M.G. (2015). *Analysis of final consumption and gross investment influence on GDP – multiple linear regression model*, Theoretical and Applied Economics, No. 3/2015 (604), Autumn, pg 137-142, ISSN 1841-8678
4. Anghelache, C., Soare, D.V., Popovici, M. (2015). *Analysis of Gross Domestic Product Evolution under the Influence of the Final Consumption*, Theoretical and Applied Economics, Volume XXII, No.4 (605), Winter, pp. 45-52, ISSN 1841-8678
5. Anghelache, C., Manole, A., Anghel M.G. (2015). *Analysis of final consumption and gross investment influence on GDP – multiple linear regression model*, Theoretical and Applied Economics, No. 3/2015 (604), Autumn, Pages: 137-142
6. Anghelache, C., Anghel, M.G. (2015). *GDP Analysis Methods through the Use of Statistical – Econometric Models,* „ECONOMICA" Scientific and didactic journal, nr. 1 (91), Chişinău, Republica Moldova, Pages: 124-130
7. Anghelache, C., Anghel, M.G. (2015). *Model of Analysis of the Dynamics of the DFI (DFI) Sold Correlated with the Evolution of the GDP at European Level*, Romanian Statistical Review Supplement, No. 10, Pages: 79-85
8. Benjamin, C., Herrard, N., Houée-Bigot, M., Tavéra, C. (2012). *Forecasting with an Econometric Model*, Springer, ISBN 978-3-642-11648-3
9. Capelli, C., Vaggi, G. (2013). *A better indicator of standards of living: The Gross National Disposable Income,* University of Pavia, Department of Economics and Management in DEM Working Papers Series with number 062.
10. Dumitrescu, D., Anghel, M.G, Anghelache, C. (2015). *Analysis Model of GDP Dependence on the Structural Variables*, Theoretical and Applied Economics, Volume XXII, No.4 (605), Winter, Pages: 151-158
11. De Michelis, N., Monfort, P. (2008). *Some reflections concerning GDP, regional convergence and European cohesion policy*, Regional Science Policy & Practice, Volume (Year): 1 (2008), Issue (Month): 1 (November), Pages: 15-22
12. Hulten, Ch., Schreyer, P. (2010). *GDP, Technical Change, and the Measurement of Net Income: the Weitzman Model Revisited*, NBER Working Paper No. 16010 Issued in May 2010
13. Jones, Ch., Klenow, P. (2010). *Beyond GDP? Welfare across Countries and Time*, NBER Working Paper No. 16352
14. Macdonald, R. (2010). *Real Gross Domestic Income, Relative Prices and Economic Performance Across the OECD*, Statistics Canada, Analytical Studies Branch in Economic Analysis (EA) Research Paper Series with number 2010059e.
15. Ramcharan, R. (2007). *Does the Exchange Rate Regime Matter for Real Shocks? Evidence from Windstorms and Earthquakes*, Journal of International Economics, Volume 73, No. 1, Pages: 31–47
16. Vintrová, R. (2005). *What the GDP Indicator Does Not Reveal in Economic Analyses (in English)*, Charles University Prague, Faculty of Social Sciences in its journal Finance a uver - Czech Journal of Economics and Finance, Volume (Year): 55 (2005), Issue (Month): 11-12 (November), pp. 578-594