"Dimitrie Cantemir" Christian University
Knowledge Horizons - Economics
Volume 11, No. 2, pp. 44 – 51
P-ISSN: 2069-0932, E-ISSN: 2066-1061
© 2019 Pro Universitaria
www.orizonturi.ucdc.ro

THE CLASSIC REGRESSION MODEL (USING TESTS T AND F)

Liviu Constantin STOICA

PhD, Academy of Economic Studies, Bucharest, Romania, Email: stoica.liviu.constantin@gmail.com

Abstract *In this article entitled "The classic Regression Model," the study was conducted in three parts. In the first part we introduced an introduction to the regression, after which in the second part I studied linear linear regression. In this part we studied the slope and the least squares method with which we made the regression equations. In the last part we made a demonstration application on regression, an application that was made with SPSS.*

Key words:
links, slope, regression, statistics, variables
JEL Codes:
M1

1. INTRODUCTION

Social and economic phenomena in general do not evolve independently; they are related to other economic variables. This gives the possibility that, using the knowledge of the level of a certain variable, it can be predicted the level of another variable with which it is in a certain dependence. For example, the volume of sales of goods and services can be predicted on the per capita income of the population.

The regression analysis studies the dependence between a resolvable variable (y) and one or more independent variables (x). The resultative feature is also called the dependent, endogenous or effect characteristic, and the independent feature is also called the factorial, exogenous or cause characteristic.

Regression shows how a variable is dependent on another variable (or other variables). In the study of economic dependencies we must take into account that socio-economic phenomena are complex phenomena, under the influence of many factors, some essential, others happening, with different action and intensity, in a different direction.

Statistical links can be classified according to several criteria, namely:

a) After the number of independent characteristics taken into consideration:
   - Simple connections when studying the dependence between a dependent characteristic (y) and an independent characteristic (x), y = f (x). Example: Dependency between volume of goods (y) and sales (x).

- Multiple links when studying the dependency between a dependent feature (y) and two or more independent features. So $y = f(x_1, x_2, ..., x_n)$. For example, the dependence between the volume of goods (x_1), the number of sales and the commercial area (x_2).

b) After the links are directed, they can be:
- direct links when the dependent characteristic changes in the same way as the independent feature: if x increases, y increases; if x decreases, y decreases.
- Reverse links when the dependent feature changes in the opposite direction to changing the independent feature. If x increases, y decreases; if x decreases, y increases.

c) After the analytical expression of the links, they can be:
- Linear links - those dependencies that can be expressed using the linear function.
- Non-linear (curvilinear) links - those dependencies that can be expressed using non-linear functions (parabola, hyperbolic, exponential function, etc.).

d) According to the way in which the variables are linked in time, we encounter:
- Synchronous (concomitant) links, in which the modification of a variable occurs approximately with the change of the other / others;
- Asynchronous links, in which the effect manifests after a period of time from the manifestation of the cause.

2. SIMPLE LINE REGRESION

The simple regression method involves explaining a resolving variable Y based on a factorial variable, using a model named regression function. In such a model, the notion of causality is used, ie the changes in the independent variable determine changes in the dependent variable. Exact relationships between phenomena and socio-economic processes can not be described, no matter how many factorial features we consider, due to random phenomena that can not be modeled or explained. These unexplained variations in the model, caused by non-included variables, form the random (residual) error.

We will consider a linear function to express the link between the two variables:

$y_i = a + bx_i + e_i$, with the predictable component:

$$\hat{y}_i = a + bx_i$$

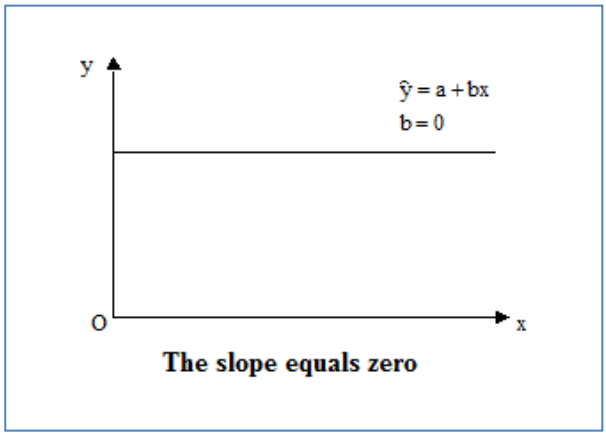
where a and b are the coefficients of the regression function, so e_i the residual value for the unit i.

$e_i = y_i - (a + bx_i)$, that is, the measure of distance from a point (x_i, y_i) to the regression line.

The parameter "a" expresses the value of y when, therefore, is the intersection of the straight with the OY axis and bears the name of intercept.

The parameter "b" is called a regression or slope coefficient and is of great importance in regression analysis.

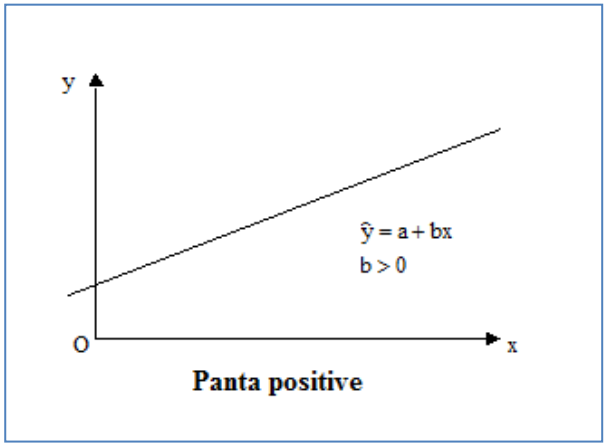
If b = 0, the variable y does not depend on x, so the two variables are independent.



The slope equals zero

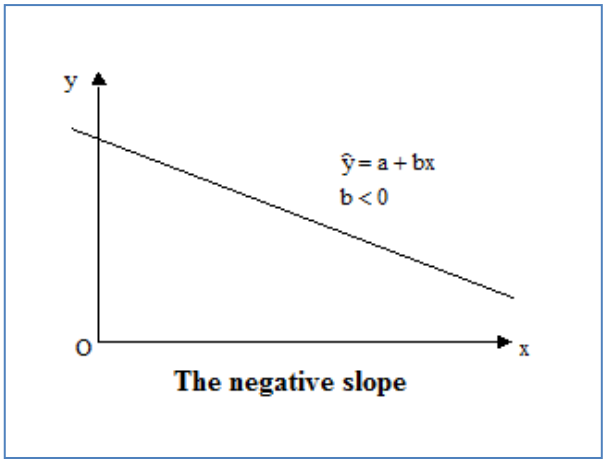
If, the two variables are dependent, namely:

- if b> 0 the link is direct;



Panta positive

- If b <0 the link is inverse.



The negative slope

The magnitude of the coefficient b, the slope with the geometric sense, shows how much y changes when the variable x changes with a unit.

Estimation of parameters a and b is typically accomplished using the least squares method (MCMMP), ie minimizing the sum of squares of e_i deviations based on (x_i, y_i) values observed in a n-volume sample. The lower the e_i value with the closer the $\hat{y}_i$ is to the observed y_i value.

MCMMP assumes the following assumptions:

1) the values x_i and y_i were obtained without observation or measurement errors;

2) the variables are independent of each other;

3) the perturbation variable follows a normal distribution law.

We will have:

$$\sum e_i^2 = \sum (y_i - \hat{y})^2$$

$$\sum e_i^2 = \sum (y_i - \hat{a} - \hat{b}x)^2$$

Estimates of $\hat{a}$ and $\hat{b}$ are obtained on the basis of the first order partial derivatives of the function:

$$\begin{cases} \frac{\delta f}{\delta a} = 2\sum\left(y_j - \hat{a} - \hat{b}x\right)(-1) = 0 \\ \frac{\delta f}{\delta b} = 2\sum\left(y_j - \hat{a} - \hat{b}x\right)(-x) = 0 \end{cases}$$

This leads to the following system of simultaneous normal equations:

$$\begin{cases} n\hat{a} + \hat{b}\sum x = \sum y \\ \hat{a}\sum x + \hat{b}\sum x^2 = \sum xy \end{cases}$$

By solving the system by Cramer we get:

$$\hat{a} = \frac{\begin{vmatrix} \sum y & \sum x \\ \sum xy & \sum x^2 \end{vmatrix}}{\begin{vmatrix} n & \sum x \\ \sum x & \sum x^2 \end{vmatrix}} = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - \left(\sum x\right)^2}$$

$$\hat{b} = \frac{\begin{vmatrix} n & \sum y \\ \sum x & \sum xy \end{vmatrix}}{\begin{vmatrix} n & \sum x \\ \sum x & \sum x^2 \end{vmatrix}} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

Based on the $\hat{a}$ and $\hat{b}$ coefficients, the regression equation for each magnitude of the x characteristic is determined, which are also called the theoretical values of the y characteristic against x. The operation of replacing real terms with values of regression equations is called adjustment. Thus, $\sum y_i = \sum \hat{y}_i$ will be obtained.

To characterize the intensity of the relationship between variables y and x, we use the correlation, which shows the degree to which a variable is dependent on another variable (or other variables).

3. DEMONSTRATIVE APPLICATION REGRESION

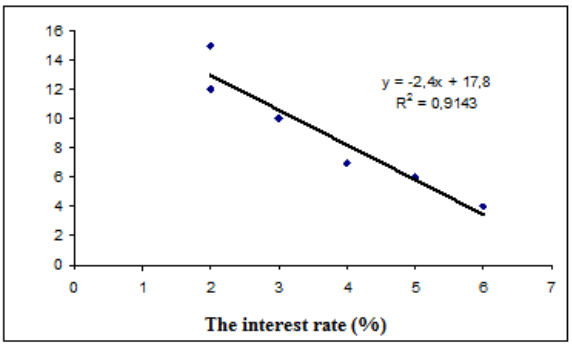
Suppose we have information about the investment and the interest rate.

The investment million	The interest rate
10	3
7	4
6	5
12	2
4	6
15	2

Table 1. Investment data and interest rate

We want to express, through an econometric model using the data in Table 1, the link between the investment and the interest rate; then, once the validated model, we can use it to make predictions at a macroeconomic level.

We only have one independent variable. It is useful to first make an XY graphical representation. The result with the Excel application is shown in Figure 1.



All clues are in the direction of using a classical regression model (the dependence seems linear, errors seem to have constant dispersion, the free term seems to be different from zero). Using a classic software, such as the Regression module in Excel, we get the following results:

A.Bonitatea modelului

SUMMARY OUTPUT

Regression Statistics		
Multiple R	R	0,956183
R Square	R^2	0,914286
Adjusted R Square	$\bar{R}^2$	0,892857
Standard Error	$s_{\hat{a}}$	1,341641
Observations	n	6

Figure 1 - Statistical regression (Summary Output)

ANOVA

	df	SS		MS		Fcalc		Significance F
Regression	$k=1$	76,8	$SSR=\sum_i(\hat{y}_i-\bar{y})^2$	76,8	$MSR=\frac{SSR}{k}$	42,66667	$MSR=\frac{SSR}{k}:\frac{n-k-1}{SSE}$	0,002838
Residual	$n-k-1=4$	7,2	$SSE=\sum_i(y_i-\hat{y}_i)^2$	1,8	$MSE=\frac{SSE}{n-k-1}$			
Total	$n-1=5$	84	$SST=\sum_i(y_i-\bar{y})^2$					

Figure 2 - The Anova test
where k = the number of exogenous variables in the model

	Coefficients		Standard Error		t Stat		P-value	Lower 95%		Upper 95%	
Intercept	17,8	$\hat{a}$	1,454304	$s_{\hat{a}}$	12,23953	$t^1_{calc}=\frac{\hat{a}}{s_{\hat{a}}}$	0,000256	13,7622	$\hat{a}-t_{\alpha/2,n-1}\cdot s_{\hat{a}}$	21,8378	$\hat{a}+t_{\alpha/2,n-1}\cdot s_{\hat{a}}$
Rata dobanzii %	-2,4	$\hat{b}$	0,367423	$s_{\hat{b}}$	-6,53197	$t^2_{calc}=\frac{\hat{b}}{s_{\hat{b}}}$	0,002838	-3,42013	$\hat{b}-t_{\alpha/2,n-1}\cdot s_{\hat{b}}$	-1,37987	$\hat{b}+t_{\alpha/2,n-1}\cdot s_{\hat{b}}$

Figure 3 - Parameter estimation

Let's look at the results in the boxes shown in Table 2 one at a time.

Box A provides information about the indicators that reveal the model's robustness or how good the model we are analyzing.

These indicators are: the correlation ratio (Multiple R), the determination coefficient (R Square), and the adjusted $\bar{R}^2$ (Adjusted R Square) determination coefficient. The higher R^2 and $\bar{R}^2$ values are closer to 1, the better the regression.
Interpreting results in the SUMMARY OUTPUT table:

- R = 0.956183 shows that there is a strong link between the level of investments and the interest rate level.
- R^2 = 0.914286 shows that 91.42% of the investment variation is explained by the interest rate, so the interest rate is a determinant factor included in the model.
- Average error deviation of error s_u = 1.341641. If this indicator is zero it means that all points are on the right of regression.

For our application, as all listed rating ranges are close to 1, we can conclude that the linear linear regression model is good.

Box B refers to the decomposition of the total variance (SST) of the dependent variable into two components: regression variant (SSR) and unexpressed variance (SSU) or residual variance. Here we also identify degrees of freedom associated with decomposition, we have k regressions in the model and in the observations, we have the equality. There are two important cells in this box that we need to be aware of: F and Significance F. The values in these cells give us important elements underlying the validation of the regression model (in its totality).

They provide us with information about the calculated value of the test statistic F and the error that we can do when we reject the regression model as inappropriate. The decision rule for accepting the model is: high values for test statistic F and low values for Significance F. Significance F is the value of the error we make by rejecting the null hypothesis when it is actually true.

On our data, as F = 42,66667 is a high value and Significance F = 0,002838, so a very small value, we accept that the chosen pattern adjusts well the data in the sample.Significance F must generally be less than 5% (0.05).

Box C gives us information about the coefficients of the regression model in the Coefficients column, standard coefficient errors in the Standard Error column, elements for applying the t-Student significance test for each coefficient (t Stat columns and P-value). So, the value in the t-Stat column is obtained by dividing the estimated value for the estimator's standard error for each model estimator and thus we obtain the calculated value of the t test for each model estimator. This calculated value is compared to the table value taken from the distribution table Student.

- Intercept is the free term, so the coefficient $\hat{a}$ is 17.8. The free term is the point where the independent variable (exogene) is 0. So the investment level is 17.8 if the interest rate is 0. Because t_{a_0} = 12,2395 and the P-

value significance threshold is 0.000256 < 0.05 means that this coefficient is significant. The confidence interval for this parameter is 13,7622 $\leq \alpha \leq$ 21,8378

- The coefficient $\hat{b}$ is 2.4, which means that the increase of the interest rate by one percentage point, the level of investments will increase by 2.4 Because t_{a1} = -6,53197 and the significance threshold P-value is 0.002838 < 0.05 means that this coefficient is significant. The confidence interval for this parameter is -3,42013 $\leq \beta \leq$ -1,37987.

About p-value

Before starting a classic statistical test, there is a question of choosing a level of significance. It expresses the maximum risk of mistakes that we are willing to accept (typically 5%, 1% or even lower) when we make the decision to reject the null hypothesis.

Modern software offers the "inverse" option. Namely, the risk of making the wrong decision based on the data we have is assessed, and it is up to each of them to take or not to take the risk. This risk based on the data appears in the tables at each significance test, and is called p (p-value).

Here, too, we have information about the confidence intervals calculated for each parameter in the regression model.

$$\hat{a} - t_{\alpha/2,n-2} \cdot \sigma_{\hat{a}} \leq a \leq \hat{a} + t_{\alpha/2,n-2} \cdot \sigma_{\hat{a}}$$

$$\hat{b} - t_{\alpha/2,n-2} \cdot \sigma_{\hat{b}} \leq b \leq \hat{b} + t_{\alpha/2,n-2} \cdot \sigma_{\hat{b}}$$

How do we analyze the information in this box?

1. For a coefficient to be significantly different from zero, so the variable regressor (var. Dependent) associated with it influences the dependent variable, we must have small values in the P-value, for example 5% or below 5% in the "t" column we have then high values in the mode). Specifically, for the free term of the model (Intercept) we have P-value = 0.043, ie we can say that if we reject the assumption that the intercept is zero, we make only a 4% error. We hereby reject this statement and accept the true hypothesis that the intercept is different from zero. (Analogously, we come to the conclusion that the slope of the regression line is statistically different from zero).
2. The last two columns give us information on 95% confidence intervals for each model coefficient. Thus, for the free (theoretical) term of the model we obtain the interval (-132.0474401, -2.5126). Analogously, for the slope of the regression equation we have the confidence interval (0.905614, 1.052304). It is very important that none of these confidence intervals contain 0, so we are encouraged to say that the model is good.

4. CONCLUSIONS

In this article, following the study of regression and implementation of the demonstration application, SPSS investment and interest rate application, we found that there is only one independent variable. At the same time, we also applied the Anova test to complete the deviation of the total variance and to apply the F test. Following the application of this model we found that if the number of exogenous variables in the model is 1 then the value of F is 0.002838. In addition to this model, we also made estimates for model parameters, standard errors, t test, and confidence intervals. As a result of this analysis we found that for the aforementioned value (F is 0.002838) on the interest rate, the coefficient is -2.4.

5. BIBLIOGRAPHY

- A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in Proc. Interspeech, 2012, pp. 22–25
- Dollar, P., Welinder, P., & Perona, P. (2010). Cascaded pose regression. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- Elad, M., & Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. IEEE Transactions on Image Processing, 15 (12), 3736–3745.
- Xiong, X., De la Torre, F. (2013) Supervised descent method and its applications to face alignment. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- X.G.Lu,Y.Tsao,S.Matsuda,andC.Hori,"Spe echenhancement based on deep denoising Auto-Encoder," in Proc. Interspeech , 2013, pp. 436–440
- Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, Fellow, IEEE, "A Regression Approach to Speech Enhancement Basedon Deep Neural Networks", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 23, NO. 1, JANUARY 2015
- Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," IEEE/ACM Trans. Speech Audio Process.,vol. 22, no. 12, pp. 1849–1858, Dec. 2014