

THE PREDICTION OF ECONOMIC EVOLUTION THROUGH REGRESSION AND EXTRAPOLATION

Lecturer Sion Beatrice, Ph.D
Professor Cezar Mihălcescu, PhD
Assistant Alexandra Mărginean, Ph.D¹

ABSTRACT

Predictive methods aim at the anticipation of future decisions based on the already existing information, in order to envisage certain scenarios.

Regression is a method of statistical analysis through which, once certain values are assigned to a dependant variable y and to one or more independent variables x^j ($j=1, \dots, m$), a simple expression of the function that expresses the connection between them is sought.

Key words: prediction, linear regression, square regression, cubic regression

1. INTRODUCTION

In technical terms, the connection between two or more variables is called a **correlation**, and establishing the type of connection that a **dependent** variable has with one or more **independent** variables is made through **regression analysis**.

Regression can be of three types: linear regression, square regression and cubic regression. Linear regression through the method of the least squares is the most widespread. It is the method called “regression”, “linear regression”, “multiple regression” or the “least squares” when a model is built.

The purpose of multiple regression (a term used by Pearson, 1908) is to highlight the relation between a dependent variable (explained, endogenous or resultant variables) and a lot of independent variables (explanatory, factor, exogenous, predictor ones). In the use of multiple regression there is often an attempt to answer the following questions: “what is the best prediction for...?”, “what is the most efficient predictor for...?”.

2. REGRESSION METHODS

The method of multiple regression is generalized through the theory of the “general linear model”, in which there are more dependent variables simultaneously, as well as variables that are not independent from a linear point of view.

The main stages in the construction of the regression model are:

1. **Identification** – the descriptive stage in which dependent variables are identified, along with the type of relations that they express
2. **Specification** – the stage in which the best expression of the variables is sought

¹ School of Domestic and International Economy of Tourism, Romanian-American University, beatrice_sion@yahoo.com, cezar_mihalcescu@hotmail.com, marginean.alexandra.roxana@profesor.rau.ro

3. **Estimation** of parameters for the model that is to be examined
4. **Testing** of the significance of parameters that get estimated in the model
5. **Validation** of the model under analysis
6. **Use** of the model in simulation and prediction operations.

In order to actually accomplish the predictions made on the basis of the regression analysis, one needs to go through the following stages:

- **Formulation of the problem** – the decision-making party (the manager) must define the problem that needs to be analyzed, in terms of the variables that have to be explained and the values of which are to be predicted. In this incipient formulation, the decisional problem is described and we identify the variables for which predictions will be made, along with the variables that are dependent on them.
- **Choosing the economic indicators** – after the identification of independent variables, we search for other adjacent factors that are liable to influence the respective variable and that can be included in the regression equation.
- **The first analysis of the regression equation** actually represents a statistical analysis of the equation components and is performed automatically, one of the results being to establish the correlation matrix with the help of the Microsoft Excel 2014 tools.
- **The analysis of the simple correlation matrix** is performed to the purpose of choosing the variables that need to be part of the regression equation. In this analysis, we must identify those variables that are strongly correlated with the dependent variable, but superficially correlated among themselves. At the end of this stage, 3 or 4 regression equations are taken into account for the analysis.
- **The choice of a regression equation from the ones that we have identified** – based on the available data, the computer will determine not only the regression coefficients, but also the elements that allow the test of their significance. Only the significant equations will be kept.
- **The check of the regression conditions validity.**
- **Preparing the prediction** – once we have chosen a regression equation whose correlation coefficient value is big enough and which is adequate from the perspective of the significance test, the decision-making party (the manager) can use this equation as a basis for analysis for the prediction that he intends to make. At this point, a trust interval for individual predictions needs to be established, as well as the precision of the value of every independent variable.

The main advantage of regression analysis consists in the fact that it is a statistical method, which presupposes an estimation of the degree of precision and significance, and it can be used in all types of causal relations on condition that the variable that we have in view depends on independent variables.

The fact that it is a statistical method is also its **main disadvantage**. That is why many decision-makers avoid this type of analysis. Another downside is the big volume of data and the costs for data collection in order to establish the initial regression equation and for the analysis of its validity in time (if a modification in the causal relation between an

independent variable and the dependent one appears, the collection of new data is necessary, along with the redefinition of the regression equation).

EXCEL offers more statistical functions to be used in the making of predictions within linear regression: FORECAST, TREND, LINEST. There are also not only functions for other types of regression (for example, the LOGEST and GROWTH functions for exponential regression), but also the possibility to make predictions through graphs (by attaching tendency curves to graphic representations).

Regression is a method of statistical analysis through which, by the assignation of determined values to a dependent variable y and to one or more independent variables x^j ($j=1, \dots, m$), a simple expression of the function that expresses the connection between them is sought.

- By **extrapolation**, based on this expression and on other values of independent variables, the corresponding values for the dependent variable are sought. If statistical data expresses an evolution in time, then the calculation of future values of the dependent variable y is called **prognosis/prediction**, and that of past values – **retrognosis**.
- In regression analysis we consider that the dependent variable y is a function of one or more independent variables of the type: $y = f(x_1, x_2, \dots, x_n) + \mathcal{E}$,
- $\mathcal{E}$ represents the deviation of the theoretical variable from the experimental one.
- In regression analysis, we aim to determine the function f , so that $\mathcal{E}$ should be a variable with a null average and minimum dispersion.
- The easiest model to follow is simple regression: $y = ax + b$.
- If this model is not satisfactory, we can use the multilinear regression of the type: $y = a_1x + a_2x^2 + a_3x^3 + \dots + a_mx^m + b$, where
- a_j ($j=1, \dots, m$) and b need to be determined.
- The description of the relation that could exist between the two variables (x and y) analyzes if the ascending tendency of one entails an ascending/descending tendency of the other variable.
- If $m=1$ – **linear dependence**, and the regression function is represented by a line whose angle is given by a_j , b gives the ordinate of the intersection point of the line with the oy axis.
- Various other dependence functions can be reduced to the multilinear function through various transformations. For regression, the most widely used method is that of the least squares, through which the minimization of the sum of the squares of distances between given values for y and those calculated through the regression function is sought.
- The final aim is prediction, under the condition that the analysis is possible, the two variables being indeed correlated; the correlation method is noted with R , which takes values between 1 and -1. The closer the value is to 1 or -1, the more relevant the regression. If $R=0$, it means that the two variables are unassociated and do not influence each other.

3. CASE STUDY

In this case study, we are interested in the evolution of sales of towel manufacturing companies, according to the values of hotel construction in a country (Romania) and to other construction contracts that have been made so far.

The data known from the hypothesis:

- The values of the amount of towel sales between 1997-2013
- The values of the hotel constructions between 1997-2013
- The values of the construction contracts closed between 1997-2013
- The probable values of the hotel constructions in the next 5 years 2014-2018
- The probable values of the construction contracts in the next 5 years 2014-2018.

What is needed: to calculate the evolution of towel sales between 2014-2018.

Independent variables: x_1 -values of the towel manufacturing (already made and probable in the next 5 years)

The regression function cannot be found in ribbon, it can be installed from the menu File-Options- Add Ins-Analysis ToolPak. As a result of the installation of the analysis tool, the Regression function is installed in the menu Data-Data Analysis.

The main page, where the data from the hypothesis has already been installed, has been called **ENUNCIATION**.

	A	B	C	D	E
1	Evoluția vânzărilor de prosoape				
2					
3	anul	nr. crt.	valoarea productiei prosoape, x1	contracte constructii hoteluri, x2	vanzari prosoape, y
4	1997	1	3909	9430	280
5	1998	2	5119	10360	281.5
6	1999	3	6666	14500	337.4
7	2000	4	5338	15750	404.2
8	2001	5	4321	16780	402.1
9	2002	6	6117	17440	452
10	2003	7	5559	19770	431.7
11	2004	8	7920	23760	582.3
12	2005	9	5816	31610	596.6
13	2006	10	6113	32170	620.8
14	2007	11	4258	35090	513.6
15	2008	12	5591	36042	606.9
16	2009	13	6675	36580	629
17	2010	14	5543	37140	602.7
18	2011	15	6933	43100	656.7
19	2012	16	7638	45620	778.5
20	2013	17	7752	47380	877.6
21	2014	18	6400	48510	
22	2015	19	7900	51230	
23	2016	20	8400	57470	
24	2017	21	8600	61030	
25	2018	22	8900	66250	

Figure 1: The enunciation of the case study: the evolution of towel sales

Source: Microsoft Office- Excel

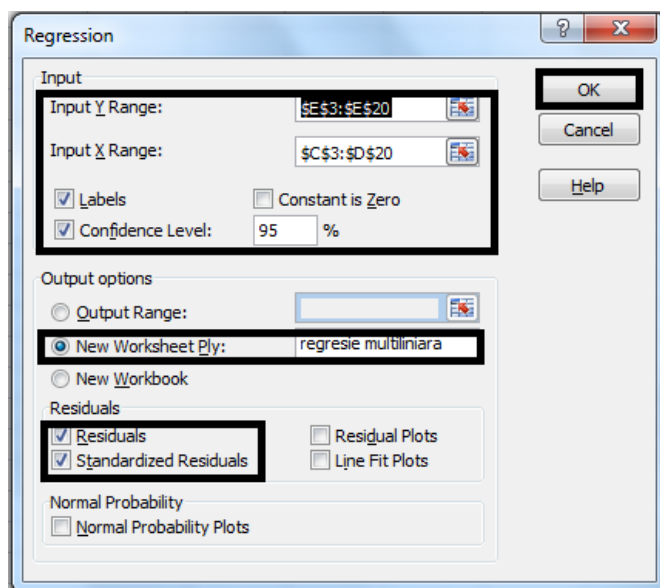


Figure 2: The use of the regression function

Source: Microsoft Office- Excel

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2	Regression Statistics								
3	Multiple R	0.970107947							
4	R Square	0.941109429							
5	Adjusted R Square	0.932696491							
6	Standard Error	43.23999635							
7	Observations	17							
8	ANOVA								
9		df	SS	MS	F	Significance F			
10	Regression	2	418305.6168	209152.8084	111.86453	2.45652E-09			
11	Residual	14	26175.76198	1869.697284					
12	Total	16	444481.3788						
13									
14									
15		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
16	Intercept	22.74934593	54.70543435	0.415851664	0.68382262	-94.58214106	140.080833	-94.5821411	140.080833
17	valoarea productiei prosoape, x1	0.035561092	0.010596264	3.356002833	0.00470741	0.012834366	0.05828782	0.012834366	0.05828782
18	contracte constructii hoteluri, x2	0.010720475	0.001016295	10.54859015	4.8018E-08	0.00854074	0.01290021	0.00854074	0.01290021
19									
20	RESIDUAL OUTPUT								
21	Observation	Predicted vanzari prosoape, y	Residuals	Standard Residuals					
22	1	262.8517366	17.14826338	0.423965454					
23	2	315.8506997	-34.35069974	-0.849270254					
24	3	415.2464768	-77.84647685	-1.924639023					
25	4	381.4219413	22.7780587	0.563153818					
26	5	356.2984007	45.80159932	1.132376813					
27	6	427.2416352	24.7583648	0.612113958					
28	7	432.3772537	-0.677253719	-0.016744097					

Figure 3: The Excel-generated report as a result of regression analysis

Source: Microsoft Office- Excel

In order to interpret, economically and statistically, the results generated by the regression function of the Excel and their relevance, the main instruments of the analysis and their explanation are presented in order to understand the significance of these results.

Significance F – displays the critical probability of the test, so that if the column $Sig. < \alpha$, the hypothesis of lack of significance of independent variables is rejected, in favor

of the hypothesis that the regression model is a salient one. We also say that the test is one of significance over **R² (Adjusted R Square)**.

Standard Error – the standard error of the coefficient (the standard deviation in the coefficient random distribution)

R-Square (R²) – explains x% of the variation of x in relation to y

t Stat – the statistics of the coefficient significance test. If the given values are on the regression curve, then $R^2=1$, meaning a perfect correlation. The closer R^2 is to 1, the more adequate the regression equation.

The F test indicates if the whole regression equation is significant. The calculated F value is F_c . The theoretical value of the inverse of the statistical distribution of F, noted as F_t , takes determined values that are found in statistical tables from specialized textbooks. If $F_c > F_t$, then the regression equation is significant (namely, it checks the model validity).

P-value – the probability for the hypothesis that the estimated parameter should be equal to 0; if P-value is smaller than the significance threshold, then we reject this hypothesis.

Multilinear regression of the type: $y = a_1x + a_2x^2 + a_3x^3 + \dots + a_mx^m + b$, where a_j ($j=1, \dots, m$) and b need to be determined.

=B\$17*enunt!C4+'summary output1'!\$B\$18*enunt!D4+'summary output1'!\$B\$16

anul	vanzari prosoape, y	predictia
1997	280	262.8517
1998	281.5	315.8507
1999	337.4	415.2465
2000	404.2	381.4219
2001	402.1	356.2984
2002	452	427.2416
2003	431.7	432.3773
2004	582.3	559.1117
2005	596.6	568.4469
2006	620.8	585.0120
2007	513.6	550.3500
2008	606.9	607.9588
2009	629	652.2746
2010	602.7	618.0229
2011	656.7	731.3469
2012	778.5	783.4331
2013	877.6	806.3551
2014		770.3906
2015		852.8919
2016		937.5682
2017		982.8453
2018		1,049.4746

Figure 4: Prediction calculation

Source: Microsoft Office- Excel

As a result of making the prediction, whose equation is $y=35.56x+10.72x^2+22.749$, the graph from figure 5 was drawn, in which we notice a parallel between towel sales and the Excel-generated prediction; as we can see from the graph, the prediction is quite faithful to reality, and the prediction for the next 5 years is optimistic, as it predicts growth from one year to another.

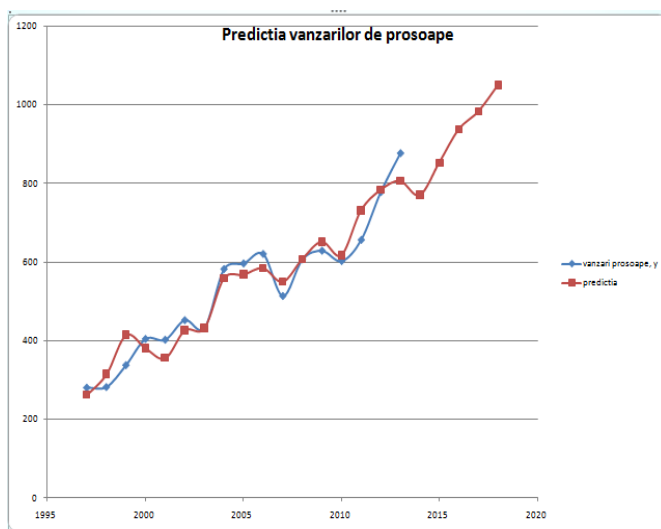


Figure 5: Graph with towel sales prediction

Source: Microsoft Office- Excel

a) **Linear regression:** $y=ax+b$, a = interceptor, b =the regression coefficient/
the angle of the regression line

	A	B	C	D	E
1	Evolutia costurilor unei agentii de turism				
	nr. crt.	valoarea productiei prosoape, x1	$x2=(x1^2)$	$x3=(x1^3)$	cost prod, y
2					
3	1	390	152100	59319000	233000
4	2	511	261121	133432831	333000
5	3	666	443556	295408296	456009
6	4	533	284089	151419437	345890
7	5	432	186624	80621568	567789
8	6	611	373321	228099131	234234
9	7	555	308025	170953875	123890
10	8	792	627264	496793088	678890
11	9	581	337561	196122941	789890
12	10	611	373321	228099131	890678
13	11	425	180625	76765625	678567
14	12	559	312481	174676879	456345
15	13	667	444889	296740963	345577
16	14	554	306916	170031464	797867
17	15	693	480249	332812557	565656
18	16	763	582169	444194947	677890
19	17	775	600625	465484375	789678
20	18	640	409600	262144000	567567
21	19	790	624100	493039000	456456
22	20	840	705600	592704000	456456
23	21	860	739600	636056000	678678
24	22	890	792100	704969000	567678

Figure 6: Calculation of the square and cubic regression

Source: Microsoft Office- Excel

From the linear regression summary output, the result is the function $y=379x+287774$. R is almost 0, so the implication is that the correlation between coefficients is almost inexistent. Thus, we will renounce to calculate the prediction, because it will bear no relevance, and the square regression will be resorted to instead.

Figure 7: Generation of results for linear regression

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.265							
R Square	0.070							
Adjusted R Square	0.024							
Standard Error	204482.028							
Observations	22							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	63306815415	63306815415	1.51	0.23279859			
Residual	20	8.36258E+11	41812899620					
Total	21	8.99565E+11						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	287774	202769	1	0	-135195	710743	-135195	710743
valoarea productiei prosoape, x1	379	308	1	0	-264	1022	-264	1022

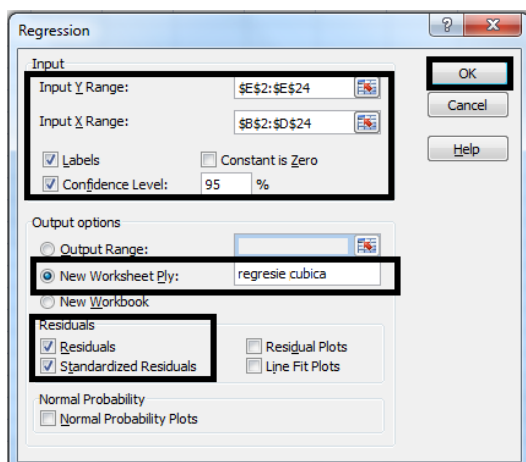
Source: Microsoft Office- Excel

b) Square regression: it has the syntax $y=a_1x+a_2x^2+b$, which will be used to calculate the prediction in case R Square has a relevant value, namely a value that is as close to 1 as possible; since the value is very close to 0, the equation is not relevant and is not faithful to reality; thus, we will renounce the calculation of the square regression and we will try the cubic regression.

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2	Regression Statistics			$y=805*x+0*x^2+156960$					
3	Multiple R	0.27		$y=805*x+156960$ ecuatia nu este patratica					
4	R Square	0.07							
5	Adjusted R Square	-0.03							
6	Standard Error	209669.95							
7	Observations	22							
8									
9	ANOVA								
10		df	SS	MS	F	Significance F			
11	Regression	2	64296499824	32148249912	0.731282083	0.494354729			
12	Residual	19	8.35268E+11	43961489895					
13	Total	21	8.99565E+11						
14									
15		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
16	Intercept	156960	896299	0	1	-1719014	2032935	-1719014	2032935
17	valoarea productiei prosoape, x1	805	2857	0	1	-5174	6785	-5174	6785
18	$x2=(x1^2)$	0	2	0	1	-5	4	-5	4

Figure 8: Generation of results for square regression

c) Cubic regression: it has the syntax: $y=a_1x+a_2x^2+a_3x^3+b$



Source: Microsoft Office- Excel

Figure 10: Generation of results for cubic regression

1	SUMMARY OUTPUT								
2									
3		Regression Statistics							
4	Multiple R	0.27							$y = -3389x_1 + 7x_2 + 0x_3 + 982871$
5	R Square	0.07							$y = -3389x_1 + 7x_2 + 982871$
6	Adjusted R Square	-0.08							Regresia nu este cubica pt ca coeficientul lui x3 este 0
7	Standard Error	215162.28							
8	Observations	22							
9									
10	ANOVA								
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
12	Regression	3	66258262591	22086087530	0.477074827	0.70218119			
13	Residual	18	8.33307E+11	46294808068					
14	Total	21	8.99565E+11						
15									
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	982871	4116217	0	1	-7664981	9630722	-7664981	9630722
18	valoarea productiei prosoape, x1	-3389	20583	0	1	-46632	39855	-46632	39855
19	x2=(x1^2)	7	33	0	1	-63	76	-63	76
20	x3=(x1^3)	0	0	0	1	0	0	0	0

Source: Microsoft Office- Excel

As we can notice in figure 10, R Square has the value 0,07, which is close to 1, which leads us to the conclusion that both the square regression and the cubic one are not faithful to reality and we will renounce to the calculation of the prediction. The only regression that has brought relevant results has been the linear regression.

4. CONCLUSION

Statistical analysis instruments, such as regression, help the user make predictions. Although there are no perfect relations in the real world, through regression, predictions can be made of a variable, function of the value of another. Prediction is the process of estimating the value of a variable by knowing the value of another.

Regression is tightly connected with the concept of correlation. A powerful association between two elements leads to an increase in the precision of a variable prediction function of another.

BIBLIOGRAPHY

Sarbu C., Jantschi I., Statistic validation and evaluation of the analytic methods by comparative studies. I. Analytical methods validation by regression analysis, ISSN 0034-7752, Roumain

Breaz N., Aldea M., On the smoothing spline regression models, No 15/2008

Marinoiu C., Choosing a smoothing parameter for a curve fitting by minimizing the expected prediction error, Acta Universitatis Apulensis, Mathematics- Informatics, no 5, 91-96, 2003

<http://statisticasociala.tripod.com/regresie.htm>

http://www.mpopa.ro/statistica_licenta/sem_2/spss_08_regresia_simpla.pdf