REAL ESTATE VALUATION MODELING WITH SIMPLE LINEAR REGRESSION

Irina Ana-Maria MIHĂILESCU, Assist. Lect. PhD stud. Eng., TUCEB – Faculty of Geodesy, inamihailescu@gmail.com

Abstract: *This paper presents an example of simple linear regression modeling applied in real estate valuation and its purpose is to show what are the main ideas and steps in developing such a valuation model. To begin with, it underlines the importance of statistics and the role it plays in real estate valuation, presents the three types of Automated Valuation Models (AVM) structures (additive-linear, multiplicative and hybrid-nonlinear) and continues with the main characteristics of simple linear regression, multiple linear regression and nonlinear regression. The application within this paper shows step by step how to establish the regression equation, to estimate and test the equation parameter, to analyze the variance and the equation's residuals.*

Keywords: *property valuation, real estate valuation, statistics*

1. AVMs developing process

Models generally are representations of how things or real phenomena look and function. In particular, each real estate assessment approach, method or technique can be modeled separately as the whole valuation process. Statistical modeling is used for many years by expert valuers, proving to be a very useful tool. The results of statistical modeling of valuation are called Automated Valuation Models (AVM) and their applicability in valuing both individual property and mass appraisal especially is high. Basically, AVM's are software based on mathematical formulas that produce an estimator of market value, based on market analysis, property characteristics, market conditions, using information collected in advance. AVM credibility and accuracy of its results depend on the data used in valuation and also on experience and training of the person or team that develops the model.

The process in which an AVM can be effectively achieved is very complex, involving the following steps: identification of subject property, setting the limiting conditions, data management and analysis of their quality, setting the model specification, stratification, model calibration, testing and quality assurance of the application, reconciliation of the value obtained and final reconciliation. The first steps are identical to those of the classic assessment and establishing model specifications refers to the process of designing AVM's structure and consists in defining its shape and in integration of valuation theory, economic analysis and market factors in order to select the appropriate variables. AVM structure may be additive (linear), multiplicative or hybrid (nonlinear). These are set out below for the three valuation approaches, compared with conventional valuation models.

For sales comparison approach:

- *classical model*: $VP_S = P_C + Cor_C$

in which: VP_S - market value of subject property, P_C - selling price of comparable property, Cor_C - adjustments applied to comparable.

- in *additive model* contribution of independent variables is summed (a), in *multiplicative model* their contribution is multiplied (b) and the *hybrid model* is a combination of additive and multiplicative models (c):

$$VP_S = B_0 + B_1 \cdot x_1 + B_2 \cdot x_2 + ... \quad (a)$$

$$VP_S = B_0 \cdot x_1^{B1} \cdot x_2^{B2} \cdot ...... \quad (b)$$

$$VP_S = \pi GQ \cdot (\pi CQ \cdot \Sigma CA + \pi TQ \cdot \Sigma TA + \Sigma A) \quad (c)$$

in which: VP_S - market value of subject property (dependent variable), B_0 - constant of the model, x_i - independent variables, B_i - independent variables coefficients, πGQ - multiplied general qualitative variables, πCQ - multiplied building qualitative variables, ΣCA - sum of building additive variables, πTQ - multiplied land qualitative variables, ΣTA - sum of land additive variables, ΣA - sum of other variables, respectively.

Multiplicative model is much difficult to calibrate because the variables must be transformed into logarithmic form, but the advantage is that it allows the application of corrections proportionally to the subject property value.

For cost approach:

- *classical model:* $VP_S = VT_S + (CN - \Sigma D)$
- *hybrid model:* $VP_S = \pi GQ \cdot [(1 - CQ_D) \cdot CN - VT_S]$

in which: VP_S - market value of the subject property land, CN - replacement cost for the subject property building, ΣD - sum of depreciations of the subject property building, CQ_D - building qualitative variable, respectively.

For income capitalization approach:

- *classical model* for income capitalization method: $VP_S = \frac{VNE}{c}$
- *hybrid model*: $VP_S = (\sum C - P - Ch) \cdot c^{-1}$

in which: VNE - net income, $\sum C$ - sum of rents in one year, *P*- vacancy losses, *Ch* – operating expenses, *c* - capitalisation rate.

The most used calibration methods of the established model are based on statistical methods such as simple and multiple linear regression or nonlinear regression, but it can be made by methods recently developed or taken from other areas and adapted, such as Adaptive Estimation Procedure (AEP) taken from engineering field, Artificial Neural Networks (ANN) originally applied for biological processes and Time Series Analysis, each of these calibration procedures having advantages and disadvantages. In practice, stages of model specification and calibration are carried out iteratively: specification of model - testing the model specifications through calibration – adjustment of model specifications - testing the adjusted specifications, repeating these operations until they gain the designed confidence level.

Testing the model and assuring its quality are needed to determine the scope of previously calibrated model and/ or if are needed other specifications. All data used to specify and calibrate the model must pass a series of tests, including: the amount of data must be sufficiently large to produce models with a reasonable degree of prediction of market value, transaction data must reflect the maximum conditions for transactions at market value, subjective data should be consistent in the entire population of properties for which model is used, is essential for the model quality that data concerning property characteristics to have a high degree of accuracy; checked in the field, these data should be accurate within a percentage of 95%.

2. Statistical regression in valuation

Regarding the valuation process, statistical regression has two major advantages: it can be used to value a large number of properties at the same time, quickly and economically, and besides it helps estimate their values, it also explains it, situation which is not achievable through classical methods of valuation.

Its limitations are evidenced by the fact that much more data is needed than for classical approaches application, for example, when subject property is valued through sales comparison classical approach, may be used only 5-10 comparables, but for statistical methods there are needed 5 to 50 times the number of independent variables, amount of data which is difficult to obtain. It is also important to pay particular attention to variables used in the regression model because some variables may be associated with the value of a property but does not necessarily mean they causes it. Regression analysis determines and explains the associations between variables, but not the causes too. Also, a regression equation should be used for the sample for which it was developed, not for another sample, and, moreover, because very few relationships remain stable over time, the database must be continuously updated to capture any changes in the temporal relationship between variables.

2.1. Simple linear regression

It is used to determine the dependent variable (endogenous) denoted *y* using a single independent variable (exogenous) *x*. The relationship is positive if the value of *y* increases with increasing value of x and is negative if the value of y decreases with increasing value of *x*. It is considered that a change of independent variable will be reflected in a proportional change of the dependent variable.

Simple linear regression model is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,, \quad i = \overline{1,n}$$

in which: β_0 - the constant (intercept), i.e. the value of *y* as $x = 0$, β_1 - slope of the regression line, ε_i - residual variable, which quantifies the influence of other random variables on *y* and has the expression $\varepsilon_i = y_i - \hat{y}_i$, where $\hat{y}_i$ is the estimated value of the dependent variable through regression model x_i.

Simple linear regression model takes into account a number of assumptions such as: parameter estimation is based on a sample (x_i,y_i) which represent couples of values of independent and dependent variables, error variance is constant, residual variable ε_i is of zero average, between residual terms *εi* there are no correlations, residual variables *εi* are not correlated with the independent variable *x*, residual variables *εi* are normally distributed. If these assumptions are met and there is a linear dependence relationship between variables then this relationship is expressed mathematically by the relationship above.

For simple linear regression analysis, the unknown parameters of the model, β_0 and β_1, are estimated and also *εi* errors (e.g. the method of least squares), then the regression model is tested, the correlation between the two variables is determined and then predictions can be made.

Establishing equation parameters with least square method:

the slope: $$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$ and the intercept: $$b_0 = \bar{y} - b_1\bar{x}$$

Simple linear regression model is useful only if the two variables *x* and *y* are in a significant relationship. To judge the significance of this relationship may be used t test for independent testing the significance of slope β_1 or intercept β_0 of the regression line, or F-test for simultaneously or separately testing of both parameters of the equation.

2.2. Multiple linear regression

Market value of a property depends on several independent variables and therefore in valuation practice is used multiple linear regression. It expresses the relationship between the dependent variable y and at least two independent variables x_i , and its model is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_m x_{mi} + \varepsilon_i$$

in which: β_0 – the intercept, $\beta_1, \beta_2, \ldots, \beta_m$ – multiple regression parameters, ε_i – residual variable

Or in matrix format: $Y = XB + E$

where:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & X_1 & X_2 & \cdots & X_m \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{m1} \\ 1 & x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{mn} \end{bmatrix},$$

$$E = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \text{ and } B = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{m1} & \beta_{m2} & \cdots & \beta_{mn} \end{bmatrix}$$

B - multiple linear regression parameters matrix, E - regression residuals vector.

The assumptions taken into account to achieve multiple linear regression model are, as with simple regression: multiple regression residual variables are of zero average, residual variables are not correlated among themselves, residual variables are statistically independent of variables *xi*, the residual variable vector ε is normally distributed.

Establishing equation parameters with least square method:

$$B = \left(X^T X\right)^{-1} X^T Y$$

Testing significance of multiple linear regression relationship can be also performed with Student and Fisher statistical tests.

2.3. Nonlinear regression

If following a graphical representation of data through a scatter plot we observe that the points are not grouped along a straight line, then we assume that regression is not linear and a nonlinear regression solution may be applying. Nonlinear regression is used to describe the relationship between two variables, the dependent variable y and independent variable x, considering that other factors have a constant and negligible influence on y. There are several mathematical models which can be used in nonlinear regression: exponential model (d) logarithmic model (s), second-degree parabolic model (f) and others.

$$Y = a \cdot b^x + e \qquad (d)$$

$$Y = a + b \cdot \lg x_i + e \qquad (e)$$

$$Y_i = a + b \cdot x_i + c \cdot x_i^2 + e \qquad (f)$$

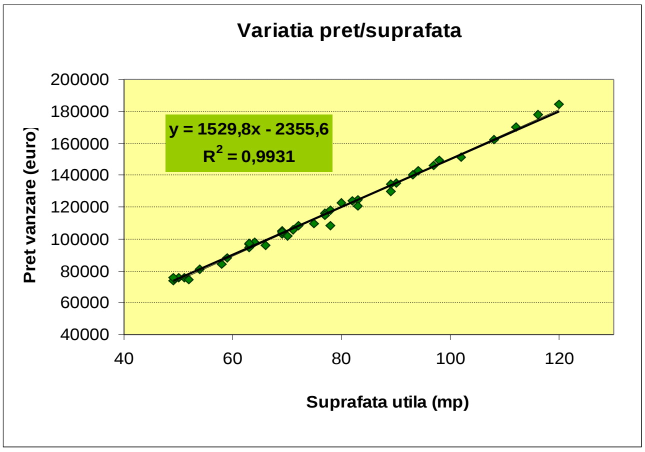
3. Valuation modeling with simple linear regression

Number of used properties - 40 (minimum number for a sample of data - 30), apartments with 2 and 3 bedrooms, located in the center of the capital, usable surface varies between 49-120 sqm., selling prices vary in between 73 500 € and 184 000 €.

3.1. Establishing simple linear regression equation

It is used a scatter plot and the trend line is set in order to examine whether there is dependence between sale prices and usable areas of apartments and parameter b0 is tested to determine its significance.

Fig. 1. Scatter plot – price/surface



- there is a linear relationship between the two variables that must be mathematically analyzed,
- the simple linear regression model established for the market data is:
  $Val_{propr} = 1529{,}8\ € \cdot S_{usable} - 2355{,}6\ €$
- intercept b_0 has a negative value (-355.6 €) which means that a property with a usable area of 0 sqm would have a sale price of - 2355 €, a meaningless conclusion.

Table.1. Testing b_0 parameter

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept b_0	-2355,6	1641,9	-1,434616	0,159576	-5679,5767	968,3939

- testing the significance of intercept by Student test (Table 1), we get $|t_{b0}|$= 1,434, critical value of the tables is $t_{\alpha/2}$= 2,024 (for α = 0.05 and 38 degrees of freedom), so $|t_{b0}|<t_{\alpha/2}$, resulting that the regression line intercept is not significant for the established model
- analyzing P-value (observed level of significance - the probability that the test result is incorrect), we obtained intercept P-value of 0.15957, higher than the critical value of 0.05, so the regression line intercept is not significant,
- so I excluded the intercept from the initial single regression equation and obtained the following model, to be further examined and tested:

$$Val_{propr} = 1500{,}9\ (\text{euro}) \cdot S_{usable}$$

3.2. Regression analysis

Table 2. Regression analysis

Regression Statistics	
Multiple R	0,999786
R Square	0,999572
Adjusted R Square	0,973931
Standard Error	2493,98
Observations	40

- the linear correlation coefficient r (Multiple R) has the value of 0.99978, close to 1, supporting the conclusion made earlier that between the two variables is a strong direct linear dependence. The impact of independent variable (usable area) on the selling price is significant.

- the coefficient of determination r^2 (R Square) is 0.999572, indicating that 99.9% of sales prices can be explained by changes in usable area and the rest is explained by the influence of other factors such as the flat level (floor), the degree of improvement, its age and others. That is, the obtained model corresponds to a percentage of 99.9% of the analyzed properties.
- standard error is 2,493 €, which means that market value predicted by the model $\hat{y}$ can be different with 2,493 € from the average value of sold apartments.

3.3. Computing and testing parameter b_1

Table 3. Computing and testing parameter b_1

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept b_0	0	-	-	-	-	-
Usable area b_1 (mp)	1500,938	4,972125	301,8704	2,59E-67	1490,8809	1510,995

- the value of parameter b_1 is positive so with increasing usable area of the apartment, also increases the property market value.
- the value of b_1 is 1500.9 € which indicates that for each additional square meter of usable area are added 1500.9 € to the property market value.
- testing the significance of the two parameters is performed with Student test; for the slope we obtained t_{b1}= 301,87, value compared with the critical value $t_{\alpha/2}$=2,024 (for α = 5% and 38 degrees of freedom). Because $t_{b1} > t_{\alpha/2}$, we concluded that the hypothesis $H_0 : \beta_1 = 0$ can be rejected in favor of $H_1 : \beta_1 \neq 0$, so the slope of the regression line is correctly determined.
- P-value for the slope is $2,59 \cdot 10^{-67}$, much less than the limit of 0.05, so the slope of the regression line is correctly specified.
- a confidence interval was computed for the slope $\left| b_1 \pm t_{\alpha/2} \cdot s_{b_1} \right|$ resulting a min_{int} = 1,490 € and max_{int}=1511 €. This confidence interval helps us to be 95% confident that a property which is bigger with 1 sqm of usable area than the usable area of another property, has its market value higher with an amount between minimum 1,490 € and maximum 1,511 €.

3.4. Variance analysis

For the variance analysis is used Fisher-Snedecor test. We set a confidence level of 95% (significance level α = 5%) and 38 degrees of freedom (for both parameters).

Table 4.Variance analysis

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	5,667E+11	5,667E+11	91125,79	7,73E-66
Residual	39	242578629,2	6219964,852		
Total	40	5,670E+11			

- the calculated value of statistical index F is 91125.79, higher than the critical value F_α=4.09 extracted from critical values tables corresponding to df_1 = 1 and df_2 = 38.

- thus, the null hypothesis $H_0 : \beta_1 = 0$ is rejected in favor of alternative hypothesis $H_1 : \beta_1 \neq 0$ and indicate that the linear regression model is valid in statistical terms and probability of mistakes is very small, $7,736 \cdot 10^{-66}$.

3.5. Residuals analysis

With this analysis, the validity of statistical assumptions listed above is checked. Very rare these assumptions are perfectly met. Important is to detect if there is a serious offense against them. Note that only one residual value is too high, the one corresponding to the property number 36.

a. Checking normality assumption - I present residual values histogram (fig. 2), normal probability chart for residual values (fig. 3) and steam-and-leaf (table 5) with related statistical indicators.

Fig. 2. Residuals histogram

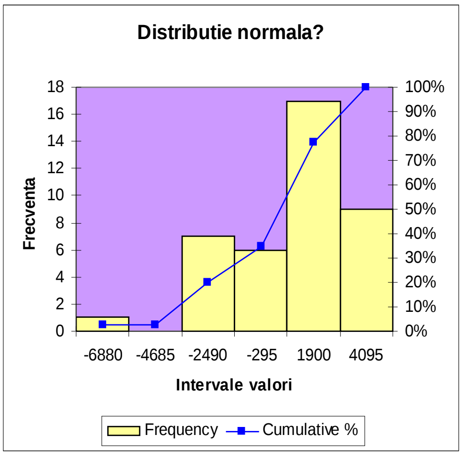


Table 5. Residuals Steam-and-leaf

Stem unit:	1000
-9	1
-8	
-7	
-6	
-5	
-4	1
-3	6 5 1 1 1 1
-2	1
-1	
-0	6 6 6 6 5 1 1 1 0
0	0 1 4 4 4 4 5 9 9 9
1	0 4 9 9 9 9 9
2	4 4 5
3	9
4	1

Statistics	
Sample Size	40
Mean	-132,274
Median	262,2959
Std. Deviation	2490,385
Minimum	-9073,16
Maximum	4087,439

- these three elements visual indicate that errors are not perfectly distributed as standard normal curve: histogram and steam-and-leaf are Gaussian bell shaped, but slightly asymmetric to the left, indicating departure from normality hypothesis.
- as normal probability graph is not a straight line, this is a possible indication that the normality assumption is violated. To check normality distribution is used Kolmogorov – Smirnov test.

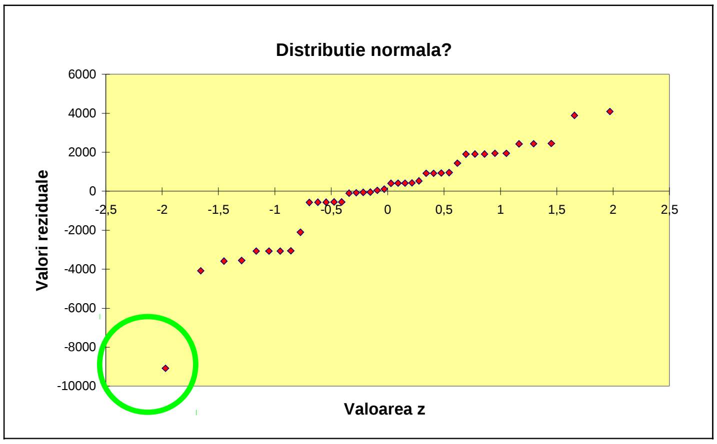


Fig. 3. Normal probabilities chart

Table 6. K-S test

One-Sample Kolmogorov-Smirnov Test		
N		40
Normal Parameters[a,b]	Mean	-132.2565
	Std. Deviation	2490.36519
Most Extreme Differences	Absolute	.205
	Positive	.100
	Negative	-.205
Kolmogorov-Smirnov Z		1.296
Asymp. Sig. (2-tailed)		.070

a. Test distribution is Normal.

b. Calculated from data.

Kolmogorov-Smirnov test:

H_0: sig > 0,05 – the series is normally distributed
H_1: sig < 0,05 – the series is not normally distributed

- for α = 0.05 and a series of 40 data, we find KS test critical value d = 0.070, so sig> 0.05, the series is normally distributed.

a. Checking the assumption of residual values ε_i constant variance

In order to perform this check are used scatter plots of residual values against usable areas (fig. 4) and against predicted sales prices $\hat{y}$ (fig. 5).

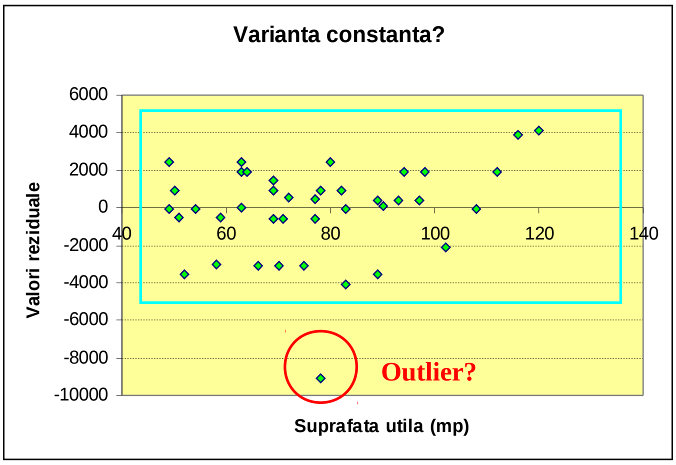


Fig. 4. Scatter plot εi vs. usable area

- the graph in fig. 4, which presents the errors dispersion depending on the usable surface, fall in a band, suggesting that the dispersion error is constant.

We wonder if the point encircled in both graphs (corresponding to the property number 36), with a 7.7% relative residual value, is an outlier - aberrant value. To verify the existence of outliers in a data series can be applied Grubbs test and if such values exist, they must be eliminated from the series of data and statistical analysis iteratively rerun.

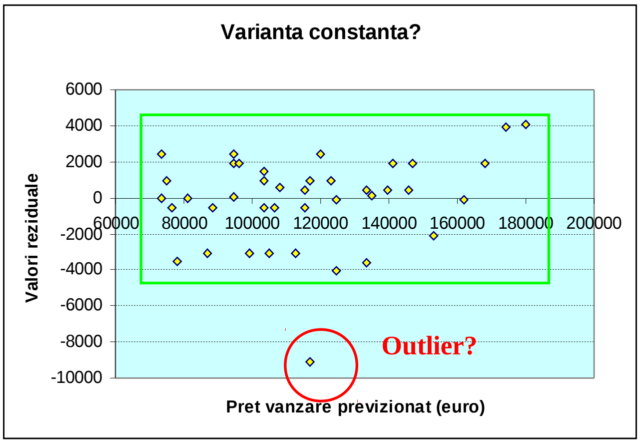


Fig. 5. Scatter plot εi vs. predicted prices

- fig. 5, representing the scattering of errors based on sales prices predicted by the regression model is similar, emphasizing the earlier conclusion.

Grubbs test:

H_0: $u < u_{critical}$ – there are no outliers in the data series

H_1: $u \geq u_{critical}$ – at least one aberrant value in the data series

We calculated the value of Grubbs statistical index $u = \left| \frac{x_a - \bar{x}}{s} \right|$ for all residual values, with s = 2493.9857. From critical values tables corresponding to α = 0.05 and a series of 40 data, we extracted $u_{critical}$ = 2.87 , so the residual value |u|=3,6380 belonging to the property number 36 is an outlier and should be excluded from the data series.

c. Checking the errors independence assumption

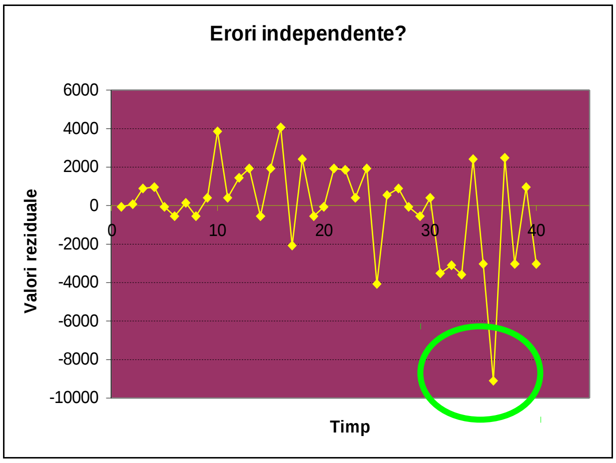


Fig. 6. Scatter plot residuals vs. time

- this hypothesis is verified using a scatter plot of residual values against time factor.
- the shape of the obtained graph suggests a positive auto correlation of errors because the line has a cyclical pattern.
- for further analysis is required the Durbin-Watson statistical test which measures the correlation between one residual value and the previous residual value (auto correlation).

T ab.7. Durbin-Watson index

Durbin-Watson Calculations	
Sum of Squared Difference of Residuals	517738132,3
Sum of Squared Residuals	230115307,7
Durbin-Watson Statistic	2,249907394

Durbin-Watson test:

H_0:$d>d_{U,\alpha}$ -residual values are not auto correlated

H_1:$d<d_{U,\alpha}$ -residual values are auto correlated

- statistical indicator value is d = 2.25. Using tables for critical values we extracted the values for rejecting the hypothesis, sample size being 40, α = 5% and k = 1 (one independent variable)

$d_{L,\alpha}$= 1,44 şi $d_{U,\alpha}$= 1,54.

- because $d>d_{U,\alpha}$, residual values are not auto correlate so they are independent.

Following are the conclusions regarding the new data series, comprising 39 properties remaining after excluding property number 36.
1st conclusion: the simple linear regression equation corresponding to the series of 39 new properties: $Val_{propr} = 1503{,}8\ € \cdot S_{usable}$
2nd conclusion: there is still a strong direct linear dependence between usable area and selling price variable: linear correlation coefficient r = 0.999857 and coefficient of determination r^2=0,999715, higher than the original series of 40 properties and standard error is smaller.
3rd conclusion: slope b_1 is positive 1503.8 €: an addition of 1 sqm usable area means a 1503,8 € growth of market value of the property and is properly determined: t_{b1} =365,17 more than $t_{critical}$ = 2.024 and P-value = 5.6 * $5{,}6 \cdot 10^{-69}$ is less than the critical threshold of 0.05 (Student test)
4th conclusion: confidence interval: min_{int} = 1,495 € and max_{int} = 1,512 €, so we can be 95% confident that a property whose usable area is greater with 1sqm comparing with other usable areas has a market value higher with not less than 1,495 € but not more than 1,512 €;
5th conclusion: linear regression model is statistically valid and the probability of mistaking is extremely small $2{,}05 \cdot 10^{-67}$: F=133351,3 much higher than the critical value F_α=4,09 extracted from tables (Fisher test)
6th conclusion: the series of 39 observations is normally distributed: for α = 0.05, sig = 0.227 therefore greater than the critical value 0.05 (Kolmogorov-Smirnov test)
7th conclusion: the series of 39 data values are not anomalous: $u_{critical}$=2,86, higher than the values corresponding residuals (errors) for all 39 data (Grubbs test)
8th conclusion: residual values are independent: statistical indicator value is d=2,446 greater than $d_{U,\alpha}$= 1,54 ($d_{L,\alpha}$= 1,43) (Durbin-Watson test).

4. Conclusions

Simple linear regression model meets all the assumptions of this type of regression and it can be used to assess real estate only on its usable area: $Val_{propr} = 1503{,}8\ € \cdot S_{usable}$

In classical approach of sales comparison the corrections are calculated separately and applied to the selling price of comparables according to the elements of comparison, step by step, in a certain order. Through statistical modeling, these corrections are inserted directly into the model therefore enabling simultaneous correction of these elements of comparison. In cost approach are determined the land value, the cost of new of the building and all its depreciations and in income capitalization approach is necessary to determine the potential income of the subject property, rents, operating expenses, vacancy losses, capitalization rates or factors. Any of these elements can be determined through statistical modeling, considering such factors as dependent variables and factors influencing them as independent variables.

Thus, among the advantages of assessing real estate through statistical modeling are included: allows the use of unlimited data sets, the use of a large number of variables if these are necessary and relevant, allows calculation of the accuracy in estimating the value, which may be exceptionally good when sufficient market data are used, ensure uniform valuation, model can be used both for individual assessments and for mass appraisal, allows the automation of valuation process, short time for processing, data analysis and establishing the conclusion of the valuation, objectivity and efficiency. Those can only be achieved based on a very serious and coherent mathematical and statistical judgement and the must model be tested before actually used in the valuation. A disadvantage is that statistical modeling can not be applied to small markets with few transactions or to non-standard markets.

5. Bibliography

1. *Business Statistics in Practice, Bruce L. Bowerman, Richard T. O'Connell, McGraw Hill, fourth edition, 2007*
2. *Practical Applications in Appraisal Valuation Modeling, Statistical Methods for Real Estate Practitioners, M.S. Kane, M.R. Linné, J.A. Johnson, Appraisal Institute, 2004*
3. *Standard on Automated Valuation Models (AVMs), IAAO, 2003*
4. *Statistical Analysis, S.B. Richmond, second edition, 1964; Statistică aplicată pentru modelarea evaluării imobiliare, Iuliu Paşca, Editura Politehnică, 2008*
5. *Statistică: noțiuni fundamentale şi aplicații, Vergil Voineagu, Dana Colibabă, Giani Grădinaru, Editura ASE, 2002*
6. *Statistică teoretică şi economică – note de curs, s.l.drd. Silvia Elena Cristache, Editura ASE, 2002*
7. *Statistică teoretică şi economică - teorie şi aplicații, Constantin Anghelache, Editura Economică, 2004*
8. *Statistică: teorie şi aplicații, Tudorel Andrei, Stelian Stancu, Daniel Traian Pele, Editura Economică, ediția a doua, 2002.*