Regression and correlation analysis – regresná a korelačná analýza

Definition:
Regression and correlation analysis investigates the relationships between two or more quantitative statistical attributes (bivariate data). Its aim is to find a suitable **regression line**, **parameter estimates** (regression analysis) and to **measure the goodness of fit** (correlation analysis).
Regression analysis is a statistical procedure that can be used to develop a mathematical equation showing how variables are related. Correlation analysis is a procedure for determining the extent to which the variables are linearly related. If such a relationship exists, correlation analysis is used for providing a measure of the relative strength of the relationship.

Note:
In simple correlation and regression studies, data are collected on two quantitative variables to determine whether a relationship exists between the two variables. To graphically analyze the data, we can display the data on a two-dimensional graph. Such plots are called **scatter plots**. The variable along the vertical axis is called the **dependent variable**, and the variable along the horizontal axis is called the **independent variable**. The variable that is being predicted by the mathematical equation is called the **dependent variable**. The variable(s) being used to predict the value of the dependent variable are called the **independent variables**.
Notation: We will let **y** represent the **dependent variable**, and we will let **x** represent the **independent variable**.

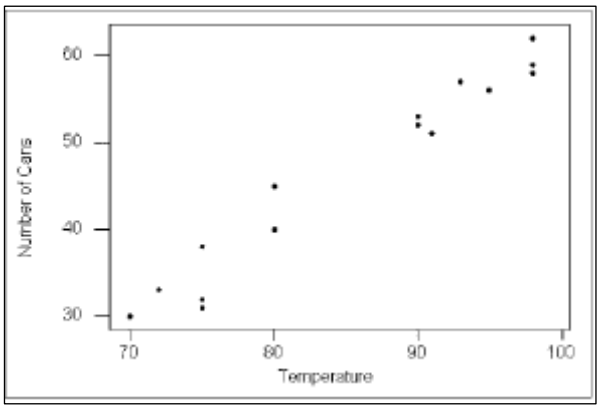
In multiple correlation and regression studies, data are collected on more than two quantitative variables (one dependent variable and more than one independent variable) to determine whether a relationship exists between these variables.

Deterministic model is a relationship between an independent variable and a dependent variable whereby specifying the value of the independent variable allows one to compute exactly the value of the dependent variable (i.e. predicted values).

Probabilistic model is a relationship between an independent variable and a dependent variable in which specifying the value of the independent variable is not sufficient to allow determination of the value of the dependent variable (i.e. predicted values).

Scatter plot – graf závislostí

Definition:
A scatter plot is a graph of the ordered pairs (x, y) of values for the independent variable x and the dependent variable y.



Scatter plot
y (dependent variable) – number of cans
x (independent variable) – temperature

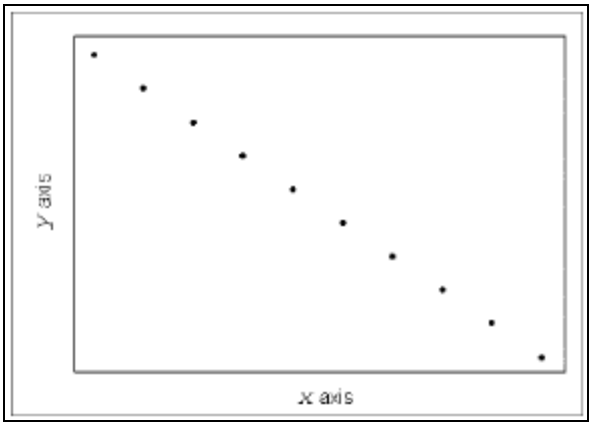
Two variables are said to be **positively related** if larger values of one variable tend to be associated with larger values of the other.

Two variables are said to be **negatively related** if larger values of one variable tend to be associated with smaller values of the other.

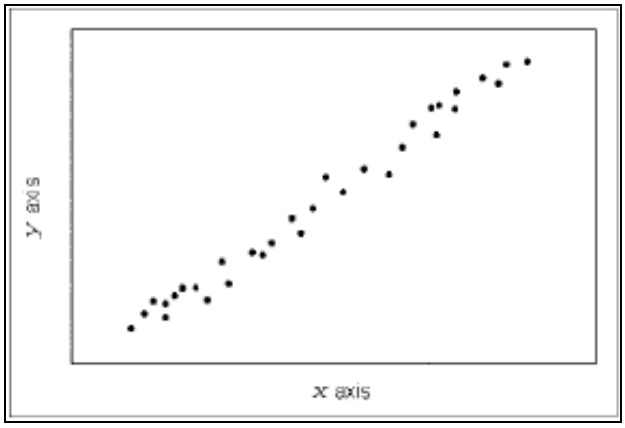
If the data are on the straight line, there is a **perfect association** (positive or negative) between the variable(s).

Scatter plots can display various patterns:

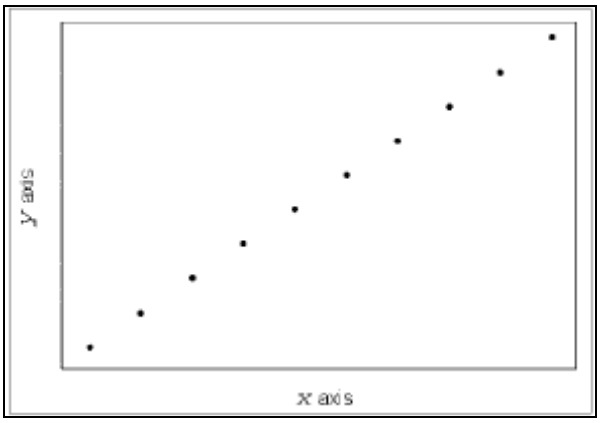
- linear – data are displayed in the scatter plot in a linear form
- nonlinear - data are displayed in the scatter plot in a nonlinear form



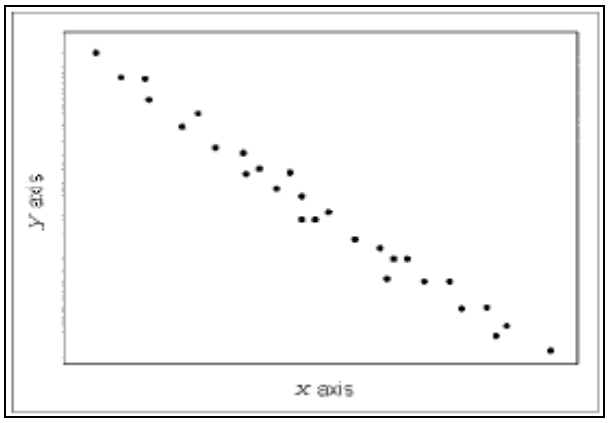
Perfect negative association



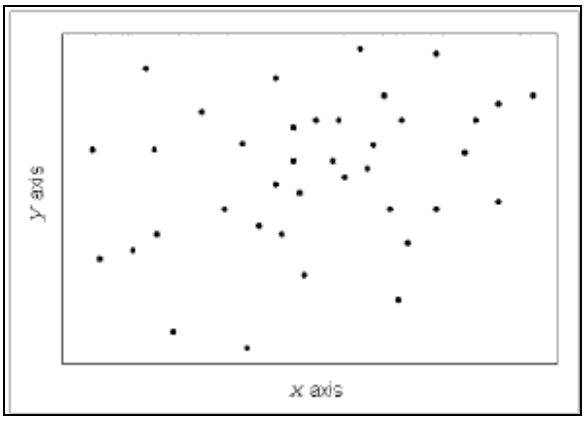
Very strong positive association



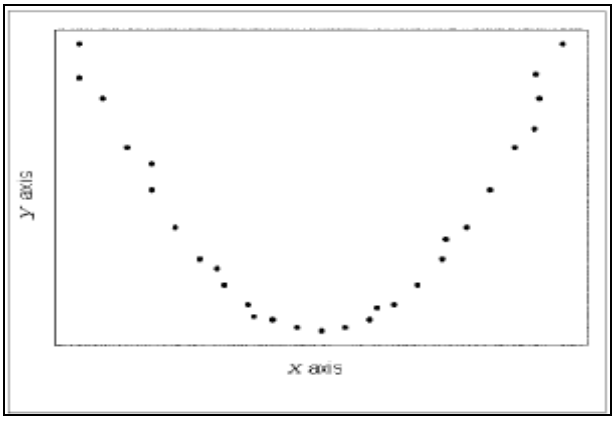
Perfect positive association



Very strong negative association



No association



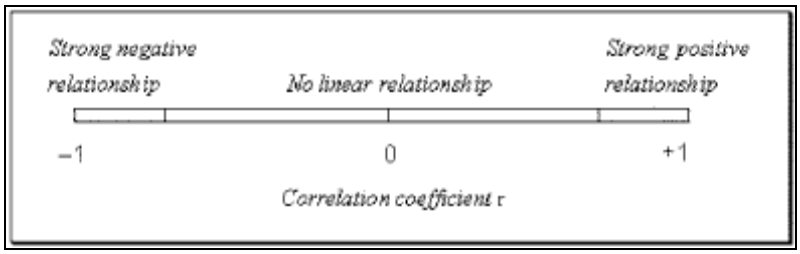
Nonlinear association

Correlation coefficient – korelačný koeficient

Definition:
Correlation coefficient is a numerical measure of the association between two variables. It measures the strength and direction of a relationship between two variables using. It is denoted by the letter **r** (sample correlation coefficient) or ρ (population correlation coefficient).

Properties of the correlation coefficient:

- The range of the correlation coefficient is from -1 to +l.
- If there is a perfect **positive** linear relationship between the variables, the value of **r** will be equal to **+l.**
- If there is a perfect **negative** linear relationship between the variables, the value of **r** will be equal to **-1.**
- If there is a **strong positive** linear relationship between the variables, the value of **r** will **be close to +l**.
- If there is a **strong negative** linear relationship between the variables, the value of **r** will be **close to -1**.
- If there is little or **no linear** relationship between the variables, the value of **r** will be **close to 0**.



Coefficient of determination – koeficient (index) determinácie

Definition:
The coefficient of determination measures the proportion of the variability in the dependent variable (y variable) that is explained by the regression model through the independent variable (x variable). It is a measure of the goodness of fit for the estimated regression model.

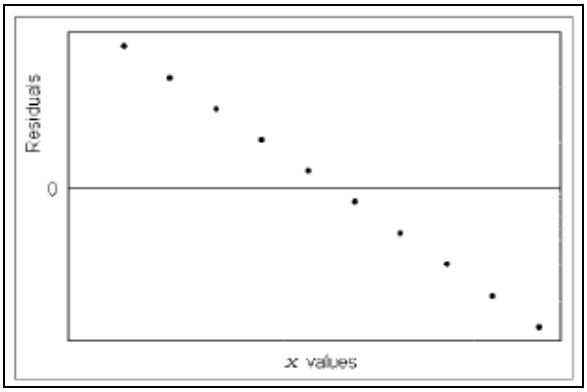
Properties of the coefficient of determination:

- The coefficient of determination is obtained by squaring the value of the correlation coefficient.
- The symbol used is R^2.
- Note that $0 \le R^2 \le 1$.
- R^2 values **close to 1** would imply that the model is explaining **most of the variation** in the dependent variable and may be a very useful model.
- R^2 values **close to 0** would imply that the model is explaining **little of the variation** in the dependent variable and may not be a useful model.

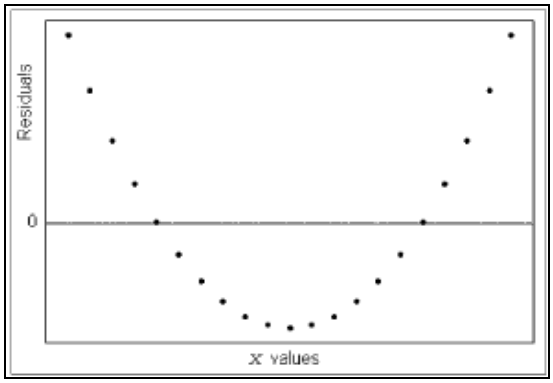
Residual plots – grafy rezíduí

Definition:
Residuals are just errors. In particular, a residual is the difference between an actual observed y value and the corresponding predicted y value, i.e. $e_i = Y_i - \hat{Y}_i$. Standardized residual is the value obtained by dividing the residual by its standard deviation.

Note:
Plots of residuals may display patterns that would give some idea about the appropriateness of the model. If the functional form of the regression model is incorrect, the residual plots constructed by using the model will often display a pattern. The pattern can then be used to propose a more appropriate model.



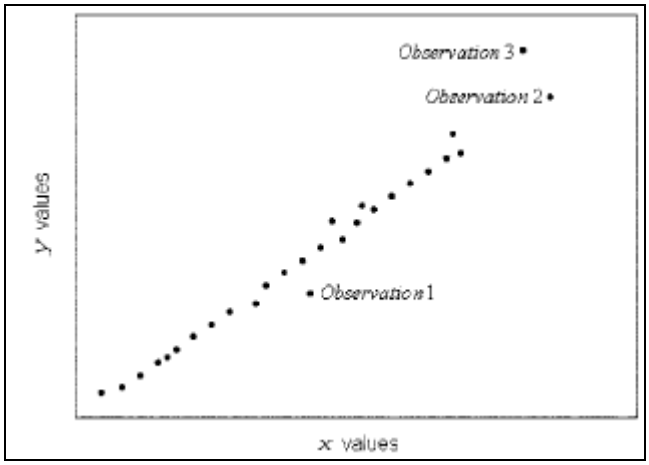
Linear residual plot



Nonlinear residual plot

Outliers and Influential points – outliery a vplyvné body

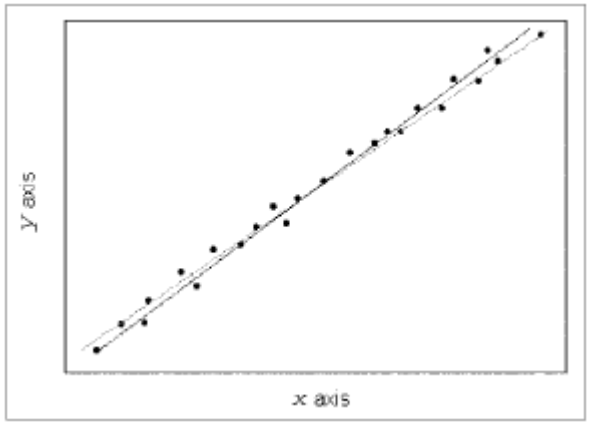
Definition:
A value that is well separated from the rest of the data set is called an outlier. With respect to the line of best fit, an outlier is an observation with a large absolute residual value. That is, an outlier will fall far from the regression line and will not follow the pattern of the linear relationship expressed by the line of best fit. An observation that causes the values of the slope and the intercept in the line of best fit to be considerably different from what they would be if the observation were removed from the data set is said to be influential.



Plot illustrating outliers and influential points

Least-Squares Regression Line – regresná priamka MNŠ

In investigating the relationship between two variables, the first thing one should do is to prepare a scatter plot after the data are collected. From the plot, one can observe any pattern. If the correlation coefficient is reasonably large (positive or negative), the next step would be to fit the regression line which best fits or models the data (line of best fit).



Line of best fit

Regression analysis allows us to determine which of the two lines best represents the relationship. In elementary statistics, the equation of the regression line is usually written as $\hat{y} = ax + b$, where a is the **slope**, b is they **intercept**, and $\hat{y}$ is read as "y hat," and it gives the predicted y value for a given x value. Least-squares analysis allows us to determine values for a and b such that the equation of the regression line best represents the relationship between the two variables by minimizing the error sum of squares-that is, by minimizing $\sum(y_i - \hat{y}_i)^2$, where $(y_i - \hat{y}_i)$ is the error for a given y value. This regression line is usually called the line of best fit. We usually refer to this type of regression analysis as **simple regression analysis**, since we are dealing only with straight-line models involving one independent variable. If there are more than one independent (x) variable then this type of regression analysis is known as **multiple regression analysis**. The equations that one can use to compute the values for a and b are:

$$a = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

Note: Least squares method is the approach to develop the estimated regression equation which minimizes the sum of squared residuals and at the same time requires the sum of residuals to be zero.

Elaborated by: Ing. Martina Majorová, Dept. of Statistics and Operations Research, FEM SUA in Nitra
Reference: JAISINGH, L.: Statistics for the Utterly Confused

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,914612315
R Square	0,836515687
Adjusted F	0,82393997
Standard E	13,20593358
Observatio	15

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	11600,57647	11600,57647	66,51833	1,81E-06
Residual	13	2267,156863	174,3966817		
Total	14	13867,73333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95,0%*	*pper 95,0%*
Intercept	32,46813725	11,78412046	2,755244853	0,016373	7,010093	57,92618	7,010093	57,92618
X Variable	14,60294118	1,790480761	8,155877178	1,81E-06	10,73484	18,47104	10,73484	18,47104

Regression analysis report (simple regression where monthly sales for certain goods represent a dependent variable and advertising costs represent an independent variable)

Interpretation of simple regression analysis report:

Regression statistics – ukazovatele tesnosti závislosti a kvality modelu

Multiple R – (jednoduchý) korelačný koeficient
Multiple R=0.9146, i.e. there is a very strong relationship between the dependent and the independent variable (even a small change in the values of the independent variable will greatly affect the values of the dependent variable). The closer the value to 1 is, the better it is.

R Square – index (koeficient) determinácie
R Square=0.8365, i.e. approximately 83.65% of variability of the dependent variable is expressed by the regression model through the independent variable. The closer the value to 1 is, the better it is.

Adjusted R Square – korigovaný (upravený) koeficient determinácie
Adjusted R Square=0.8239; it is a measure of the goodness of fit for the estimated regression equation (as R Square) which accounts for the number of independent variables for the model (that is why its value is smaller than the value of R Square)
Note: if the value of R Square is small and the model contains a large number of independent variables, the adjusted coefficient of determination can take on negative value

Standard Error – štandardná chyba modelu
Standard Error=13.2056; it is an error which is still present in each regression model due to a human factor; the smaller the standard error is, the better it is

Observation – počet pozorovaní
Observation=15, i.e. there were 15 observations both for the dependent variable and for the independent variable

ANOVA – výstup pre analýzu rozptylu

Regression – vysvetlená variabilita
Residual – nevysvetlená (reziduálna) variabilita
Total – celková variabilita
SS (sum of squares) – súčet štvorcov
MS (mean squares) – priemer štvorcov

F – testovacie kritérium F testu

Significance F - teoretická hladina významnosti, pomocou ktorej vyhodnocujeme test (Significance F is compared with the level of significance – alpha):

if Significance F > 0,05 → the regression model as a whole is not statistically significant (–)
if Significance F < 0,05 → the regression model as a whole is statistically significant at the 0.05 level of significance (+)
if Significance F < 0,01 → the regression model as a whole is statistically significant at the 0.01 level of significance (++)

Significance F =1,81E-06 (i.e. $1{,}81 \cdot 10^{-6}$)
Significance F < 0,01 → the regression model as a whole is statistically significant at the 0.01 level of significance (++)

Parameter estimates and their significance
Coefficients – vypočítané koeficienty (parametre) regresného modelu
Intercept – lokujúca konštanta, absolútny člen, aditívna premenná (všetko synonymické výrazy); it is the intersection of y (dependent variable) when x (independent variable) is equal to zero
Intercept=32.468, i.e. sales for certain goods are equal to 32.4mil.SKK at zero advertising costs
X Variable1 – regresný koeficient; it is a slope (a change in dependent variable divided by a change in independent variable)
X Variable1=14.6, i.e. if we increase advertising costs **by a single unit** (1 mil. SKK) then sales for certain goods will increase by 14.6mil.SKK

Standard error – štandardná chyba jednotlivých parametrov (it should not be more than half the value of the referring coefficient)

tStat – test statistics for t-test in which we analyze the significance of referring coefficients

p-value – probability value; p-value is compared with the level of significance – alpha:

if p-value > 0,05 → the referring coefficient is not statistically significant (–)
if p-value < 0,05 → the referring coefficient is statistically significant at the 0.05 level of significance (+)
if p-value < 0,01 → the referring coefficient is statistically significant at the 0.01 level of significance (++)

p-value (Intercept) =0,016
p-value < 0,05 → the referring coefficient is statistically significant at the 0.05 level of significance (+)

p-value (X Variable1)=1,81E-06 (i.e. $1{,}81 \cdot 10^{-6}$)
p-value < 0,01 → the referring coefficient is statistically significant at the 0.01 level of significance (++)

Lower 95%, Upper 95% - 95% interval estimates of parameters (coefficients); value of the refering coefficient must lie in the given interval

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,823436
R Square	0,678048
Adjusted F	0,624389
Standard E	6,055902
Observatio	15

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	2	926,8458826	463,4229	12,6363	0,001114
Residual	12	440,0874507	36,67395		
Total	14	1366,933333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95,0%*	*pper 95,0%*
Intercept	-22,8283	11,93402799	-1,91287	0,079922	-48,8303	3,173714	-48,8303	3,173714
X Variable	138,1401	46,18926004	2,990742	0,011259	37,50239	238,7779	37,50239	238,7779
X Variable	1,387386	0,490686652	2,827438	0,015248	0,318272	2,456501	0,318272	2,456501

Regression analysis report (multiple regression where monthly sales for certain goods represent a dependent variable and advertising costs, number of agents represent the independent variables)

Interpretation of multiple regression analysis report is the same as interpretation of simple regression analysis report but there is one difference when interpreting the regression coefficients (slopes)

X Variable1=138.1; i.e. if we increase advertising costs **by a single unit (1mil.SKK) and other independent variable(s) are held constant (ceteris paribus)** then sales for certain goods will increase by 138.1mil.SKK
X Variable2=1.38; i.e. if we increase number of agents **by a single unit (1 person) and other independent variable(s) are held constant (ceteris paribus)** then sales for certain goods will increase by 1.38mil.SKK

Closing remarks:

y	$x_1, x_2, \ldots, x_k$
(a) Predictand	Predictors
(b) Regressand	Regressors
(c) Explained variable	Explanatory variables
(d) Dependent variable	Independent variables
(e) Effect variable	Causal variables
(f) Endogenous variable	Exogenous variables
(g) Target variable	Control variables

Synonyms