

**SKÚMANIE ŠTATISTICKEJ
ZÁVISLOSTI DVOCH
KVANTITATÍVNYCH ZNAKOV
(PÁROVÁ LINEÁRNA REGRESIA)**

Príčinné (kauzálne) závislosti javov:

- **jednostranná** závislosť – vzťah medzi príčinou a účinkom je jednostranný;

výživa obyvateľstva a stupeň ekonomického rozvoja krajiny

- **obojsstranná** závislosť – príčina a účinok na seba navzájom pôsobia a navzájom sa ovplyvňujú.

váha a veľkosť balíkov expedovaných z výrobnéj haly

cena auta a jeho výbava

zdanlivá závislosť – vzťah medzi javmi nie je dôsledkom ich vzájomnej príčinnej súvislosti, ale pôsobenia ďalšieho javu.

nezamestnanosť Slovenska a Španielska

Ak skúmame závislosti medzi kvantitatívnymi znakmi, hovoríme o **korelačnom počte**; pri skúmaní závislostí medzi kvalitatívnymi znakmi hovoríme o **meraní ich asociácie**.

regresná analýza – chceme charakterizovať priebeh závislosti, t.j. odhadnúť **funkčný vzťah**, podľa ktorého sa mení závislá premenná pri zmenách nezávislej premennej.

viacnásobnú (mnohonásobnú) koreláciu a regresiu – viac ako jedna nezávislá premenná.

úplná nezávislosť úplná funkčná závislosť

- kovarianciou;
- korelačným koeficientom.

Korelácia neimplikuje kauzalitu!

Korelačná párová analýza

Kovariancia určuje, či sú premenné X a Y vo vzájomnej lineárnej závislosti.

Medzi premennými X a Y vo výberovom súbore platí:

$$cov(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y}) ,$$

kde x_i, y_i sú hodnoty premenných X a Y v i -tom pozorovaní (na i -tej štatistickej jednotke) a $\bar{x}$ a $\bar{y}$ sú aritmetické priemery premenných X a Y .

Vlastnosti kovariancie:

- $cov(x, y) = cov(y, x) \in R$;
- $cov(x, y) = 0$ reprezentuje lineárnu nezávislosť premenných X a Y , nehovorí nič o iných typoch závislostí;
- $cov(x, y) > 0$, ak medzi premennými X a Y existuje **priama** lineárna závislosť (hodnoty premennej X a Y narastajú súčasne);
- $cov(x, y) < 0$, ak medzi premennými X a Y existuje **nepriama** lineárna závislosť (s nárastom hodnôt premennej X hodnoty premennej Y klesajú).

Nevýhodou kovariancie je, že závisí od mierky, v ktorej sú vyjadrené hodnoty premenných X, Y .

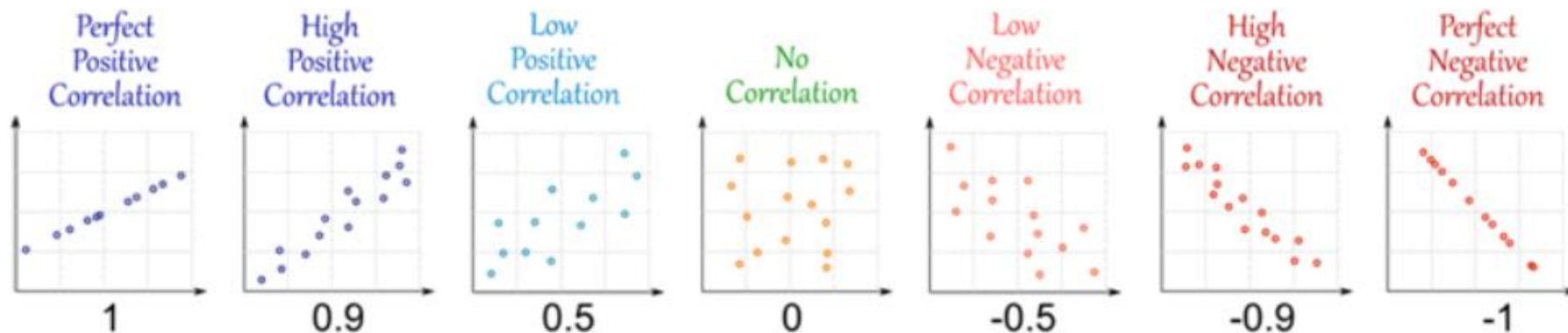
(Výberový) koeficient korelácie $r_{xy} = \frac{cov(x,y)}{s_x \cdot s_y}$ je normovanou mierou vzájomnej lineárnej závislosti – mierou sily, tesnosti lineárneho vzťahu medzi premennými X, Y .

Vlastnosti korelačného koeficientu sú:

- $r_{xy} = r_{yx}$; $r_{xy} \in \langle -1, 1 \rangle$;
- $r_{xy} = 0$ ($cov(x, y) = 0$), ak sú premenné X a Y lineárne nezávislé (nekorelované);
- $r_{xy} > 0$, ak medzi premennými X a Y existuje priama lineárna závislosť;
- $r_{xy} < 0$, ak medzi premennými X a Y existuje nepriama lineárna závislosť;
- ak $|r_{xy}| < 0,3$ hovoríme, že medzi premennými X a Y je malá lineárna závislosť;
- ak $0,3 \leq |r_{xy}| \leq 0,8$, potom medzi X a Y je mierna (stredná) lineárna závislosť;
- ak $|r_{xy}| > 0,8$, potom medzi premennými X a Y je silná (tesná) lineárna závislosť;
- $r_{xy} \rightarrow +1$, znamená vysokú (tesnú) priamu lineárnu závislosť medzi X a Y ;
- $r_{xy} \rightarrow -1$, znamená vysokú (tesnú) nepriamu lineárnu závislosť medzi X a Y ;
- $|r_{xy}| = 1$, ak medzi premennými X a Y existuje priama lineárna závislosť (namerané hodnoty ležia na modelovej priamke)- deterministická závislosť.

Hodnota korelačného koeficientu r_{xy} je nezávislá na merných jednotkách!

Korelácia môže byť priama (pozitívna) alebo nepriama (negatívna):



<https://www.mathsisfun.com>

Štatistické funkcie v Exceli:

hodnotu kovariancie určíme `=COVARIANCE.P (pole1; pole2)`,

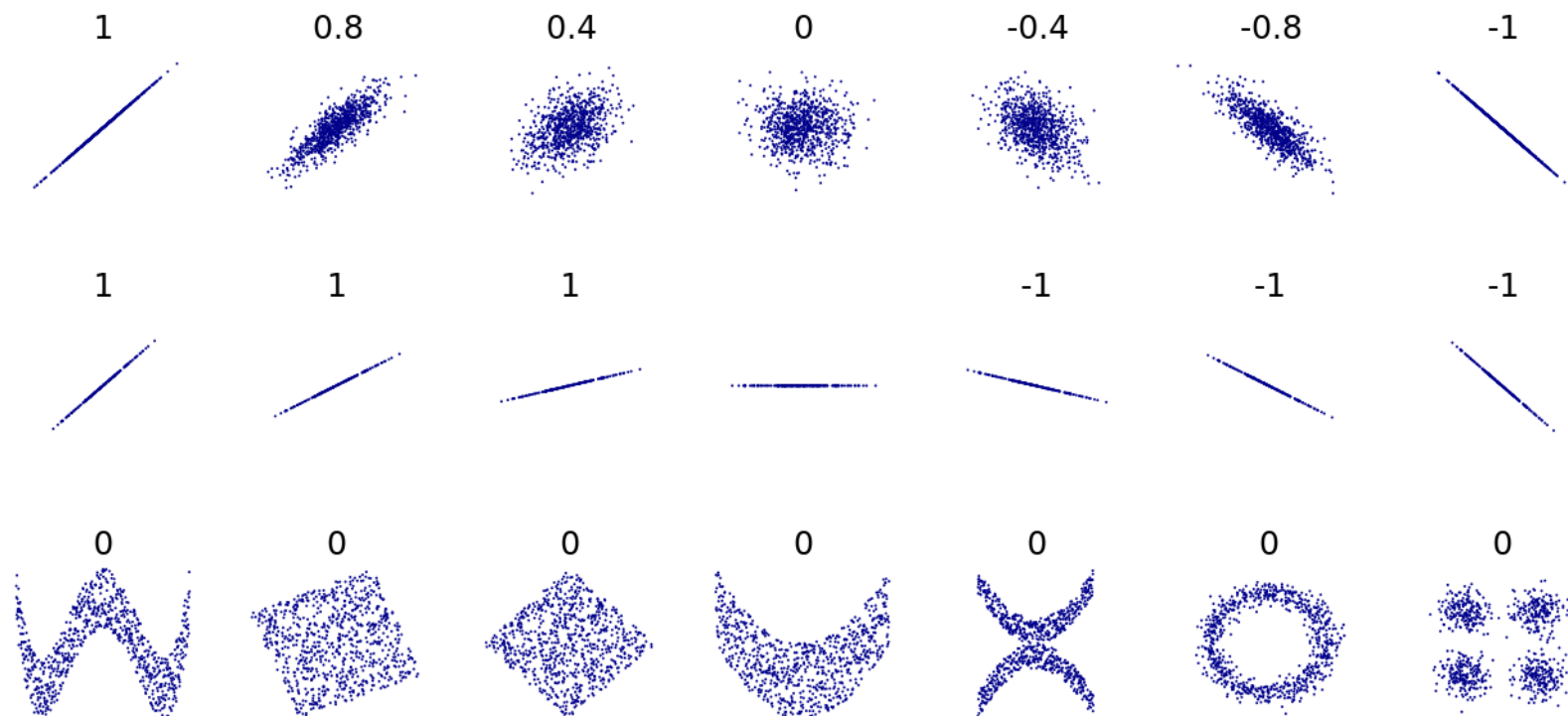
hodnotu korelačného koeficientu `=CORREL (pole1; pole2)` alebo

`=PEARSON(pole1; pole2)`,

kde pole1 a pole2 predstavujú oblasť 1. premennej a oblasť 2. premennej.

Na poradí zadávaných oblastí nezáleží.

Pozor, korelačný koeficient hovorí iba o **miere tesnosti lineárnej závislosti** náhodných premenných X a Y ! Nič nehovorí o iných typoch závislostí.



Zdroj: https://commons.wikimedia.org/wiki/File:Correlation_examples2.svg

Regresná párová analýza

sa zaoberá tvarom (charakterom) závislosti dvoch náhodných premenných X a Y ; úlohou je nájsť vhodný matematický model, ktorý vyjadruje závislosť medzi závislou (vysvetľovanou) premennou Y a nezávislou (vysvetľujúcou) premennou X , t.j. zvoliť vhodný typ regresnej funkcie a následne odhadnúť jej parametre, pričom východiskom sú empirické (výberové) údaje.

Regresná lineárna funkcia môže byť ľubovoľná funkcia $Y = f(X, B_0, B_1)$.

Lineárny vzťah medzi X a Y vyjadríme funkčnou závislosťou $Y = B_0 + B_1X$.

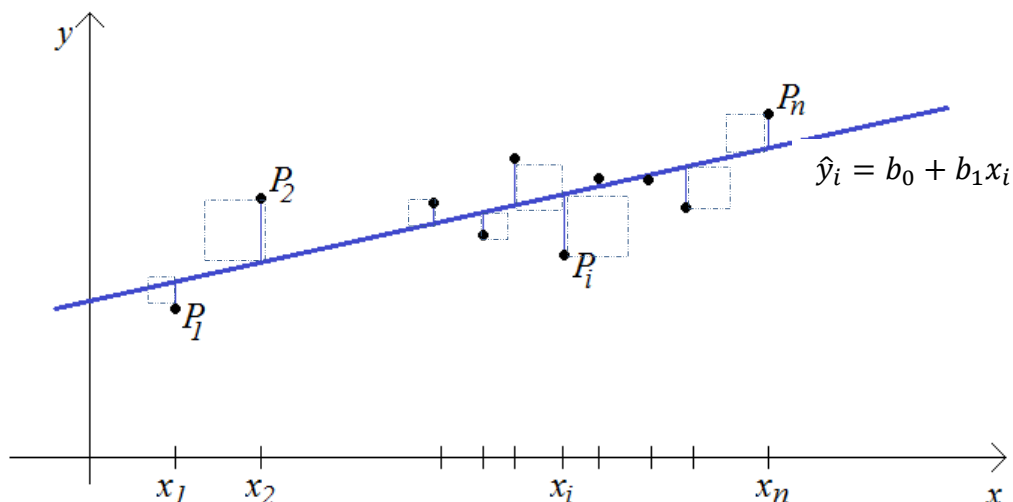
Bodový odhad označíme $\hat{y}_i = b_0 + b_1x_i$.

MNŠ hľadáme minimum súčtu reziduálnych štvorcov (štvorcov odchýlok nameraných a modelovaných hodnôt).

Hľadáme lokálne minimum funkcie

$$G(b_0, b_1) = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \\ \sum_{i=1}^n (y_i - (b_0 + b_1x_i))^2 = \sum_{i=1}^n (y_i - b_0 - b_1x_i)^2 .$$

Metóda najmenších štvorcov (MNŠ)



Hľadáme minimálny rozdiel medzi skutočne nameranou (empirickou) hodnotou závislej premennej y_i a jej vyrovnanou (teoretickou, odhadovanou, modelovanou) hodnotou $\hat{y}_i$.

$$G(b_0, b_1) = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

Z nutnej podmienky existencie extrému

$$\frac{\partial G}{\partial b_0} = \sum_{i=1}^n 2(y_i - b_0 - b_1 x_i) \cdot (-1) = 0$$

$$\frac{\partial G}{\partial b_1} = \sum_{i=1}^n 2(y_i - b_0 - b_1 x_i) \cdot (-x_i) = 0$$

$\Rightarrow$

systém normálnych rovníc

$$nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i;$$

$$b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

Výpočet neznámych parametrov b_0 a b_1 v regresnej funkcii $\hat{y}_i = b_0 + b_1 x_i$

- riešenie systému normálnych rovníc (napr. pomocou Cramerovho pravidla)

$$b_0 = \frac{D_1}{D}, b_1 = \frac{D_2}{D}, \text{ kde } D \neq 0$$

- riešenie pomocou maticovej rovnice

matica plánu $X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, vektor odhadov parametrov $B = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$, vektor meraní $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$

$$XB = Y \quad / \cdot X^T \text{ zľava}$$

$$X^T XB = X^T Y$$

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

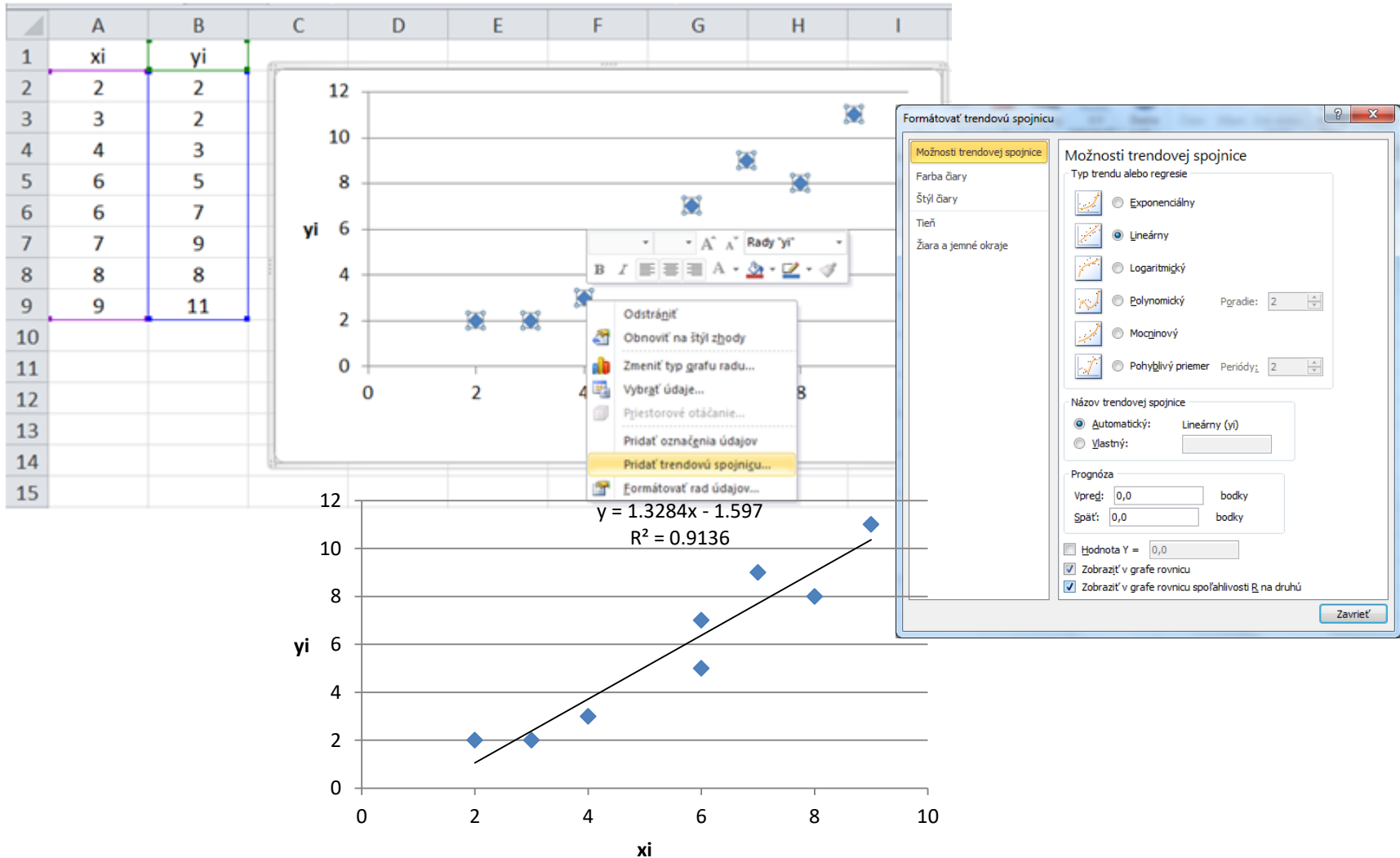
$$B = (X^T X)^{-1}X^T Y$$

X^T je transponovaná matica

$X^T X$ je štvorcová matica, ak $\exists$ k nej $(X^T X)^{-1}$

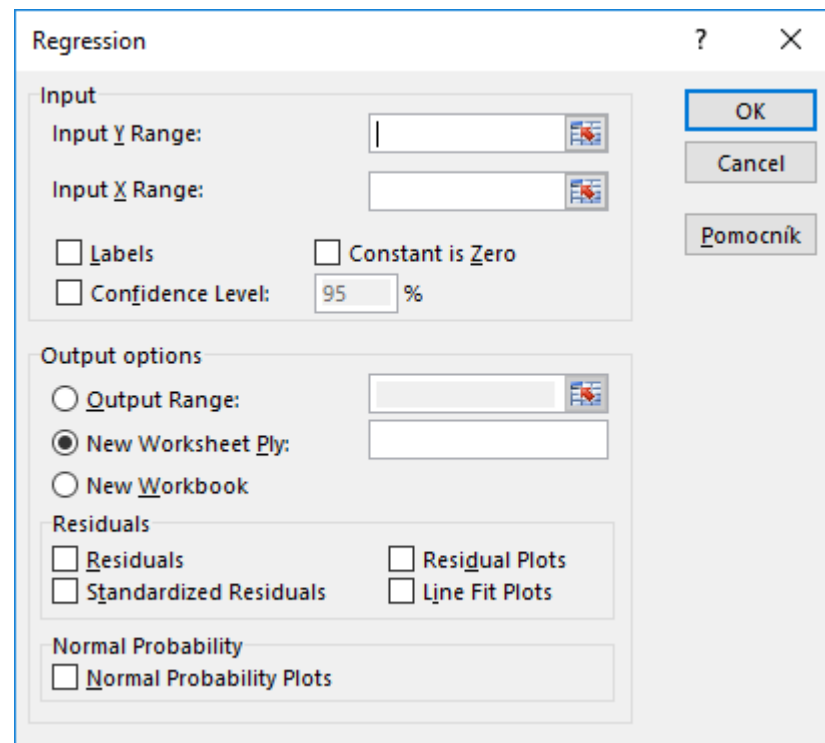
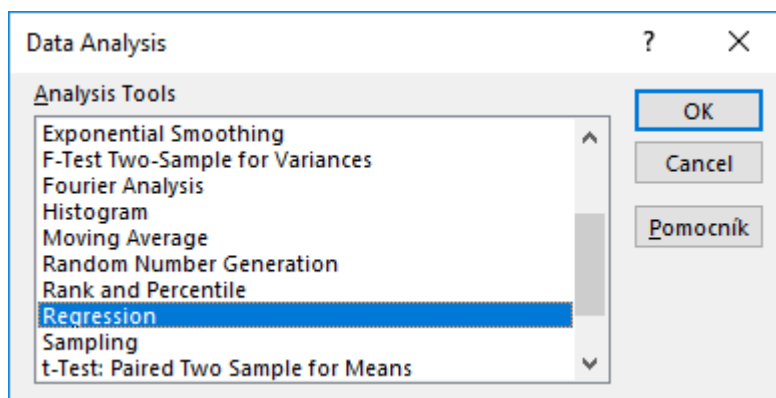
Riešenie v Exceli

- grafické riešenie – vyrovňavajúcu regresnú priamku preložíme bodmi VS



Riešenie v Exceli

- regresná analýza ako súčasť analýzy dát



- preddefinované štatistické funkcie v Exceli

$b_0 = \text{INTERCEPT}(\text{známe_y}; \text{známe_x}),$

$b_1 = \text{SLOPE}(\text{známe_y}; \text{známe_x}),$ dôležité je poradie zadávaných premenných

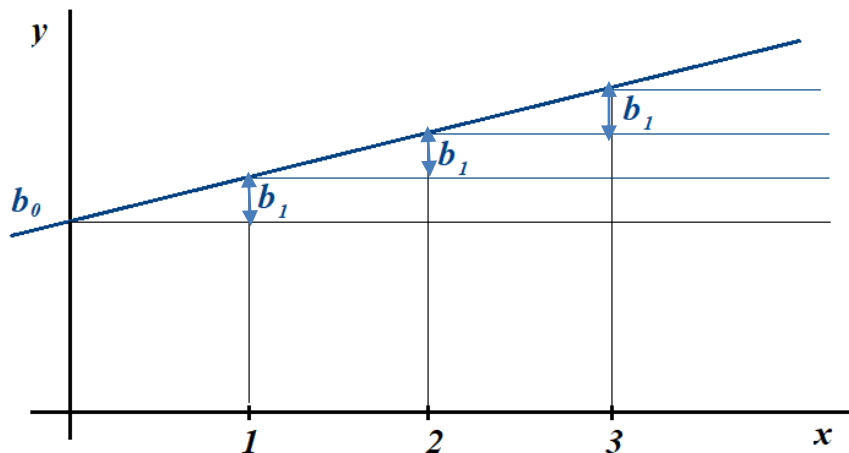
Interpretácia parametrov

Parameter b_0 nazývame **lokujúca konštanta**, znamená hodnotu premennej Y za predpokladu, že premenná X sa rovná nule.

Geometrický význam: priesečník priamky s y-ovou osou

Parameter b_1 nazývame **regresný koeficient**, znamená priemerný prírastok (úbytok) premennej Y zodpovedajúci zmene premennej X o jednu jednotku.

Geometrický význam: smernica regresnej priamky, t.j. tangens orientovaného uhla, ktorý priamka zviera s kladnou x-ovou poloosou

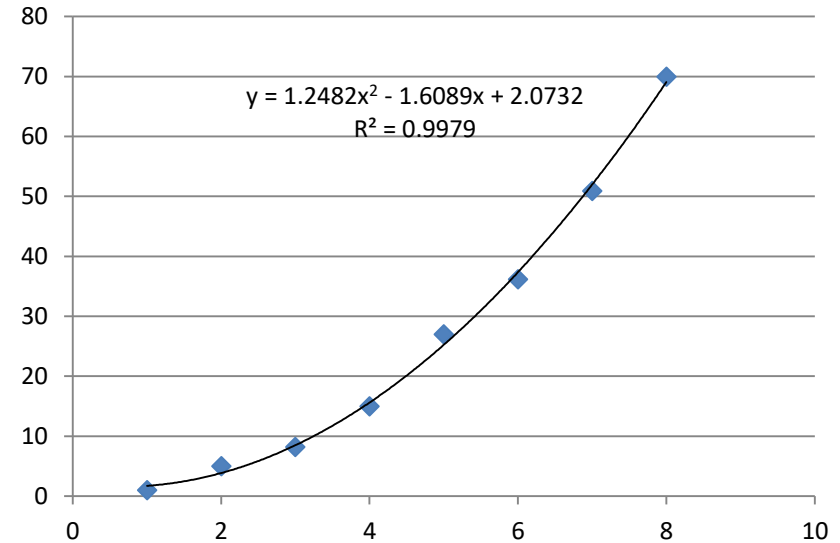


Regresný koeficient b_1 má vždy rovnaké znamienko ako koeficient korelácie r_{xy} .
Regresná priamka prechádza bodom $[\bar{x}, \bar{y}]$.

Iné typy regresných funkcií – kvadratická regresia

Systém normálnych rovníc

$$\begin{aligned}nb_0 + b_1 \sum_{i=1}^n x_i + b_2 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i; \\b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 + b_2 \sum_{i=1}^n x_i^3 &= \sum_{i=1}^n x_i y_i \\b_0 \sum_{i=1}^n x_i^2 + b_1 \sum_{i=1}^n x_i^3 + b_2 \sum_{i=1}^n x_i^4 &= \sum_{i=1}^n x_i^2 y_i\end{aligned}$$



Iné typy regresných funkcií – transformácia na lineárnu regresia

Mnohé funkcie, ktoré nie sú lineárne, možno transformáciou na lineárne upraviť a potom použiť MNŠ

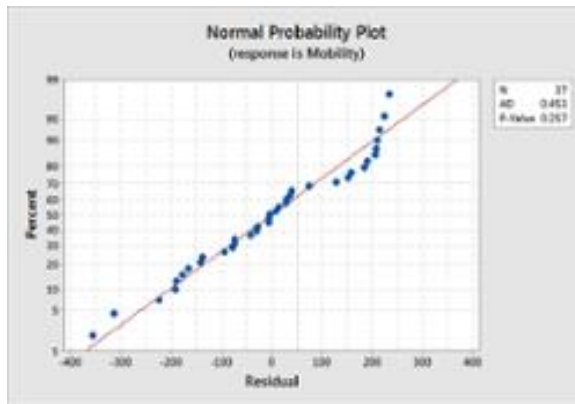
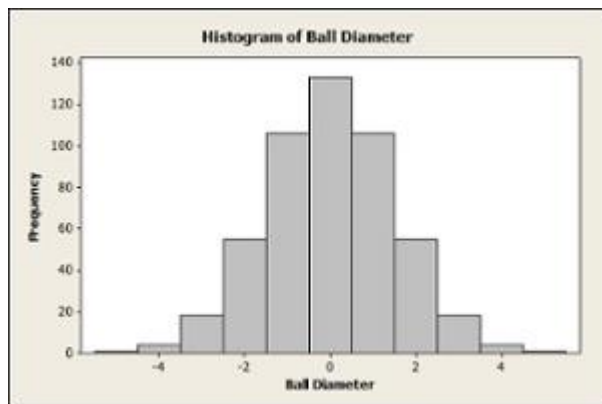
- $y = a \cdot e^{bx} \rightarrow \ln y = \ln a + bx \rightarrow z = A + B \cdot u$
pôvodné parametre $a = e^A, b = B$
- $y = a \cdot x^b \rightarrow \ln y = \ln a + b \cdot \ln x \rightarrow z = A + B \cdot u$
pôvodné parametre $a = e^A, b = B$
- $y = a \cdot b^x \rightarrow \ln y = \ln a + \ln b \cdot x \rightarrow z = A + B \cdot u$
pôvodné parametre $a = e^A, b = e^B$
- $y = a + \frac{b}{x} \rightarrow \frac{1}{x} = u \rightarrow z = A + B \cdot u$
pôvodné parametre $a = A, b = B$

Grafická analýza vhodnosti regresného modelu

Požiadavky na reziduá pre „dobrý“ regresný model:

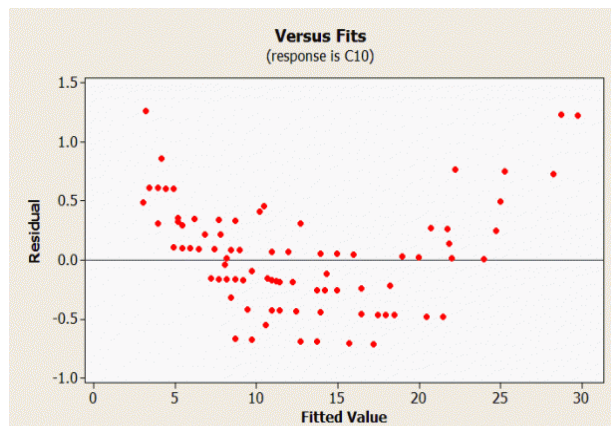
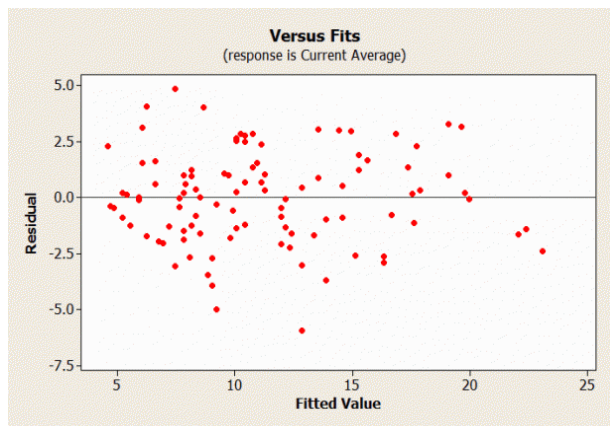
1. $N(0;\sigma^2)$
2. homoskedasticita
3. nezávislosť

- **kontrola normality reziduí** $e_i = y_i - \hat{y}_i$
histogram štandardizovaných reziduí a normálny pravdepodobnostný graf



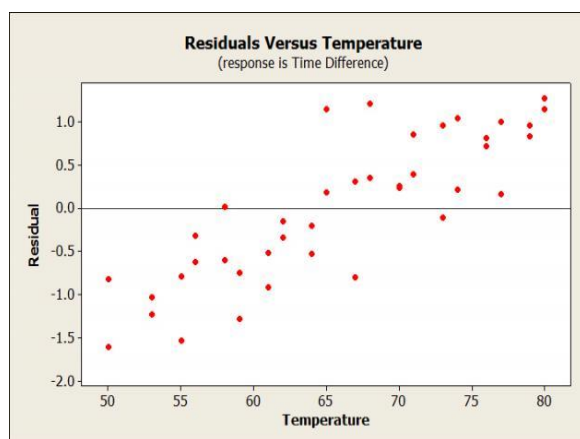
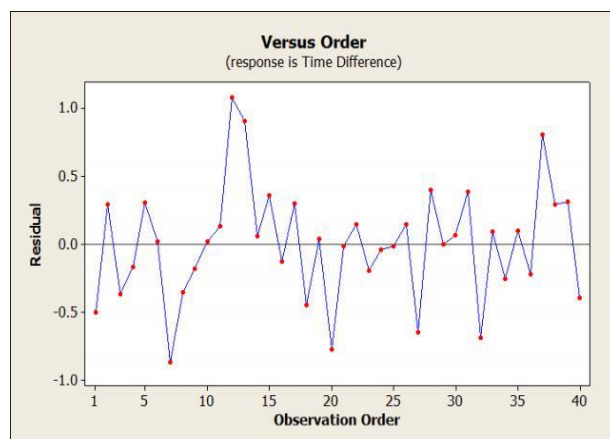
Grafická analýza vhodnosti regresného modelu

- graf zobrazujúci vzťah reziduí a vyrovnaných hodnôt premennej Y



residuals versus fit

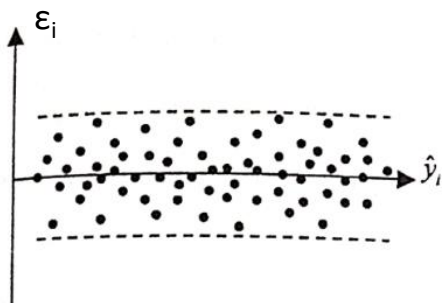
- graf zobrazujúci vzťah reziduí a hodnôt premennej X
správanie sa reziduí v čase by malo byť konštantné



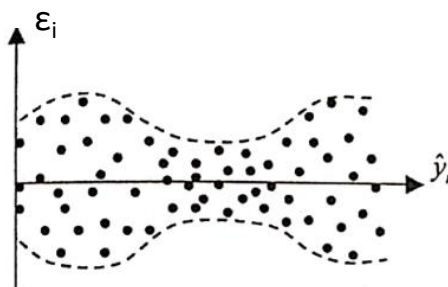
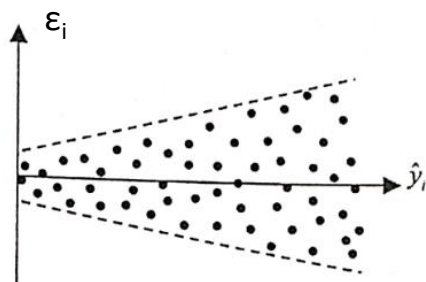
residuals versus order

Rôzne rozloženie štandardizovaných (studendizovaných) reziduí vzhľadom na hodnoty regresnej funkcie

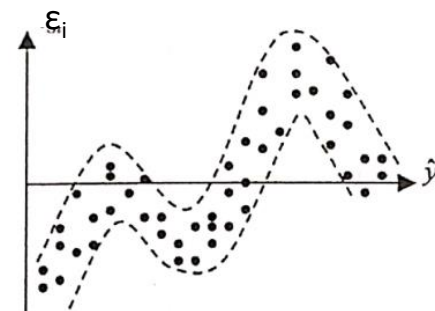
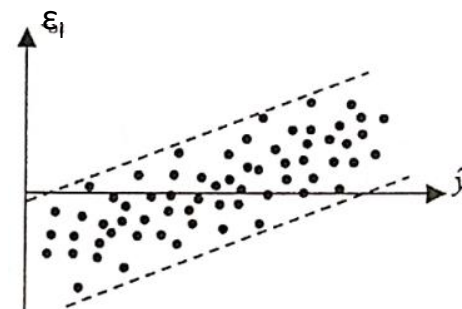
a) reziduá neprejavujú systematické tendencie a spadajú do horizontálneho pásu sústredeného okolo nuly



b) na rôznych úrovniach odhadnutých hodnôt Y je rôzna rozptýlenosť reziduí (porušenie homoskedaticity náhodnej zložky)



d) v správaní reziduí pozorujeme systematické tendencie (porušenie nezávislosti náhodných chýb)

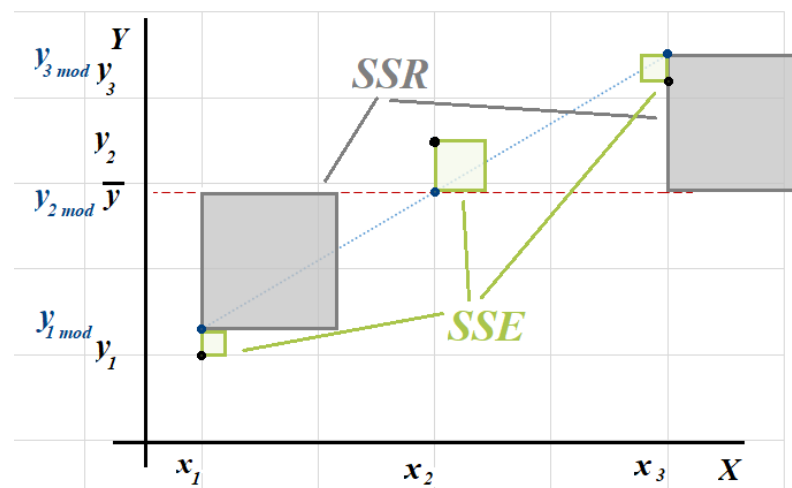
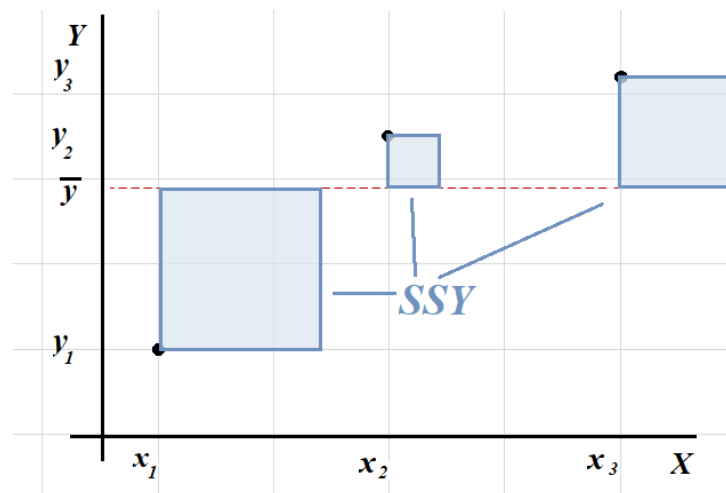
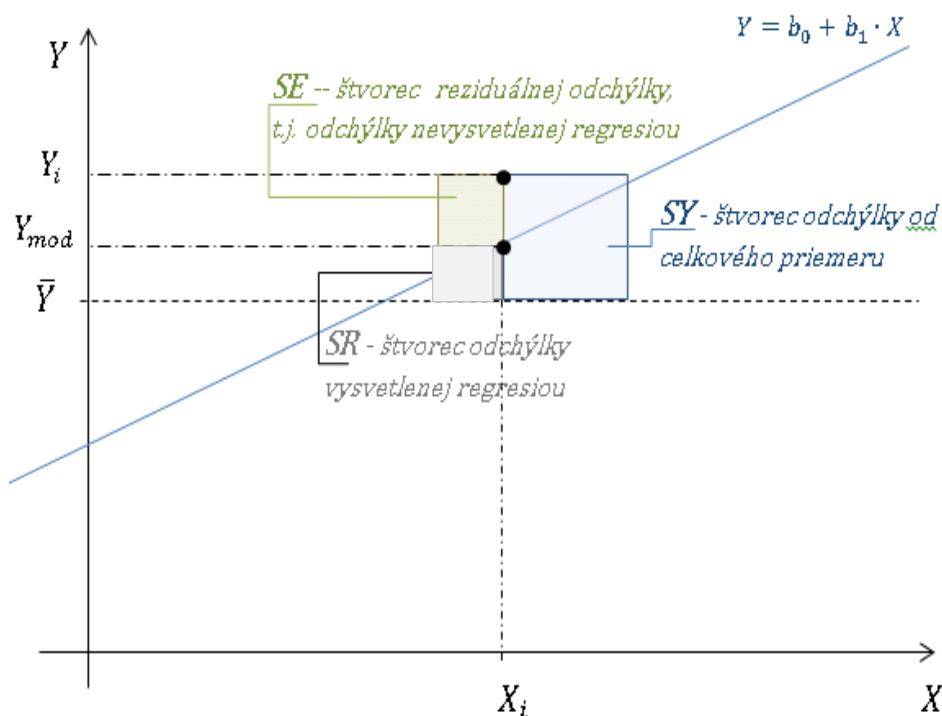


Rozklad celkovej variability na zložky reziduálnu a regresnú

$$SSY = SSE + SSR$$

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

(pre prípad i-teho pozorovania)



SSY - celková variabilita premennej Y - celkový súčet štvorcov - sum of squares total

SSR - variabilita vysvetlená regresným modelom - sum of squares due to regression

SSE - reziduálny súčet štvorcov - variabilita nevysvetlená regresným modelom - sum of squares due to error

Porovnanie rôznych regresných modelov

Pri riešení regresnej úlohy prichádza do úvahy viacero regresných funkcií, preto sa skúma, ktorý z týchto modelov najlepšie vystihuje závislosť medzi premennými X a Y .

To sa dá merať rôznymi charakteristikami:

- súčtom reziduálnych štvorcov $SSE = \sum_i (y_i - \hat{y}_i)^2$ – cieľom MNŠ bolo minimalizovať jeho hodnotu, t.j. minimalizovať variabilitu nevysvetlenú regresným modelom,
- reziduálnym rozptylom $s_{rez}^2 = \frac{SSE}{n-2}$,
- štandardnou odchýlkou rezíduí $s_{rez} = \sqrt{s_{rez}^2}$,
- koeficientom determinácie R^2 ,
- preveriť rôznymi testami.

Z dvoch regresných modelov, ktoré by teoreticky prichádzali do úvahy, je lepší na základe analýzy rezíduí ten, ktorý má menšie hodnoty SSE , s_{rez}^2 a s_{rez} a ten, ktorý má väčšiu hodnotu koeficientu determinácie R^2 .

Koeficient (index) determinácie (R-square)

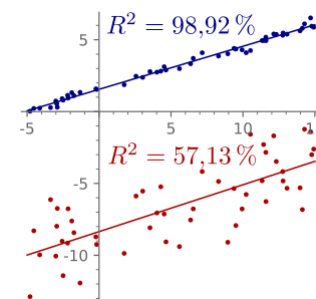
ukazuje, koľko percent celkovej variácie (rozptylu) premennej Y je vysvetlenej modelom. Je definovaný podielom variability vysvetlenej modelom k celkovej variabilite

$$R^2 = \frac{SSR}{SSY} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}; R^2 = 1 - \frac{SSE}{SSY}; R^2 \in \langle 0,1 \rangle; R^2 = r_{xy}^2$$

Čím viac sa R^2 blíži k 1, tým je praktické použitie modelu rozumnejšie.

$R^2 < 0,7$ - použitie regresného modelu má teoretický význam.

$R^2 > 0,9$ - „slušný“ model pre praktické využitie.



Platí, že čím väčší je podiel variability SSR na celkovej variabilite SSY , tým je závislosť X a Y silnejšia - z dvoch modelov „kvalitnejší“ ten, ktorý má väčší koeficient determinácie.

R^2 vyjadruje stupeň príčinnej závislosti Y od X , hovorí, nakoľko je nami nájdený model „dobrý“, t.j. vyjadruje, aký podiel variability závislej premennej Y regresný model vysvetľuje lineárnou závislosťou od variability nezávislej premennej X .

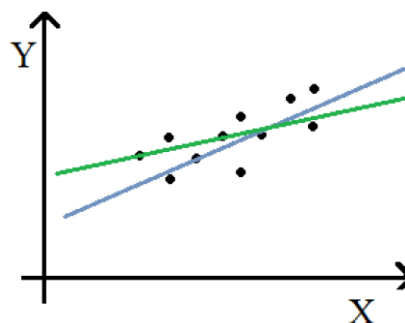
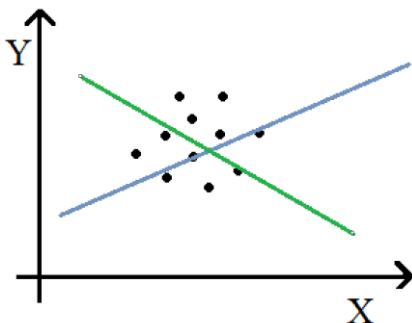
Upravený (korigovaný) koeficient determinácie (Adjusted R-square)

$$R_{adj}^2 = 1 - (1 - R^2) \frac{n-1}{n-p-1}, \text{ kde } n \text{ je počet pozorovaní, } p \text{ je počet vysvetľujúcich premenných}$$

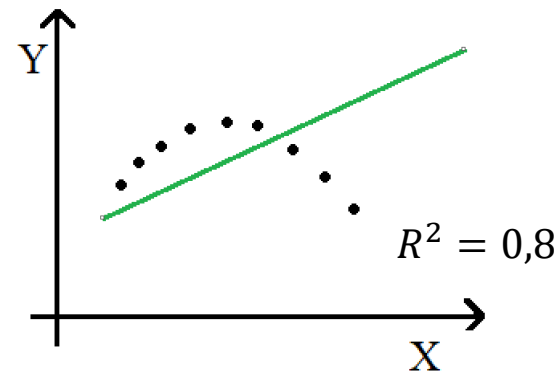
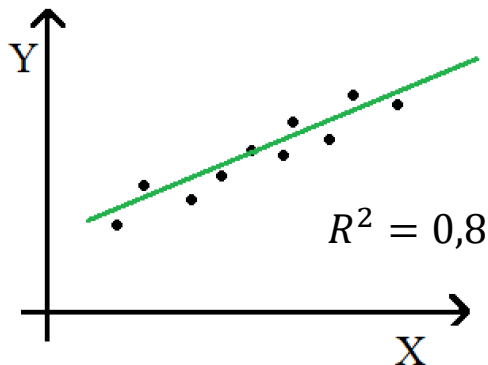
Táto charakteristika penalizuje model za počet vysvetľujúcich premenných – môže sa použiť aj na porovnanie kvality modelov s rôznym počtom vysvetľujúcich premenných; výrazný rozdiel medzi R^2 a R_{adj}^2 indikuje, že do modelu bolo zahrnutých príliš veľa premenných; môže mať zápornú hodnotu

Nedostatky odhadu lineárnej regresie

- **Vplyv variability nezávislej premennej na odhad regresného koeficientu** - Malá rozptýlenosť premennej X má za následok nespoľahlivý odhad regresného koeficientu (väčšia variabilita premennej X zabezpečí spoľahlivejší odhad koeficientu b_1)



- **Preceňovanie charakteristík kvality** - Posudzovať kvalitu modelu len na základe mier kvality odhadu (korelačný koeficient, koeficient determinácie) môže byť zavádzajúce.



Nedostatky odhadu lineárnej regresie

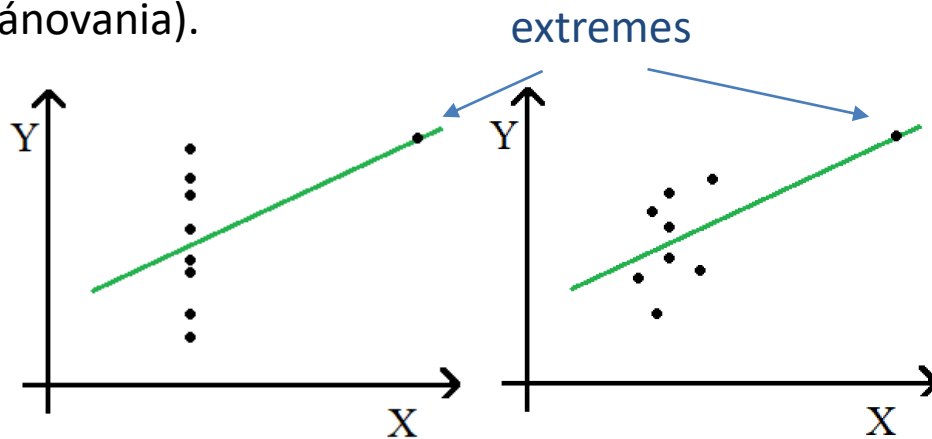
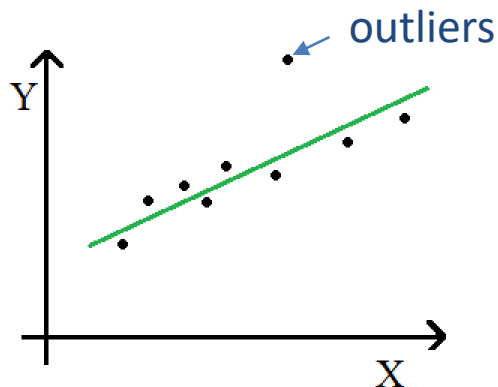
- **Odľahlá hodnota v nameraných dátach**

Odľahlý bod je taký, ktorý leží mimo základnej konfigurácie bodov v grafe. Odľahlý bod nazývame vplyvným, keď sa po jeho odstránení podstatne zmení poloha regresnej krivky.

Odľahlé body sa rozdeľujú na:

A. Vybočujúce hodnoty (**outliers**) sú vysoké alebo nízke hodnoty y_i , ktoré sa zásadne líšia od ostatných hodnôt y . Prítomnosť takýchto bodov významne zvyšuje hodnotu reziduálneho rozptylu, čím významne zhoršuje presnosť predikčných vlastností modelu (tieto body je potrebné identifikovať, zistiť príčiny odľahlosti a znovu premerať podozrivý bod, prípadne vylúčiť).

B. Extrémne hodnoty (**extremes**) sú mimoriadne hodnoty x_i , ktoré predstavujú zásadne odlišnú kombináciu hodnôt x . Body, ktoré sú odľahlé v smere x sú často vplyvné (sú dôsledkom nevhodného plánovania).



Použitie regresnej priamky v praxi

Rovnica regresnej priamky umožňuje predikovať - odhadovať hodnoty premennej Y , ak poznáme hodnotu premennej X .

V Exceli je bodový odhad hodnoty závislej premennej Y pre konkrétnu hodnotu premennej X (z intervalu skúmaných hodnôt) výstupom funkcie

= FORECAST (x;známe_y; známe_x),

kde x je ľubovoľná hodnota premennej X pre ktorú odhadujeme hodnotu Y , známe y a známe x sú namerané údaje, cez ktoré sme prekladali regresnú priamku.

Praktickú vypovedaciu schopnosť majú len tie hodnoty x , ktoré sú z intervalu nameraných hodnôt $\langle x_{min}, x_{max} \rangle$, prípadne „kúsok“ pod minimom alebo nad maximom.

Výstup analytického nástroja Regresia (Regression)

11

SUMMARY OUTPUT

12

13

Regression Statistics

14

Multiple R0,955840586

15

R Square0,913631226

16

Adjusted R Square0,89923643

17

Standard Error1,078971318

18

Observations8

19

20

ANOVA

21

22

Regression

23

Residual

24

Total

25

26

27

Intercept

28

xi

29

30

31

32

RESIDUAL OUTPUT

33

34

35

36

37

38

39

40

41

42

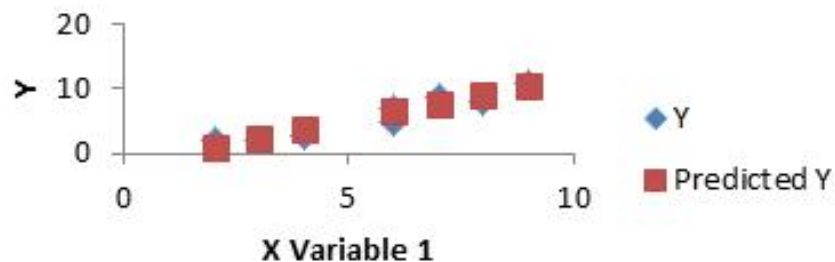
R²

Štatistická významnosť modelu - celkový F test (test linearity)

Štatistická významnosť regresného koeficientu (parametra lin. modelu)

testy

X Variable 1 Line Fit Plot



Celkový F-test

Test štatistickej významnosti regresného modelu

H_0 : Lineárny model nie je štatisticky významný (X a Y nie sú lineárne závislé).

H_1 : Lineárny model je štatisticky významný (X a Y sú lineárne závislé).

je založený na rozklade súčtu štvorcov odchýlok SSY na:

1. SSE (sum of squares due error) – súčet reziduálnych štvorcov
2. SSR (sum of squares due to regression) – suma štvorcov odchýlok vysvetlenej regresiou

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

↑ SSY ↑ SSE ↑ SSR

y_i je nameraná hodnota, $\hat{y}_i$ je modelovaná hodnota, $\bar{y}$ je priemer hodnôt závislej premennej Y .

Testovacia štatistika $F = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{s_R^2}{s_E^2} = \frac{MSR}{MSE} \sim F(1; n - 2)$

MSR je priemerný štvorec regresie (mean square of regression) a MSE je priemerný štvorec chýb (mean square of errors).

F-test je jednostranný (pravostranný) test.

Individuálny t-test o lineárnej nezávislosti X a Y

Regresný koeficient B_1 je smernicou priamky, vyjadruje priemernú zmenu Y pri zmene X o jednu jednotku.

Ak $B_1 = 0$, potom regresná priamka je rovnobežná s x-ovou osou, t.j. po zmene nezávislej premennej X sa hodnoty závislej premennej Y nemenia, t.j. medzi X a Y neexistuje lineárna závislosť.

Testujeme hypotézy

$H_0: B_1 = 0$ (X a Y nie sú lineárne závislé).

$H_1: B_1 \neq 0$ (X a Y sú lineárne závislé).

Testovacia štatistika $t = \frac{b_1}{s_{rez}} \sqrt{\sum_i (x_i - \bar{x})^2} = \frac{b_1}{\frac{s_{rez}}{\sqrt{\sum_i (x_i - \bar{x})^2}}} = \frac{b_1}{SE(b_1)} \sim t(n - 2)$, kde

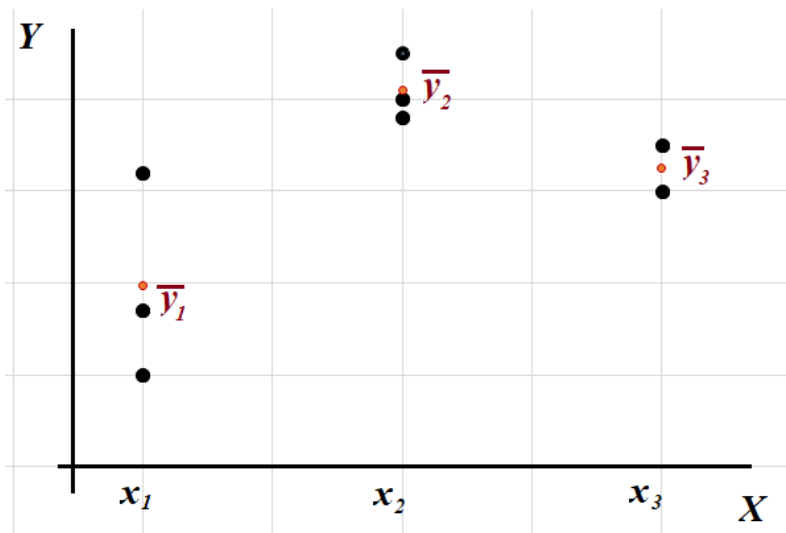
b_1 je výberový regresný koeficient,

$SE(b_1)$ je štandardná chyba odhadu koeficientu b_1 ,

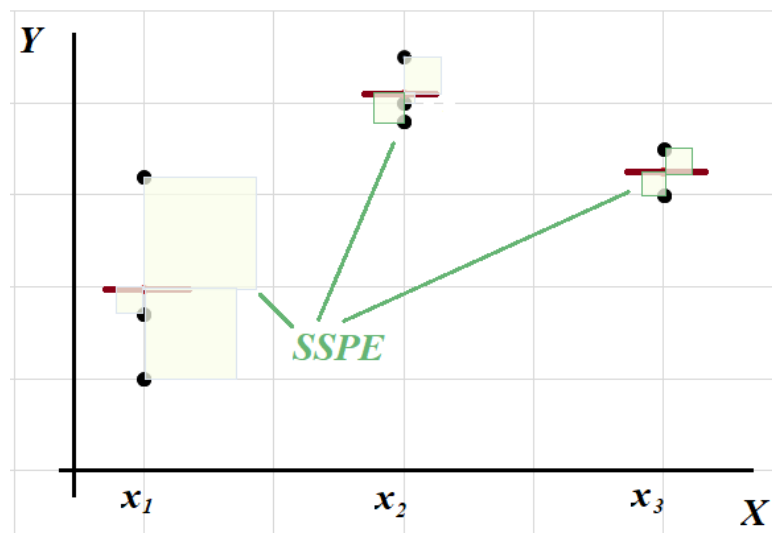
s_{rez} je štandardná odchýlka rezíduí.

Rozklad súčtu reziduálnych štvorcov na zložky Pure Error a Lack of Fit

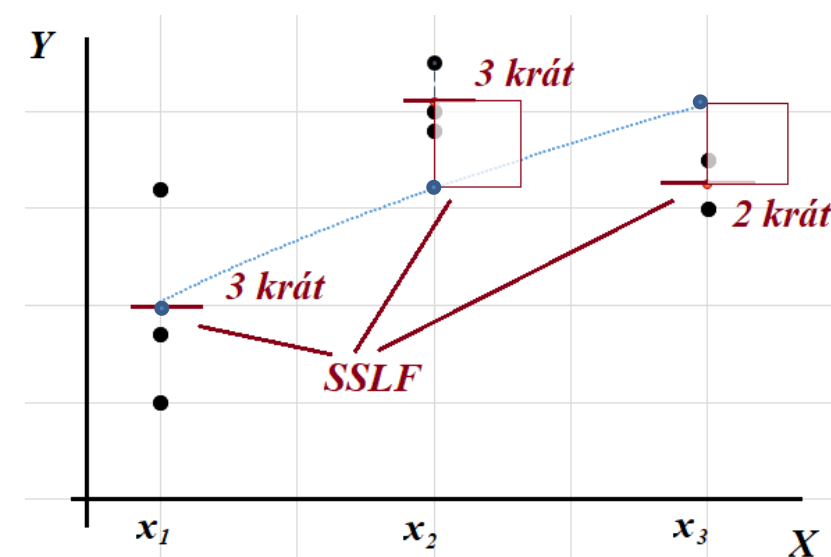
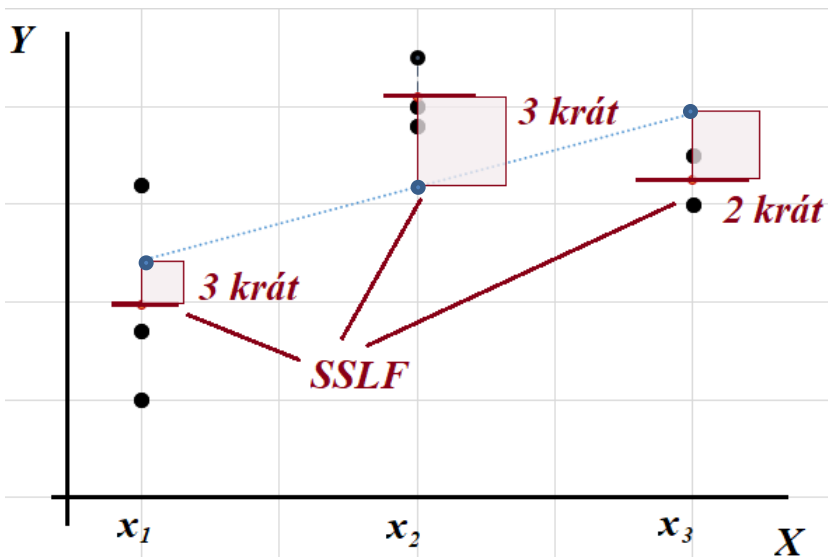
$$SSE = SSPE + SSLF$$



regresným modelom je priamka



regresným modelom je parabola



Lack of fit („nedostatok vhodnosti modelu“)

Test adekvátnosti modelu – vyžaduje opakované merania v bodoch x_i a je založený na rozklade súčtu reziduálnych štvorcov SSE na:

1. **SSPE** (sum of squares due to pure error) - chybu pri meraní (čistú chybu)
2. **SSLF** (sum of squares due to lack of fit) - chybu modelu, ktorá vyjadruje mieru nezhody empirického modelu s teoretickým.

$$\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^m \sum_{j=1}^{n_i} (\bar{y}_i - \hat{y}_i)^2$$

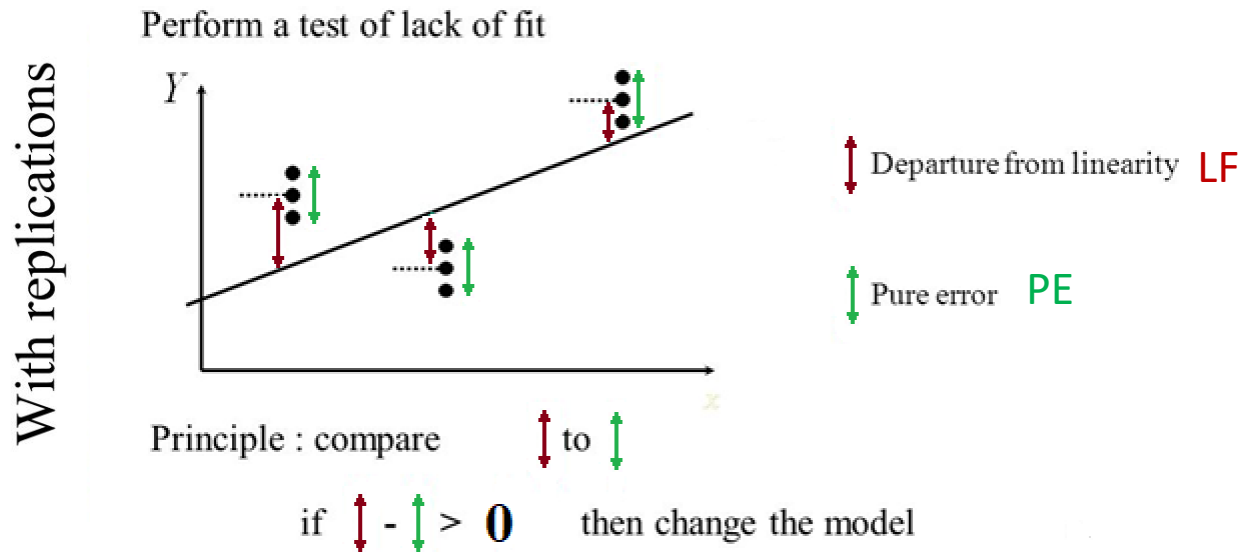
↑**SSE** ↑**SSPE** ↑**SSLF**

kde y_{ij} označuje j -te pozorovanie pre x_i kde $i = 1, 2, \dots, m$ a $j = 1, 2, \dots, n_i$ pričom n_i je počet opakovaní x_i a m je počet úrovní premennej X (jej rôznych hodnôt);

Pre rovnaký počet opakovaných meraní $SSLF = \sum_{i=1}^n n_i (\bar{y}_i - \hat{y}_i)^2$

Regresný model považujeme za adekvátny (dostatočný), ak sa hodnoty SSPE a SSLF štatisticky významne nelíšia. Ak je rozdiel medzi SSLF a SSPE významný, považujeme model za nedostatočný.

Nechceme, aby suma štvorcov lack of fit bola „veľká“ zložka reziduálnej chyby.



Testujeme hypotézy

H_0 : $SSLF - SSPE = 0$ (lineárny model je dostatočný, je adekvátny – no lack of fit).

H_1 : $SSLF - SSPE > 0$ (lineárny model nie je dostatočný, nie je adekvátny – lack of fit).

$$\text{Testovacia štatistika } F = \frac{\frac{SSLF}{df_{LF}}}{\frac{SSPE}{df_{PE}}} = \frac{s_{LF}^2}{s_{PE}^2} = \frac{MSLF}{MSPE} \sim F(df_{LF}, df_{PE})$$

$$df_E = n - p, df_{PE} = \sum_{i=1}^m (n_j - 1), df_{LF} = df_E - df_{PE} \text{ a}$$

p je počet parametrov v regresnej funkcii (pri lineárnom modeli 2),
 n je počet nameraných dvojíc hodnôt, n_j počet meraní pre rovnaké x_i ,
 m je počet rôznych hodnôt x_i .

Regression Analysis: length versus age

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	32966	32965.8	210.66	0.000
age	1	32966	32965.8	210.66	0.000
Error	76	11893	156.5		
Lack-of-Fit	4	3080	770.0	6.29	0.000
Pure Error	72	8813	122.4		
Total	77	44859			

Ak zamietame H_0 o adekvátnosti modelu v prospech alternatívy H_1 , že model nie je adekvátny, je potrebné zostrojiť zložitejší model – pridáme ďalšie premenné, vyššie mocniny premenných, ich interakcie...

Príklad

U deviatich náhodne vybraných otcov bola zistená ich výška a výška ich dospelých prvorodených synov s týmito výsledkami:

výška otca (x_i)	174	180	176	168	182	188	176	177	174
výška syna (y_i)	177	182	176	173	180	191	179	181	176

Odhadnite koeficienty regresnej priamky.

Príklad

V tabuľke sú výsledky meraní otáčok (min^{-1}) a výkonu motora (kW) :

otáčky (min^{-1})	2000	2500	3000	3500	4000
výkon (kW)	29	43	55	64	71

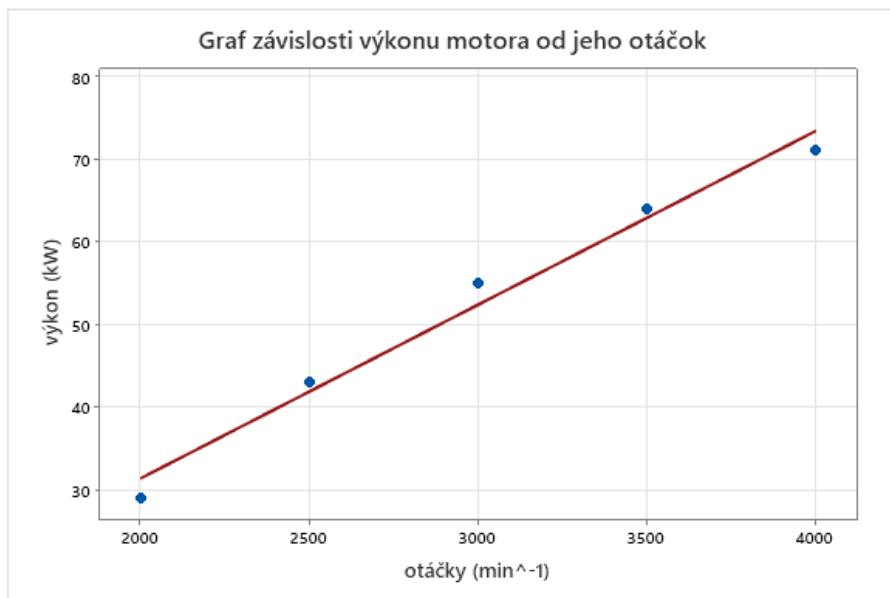
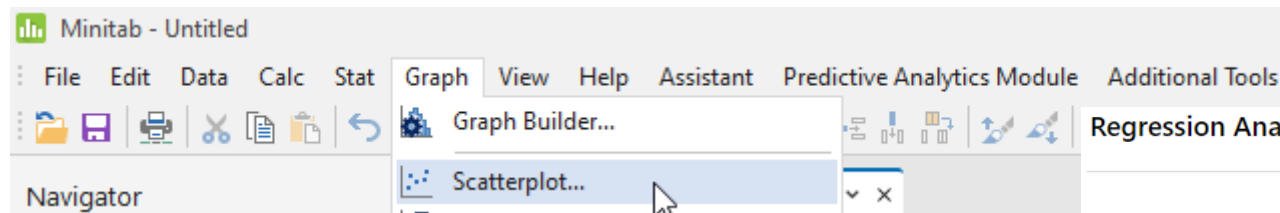
$$\text{výkon [kW]} = \frac{\text{moment [Nm]} \cdot \text{otáčky [1/min]}}{9\,549}$$

Určte výberový korelačný koeficient výkonu motora a jeho otáčok. Odhadnite koeficienty regresnej priamky.

Riešenie príkladu v Minitabe

otáčky (min ⁻¹)	2000	2500	3000	3500	4000
výkon (kW)	29	43	55	64	71

$$\text{výkon [kW]} = \frac{\text{moment [Nm]} \cdot \text{otáčky [1/min]}}{9\,549}$$



Regression Analysis: výkon (kW) versus otáčky (min⁻¹)

Regression Equation

$$\text{výkon (kW)} = -10,60 + 0,02100 \text{ otáčky (min}^{-1}\text{)}$$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-10,60	5,12	-2,07	0,130	
otáčky (min ⁻¹)	0,02100	0,00166	12,64	0,001	1,00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2,62679	98,16%	97,54%	92,21%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1102,50	1102,50	159,78	0,001
otáčky (min ⁻¹)	1	1102,50	1102,50	159,78	0,001
Error	3	20,70	6,90		
Total	4	1123,20			



Navigator

Scatterplot of výkon (kW)

Regression Analysis: výkon

Basic Statistics

Regression

ANOVA

DOE

Control Charts

Quality Tools

Reliability/Survival

Predictive Analytics

Multivariate

Time Series

Fitted Line Plot...

Regression

Nonlinear Regression...

Stability Study

Orthogonal Regression...

Partial Least Squares...

Binary Fitted Line Plot.

Fit Regression Model...

Best Subsets...

Predict...

Factorial Plots...

Contour Plot...

Surface Plot...

Fit Regression Model

Model the relationship between categorical or continuous predictors and one response. Easily include interaction and polynomial terms, or transform the response if needed.

Regression

C1 otáčky (min⁻¹)

C2 výkon (kW)

C3 FITS

C4 RESI

C5 SRES

C6 COEF

Responses:

'výkon (kW)'

Continuous predictors:

'otáčky (min⁻¹)'

Regression: Results

Display of results: Simple tables

☒ Method☒ Analysis of variance☒ Model summary☒ Coefficients: Default☒ Regression equation: Separate☒ Fits and diagnostics: Only☐ Durbin-Watson statistic

Help

Regression: Graphs

Effects Plots

☐ Pareto

Residuals for plots: Standardized

Residuals plots

☐ Individual plots☐ Histogram of residuals☐ Normal probability plot of residuals☐ Residuals versus fits☐ Residuals versus order☒ Four in one

Residuals versus the variables:

Select

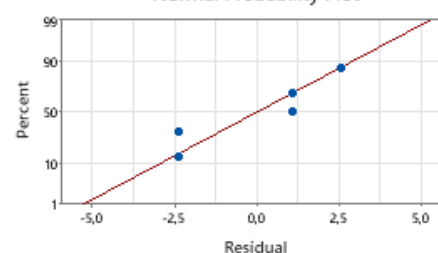
Help

OK

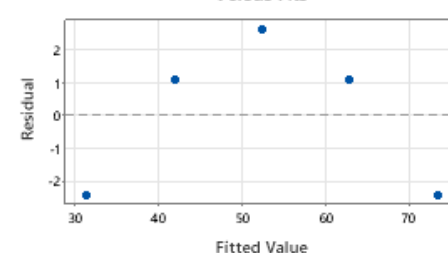
Cancel

Residual Plots for výkon (kW)

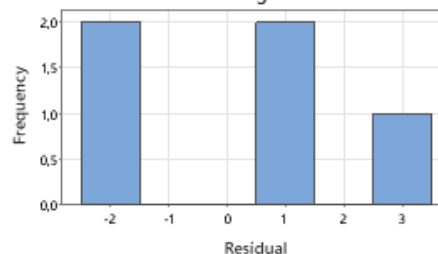
Normal Probability Plot



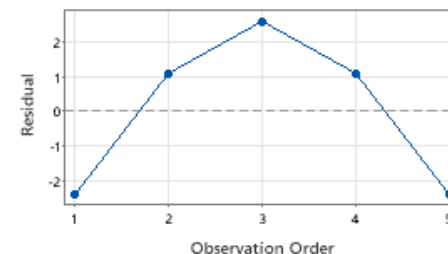
Versus Fits



Histogram



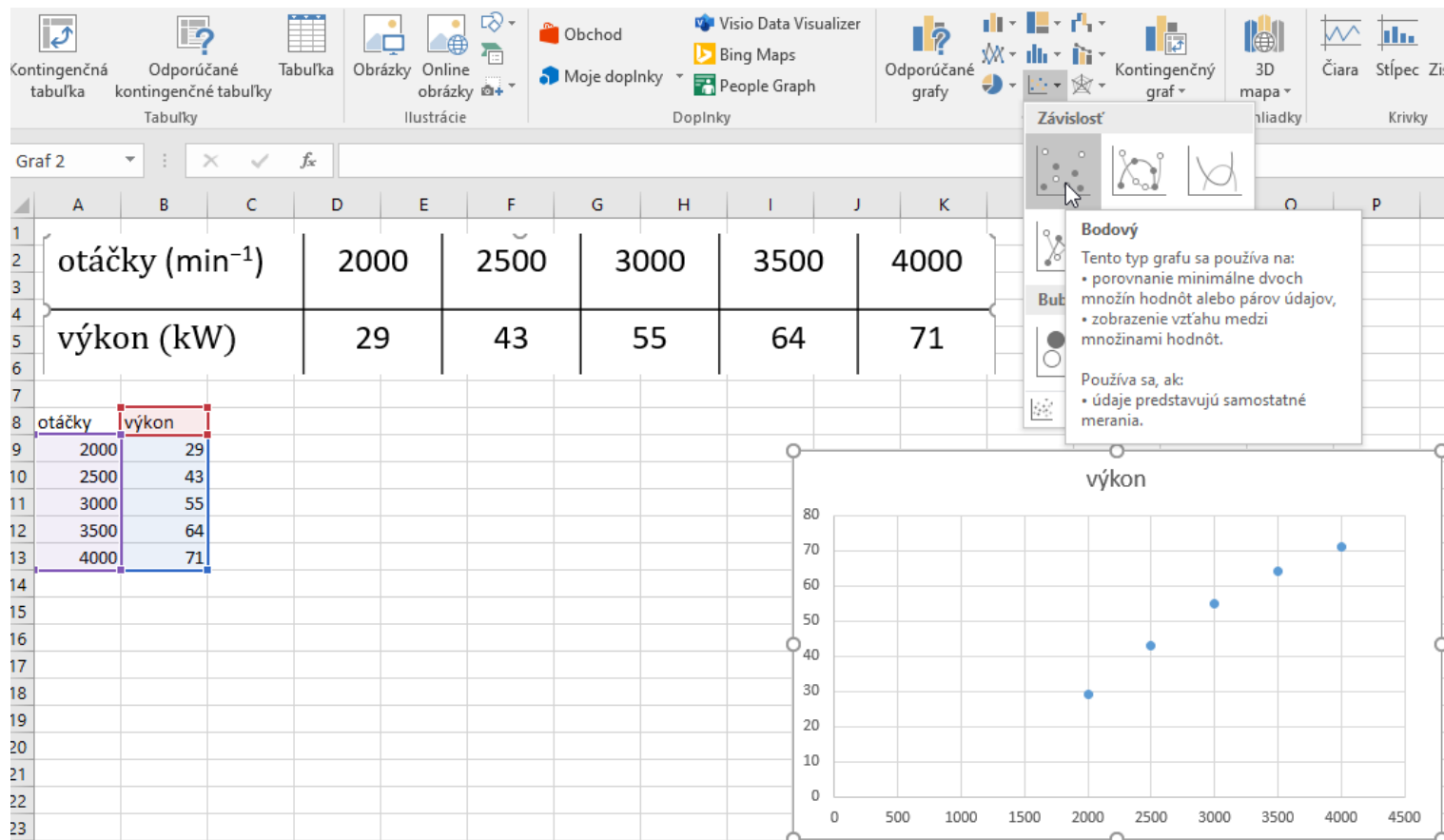
Versus Order



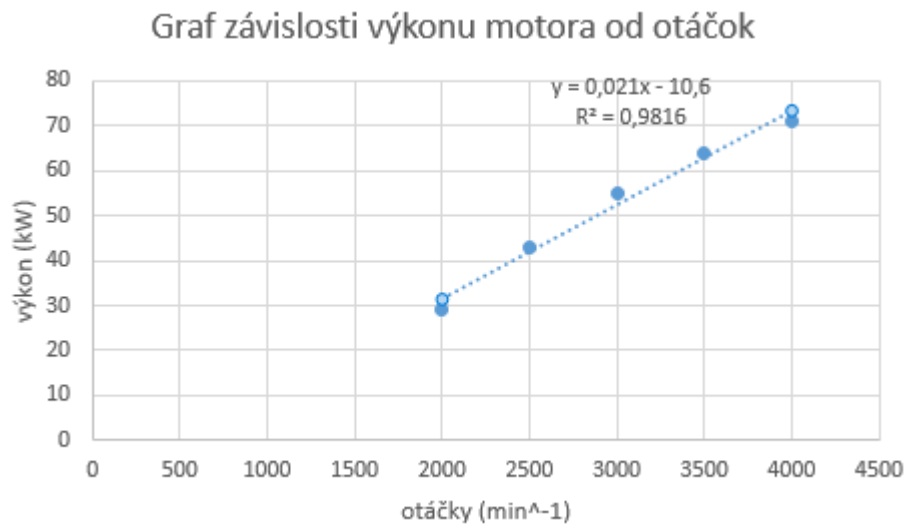
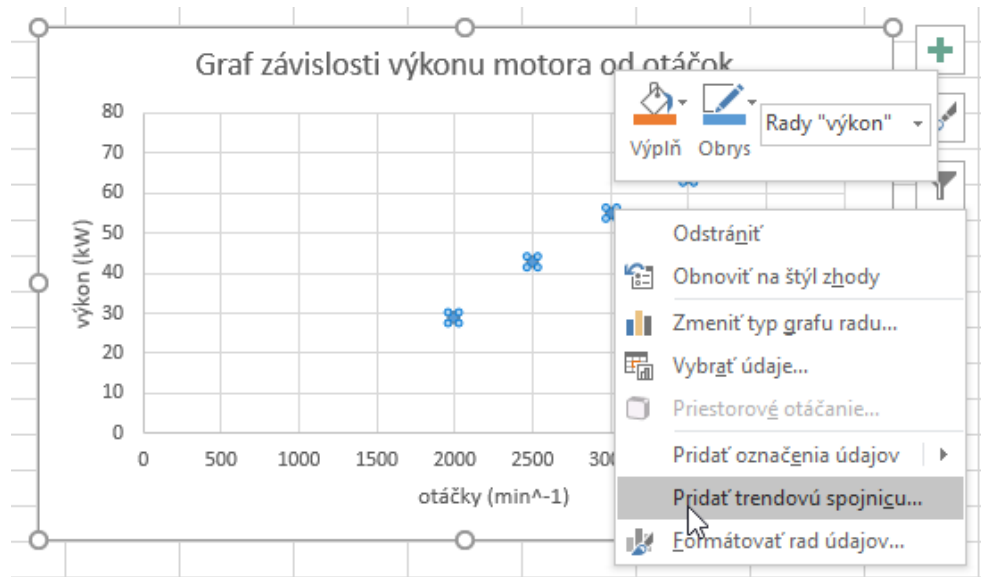
Riešenie príkladu v Exceli

otáčky (min^{-1})	2000	2500	3000	3500	4000
výkon (kW)	29	43	55	64	71

$$\text{výkon [kW]} = \frac{\text{moment [Nm]} \cdot \text{otáčky [1/min]}}{9\,549}$$



Riešenie príkladu v Exceli



Formátovať trendovú spojnicu

Možnosti trendovej spojnice ▾

Možnosti trendovej spojnice

- ☐ Exponenciálna
- ☒ Lineárna
- ☐ Logaritmická
- ☐ Polynomická
- ☐ Mocninová
- ☐ Pohyblivý priemer

Poradie 2

Obdobie 2

Názov trendovej spojnice

- ☒ Automaticky
- ☐ Vlastný

Prognóza

Dopredu 0,0 periódy

Späť 0,0 periódy

0,0

☐ Nastaviť bod pretnutia

☒ Zobraziť v grafe rovnicu

☒ Zobraziť v grafe rovnicu spoľahlivosti R na druhú