

REGRESNÍ ANALÝZA

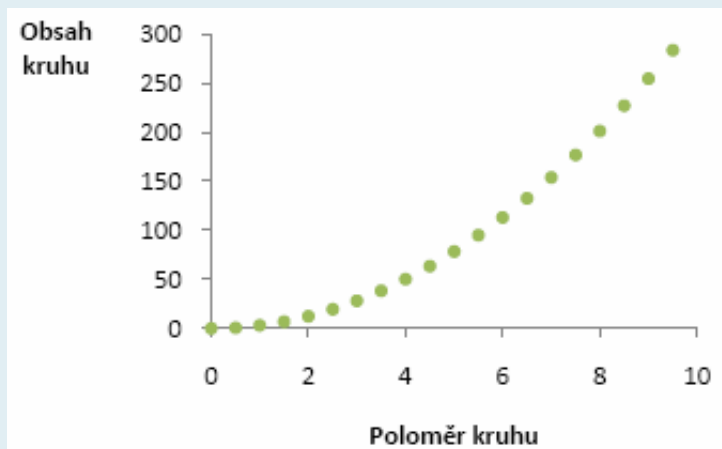


13. cvičení

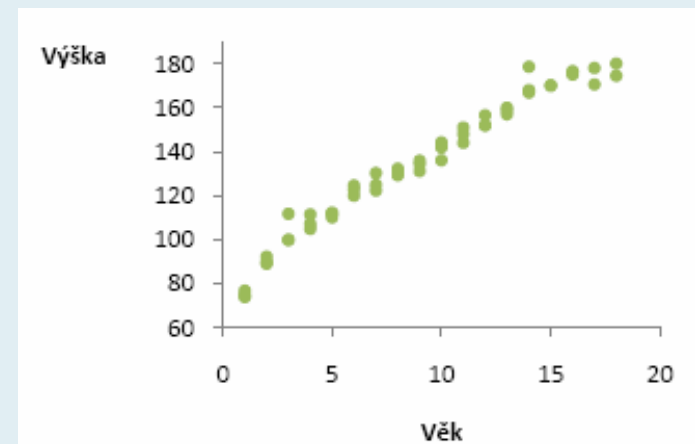
Závislost náhodných veličin

Závislost mezi kvantitativními proměnnými X a Y :

- ❖ **Funkční závislost** – hodnotami nezávisle proměnných je jednoznačně dána hodnota závislé proměnné. $Y=f(X)$
- ❖ **Stochastická (volná) závislost** – systematický pohyb jedné veličiny při růstu či poklesu druhé veličiny (studujeme prostřednictvím korelační a regresní analýzy)



Korelační pole pro funkční závislost



Korelační pole pro stochastickou závislost

Korelační a regresní analýza

Vyhodnocují vztah mezi spojitými veličinami.

❖ **Nekauzální vztahy** (neznáme příčinu a důsledek) – korelační analýza

❖ **Kauzální vztahy** (víme co je příčina a co je následek) – korelační i regresní analýza.

Nezávisle proměnná (vysvětlující proměnná) = vysvětluje chování závislé proměnné (příčina)

Závisle proměnná (vysvětlovaná proměnná) = proměnná jejíž chování se snažíme vysvětlit (následek)

Typy regrese

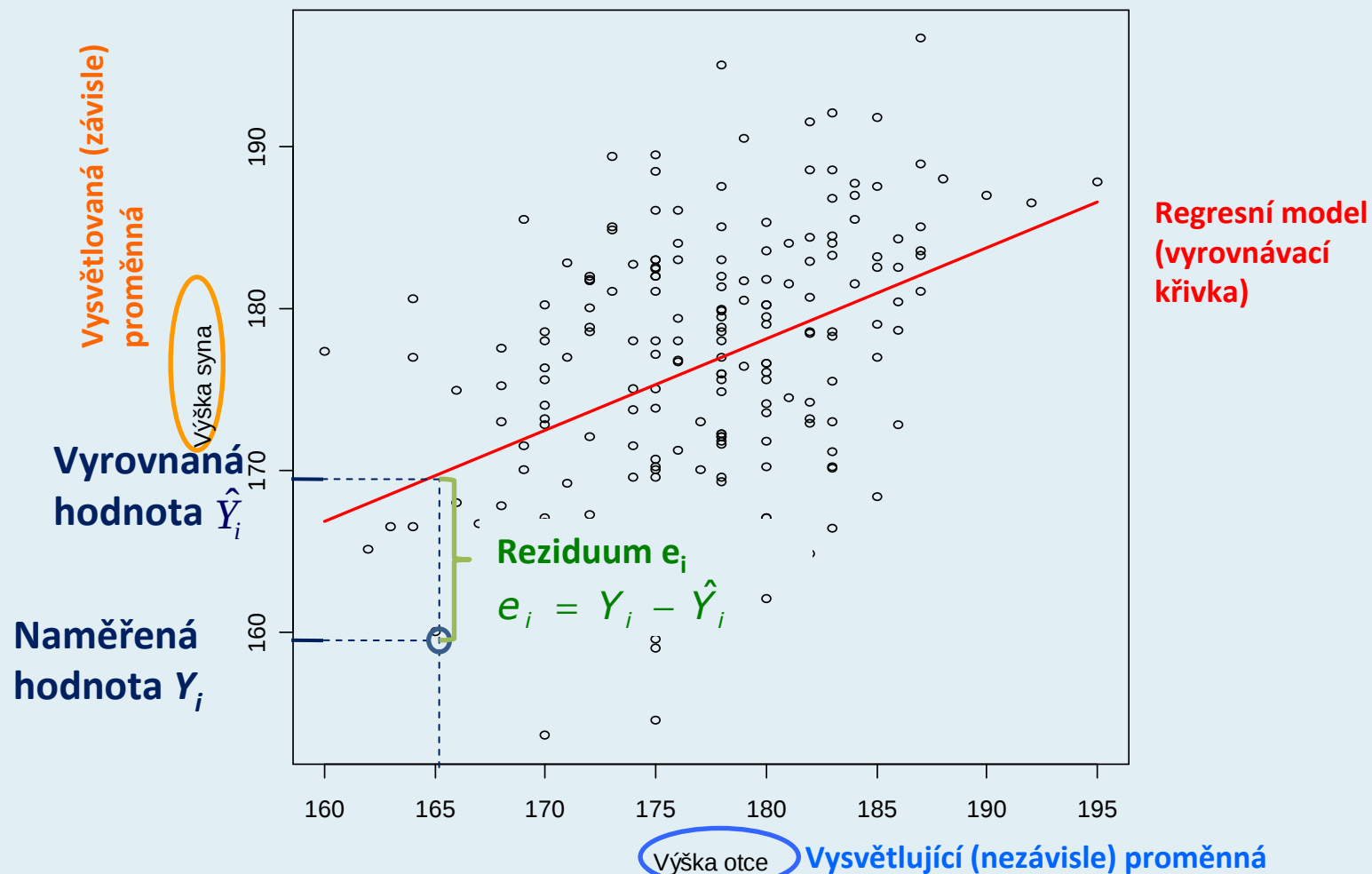
- ❖ **Lineární regrese** - pro popis závislosti veličin využívá funkce lineární v parametrech (např. $Y = \beta_0 + \beta_1 X$), resp. funkce, které lze na lineární v parametrech převést pomocí vhodné transformace (např. $Y = \beta_0 X^{\beta_1}$)
- ❖ **Nelineární regrese** - pro popis závislosti veličin využívá funkce nelineární v parametrech (tyto funkce nelze na lineární v parametrech převést pomocí žádné transformace – např.: $Y = \beta_0 \beta_1 X$)

Typy regrese

- ❖ **Jednoduchá regrese** - studuje kauzální závislost dvou veličin (velikost syna na velikosti otce)
- ❖ **Vícenásobná regrese** - studuje kauzální závislost jedné veličiny na alespoň dvou dalších veličinách (velikost syna na velikosti otce a matky)

Korelační pole

= Zakreslená data do bodového pole.



Jednoduchý lineární regresní model

Vyrovňovací křivka:

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

Parametry modelu

Reziduum
Náhodná složka –
popisuje vliv
náhodných nebo
nepozorovaných
regresorů a vliv
náhody

Předpoklady jednoduchého reg. modelu

- ❖ Lineární regresní model je lineární v parametrech.
- ❖ Parametry modelu β_i mohou nabývat libovolných hodnot.
- ❖ Normalita náhodné složky (reziduí e_i).
- ❖ Nulová střední hodnota náhodné složky (reziduí) – $E(e_i) = 0$.
- ❖ Rozptyl náhodné složky (reziduí e_i) je konstantní.
- ❖ Kovariance náhodné složky je nulová - $Cov(e_i, e_j) = 0$ pro každé $i \neq j$, kde $i, j = 1, 2, \dots, n$.

Podmínky lineárního regresního modelu je nutno v rámci regresní analýzy ověřit.

Postup při regresní analýze

1. Exploratorní analýza korelačního pole (případný odhad typu regresní funkce, identifikace vlivných bodů)
2. Odhad koeficientů regresní funkce (aplikace vyrovnávacího kritéria)
3. Verifikace modelu - vychází z ověření předpokladů jednoduchého regresního modelu. Většina ověřovacích metod je založena na studiu chování reziduí.
 1. Celkový F-test
 2. Dílčí t-testy
 3. Index determinace
 4. Autokorelace reziduí
 5. Testy reziduí
4. Predikce (pás spolehlivosti, pás predikce)

1. Exploratorní analýza korelačního pole

- ❖ Odhad typu regresní funkce (pokud není znám)
- ❖ Identifikace vlivných bodů (pozor na body signalizující chybějící část populace ve výběru)

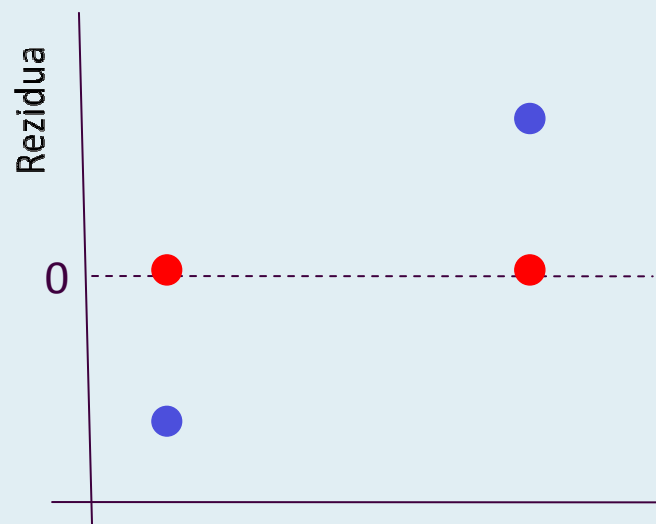
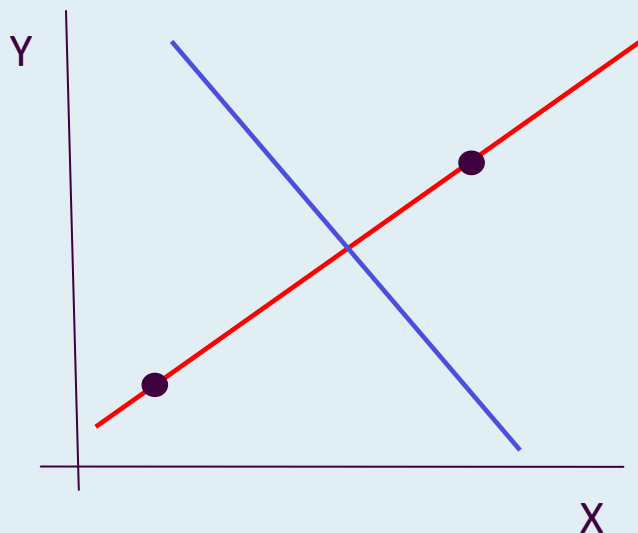
V appletu Regrese

<http://mi21.vsb.cz/modul/uvod-do-statistiky>

sledujte vliv pozice vlivných bodů na pozici vyrovnávací přímky.

2. Odhad koeficientů regresní funkce

- ❖ Cílem je minimalizace reziduí.
- ❖ **Vyrovňovací kritéria** - kritéria pomocí nichž volíme nejvhodnější způsob odhadu parametrů regresní funkce.
- ❖ Proč nestačí minimalizovat součet reziduí?

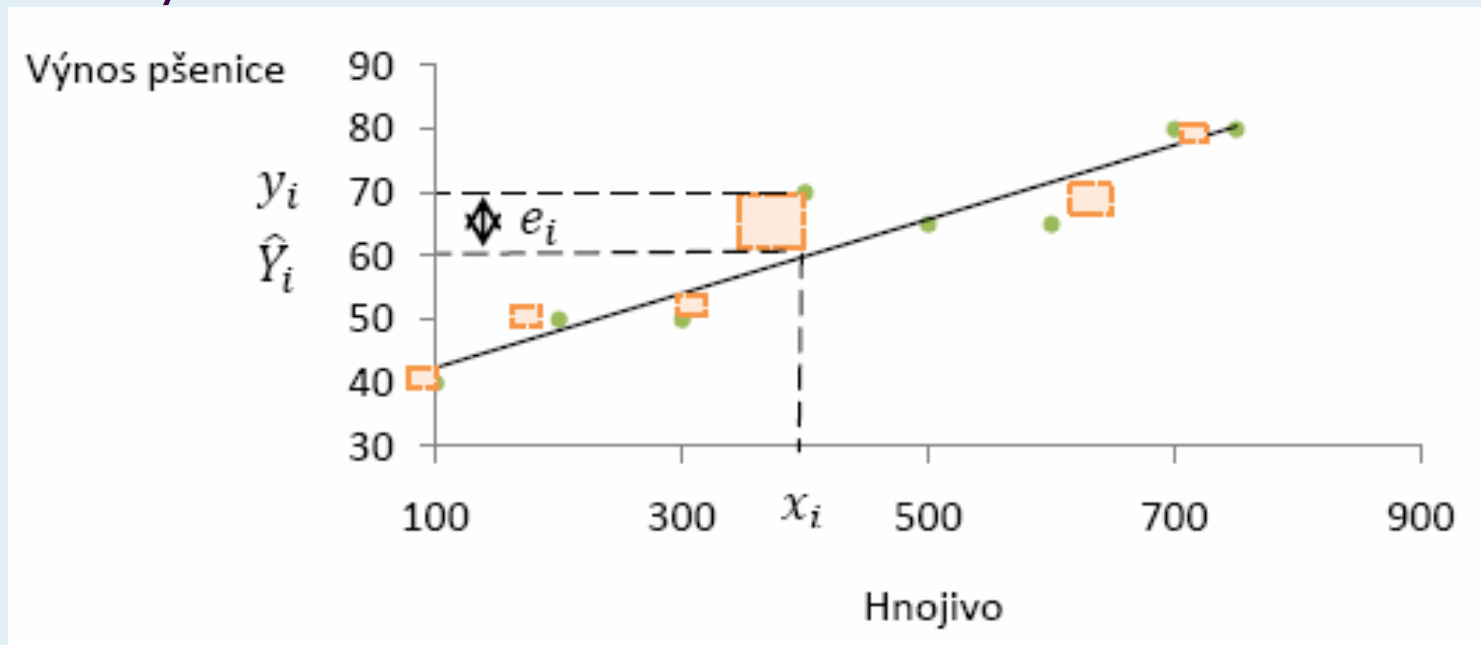


- ❖ Mohlo by dojít k tomu, že součet reziduí je nulový, přestože jednotlivá rezidua jsou „velká“.
- =>

2. Odhad koeficientů regresní funkce Metoda nejmenších čtverců

=> Metoda nejmenších čtverců:

- ❖ Minimalizuje součet čtverců reziduí.
- ❖ Nejpoužívanější vyrovnávací kritérium pro lineární regresní modely.



Vizualizace principu metody nejmenších čtverců

2. Odhad koeficientů regresní funkce

Metoda nejmenších čtverců pro přímku

Regresní přímka:

$$EY_i = \beta_0 + \beta_1 \cdot x_i$$

Bodový odhad regresní přímky:

$$\hat{Y}_i = b_0 + b_1 \cdot x_i$$

Součet čtverců reziduí:

$$\begin{aligned} \varphi &= \sum_{i=1}^n (e_i)^2 = \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - b_0 - b_1 x_i)^2 \end{aligned}$$

Minimalizace $\varphi(b_0, b_1)$:

$$\frac{d\varphi}{db_0} = -2 \sum_{(i)} (Y_i - b_0 - b_1 \cdot x_i) = 0$$

hledáme stacionární
body, tj. body, ve
kterých jsou
parciální derivace
nulové

$$\frac{d\varphi}{db_1} = -2 \sum_{(i)} [(Y_i - b_0 - b_1 \cdot x_i) \cdot (x_i)] = 0$$

2. Odhad koeficientů regresní funkce

Metoda nejmenších čtverců pro přímku

$$\frac{d\varphi}{db_0} = -2 \sum_{(i)} (Y_i - b_0 - b_1 \cdot x_i) = 0$$

$$\frac{d\varphi}{db_1} = -2 \sum_{(i)} [(Y_i - b_0 - b_1 \cdot x_i) \cdot (x_i)] = 0$$

$$\sum_{(i)} Y_i - nb_0 - b_1 \sum_{(i)} x_i = 0$$

$$\sum_{(i)} x_i Y_i - b_0 \sum_{(i)} x_i - b_1 \sum_{(i)} x_i^2 = 0$$

$$b_0 = \frac{\sum_{(i)} Y_i}{n} - b_1 \frac{\sum_{(i)} x_i}{n} = \bar{Y} - b_1 \cdot \bar{x}$$

$$b_1 = \frac{n \sum_{(i)} x_i Y_i - \sum_{(i)} x_i \sum_{(i)} Y_i}{n \sum_{(i)} x_i^2 - \left(\sum_{(i)} x_i \right)^2}$$

2. Odhad koeficientů regresní funkce

Význam bodových odhadů jednotlivých koeficientů lineární regrese:

- ❖ b_0 – odhaduje hodnotu závisle proměnné za předpokladu, že hodnoty všech regresorů jsou nulové (např. výnos pšenice pokud nepoužíváme žádné hnojivo)
- ❖ b_1 – odhaduje závisle proměnnou v případě, kdy se regresor zvýší o 1 (např. navýšení výnosu pšenice při zvýšení množství hnojiva o 1kg/m^2)

3. Verifikace modelu – celkový F-test

Testuje, jestli je vysvětlovaná proměnná lineární kombinací vybraných funkcí vysvětlující proměnné.

Nulová a alternativní hypotéza:

$$H_0: \beta_1 = \dots = \beta_k = 0$$

$$H_A: \overline{H_0}$$

Testová statistika:
$$F = \frac{\frac{SS_{\hat{Y}}}{k}}{\frac{SS_R}{n - k - 1}} = \frac{\frac{\sum_{(i)} (\hat{Y}_i - \bar{Y})^2}{k}}{\frac{\sum_{(i)} (Y_i - \hat{Y}_i)^2}{\underbrace{n}_{\text{počet pozorování}} - \underbrace{k}_{\text{počet regresorů}} - 1}} \rightarrow F(k; n - k - 1)$$

Výpočet p-value:
$$p - value = 1 - F(x_{OBS})$$

3. Verifikace modelu – celkový F-test

Výstupem testu je opět tabulka ANOVA:

Zdroj rozptýlenosti	Součet čtverců	Stupně volnosti (DF)	Průměrný čtverec	Testová stat. F	P-value
Model	$SS_{\hat{Y}} = \sum_{(i)} (\hat{Y}_i - \bar{Y})^2$	k	$MS_{\hat{Y}} = \frac{SS_{\hat{Y}}}{k}$	$F = \frac{MS_{\hat{Y}}}{MS_R}$	$1 - F(x_{OBS})$
Náhodná složka (Rezidua)	$SS_R = \sum_{(i)} (Y_i - \hat{Y}_i)^2$	$n-k-1$	$MS_R = \frac{SS_R}{n-k-1}$		
Celkový	$SS_Y = \sum_{(i)} (Y_i - \bar{Y})^2$	$n-1$			

3. Verifikace modelu – dílčí t-testy

Dílčí t-testy jsou testy o hodnotách jednotlivých parametrů regresní funkce a umožňují testovat oprávněnost setrvání příslušné funkce vysvětlující proměnné v regresním modelu. (Testujeme pro $i=0, 1, \dots, k$)

Nulová a alternativní hypotéza:

$$H_0: \beta_i = 0$$

$$H_A: \beta_i \neq 0$$

Testová statistika:
$$\frac{b_i - \beta_i}{s_{b_i}} \rightarrow t_{n-k+1}$$

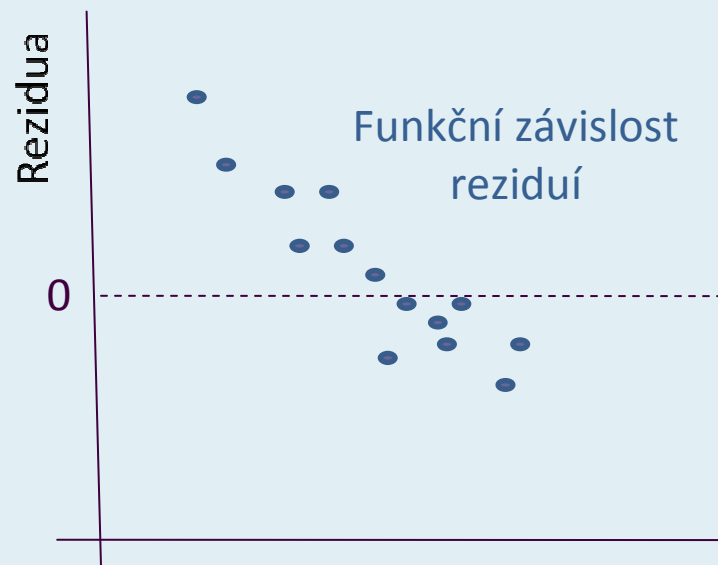
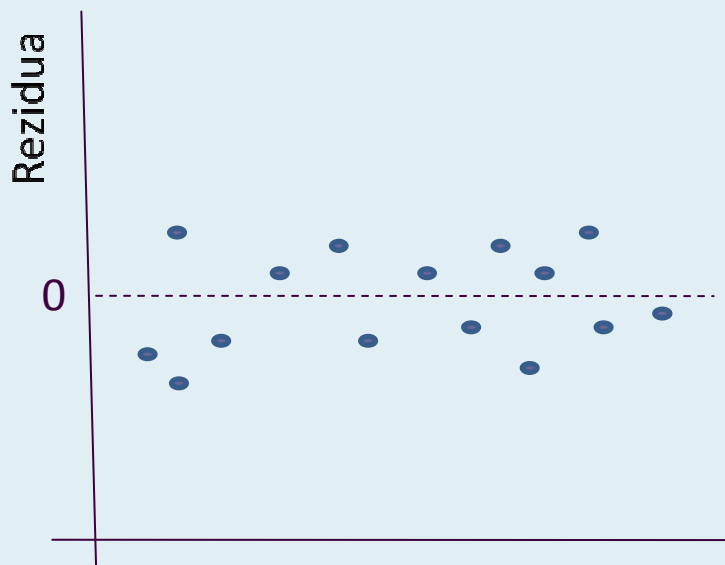
3. Verifikace modelu – index determinace R^2

- ❖ Udává **kvalitu** regresního modelu, tj. jaká část rozptylu vysvětlované proměnné je vysvětlena modelem.
- ❖ Nízká hodnota R^2 , nemusí znamenat nízký stupeň závislosti mezi proměnnými, ale může signalizovat chybnou volbu typu regresní funkce.

$$R^2 = \frac{SS_{\hat{Y}}}{SS_Y} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

3. Verifikace modelu – autokorelace reziduí

- ❖ Na základě předpokladu lin. reg. modelu, že kovariance reziduí je nulová, je zřejmé, že rovněž autokorelace reziduí musí být nulová. Lze tedy předpokládat, že na grafu reziduí nesmí být patrná žádná funkční závislost.



3. Verifikace modelu – testy reziduí

❖ **Test normality reziduí**

(Test dobré shody, Kolmogorovův-Smirnovův test, Shapiro-Wilkův test, ...)

❖ **Test homoskedasticity reziduí**

(velmi obtížný, není součástí většiny statistického software)

❖ **Test nulové střední hodnoty reziduí**

(jednovýběrový t-test)

Výstup regrese ve Statgraphicsu

Regression Analysis - Linear model: $Y = a + b \cdot X$					
Dependent variable: Snížení hladiny cukru					
Independent variable: Množství insulinu					
Parameter	Estimate	Standard Error	T Statistic	P-Value	
Intercept	-15,9643	11,1856	-1,42721	0,2034	
Slope	0,161429	0,0324596	4,97321	0,0025	
Analysis of Variance					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	2736,21	1	2736,21	24,73	0,0025
Residual	663,786	6	110,631		
Total (Corr.)	3400,0	7			
Correlation Coefficient = 0,897089					
R-squared = 80,4769 percent					
R-squared (adjusted for d.f.) = 77,223 percent					
Standard Error of Est. = 10,5181					
Mean absolute error = 7,42857					
Durbin-Watson statistic = 2,78236 (P=0,0340)					
Lag 1 residual autocorrelation = -0,391276					

$\text{Snížení hladiny cukru} = -15,9643 + 0,161429 \cdot \text{Množství insulinu}$

Typ modelu, rovnice vyrovnávací funkce

Závisle a nezávisle proměnná

Bodové odhady koeficientů regresní přímky

Bodové odhady směrodatných odchylek koeficientů regresní přímky

Výsledky dílčích t-testů

Součty čtverců pro model, reziduální a celkový

Reziduální výběrový rozptyl

Výsledek F-testu pro regresi

Korelační koeficient

Koeficient determinace

Výběrová reziduální směrodatná odchylka

Rovnice vyrovnávací přímky

Test

Vyberte správný výraz:

- a) Kolmogorovův-Smirnovův test ve své základní podobě (*lze, nelze*) použít pro testování normality.
- b) Použijeme-li χ^2 test dobré shody pro ověření toho, zda je klasická šestistěnná hrací kostka „férová“, pak má v případě platnosti nulové hypotézy testová statistika rozdělení s (4; 5; 6) stupni volnosti.

Vyberte správný výraz:

- a) Kolmogorovův-Smirnovův test ve své základní podobě (*lze*, *nelze*) použít pro testování normality.
- b) Použijeme-li χ^2 test dobré shody pro ověření toho, zda je klasická šestistěnná hrací kostka „férová“, pak má v případě platnosti nulové hypotézy testová statistika rozdělení s (4; 5; 6) stupni volnosti.

Vyberte správný výraz:

- a) Kolmogorovův-Smirnovův test ve své základní podobě (*lze*, *nelze*) použít pro testování normality.
- b) Použijeme-li χ^2 test dobré shody pro ověření toho, zda je klasická šestistěnná hrací kostka „férová“, pak má v případě platnosti nulové hypotézy testová statistika rozdělení s (*4*; *5*; *6*) stupni volnosti.

Vyberte správný výraz:

- c) Pro úplně specifikovaný test dobré shody se spojitým rozdělením je vhodnější použít (χ^2 test dobré shody, Kolmogorovův-Smirnovův test).
- d) Chceme-li pro ověření shody mezi teoretickým a empirickým rozdělením použít test dobré shody, musí být všechny (*pozorované, očekávané*) četnosti jednotlivých variant, resp. třídících intervalů, větší než 5.

Vyberte správný výraz:

- c) Pro úplně specifikovaný test dobré shody se spojitým rozdělením je vhodnější použít (χ^2 *test dobré shody*, *Kolmogorovův-Smirnovův test*).
- d) Chceme-li pro ověření shody mezi teoretickým a empirickým rozdělením použít test dobré shody, musí být všechny (*pozorované, očekávané*) četnosti jednotlivých variant, resp. třídících intervalů, větší než 5.

Vyberte správný výraz:

- c) Pro úplně specifikovaný test dobré shody se spojitým rozdělením je vhodnější použít (χ^2 *test dobré shody*, *Kolmogorovův-Smirnovův test*).
- d) Chceme-li pro ověření shody mezi teoretickým a empirickým rozdělením použít test dobré shody, musí být všechny (*pozorované*, *očekávané*) četnosti jednotlivých variant, resp. třídících intervalů, větší než 5.

Vyberte správný výraz:

- e) Čím členitější je mozaikový graf, tím (*slabší, silnější*) závislost mezi veličinami v kontingenční tabulce pozorujeme.
- f) Analyzujeme-li závislost v kontingenční tabulce, která má 4 řádky a 5 sloupců, pak χ^2 test nezávislosti můžeme použít, pokud alespoň (4; 10; 16; 20) očekávaných četností je větších než 5 a ostatní nejsou menší než (0; 1; 2).
- g) Koeficient kontingence (*se vyskytuje v intervalu (0;1); může nabývat hodnot větších než 1*).

Vyberte správný výraz:

- e) Čím členitější je mozaikový graf, tím (*slabší*, *silnější*) závislost mezi veličinami v kontingenční tabulce pozorujeme.
- f) Analyzujeme-li závislost v kontingenční tabulce, která má 4 řádky a 5 sloupců, pak χ^2 test nezávislosti můžeme použít, pokud alespoň (4; 10; 16; 20) očekávaných četností je větších než 5 a ostatní nejsou menší než (0; 1; 2).
- g) Koeficient kontingence (*se vyskytuje v intervalu (0;1); může nabývat hodnot větších než 1*).

Vyberte správný výraz:

- e) Čím členitější je mozaikový graf, tím (*slabší*, *silnější*) závislost mezi veličinami v kontingenční tabulce pozorujeme.
- f) Analyzujeme-li závislost v kontingenční tabulce, která má 4 řádky a 5 sloupců, pak χ^2 test nezávislosti můžeme použít, pokud alespoň (*4*; *10*; *16*; *20*) očekávaných četností je větších než 5 a ostatní nejsou menší než (*0*; *1*; *2*).
- g) Koeficient kontingence (*se vyskytuje v intervalu (0;1); může nabývat hodnot větších než 1*).

Vyberte správný výraz:

- e) Čím členitější je mozaikový graf, tím (*slabší*, *silnější*) závislost mezi veličinami v kontingenční tabulce pozorujeme.
- f) Analyzujeme-li závislost v kontingenční tabulce, která má 4 řádky a 5 sloupců, pak χ^2 test nezávislosti můžeme použít, pokud alespoň (*4*; *10*; *16*; *20*) očekávaných četností je větších než 5 a ostatní nejsou menší než (*0*; *1*; *2*).
- g) Koeficient kontingence (*se vyskytuje v intervalu (0;1); může nabývat hodnot větších než 1*).

Vyberte správný výraz:

- e) Čím členitější je mozaikový graf, tím (*slabší*, *silnější*) závislost mezi veličinami v kontingenční tabulce pozorujeme.
- f) Analyzujeme-li závislost v kontingenční tabulce, která má 4 řádky a 5 sloupců, pak χ^2 test nezávislosti můžeme použít, pokud alespoň (*4*; *10*; *16*; *20*) očekávaných četností je větších než 5 a ostatní nejsou menší než (*0*; *1*; *2*).
- g) Koeficient kontingence (*se vyskytuje v intervalu (0;1)*; *může nabývat hodnot větších než 1*).

Vyberte správný výraz:

- h) (*Kontingenční, Asociační*) tabulka je speciálním případem (*kontingenční, asociační*) tabulky.
- i) Je-li odhad relativního rizika $\widehat{RR}=1,2$, pak (*mezi znaky v asociační tabulce existuje závislost, mezi znaky v asociační tabulce neexistuje závislost, o závislosti znaků v asociační tabulce musí rozhodnout test*).
- j) Kvalita 50 různých výukových materiálů byla dvěma odborníky hodnocena na stupnici od 1 do 5. Vhodnou mírou závislosti mezi hodnocením jednotlivých odborníků je (*Pearsonův, Spearmanův*) korelační koeficient.

Vyberte správný výraz:

- h) (*Kontingenční, Asociační*) tabulka je speciálním případem (*kontingenční, asociační*) tabulky.
- i) Je-li odhad relativního rizika $\widehat{RR}=1,2$, pak (*mezi znaky v asociační tabulce existuje závislost, mezi znaky v asociační tabulce neexistuje závislost, o závislosti znaků v asociační tabulce musí rozhodnout test*).
- j) Kvalita 50 různých výukových materiálů byla dvěma odborníky hodnocena na stupnici od 1 do 5. Vhodnou mírou závislosti mezi hodnocením jednotlivých odborníků je (*Pearsonův, Spearmanův*) korelační koeficient.

Vyberte správný výraz:

- h) (*Kontingenční, Asociační*) tabulka je speciálním případem (*kontingenční, asociační*) tabulky.
- i) Je-li odhad relativního rizika $\widehat{RR}=1,2$, pak (*mezi znaky v asociační tabulce existuje závislost, mezi znaky v asociační tabulce neexistuje závislost, o závislosti znaků v asociační tabulce musí rozhodnout test*).
- j) Kvalita 50 různých výukových materiálů byla dvěma odborníky hodnocena na stupnici od 1 do 5. Vhodnou mírou závislosti mezi hodnocením jednotlivých odborníků je (*Pearsonův, Spearmanův*) korelační koeficient.

Vyberte správný výraz:

- h) (*Kontingenční, Asociační*) tabulka je speciálním případem (*kontingenční, asociační*) tabulky.
- i) Je-li odhad relativního rizika $\widehat{RR}=1,2$, pak (*mezi znaky v asociační tabulce existuje závislost, mezi znaky v asociační tabulce neexistuje závislost, o závislosti znaků v asociační tabulce musí rozhodnout test*).
- j) Kvalita 50 různých výukových materiálů byla dvěma odborníky hodnocena na stupnici od 1 do 5. Vhodnou mírou závislosti mezi hodnocením jednotlivých odborníků je (*Pearsonův, Spearmanův*) korelační koeficient.