

## 12. Jednoduchá lineární regrese

### 12.1. Motivace

Cíl regresní analýzy - popsat závislost hodnot veličiny Y na hodnotách veličiny X.

Nutnost vyřešení dvou problémů:

- jaký typ funkce se použije k popisu dané závislosti;
- jak se stanoví konkrétní parametry daného typu funkce?

### 12.2. Specifikace klasického modelu lineární regrese

$Y = \beta_0 + \beta_1 X + \epsilon$ , kde

$\beta_0, \beta_1, \dots, \beta_p$  - **teoretická regresní funkce**, která lineárně závisí na neznámých regresních parametrech  $\beta_0, \beta_1, \dots, \beta_p$  a známých funkcích  $f_0(X), \dots, f_p(X)$ , které již

neobsahují neznámé parametry, tj.  $\beta_0 + \beta_1 X + \dots + \beta_p f_p(X)$ , přičemž  $f_0(X) \equiv 1$ .

Složka  $\epsilon$  - **náhodná odchylka**.

Veličina Y - **závisle proměnná (též vysvětlovaná) veličina**.

Veličina X - **nezávisle proměnná (též vysvětlující) veličina**.

Pořídíme n dvojic pozorování  $(X_1, Y_1), \dots, (X_n, Y_n)$ , pro  $i = 1, \dots, n$  platí:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

O náhodných odchylkách  $\epsilon_1, \dots, \epsilon_n$  předpokládáme, že

- $E[\epsilon_i] = 0$  (odchylky nejsou systematické)
- $D[\epsilon_i] = \sigma^2$  (všechna pozorování jsou prováděna s touž přesností)
- $Cov(\epsilon_i, \epsilon_j) = 0$  pro  $i \neq j$  (mezi náhodnými odchylkami neexistuje žádný lineární vztah)
- $\epsilon_i \sim N(0, \sigma^2)$

V tomto případě hovoříme o **klasickém modelu lineární regrese**.

### 12.3. Označení

$b_0, b_1, \dots, b_p$  - **odhady regresních parametrů**  $\beta_0, \beta_1, \dots, \beta_p$  (nejčastěji je získáme metodou nejmenších čtverců, tj. z podmínky, že výraz

$\sum_{i=1}^n \left( \hat{y}_i - \beta_0 - \beta_1 X_i - \dots - \beta_p f_p(X_i) \right)^2$  nabývá svého minima pro  $\beta_j = b_j, j = 0, 1, \dots, p$ )

$\hat{m}_X; b_0, \dots, b_p$  - **empirická regresní funkce**

$\hat{y}_i = \hat{m}_X; b_0, \dots, b_p(X_i) = \beta_0 + \beta_1 X_i + \dots + \beta_p f_p(X_i)$  - **regresní odhad i-té hodnoty veličiny Y** (i-tá

predikovaná hodnota veličiny Y)

$\epsilon_i = Y_i - \hat{y}_i$  - **i-té reziduum**

$$S_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2 - \text{reziduální součet čtverců}$$

$$s^2 = \frac{S_E}{n-2} - \text{odhad rozptylu } \sigma^2$$

$$S_R = \sum_{i=1}^n (\hat{y}_i - \bar{m}_2)^2 - \text{regresní součet čtverců } (\bar{m}_2 = \frac{1}{n} \sum_{i=1}^n y_i)$$

$$S_T = \sum_{i=1}^n (y_i - \bar{m}_2)^2 - \text{celkový součet čtverců } (S_T = S_E + S_R)$$

$$ID = \frac{S_R}{S_T} - \text{index determinace } (0 < ID < 1)$$

$$ID_{adj}^2 = \frac{S_R}{S_T} \frac{n-2}{n-p-1} - \text{adjustovaný index determinace}$$

## 12.4. Maticový zápis klasického modelu lineární regrese

$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ , kde

$\mathbf{y} = (y_1, \dots, y_n)'$  - vektor pozorování závisle proměnné veličiny Y,

$\mathbf{X} = \begin{pmatrix} 1 & f_1(x_1) & \dots & f_p(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & f_1(x_n) & \dots & f_p(x_n) \end{pmatrix}$  - regresní matice

(předpokládáme, že  $h(\mathbf{X}) = p+1 > n$ )

$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$  - vektor regresních parametrů,

$\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$  - vektor náhodných odchylek.

Podmínky (a) až (d) lze zkráceně zapsat ve tvaru  $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$ .

Maticově zapsaná metoda nejmenších čtverců vede na rovnice

$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$  - systém normálních rovnic

$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$  - odhad vektoru  $\boldsymbol{\beta}$  získaný metodou nejmenších čtverců

$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$  - vektor regresních odhadů (vektor predikce)

$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$  - vektor reziduí

Vlastnosti odhadu  $\mathbf{b}$ :

- odhad  $\mathbf{b}$  je lineární, neboť je vytvořen lineární kombinací pozorování  $y_1, \dots, y_n$  s maticí vah  $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ ;

- odhad  $\mathbf{b}$  je nestranný, neboť  $E(\mathbf{b}) = \boldsymbol{\beta}$ ;

- odhad  $\mathbf{b}$  má varianční matici  $\text{var } \mathbf{b} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ ;

- odhad  $\mathbf{b} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$  vzhledem k platnosti podmínky (d);

- pro odhad  $\mathbf{b}$  platí [Gaussova - Markovova věta](#): Odhad  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$  je nejlepší nestranný lineární odhad vektoru  $\boldsymbol{\beta}$ . (Nejlepší v tom smyslu, že rozdíl varianční matice libovolného jiného nestranného odhadu vektoru  $\boldsymbol{\beta}$  a varianční matice odhadu  $\mathbf{b}$  je matice pozitivně semidefinitní.)

### 12.5. Intervaly spolehlivosti pro regresní parametry

$S_{b_j} = \sqrt{V_{jj}}$  - směrodatná chyba odhadu  $b_j$ , kde  $v_{jj}$  je j-tý diagonální prvek matice  $(X'X)^{-1}$ .

Pro  $j = 0, 1, \dots, p$  statistika  $T_j = \frac{b_j}{S_{b_j}} \sim t_{n-p-1}$ , tedy  $100(1-\alpha)\%$  interval spolehlivosti pro  $\beta_j$  má meze:  $b_j \pm t_{\alpha/2, n-p-1} \cdot S_{b_j}$ .

### 12.6. Testování významnosti modelu jako celku (celkový F-test)

Na hladině významnosti  $\alpha$  testujeme

$H_0: \beta_0, \dots, \beta_p = Q_0, \dots, Q_p$  proti  $H_1: \beta_0, \dots, \beta_p \neq Q_0, \dots, Q_p$ .

(Nulová hypotéza říká, že dostačující je model konstanty.)

Testová statistika:  $F = \frac{S_R/p}{S_E/(n-p-1)}$  má rozložení  $F(p, n-p-1)$ , pokud  $H_0$  platí.

Kritický obor:  $W = F \geq F_{\alpha, p, n-p-1}$ .

$F \in W \rightarrow H_0$  zamítáme na hladině významnosti  $\alpha$ .

Výsledky F-testu zapisujeme do tabulky analýzy rozptylu:

zdroj variability	součet čtverců	stupně volnosti	podíl	statistika F
model	$S_R$	$p$	$S_R/p$	$\frac{S_R/p}{S_E/(n-p-1)}$
reziduální	$S_E$	$n-p-1$	$S_E/(n-p-1)$	-
celkový	$S_T$	$n-1$	-	-

### 12.7. Testování významnosti regresních parametrů (dílič t-testy)

Na hladině významnosti  $\alpha$  pro  $j = 0, 1, \dots, p$  testujeme hypotézu

$H_0: \beta_j = 0$  proti  $H_1: \beta_j \neq 0$ .

Testová statistika:  $T_j = \frac{b_j}{S_{b_j}}$  má rozložení  $t(n-p-1)$ , pokud  $H_0$  platí.

Kritický obor:  $W = \{T_j \leq -t_{\alpha/2, n-p-1} \text{ nebo } T_j \geq t_{\alpha/2, n-p-1}\}$ .

$T_j \in W \rightarrow H_0$  zamítáme na hladině významnosti  $\alpha$ .

**12.8. Příklad:** U šesti obchodníků byla zjišťována poptávka po určitém druhu zboží loni (veličina X - v kusech) a letos (veličina Y - v kusech).

číslo. obchodníka	1	2	3	4	5	6
poptávka loni (X)	20	60	70	100	150	260
poptávka letos (Y)	50	60	60	120	230	320

a) Orientačně ověřte předpoklad, že data pocházejí z dvourozměrného normálního rozložení. Vypočtete výběrový koeficient korelace mezi X a Y, interpretujte jeho hodnotu a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny.

b) Předpokládejte, že závislost letošní poptávky na loňské lze vystihnout regresní přímkou. Sestavte regresní matici, vypočtete odhady regresních parametrů a napište rovnici regresní přímky. Interpretujte parametry regresní přímky.

c) Najděte odhad rozptylu, vypočtete index determinace a interpretujte ho.

d) Najděte 95% intervaly spolehlivosti pro regresní parametry.

e) Na hladině významnosti 0,05 proveďte celkový F-test.

f) Na hladině významnosti 0,05 proveďte dílčí t-testy.

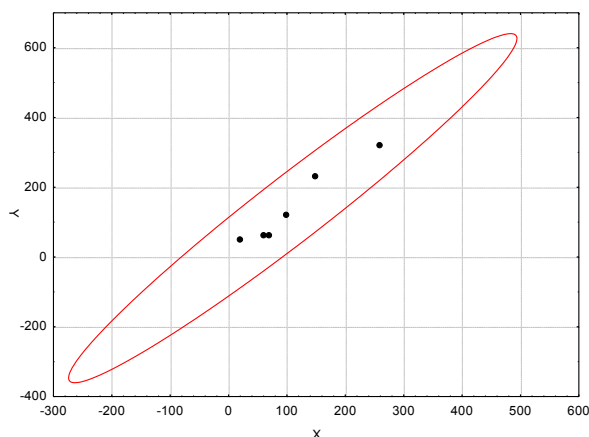
g) Vypočtete regresní odhad letošní poptávky při loňské poptávce 110 kusů.

h) Nakreslete dvourozměrný tečkový diagram s proloženou regresní přímkou.

### Řešení:

ad a) Orientačně ověřte předpoklad, že data pocházejí z dvourozměrného normálního rozložení.

Vytvoříme dvourozměrný tečkový diagram s proloženou 95% elipsou konstantní hustoty pravděpodobnosti:



Ze vzhledu diagramu je patrné, že předpoklad dvourozměrné normality je oprávněný a že mezi loňskou a letošní poptávkou existuje vcelku silná přímá lineární závislost.

Vypočtete výběrový koeficient korelace mezi X a Y, interpretujte jeho hodnotu a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny.

Výpočtem zjistíme:  $r_{12} = 0,972$ , tedy mezi poptávkou loni a letos existuje velmi silná přímá lineární závislost.

Realizace testové statistiky:  $t = \frac{r \sqrt{n}}{\sqrt{1-r^2}} = \frac{0,9726 \sqrt{26}}{\sqrt{1-0,9726^2}} = 26,9$

Kritický obor:

$$W = \left( t_{1/2}(n-2) \cup t_{1/2}(n-2, \infty), t_{0,975}(n-2) \cup t_{0,975}(n-2, \infty) \right)$$

$$= \left( 2,7764 \cup 2,7764, 2,7764 \cup 2,7764 \right)$$

Testová statistika se realizuje v kritickém oboru, hypotézu o nezávislosti veličin X a Y tedy zamítáme na hladině významnosti 0,05.

ad b) Sestavíme regresní matici.

$$\mathbf{X} = \begin{pmatrix} 1 & f_1(x_1) & \dots & f_p(x_1) \\ \dots & \dots & \dots & \dots \\ 1 & f_1(x_n) & \dots & f_p(x_n) \end{pmatrix}, \text{ tedy } \mathbf{X} = \begin{pmatrix} 1 & 20 \\ 1 & 60 \\ 1 & 70 \\ 1 & 100 \\ 1 & 150 \\ 1 & 260 \end{pmatrix}$$

Podle vzorce  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$  získáme odhady regresních parametrů.

Nejprve vypočítáme matici

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 6 & 660 \\ 660 & 10900 \end{pmatrix}$$

a k ní inverzní matici

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0,499084000305 & -0,003022000002 \\ -0,003022000002 & 0,000002 \end{pmatrix}$$

Dále získáme součin

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 840 \\ 13850 \end{pmatrix}$$

a nakonec vektor odhadů regresních parametrů:

$$\mathbf{b} = \begin{pmatrix} 0,499084000305 & -0,003022000002 \\ -0,003022000002 & 0,000002 \end{pmatrix} \begin{pmatrix} 840 \\ 13850 \end{pmatrix} = \begin{pmatrix} 0,6868 \\ 1,2665 \end{pmatrix}$$

Regresní přímka má tedy rovnici

$$y = 0,6868 + 1,2665 x.$$

Znamená to, že při nulové loňské poptávce by letošní poptávka činila 0,6868 kusů a při zvýšení loňské poptávky o 10 kusů by se letošní poptávka zvedla o 12,665 kusů.

ad c) Nyní vypočteme vektor regresních odhadů proměnné Y (vektor predikce):

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \begin{pmatrix} 1 & 20 \\ 1 & 60 \\ 1 & 70 \\ 1 & 100 \\ 1 & 150 \\ 1 & 260 \end{pmatrix} \begin{pmatrix} 0,6868 \\ 0,8934 \\ 1,2345 \\ 1,9065 \\ 3,297 \end{pmatrix} = \begin{pmatrix} 20,2 \\ 76,8 \\ 89,4 \\ 123,4 \\ 190,6 \\ 329,7 \end{pmatrix}$$

Stanovíme vektor reziduí:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} 50 \\ 60 \\ 60 \\ 120 \\ 230 \\ 320 \end{pmatrix} - \begin{pmatrix} 20,2 \\ 76,8 \\ 89,4 \\ 123,4 \\ 190,6 \\ 329,7 \end{pmatrix} = \begin{pmatrix} 29,8 \\ 16,8 \\ -29,4 \\ 73,4 \\ 39,34 \\ -9,97 \end{pmatrix}$$

Pomocí vektoru reziduí vypočteme reziduální součet čtverců:

$$S_E = \mathbf{e}'\mathbf{e} = (23,98 \quad -16,68 \quad -29,34 \quad -7,34 \quad 39,34 \quad -9,97) \cdot \begin{pmatrix} 29,8 \\ 16,8 \\ -29,4 \\ 73,4 \\ 39,34 \\ -9,97 \end{pmatrix} = 3451,11.$$

Odhad rozptylu:  $s^2 = \frac{S_E}{n-1} = \frac{3451,11}{6} = 575,185$

Dále potřebujeme celkový součet čtverců

$$S_T = (\mathbf{y} - \mathbf{m}_2)'(\mathbf{y} - \mathbf{m}_2),$$

kde  $\mathbf{m}_2$  je sloupcový vektor typu  $n \times 1$  složený z průměru  $m_2$  závisle proměnné veličiny Y. V našem případě je  $m_2 = 140$ . Po dosazení do vzorce pro celkový součet čtverců tedy dostaneme

$$S_T = (50-140, 60-140, 60-140, 120-140, 230-140, 320-140) \cdot \begin{pmatrix} 50 \\ 60 \\ 60 \\ 120 \\ 230 \\ 320 \end{pmatrix} = 61800.$$

(Celkový součet čtverců lze získat také tak, že výběrový rozptyl veličiny Y vynásobíme  $n-1$ :  $S_T = 5.12360 = 61800$ .) Regresní součet čtverců pak je:

$$S_R = S_T - S_E = 61800 - 3451,11 = 58348,89.$$

Index determinace:  $ID = \frac{S_R}{S_T} = \frac{58348,89}{61800} = 0,944$

Znamená to, že variabilita hodnot závisle proměnné veličiny je z 94,42% vysvětlena regresní přímkou.

(V případě regresní přímky platí  $ID^2 = r_{12}^2$ . V našem případě bylo zjištěno, že  $r_{12} = 0,972$ , tedy  $ID^2 = 0,9447$ .)

ad d) Vypočteme směrodatné chyby odhadů regresních parametrů  $b_0$  a  $b_1$  podle vzorce  $S_{b_j} = \sqrt{v_{jj}}$ ,  $j = 0, 1$ , kde  $v_{jj}$  je  $j$ -tý diagonální prvek matice  $(\mathbf{X}'\mathbf{X})^{-1}$ :

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0,499084 & 0,000305 \\ 0,000302 & 0,000027 \end{pmatrix}$$

Přitom si uvědomíme, že  $v_{00} = 0,499084$ ,  $v_{11} = 0,000027$

$$S_{b_0} = \sqrt{v_{00}} = \sqrt{0,499084} = 0,70642$$

$$S_{b_1} = \sqrt{v_{11}} = \sqrt{0,000027} = 0,0153$$

Stanovíme meze 95% intervalů spolehlivosti pro regresní parametry  $\beta_0$  a  $\beta_1$ .

K tomu slouží vzorec  $b_j \pm t_{p, n-p-1} \cdot S_{b_j}$ ,  $j = 0, 1$ .

95% interval spolehlivosti pro  $\beta_0$ :

$$d = -1,754 \cdot S_{b_0} = -1,754 \cdot 0,70642 = -1,239$$

$$h = +1,754 \cdot S_{b_0} = +1,754 \cdot 0,70642 = 1,239$$

Znamená to, že  $-1,239 < \beta_0 < 1,239$  s pravděpodobností aspoň 0,95.

95% interval spolehlivosti pro  $\beta_1$ :

$$d = -1,754 \cdot S_{b_1} = -1,754 \cdot 0,0153 = -0,0268$$

$$h = +1,754 \cdot S_{b_1} = +1,754 \cdot 0,0153 = 0,0268$$

Znamená to, že  $-0,0268 < \beta_1 < 0,0268$  s pravděpodobností aspoň 0,95.

ad e) Provedení celkového F-testu: na hladině významnosti  $\alpha = 0,05$  testujeme

$H_0: \beta_1 = 0$  proti  $H_1: \beta_1 \neq 0$ .

$$\text{Testová statistika } F = \frac{S_R/p}{S_E/(n-p-1)} = \frac{58348,89}{4151,11} = 14,054$$

$$\text{kritický obor: } W = F_{1, n-p-1, \infty} = F_{0,95, 1, 4, \infty} = 7,7086$$

Protože se testová statistika realizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru  $\beta_1$  (tj. směrnice regresní přímky) zamítáme na hladině významnosti 0,05. Výsledky testování významnosti modelu jako celku zapíšeme do tabulky ANOVA:

zdroj variab.	součet čtverců	stupně volnosti	podíl	statistika F
model	$S_R = 58348,89$	$p = 1$	$S_R/p = 58348,89$	58,349
reziduální	$S_E = 3415,11$	$n-p-1 = 4$	$S_E/(n-p-1) = 853,78$	-
celkový	$S_T = 61800$	$n-1 = 5$	-	-

ad f) Provedení dílčích t-testů:

Na hladině významnosti  $\alpha = 0,05$  testujeme  $H_0: \beta_0 = 0$  proti  $H_1: \beta_0 \neq 0$ .

Testová statistika:  $t_0 = \frac{b_0}{s_{b_0}} = \frac{0,686}{0,209} = 3,28$

kritický obor:

$$W = \left( -\infty, -t_{1-\alpha/2}(n-p-1) \right) \cup \left( t_{1-\alpha/2}(n-p-1), \infty \right) = \left( -\infty, -2,7764 \right) \cup \left( 2,7764, \infty \right)$$

Protože se testová statistika nerealizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru  $\beta_0$  (tj. posunutí regresní přímky) nezamítáme na hladině významnosti 0,05.

Ke stejnému výsledku dospějeme, podíváme-li se na 95% interval spolehlivosti pro  $\beta_0$ . Vypočítali jsme, že  $-56,63 < \beta_0 < 58$  s pravděpodobností aspoň 0,95.

Protože tento interval obsahuje 0, hypotézu  $H_0: \beta_0 = 0$  nezamítáme na hladině významnosti 0,05.

Na hladině významnosti  $\alpha = 0,05$  testujeme  $H_0: \beta_1 = 0$  proti  $H_1: \beta_1 \neq 0$ .

Testová statistika:  $t_1 = \frac{b_1}{s_{b_1}} = \frac{1,2665}{0,078} = 16,23$

kritický obor:

$$W = \left( -\infty, -t_{1-\alpha/2}(n-p-1) \right) \cup \left( t_{1-\alpha/2}(n-p-1), \infty \right) = \left( -\infty, -2,7764 \right) \cup \left( 2,7764, \infty \right)$$

Protože se testová statistika realizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru  $\beta_1$  (tj. směrnice regresní přímky) zamítáme na hladině významnosti 0,05.

Ke stejnému výsledku dospějeme, podíváme-li se na 95% interval spolehlivosti pro  $\beta_1$ . Vypočítali jsme, že  $0,841 < \beta_1 < 1,692$  s pravděpodobností aspoň 0,95.

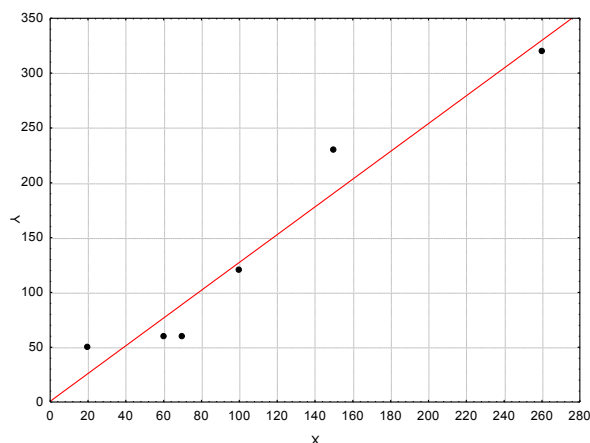
Protože tento interval neobsahuje 0, hypotézu  $H_0: \beta_1 = 0$  zamítáme na hladině významnosti 0,05.

V případě modelu regresní přímky je dílčí t-test pro parametr  $\beta_1$  ekvivalentní s celkovým F-testem.

ad g) Regresní odhad pro  $x = 110$  dostaneme pouhým dosazením do rovnice regresní přímky:  $\hat{y} = 0,86x - 166,11 = 102,76$

ad h)





## Výpočet pomocí systému STATISTICA

Vytvoříme nový datový soubor se dvěma proměnnými X a Y a 6 případy:

	X	Y
1	20	50
2	60	60
3	70	60
4	100	120
5	150	230
6	260	320

a) Orientačně ověřte předpoklad, že data pocházejí z dvourozměrného normálního rozložení. Vypočítejte výběrový koeficient korelace mezi X a Y, interpretujte jeho hodnotu a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny.

Zobrazíme dvourozměrný tečkový diagram s proloženou elipsou 95% konstantní hustoty pravděpodobnosti, s jehož pomocí posoudíme dvourozměrnou normalitu dat: Grafy – Bodové grafy – vypneme Typ proložení – Proměnné X, Y - OK .

Na záložce Detaily vybereme Elipsa Normální – OK. Ve vzniklém dvourozměrném tečkovém diagramu změním rozsah zobrazených hodnot na vodorovné a svislé ose, abychom viděli celou elipsu – viz obrázek výše.

Testování hypotézy o nezávislosti: Statistika – Základní statistiky /Tabulky - Korelační matice – OK – 2 seznamy proměnných X, Y, OK. Na záložce Možnosti zaškrtneme Zobrazit detailní tabulku výsledků – Souhrn.

Prom. \ prom.	Korelace (Tabulka1) Označ. korelace jsou významné na hlad. $p < ,05000$ (Celé případy vyhechány u ChD)										
	Prum.	Sm.Od	r(X,Y)	r <sup>2</sup>	t	p	N	Kons zav.:	Směr zav.:	Kons zav.:	Směr zav.:
X	110,0	85,32									
Y	140,0	111,1	0,971	0,944	8,269	0,001	6	0,686	1,266	5,566	0,745

Ve výstupní tabulce najdeme hodnotu výběrového korelačního koeficientu  $R_{12}$  ( $r = 0,971977$ , tzn. že mezi X a Y existuje velmi silná přímá lineární závislost), realizaci testové statistiky  $t = 8,269474$  a p-hodnotu pro test hypotézy o nezávislosti ( $p = 0,001167$ ,  $H_0$  tedy zamítáme na hladině významnosti 0,05).

b) Předpokládejte, že závislost letošní poptávky na loňské lze vystihnout regresní přímkou. Vypočítejte odhady regresních parametrů a napište rovnici regresní přímky. Interpretujte parametry regresní přímky.

Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnná X  
- OK – OK – Výpočet: Výsledky regrese.

Výsledky regrese se závislou proměnnou						
R = 0,97197702 R <sup>2</sup> = 0,94473932 Upravené						
F(1,4) = 68,384 p < 0,00117 Směrod. chyba						
N=6	Beta	Sm.chy beta	B	Sm.chy B	t(4)	Úroveň
Abs. čl.			0,686	20,64	0,033	0,975
X	0,971	0,117	1,266	0,15	8,269	0,001

Ve výstupní tabulce najdeme koeficient  $b_0$  ve sloupci B na řádce označeném Abs. člen, koeficient  $b_1$  ve sloupci B na řádce označeném X. Rovnice regresní přímky:

$$y = 0,686813 + 1,266484 x.$$

Znamená to, že při nulové loňské poptávce by letošní poptávka činila 0,6868 kusů a při zvýšení loňské poptávky o 10 kusů by se letošní poptávka zvedla o 12,665 kusů.

c) Najděte odhad rozptylu, vypočítejte index determinace a interpretujte ho.

Vrátíme se do Výsledky – vícenásobná regrese – Detailní výsledky – ANOVA.

Analýza rozptylu (Tabulka 1)				
Efekt	Součet čtverců	Průměr čtverců	F	Úroveň
Regre	58384	1 58384	68,38	0,001
Rezid	3415	4 853,		
Celk.	61800			

Odhad rozptylu najdeme na řádce Rezid., ve sloupci Průměr čtverců, tedy  $s^2 = 853,78$ .

Index determinace je uveden v záhlaví původní výstupní tabulky pod označením R<sup>2</sup>. V našem případě  $ID^2 = 0,9447$ , tedy variabilita letošní poptávky je z 94,5% vysvětlena regresní přímkou.

d) Najděte 95% intervaly spolehlivosti pro regresní parametry.

Ve výstupní tabulce výsledků regrese přidáme za proměnnou Úroveň p dvě nové proměnné dm (pro dolní meze 95% intervalů spolehlivosti pro regresní parametry) a hm (pro horní meze 95% intervalů spolehlivosti pro regresní parametry). Do Dlouhého jména proměnné dm resp. hm napíšeme:  $=v3-v4*VStudent(0,975;4)$  resp.  $=v3+v4*VStudent(0,975;4)$

Výsledky regrese se závislou proměnnou : Y (Tabulka)								
R= ,97197702 R2= ,94473932 Upravené R2= ,930924								
F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,2								
N=6	Beta	Sm.chy beta	B	Sm.chy B	t(4)	Úroveň	dm $=v3-v4$	hm $=v3+v4$
Abs.ci			0,686	20,64	0,033	0,975	-56,6	57,99
X	0,971	0,117	1,266	0,15	8,269	0,001	0,841	1,691

Vidíme, že  $-56,63 < \beta_0 < 58$  s pravděpodobností aspoň 0,95 a  $0,841 < \beta_1 < 1,692$  s pravděpodobností aspoň 0,95.

e) Na hladině významnosti 0,05 proveďte celkový F-test.

Testovou statistiku F-testu a odpovídající p-hodnotu najdeme v záhlaví výstupní tabulky regrese. Zde  $F = 68,384$ , p-hodnota  $< 0,00117$ , tedy na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti modelu jako celku. (Výsledky F-testu jsou rovněž uvedeny v tabulce ANOVA.)

f) Na hladině významnosti 0,05 proveďte dílčí t-testy.

Výsledky dílčích t-testů jsou uvedeny ve výstupní tabulce regrese. Testová statistika pro test hypotézy  $H_0: \beta_0 = 0$  je 0,033272, p-hodnota je 0,975052. Hypotézu o nevýznamnosti úseku regresní přímky tedy nezamítáme na hladině významnosti 0,05. Testová statistika pro test hypotézy  $H_0: \beta_1 = 0$  je 8,269474, p-hodnota je 0,001167. Hypotézu o nevýznamnosti směrnice regresní přímky tedy zamítáme na hladině významnosti 0,05.

g) Vypočítejte regresní odhad letošní poptávky při loňské poptávce 110 kusů.

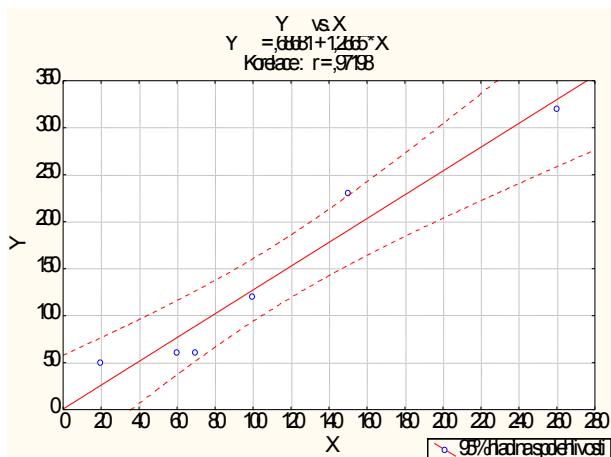
Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi Předpovědi závisle proměnné X: 110 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď.

Predpovezené hodnoty proměnné: Y			
Proměnná	B-koef.	Hodnota X	B-koef. * Hodnota X
X	1,266	110,0	139,3
Abs. chyba			0,68
Předpoklad			140,0
-95,0%L			106,8
+95,0%U			173,1

Při loňské poptávce 110 kusů je predikovaná hodnota letošní poptávky 140 kusů.

h) Nakreslete dvourozměrný tečkový diagram s proloženou regresní přímkou.

Nakreslení regresní přímky: Návrat do Výsledky: Vícenásobná regrese – Residua/předpoklady/předpovědi - Residuální analýza – Bodové grafy – Korelace dvou proměnných – X, Y – OK.



Jiný způsob: Do dvourozměrného tečkového diagramu nakreslíme regresní přímku tak, že v tabulce 2D Bodové grafy zvolíme Typ proložení: Lineární, OK.

