

## 12. Jednoduchá lineární regrese

### 12.1. Motivace

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

Nutnost vyřešení dvou problémů:

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

### 12.2. Specifikace klasického modelu lineární regrese

$Y = m(\mathbf{X}; \beta_0, \beta_1, \dots, \beta_p) + \varepsilon$ , kde

$m(\mathbf{X}; \beta_0, \beta_1, \dots, \beta_p)$  - **teoretická regresní funkce**, která lineárně závisí na neznámých regresních parametrech  $\beta_0, \beta_1, \dots, \beta_p$  a známých funkcích  $f_1(\mathbf{X}), \dots, f_p(\mathbf{X})$ , které již

neobsahují neznámé parametry, tj.  $m(\mathbf{X}; \beta_0, \beta_1, \dots, \beta_p) = \sum_{j=0}^p \beta_j f_j(\mathbf{X})$ , přičemž  $f_0(\mathbf{X}) \equiv 1$ .

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

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

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

Pořídíme n dvojic pozorování  $(\mathbf{X}_1, y_1), \dots, (\mathbf{X}_n, y_n)$  pro  $i = 1, \dots, n$  platí:

$$y_i = m(\mathbf{X}_i; \beta_0, \beta_1, \dots, \beta_p) + \varepsilon_i.$$

O náhodných odchylkách  $\varepsilon_1, \dots, \varepsilon_n$  předpokládáme, že

- a)  $E \varepsilon_i = 0$  (odchylky nejsou systematické)
- b)  $D \varepsilon_i = \sigma^2 > 0$  (všechna pozorování jsou prováděna s touž přesností)
- c)  $Cov(\varepsilon_i, \varepsilon_j) = 0$  pro  $i \neq j$  (mezi náhodnými odchylkami neexistuje žádný lineární vztah)
- d)  $\varepsilon_i \sim N(0, \sigma^2)$

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

$\sum_{i=1}^n \left( y_i - \sum_{j=0}^p \beta_j f_j(\mathbf{X}_i) \right)^2$  nabývá svého minima pro  $\beta_j = b_j, j = 0, 1, \dots, p$ )

$\hat{m}(\mathbf{X}; b_0, \dots, b_p)$  - **empirická regresní funkce**

$\hat{y}_i = \hat{m}(\mathbf{X}_i; b_0, \dots, b_p) = \sum_{j=0}^p b_j f_j(\mathbf{X}_i)$  - **regresní odhad i-té hodnoty veličiny Y (i-tá**

predikovaná hodnota veličiny Y)

$e_i = y_i - \hat{y}_i$  - **i-té reziduum**

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

$$s^2 = \frac{S_E}{n - p - 1} - \text{odhad rozptylu } \sigma^2$$

$$S_R = \sum_{i=1}^n (\hat{y}_i - m_2)^2 - \text{regresní součet čtverců} \quad (m_2 = \frac{1}{n} \sum_{i=1}^n y_i)$$

$$S_T = \sum_{i=1}^n (y_i - m_2)^2 - \text{celkový součet čtverců} \quad (S_T = S_R + S_E)$$

$$ID^2 = \frac{S_R}{S_T} = 1 - \frac{S_E}{S_T} - \text{index determinace} \quad (0 \leq ID^2 \leq 1)$$

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

## 12.4. Maticový zápis klasického modelu lineární regrese

$y = \mathbf{X}\beta + \varepsilon$ , kde

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

$\mathbf{X} = \begin{pmatrix} 1 & f_1 & \dots & f_p \\ \dots & \dots & \dots & \dots \\ 1 & f_1 & \dots & f_p \end{pmatrix}$  - regresní matice

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

$\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$  - vektor regresních parametrů,

$\varepsilon = (\varepsilon_1, \dots, \varepsilon_p)^T$  - vektor náhodných odchylek.

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

Maticově zapsaná metoda nejmenších čtverců vede na rovnice

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

$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$  - odhad vektoru  $\beta$  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  $(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$ ;

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

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

- odhad  $\mathbf{b} \sim N_{p+1}(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$  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  $\beta$ . (Nejlepší v tom smyslu, že rozdíl varianční matice libovolného jiného nestranného odhadu vektoru  $\beta$  a varianční matice odhadu  $\mathbf{b}$  je matice pozitivně semidefinitní.)

## 12.5. Intervaly spolehlivosti pro regresní parametry

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

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

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

Na hladině významnosti  $\alpha$  testujeme

$H_0: \beta_1, \dots, \beta_p = 0, \dots, 0$  proti  $H_1: \beta_1, \dots, \beta_p \neq 0, \dots, 0$ .

(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:  $W = \{F_{1-\alpha, p, n-p-1}, \infty\}$ .

$F \in W \Rightarrow H_0$  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 (dílní 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:  $W = \{(-\infty, -t_{1-\alpha/2, n-p-1}) \cup (t_{1-\alpha/2, n-p-1}, \infty)\}$ .

$T_j \in W \Rightarrow H_0$  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.

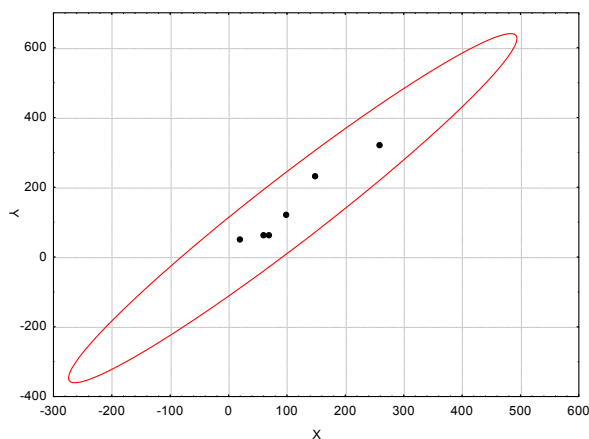
g) Vypočtete regresní odhad letošní poptávky při loňské poptávce 110 kusů.

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.

$$\text{Realizace testové statistiky: } t = \frac{r_{12} \sqrt{n-2}}{\sqrt{1-r_{12}^2}} = \frac{0,972 \sqrt{6-2}}{\sqrt{1-0,972^2}} = 3,2695 .$$

Kritický obor:

$$W = \left( -\infty, -t_{1-\alpha/2} \right) \cup \left( t_{1-\alpha/2}, \infty \right) = \left( -\infty, -t_{0,975} \right) \cup \left( t_{0,975}, \infty \right) = \\ = \left( -\infty, -2,7764 \right) \cup \left( 2,7764, \infty \right)$$

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.

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

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

Nejprve vypočítáme matici

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 6 & 660 \\ 660 & 109000 \end{pmatrix}$$

a k ní inverzní matici

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0,499084 & -0,003022 \\ -0,003022 & 0,000027 \end{pmatrix} .$$

Dále získáme součin

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 840 \\ 138500 \end{pmatrix}$$

a nakonec vektor odhadů regresních parametrů:

$$\mathbf{b} = \begin{pmatrix} 0,499084 & -0,003022 \\ -0,003022 & 0,000027 \end{pmatrix} \cdot \begin{pmatrix} 840 \\ 138500 \end{pmatrix} = \begin{pmatrix} 0,6868 \\ 1,2665 \end{pmatrix} .$$

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):

$$\hat{y} = \mathbf{Xb} = \begin{pmatrix} 1 & 20 \\ 1 & 60 \\ 1 & 70 \\ 1 & 100 \\ 1 & 150 \\ 1 & 260 \end{pmatrix} \cdot \begin{pmatrix} 0,6868 \\ 1,2665 \end{pmatrix} = \begin{pmatrix} 26,02 \\ 76,68 \\ 89,34 \\ 127,34 \\ 190,66 \\ 329,97 \end{pmatrix}.$$

Stanovíme vektor reziduí:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} 50 \\ 60 \\ 60 \\ 120 \\ 230 \\ 320 \end{pmatrix} - \begin{pmatrix} 26,02 \\ 76,68 \\ 89,34 \\ 127,34 \\ 190,66 \\ 329,97 \end{pmatrix} = \begin{pmatrix} 23,98 \\ -16,68 \\ -29,34 \\ -7,34 \\ 39,34 \\ -9,97 \end{pmatrix}.$$

Pomocí vektoru reziduí vypočteme reziduální součet čtverců:

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

$$\text{Odhad rozptylu: } s^2 = \frac{S_E}{n - p - 1} = \frac{3451,11}{6 - 1 - 1} = 862,78.$$

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

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

(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:

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

$$\text{Index determinace: } ID^2 = \frac{S_R}{S_T} = \frac{58348,89}{61800} = 0,9442.$$

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  $s_{b_j} = s \sqrt{v_{jj}}$ ,  $j = 0, 1$ , kde  $v_{jj}$  je  $j$ -tý diagonální prvek matice  $(\mathbf{X}'\mathbf{X})^{-1}$ :

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0,499084 & -0,003022 \\ -0,003022 & 0,000027 \end{pmatrix}$$

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

$$s_{b_0} = s \sqrt{v_{00}} = \sqrt{853,78} \cdot \sqrt{0,499084} = 20,6424$$

$$s_{b_1} = s \sqrt{v_{11}} = \sqrt{853,78} \cdot \sqrt{0,000027} = 0,1532$$

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

K tomu slouží vzorec  $b_j \pm t_{1-\alpha/2, n-p-1} \cdot s_{b_j}$ ,  $j = 0, 1$ .

95% interval spolehlivosti pro  $\beta_0$ :

$$d = b_0 - t_{0,975, 4} \cdot s_{b_0} = 0,6868 - 2,7764 \cdot 20,6424 = -56,63$$

$$h = b_0 + t_{0,975, 4} \cdot s_{b_0} = 0,6868 + 2,7764 \cdot 20,6424 = 58$$

Znamená to, že  $-56,63 < \beta_0 < 58$  s pravděpodobností aspoň 0,95.

95% interval spolehlivosti pro  $\beta_1$ :

$$d = b_1 - t_{0,975, 4} \cdot s_{b_1} = 1,2665 - 2,7764 \cdot 0,1532 = 0,841$$

$$h = b_1 + t_{0,975, 4} \cdot s_{b_1} = 1,2665 + 2,7764 \cdot 0,1532 = 1,692$$

Znamená to, že  $0,841 < \beta_1 < 1,692$  s pravděpodobností aspoň 0,95.

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

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

$$\text{Testová statistika } F = \frac{S_R / p}{S_E / (n - p - 1)} = \frac{58348,89 / 1}{3415,11 / (6 - 1 - 1)} = 58,384$$

$$\text{kritický obor: } w = (F_{1-\alpha/2, p, n-p-1})^\infty \cup (F_{0,95, 4, 4})^\infty = (7,7086, \infty)$$

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	$S_R = 58348,89$	$p = 1$	$S_R/p=58348,89$	68,384
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  $H_1: \beta_0 \neq 0$ .

$$\text{Testová statistika: } t_0 = \frac{b_0}{s_{b_0}} = \frac{0,6868}{20,6424} = 0,3327 ,$$

kritický obor:

$$W = \left( -\infty, -t_{1-p/2} \right) \cup \left( t_{1-p/2}, \infty \right) = \left( -\infty, -t_{0,975} \right) \cup \left( t_{0,975}, \infty \right) \\ = \left( -\infty, -2,7764 \right) \cup \left( 2,7764, \infty \right)$$

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  $H_1: \beta_1 \neq 0$ .

$$\text{Testová statistika: } t_1 = \frac{b_1}{s_{b_1}} = \frac{1,2665}{0,1532} = 8,27 ,$$

kritický obor:

$$W = \left( -\infty, -t_{1-p/2} \right) \cup \left( t_{1-p/2}, \infty \right) = \left( -\infty, -t_{0,975} \right) \cup \left( t_{0,975}, \infty \right) \\ = \left( -\infty, -2,7764 \right) \cup \left( 2,7764, \infty \right)$$

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.

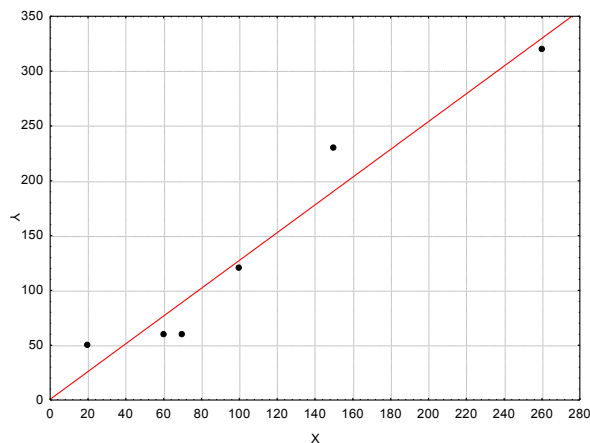
Ke stejnému výsledku dospějeme, podíváme-li se na 95% interval spolehlivosti pro  $\beta_1$ . Vypočítali jsme, že  $0,841 < \beta_1 < 1,692$  s pravděpodobností aspoň 0,95. Protože tento interval neobsahuje 0, hypotézu  $H_0: \beta_1 = 0$  zamítáme na hladině významnosti 0,05.

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:  $\hat{y} = 0,6868 + 1,2665 \cdot 110 = 140$ .

ad h)





## Výpočet pomocí systému STATISTICA

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

	1	2
	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.

Korelace (Tabulka1)											
Označ. korelace jsou významné na hlad. $p < ,05000$											
(Celé případy vynechány u ChD)											
Prom. X & prom. Y	Průměr	Sm.Odch.	r(X,Y)	r2	t	p	N	Konst. záv.: Y	Směr. záv.: Y	Konst. záv.: X	Směrnic záv.: X
X	110,0000	85,3229									
Y	140,0000	111,1755	0,971977	0,944739	8,269474	0,001167	6	0,686813	1,266484	5,566343	0,745955

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. Vypočtěte 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 : Y (Tabulka1)						
R= ,97197702 R2= ,94473932 Upravené R2= ,93092415						
F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,219						
N=6	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			0,686813	20,64236	0,033272	0,975052
X	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167

Ve výstupní tabulce najdeme koeficient  $b_0$  ve sloupci B na řádku označeném Abs. člen, koeficient  $b_1$  ve sloupci B na řádku 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, vypočtěte index determinace a interpretujte ho.

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

Analýza rozptylu (Tabulka1)					
Efekt	Součet čtv.erců	sv	Průměr čtv.erců	F	Úroveň p
Regres.	58384,89	1	58384,89	68,38420	0,001167
Rezid.	3415,11	4	853,78		
Celk.	61800,00				

Odhad rozptylu najdeme na řádku Rezid., ve sloupci Průměr čtv.erců, tedy  $s^2 = 853,78$ .

Index determinace je uveden v záhlaví původní výstupní tabulky pod označením R2. 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:  $=v_3 - v_4 * VStudent(0,975;4)$  resp.  $=v_3 + v_4 * VStudent(0,975;4)$

Výsledky regrese se závislou proměnnou : Y (Tabulka1)								
R= ,97197702 R2= ,94473932 Upravené R2= ,93092415								
F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,219								
N=6	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p	dm =v <sub>3</sub> -v <sub>4</sub> *V	hm =v <sub>3</sub> +v <sub>4</sub> *V
Abs.člen			0,686813	20,64236	0,033272	0,975052	-56,6256	57,99918
X	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167	0,841266	1,691701

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.

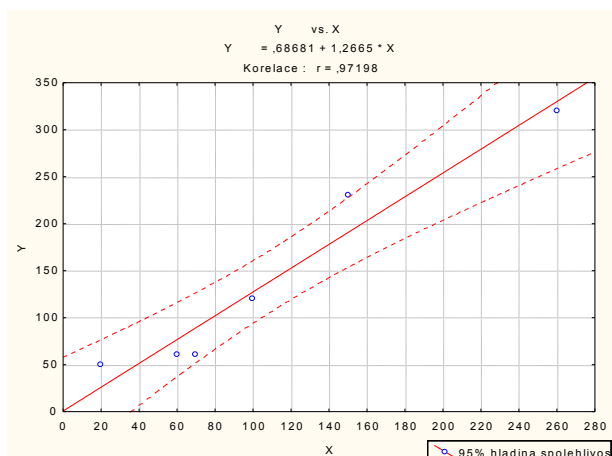
Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi Předpovědi závisle proměnné X: 110 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď.

Předpovězené hodnoty (Tabulka1)			
proměnné: Y			
Proměnná	B-váží	Hodnota	B-váží * Hodnota
X	1,266484	110,0000	139,3132
Abs. člen			0,6868
Předpověď			140,0000
-95,0%LS			106,8803
+95,0%LS			173,1197

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

**h) Nakreslete dvourozměrný tečkový diagram s proloženou regresní přímkou.**

Nakreslení regresní přímkou: 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 přímkou tak, že v tabulce 2D Bodové grafy zvolíme Typ proložení: Lineární, OK.

