Kapitola 9.: Jednoduchá lineární regresní analýza

Cíl kapitoly

Po prostudování této kapitoly budete umět
- metodou nejmenších čtverců odhadnout parametry regresní funkce
- konstruovat intervaly spolehlivosti pro regresní parametry
- testovat hypotézy o regresních parametrech
- pomocí různých kritérií posuzovat vhodnost zvolené regresní funkce

Časová zátěž

Na prostudování této kapitoly a splnění úkolů s ní spojených budete potřebovat asi 13 hodin studia.

9.1. Motivace

Cílem regresní analýzy je popsat závislost hodnot náhodné veličiny Y na hodnotách veličiny X, která může být náhodná i nenáhodná. Přitom je zapotřebí vyřešit dva problémy:

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

ad a) Při určení typu funkce je třeba provést teoretický rozbor zkoumané závislosti. Můžeme např. zkoumat závislost ceny ojetého auta (veličina Y) na jeho stáří (veličina X). Je zřejmé, že s rostoucím stářím bude klesat cena, ale není jasné, zda lineárně, kvadraticky či dokonce exponenciálně.
Vždy se snažíme o to aby regresní model byl jednoduchý, tj. aby neobsahoval příliš mnoho parametrů. Připadá-li v úvahu více funkcí, posuzujeme jejich vhodnost pomocí různých kritérií – viz dále.
Často však nemáme dostatek informací k provedení teoretického rozboru. Pak se snažíme odhadnout typ funkce pomocí dvourozměrného tečkového diagramu.
Zde se omezíme na funkce, které závisejí lineárně na parametrech $\beta_0, \beta_1, \ldots, \beta_p$.
Zvláštní pozornost budeme věnovat polynomiální funkci 1. stupně $y = \beta_0 + \beta_1 x$.

ad b) Odhady $b_0, b_1, \ldots, b_p$ neznámých parametrů $\beta_0, \beta_1, \ldots, \beta_p$ získáme na základě dvourozměrného datového souboru $\begin{pmatrix} x_1 & y_1 \\ \ldots & \ldots \\ x_n & y_n \end{pmatrix}$ metodou nejmenších čtverců, tj. z podmínky, aby součet čtverců odchylek zjištěných a odhadnutých hodnot byl minimální.

9.2. Klasický model lineární regrese

9.2.1. Popis modelu

Nechť závislost náhodné veličiny Y na veličině X je vyjádřena modelem

$$Y = m(x;\beta_0,\beta_1,\ldots,\beta_p) + \varepsilon,$$

kde $m(x;\beta_0,\beta_1,\ldots,\beta_p)$ je deterministická složka modelu a ε je náhodná složka modelu. Je to náhodná odchylka od deterministické závislosti Y na X. Popisuje závislost vysvětlované veličiny na neznámých nebo nepozorovaných veličinách a popisuje i vliv náhody. Nelze ji funkčně vyjádřit.

Jako kritérium kvality predikce závisle proměnné veličiny Y na nezávisle proměnné veličině X se obvykle volí střední kvadratická chyba predikce $E\left(\left[Y - m\left(x;\beta_0,\beta_1,\ldots,\beta_p\right)\right]^2\right)$. Lze dokázat, že střední kvadratická chyba predikce je minimální, když $m\left(x;\beta_0,\beta_1,\ldots,\beta_p\right) = E\left(Y|x\right)$, tj. závislost Y na X budeme modelovat pomocí podmíněné střední hodnoty neboli pomocí regresní funkce veličiny Y vzhledem k veličině X.

Nadále předpokládáme, že regresní funkce lineárně závisí na neznámých regresních parametrech $\beta_0,\beta_1,\ldots,\beta_p$ a známých funkcích $f_1(x),\ldots,f_p(x)$, které již neobsahují neznámé parametry, tj. $m\left(x;\beta_0,\beta_1,\ldots,\beta_p\right) = \sum_{j=0}^{p}\beta_j f_j(x)$, přičemž $f_0(x) \equiv 1$. Regresní parametry $\beta_0,\beta_1,\ldots,\beta_p$ lze interpretovat tak, že parametr β_j vyjadřuje průměrnou změnu hodnoty Y při růstu hodnoty funkce $f_j(x)$ o jednu jednotku za předpokladu, že hodnoty ostatních funkcí $f_k(x)$, k = 1, …, p, $k \neq j$, zůstanou nezměněné.

Pořídíme n dvojic pozorování $(x_1,y_1),\ldots,(x_n,y_n)$, tj. dvourozměrný datový soubor

$$\begin{pmatrix} x_1 & y_1 \\ \ldots & \ldots \\ x_n & y_n \end{pmatrix}.$$

Pro i = 1, ..., n platí: $y_i = m\left(x_i;\beta_0,\beta_1,\ldots,\beta_p\right) + \varepsilon_i$.

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

a) $E(\varepsilon_i) = 0$ (odchylky nejsou systematické)

b) $D(\varepsilon_i) = \sigma^2 > 0$ (všechna pozorování jsou prováděna s touž přesností – jsou homoskedastická)

c) $C(\varepsilon_i,\varepsilon_j) = 0$ 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.

9.2.2. Označení

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

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

$\hat{m}\left(x;b_0,\ldots,b_p\right)$ - empirická regresní funkce

$\hat{y}_i = \hat{m}\left(x_i;b_0,\ldots,b_p\right) = \sum_{j=0}^{p} b_j f_j(x_i)$ - regresní odhad i-té hodnoty veličiny Y (i-tá predikovaná hodnota veličiny Y)

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

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

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

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

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

Pro součty čtverců platí: $S_T = S_R + S_E$.

9.2.3. Maticový zápis klasického modelu lineární regrese

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

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

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

(předpokládáme, že h(**X**) = p+1 < n)

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

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

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

Např. pro regresní přímku $y = \beta_0 + \beta_1 x + \varepsilon$ má regresní matice tvar $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ a vektor regresních parametrů je $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$

Maticově zapsaná metoda nejmenších čtverců vede na rovnice

X'**Xβ** = **X**'**y** – systém normálních rovnic,

b = (**X**'**X**)$^{-1}$ **X**' **y** – odhad vektoru **β** získaný metodou nejmenších čtverců.

Pro regresní přímku získáme řešením systému normálních rovnic odhady

$b_1 = \frac{s_{12}}{s_1^2}$ a $b_0 = m_2 - b_1 m_1$ kde s_{12} je výběrová kovariance hodnot (x_i, y_i), i = 1, ..., n a s_1^2 je výběrový rozptyl hodnot $x_1, \ldots, x_n$. Regresní přímku můžeme vyjádřit ve tvaru

$y = m_2 + \frac{s_{12}}{s_1^2}(x - m_1) + \varepsilon$.

$\hat{\mathbf{y}} = \mathbf{Xb}$ – vektor regresních odhadů (vektor predikce)

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

Vlastnosti odhadu **b**:

- odhad **b** je lineární, neboť je vytvořen lineární kombinací pozorování $y_1, \ldots, y_n$ s maticí vah $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$;
- odhad **b** je nestranný, neboť E(**b**) = **β**;
- odhad **b** má varianční matici var **b** = σ^2(**X**'**X**)$^{-1}$;
- odhad **b** ~ N_{p+1}(**β**, σ^2 (**X**'**X**)$^{-1}$) vzhledem k platnosti podmínky (d);
- pro odhad **b** platí Gaussova - Markovova věta: Odhad **b** = (**X**'**X**)$^{-1}$ **X**'**y** je nejlepší nestranný lineární odhad vektoru **β**. (Nejlepší v tom smyslu, že rozdíl varianční matice libovolného jiného nestranného odhadu vektoru **β** a varianční matice odhadu **b** je matice pozitivně semidefinitní.)

9.2.4. 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

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

Řešení:

Sestavíme regresní matici.

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \text{ tedy } \mathbf{X} = \begin{pmatrix} 1 & 20 \\ 1 & 60 \\ 1 & 70 \\ 1 & 100 \\ 1 & 150 \\ 1 & 260 \end{pmatrix}.$$

Podle vzorce $\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$ 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.

Výpočet pomocí systému STATISTICA

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

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

N=6	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					
	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.

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

Na hladině významnosti α testujeme

H_0: $(\beta_1,\ldots,\beta_p)' = (0,\ldots,0)'$ proti H_1: $(\beta_1,\ldots,\beta_p)' \neq (0,\ldots,0)'$.

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

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

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

Jestliže $F \in W$, pak H_0 zamítáme na hladině významnosti α.
Výsledky F-testu zapisujeme do tabulky analýzy rozptylu:

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

Z této tabulky také můžeme snadno získat odhad rozptylu σ^2: $s^2 = \dfrac{S_E}{n-p-1}$.

9.2.6. Testování významnosti regresních parametrů (dílčí t-testy)

Na hladině významnosti α pro j = 0,1, ..., p testujeme hypotézu
H_0: $\beta_j = 0$ proti H_1: $\beta_j \neq 0$.

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

Kritický obor: $W = (-\infty, -t_{1-\alpha/2}(n-p-1)\rangle \cup \langle t_{1-\alpha/2}(n-p-1), \infty)$.

Pokud $T_j \in W$, H_0 zamítáme na hladině významnosti α.

Upozornění: Test významnosti směrnice regresní přímky (tj. test H_0: $\beta_1 = 0$ proti H_1: $\beta_1 \neq 0$) je ekvivalentní testování hypotézy o nekorelovanosti veličin X, Y (tj. testu H_0: $\rho = 0$ proti H_1: $\rho \neq 0$). Jestliže koeficient korelace veličin X, Y je blízký 0, nemá smysl počítat parametry regresní přímky.

9.2.7. Příklad
Pro zadání z příkladu 9.2.4. najděte odhad rozptylu, proveďte celkový F-test, a proveďte rovněž dílčí t-testy.

Řešení:
Nejprve vypočteme vektor regresních odhadů proměnné Y (vektor predikce):

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

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:

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

Odhad rozptylu: $s^2 = \frac{S_E}{n-p-1} = \frac{3415{,}11}{6-1-1} = 853{,}78$.

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 nx1 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

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

(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:
$S_R = S_T - S_E = 61800 - 3451{,}11 = 58348{,}89$.

Provedení celkového F-testu:
na hladině významnosti $\alpha = 0{,}05$ testujeme H_0: $\beta_1 = 0$ proti H_1: $\beta_1 \neq 0$.

Testová statistika $F = \dfrac{S_R / p}{S_E /(n-p-1)} = \dfrac{58348{,}89/1}{3415{,}11/(6-1-1)} = 68{,}384$,

kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty) = \langle F_{0{,}95}(1{,}4), \infty) = \langle 7{,}7086, \infty)$.

Protože se testová statistika realizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_1 (tj. směrnice regresní přímky) zamítáme na hladině významnosti 0,05. Výsledky testování významnosti modelu jako celku zapíšeme do tabulky ANOVA:

zdroj variability	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	$S_T = 61800$	$n-1 = 5$	-	-

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: $t_0 = \dfrac{b_0}{s_{b_0}} = \dfrac{0{,}6868}{20{,}6424} = 0{,}3327$,

kritický obor:

$$W = \langle -\infty, -t_{1-\alpha/2}(n-p-1)\rangle \cup \langle t_{1-\alpha/2}(n-p-1), \infty) = \langle -\infty, -t_{0{,}975}(4)\rangle \cup \langle t_{0{,}975}(4), \infty) =$$
$$= \langle -\infty, -2{,}7764\rangle \cup \langle 2{,}7764, \infty)$$

Protože se testová statistika nerealizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_0 (tj. posunutí regresní přímky) nezamítáme na hladině významnosti 0,05.
Ke stejnému výsledku dospějeme, podíváme-li se na 95% interval spolehlivosti pro β_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: $t_1 = \dfrac{b_1}{s_{b_1}} = \dfrac{1{,}2665}{0{,}1532} = 8{,}27$,

kritický obor:

$$W = \langle -\infty, -t_{1-\alpha/2}(n-p-1)\rangle \cup \langle t_{1-\alpha/2}(n-p-1), \infty) = \langle -\infty, -t_{0{,}975}(4)\rangle \cup \langle t_{0{,}975}(4), \infty) =$$
$$= \langle -\infty, -2{,}7764\rangle \cup \langle 2{,}7764, \infty)$$

Protože se testová statistika realizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_1 (tj. směrnice regresní přímky) zamítáme na hladině významnosti 0,05.
Ke stejnému výsledku dospějeme, podíváme-li se na 95% interval spolehlivosti pro β_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 β_1 ekvivalentní s celkovým F-testem.

Výpočet pomocí systému STATISTICA:
Abychom získali odhad rozptylu, vrátíme se do Výsledky – vícenásobná regrese – Detailní výsledky – ANOVA.

Efekt	Analýza rozptylu (Tabulka1)				
	Součet čtverců	sv	Průměr čtverců	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 čtverců, tedy $s^2 = 853{,}78$.

Testovou statistiku F-testu a odpovídající p-hodnotu najdeme v záhlaví výstupní tabulky regrese:

N=6	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					
	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

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.

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.

9.2.8. Intervaly spolehlivosti pro regresní parametry

Označme v_{jj} j-tý diagonální prvek matice $(\mathbf{X'X})^{-1}$ a $s_{b_j} = s\sqrt{v_{jj}}$ tzv. směrodatnou chybu odhadu b_j. Pro j = 0, 1, ..., p se statistika $T_j = \dfrac{b_j - \beta_j}{s_{b_j}}$ řídí rozložením $t(n-p-1)$, tedy

100(1- α)% interval spolehlivosti pro β_j má meze: $b_j \pm t_{1-\alpha/2}(n-p-1)s_{b_j}$.

(S intervaly spolehlivosti souvisí relativní chyby odhadů regresních parametrů. Získají se tak, že se vypočítá absolutní hodnota podílu poloviční šířky intervalu spolehlivosti a hodnoty odhadu. Relativní chyba odhadu by neměla přesáhnout 10%.)

9.2.9. Příklad

Pro zadání z příkladu 9.2.4. najděte 95% intervaly spolehlivosti pro regresní parametry a zjistěte relativní chyby odhadů regresních parametrů.

Řešení:

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 $(\mathbf{X'X})^{-1}$:

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

Přitom si uvědomíme, že $v_{00} = 0{,}499084$, $v_{11} = 0{,}000027$, $s^2 = 853{,}78$.

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

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

Stanovíme meze 95% intervalů spolehlivosti pro regresní parametry β_0 a β_1. K tomu slouží vzorec $b_j \pm t_{1-\alpha/2}(n-p-1)s_{b_j}$, j = 0, 1.

95% interval spolehlivosti pro β_0:

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

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

Znamená to, že -56,63 < β_0 < 58 s pravděpodobností aspoň 0,95.

Relativní chyba odhadu β_0: $\left|\frac{(58+56{,}63)/2}{0{,}6868}\right| \cdot 100\% = 8342\%$

95% interval spolehlivosti pro β_1:

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

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

Znamená to, že 0,841< β_1 < 1,692 s pravděpodobností aspoň 0,95.

Relativní chyba odhadu β_1: $\left|\frac{(1{,}692-0{,}841)/2}{1{,}2665}\right| \cdot 100\% = 33{,}6\%$.

Výpočet pomocí systému STATISTICA:

Ve výstupní tabulce výsledků regrese přidáme za proměnnou Úroveň p tři nové proměnné: dm (pro dolní meze 95% intervalů spolehlivosti pro regresní parametry), hm (pro horní meze 95% intervalů spolehlivosti pro regresní parametry) a chyba (pro relativní chyby odhadů regresních parametrů).

Do Dlouhého jména proměnné dm napíšeme:

=v3-v4*VStudent(0,975;4)

Do Dlouhého jména proměnné hm napíšeme:

=v3+v4*VStudent(0,975;4)

Do Dlouhého jména proměnné chyba napíšeme:

=100*abs(0,5*(hm-dm)/v3)

N=6	Výsledky regrese se závislou proměnnou : Prom2 (Tabulka1) R= ,97197702 R2= ,94473932 Upravené R2= ,93092415 F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,219								
	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p	dm =v3-v4*V	hm =v3+v4*	chyba =100*abs
Abs.člen			0,686813	20,64236	0,033272	0,975052	-56,6256	57,99918	8344,681
Prom1	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167	0,841266	1,691701	33,57463

Vidíme, že -56,63 < β_0 < 58 s pravděpodobností aspoň 0,95 a 0,841< β_1 < 1,692 s pravděpodobností aspoň 0,95.

Relativní chyba odhadu parametru β_0 činí 8344,68% a relativní chyba odhadu parametru β_1 činí 33,57%. V obou případech jsou chyby příliš velké.

9.2.10. Kritéria pro posouzení vhodnosti zvolené regresní funkce

a) Index determinace

Číslo $ID^2 = \frac{S_R}{S_T} = 1 - \frac{S_E}{S_T}$ se nazývá index determinace.

Vlastnosti indexu determinace:

- nabývá hodnot z intervalu $\langle 0,1 \rangle$;
- udává, jakou část variability závisle proměnné veličiny Y lze vysvětlit zvolenou regresní funkcí (často se vyjadřuje v %);
- je zároveň mírou těsnosti závislosti veličiny Y na veličině X;
- je to obecná míra, nezávislá na typu regresní funkce (lze použít i pro měření nelineární závislosti);
- je to míra, která nebere v úvahu počet parametrů regresní funkce. U regresních funkcí s více parametry vychází tedy obvykle vyšší než u regresních funkcí s méně parametry;
- tato míra není symetrická.

Za vhodnější se považuje ta regresní funkce, pro niž je index determinace vyšší. V případě, že porovnáváme několik modelů s rozdílným počtem parametrů, používáme adjustovaný index determinace: $ID_{adj}^{\ 2} = ID^2 - \frac{(1 - ID^2)p}{n - p - 1}$. Malá hodnota indexu determinace nemusí znamenat, že mezi veličinami X, Y je nízká závislost, může signalizovat nevhodnou volbu typu regresní funkce.

V případě regresní přímky je index determinace roven kvadrátu koeficientu korelace:

$ID^2 = r_{12}^{\ 2}$

b) Testové kritérium F

Za vhodnější je považována ta regresní funkce, u níž je hodnota testové statistiky

$F = \frac{S_R / p}{S_E / (n - p - 1)}$ pro test významnosti modelu jako celku vyšší.

c) Reziduální součet čtverců a reziduální rozptyl

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

Za vhodnější považujeme funkci, která má reziduální součet čtverců nižší. Reziduální součet čtverců lze použít pouze tehdy, když srovnáváme funkce se stejným počtem parametrů.

Reziduální rozptyl: $s^2 = \frac{S_E}{n - p - 1}$

Za vhodnější považujeme tu funkci, která má reziduální rozptyl nižší. Reziduální rozptyl můžeme použít vždy, bez ohledu na to, kolik parametrů mají srovnávané regresní funkce.

d) Střední absolutní procentuální chyba predikce (MAPE)

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Za vhodnější považujeme tu funkci, která má MAPE nižší.

e) Analýza reziduí

Rezidua považujeme za odhady náhodných odchylek a klademe na ně stejné požadavky jako na náhodné odchylky, tj. mají být nezávislá, normálně rozložená s nulovou střední hodnotou a konstantním rozptylem (tj. jsou homoskedastická).
Nezávislost reziduí (autokorelaci) posuzujeme např. pomocí Durbinovy – Watsonovy statistiky, která by se měla nacházet v intervalu $\langle 1{,}4;2{,}6\rangle$ (to je ovšem pouze orientační vodítko, korektní postup spočívá v porovnání této statistiky s tabelovanou kritickou hodnotou).
Normalitu reziduí ověřujeme pomocí testů normality (např. Lilieforsovou variantou Kolmogorovova – Smirnovova testu nebo Shapirovým – Wilksovým testem) či graficky pomocí N-P plotu.
Testování nulovosti střední hodnoty reziduí provádíme pomocí jednovýběrového t-testu.
Homoskedasticitu reziduí posuzujeme pomocí grafu závislosti reziduí na predikovaných hodnotách. V tomto grafu by rezidua měla být rovnoměrně rozptýlena.

9.2.11. Příklad

Pro zadání z příkladu 9.2.4. vypočtěte index determinace a interpretujte ho. Vypočtěte rovněž střední absolutní procentuální chybu predikce a najděte regresní odhad letošní poptávky při loňské poptávce 110 kusů. Proveďte analýzu reziduí. Nakreslete regresní přímku do dvourozměrného tečkového diagramu.

Řešení:

Index determinace se počítá podle vzorce $ID^2 = \frac{S_R}{S_T}$. V příkladu 9.2.7. bylo vypočteno, že regresní součet čtverců $S_R = 58348{,}89$ a celkový součet čtverců $S_T = 61800$. Index determinace: $ID^2 = \frac{58348{,}89}{61800} = 0{,}9442$.
Znamená to, že variabilita hodnot závisle proměnné veličiny je z 94,42% vysvětlena regresní přímkou.
Pro regresní přímku můžeme využít toho, že $ID^2 = r_{12}{}^2$. V našem případě zjistíme, že $r_{12} = 0{,}971977$, tedy $ID^2 = 0{,}971977^2 = 0{,}9447$.
MAPE se počítá podle vzorce $MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$. V příkladě 9.2.7. jsme vypočetli vektor reziduí $y_i - \hat{y}_i = \begin{pmatrix} 23{,}98 \\ -16{,}68 \\ -29{,}34 \\ -7{,}34 \\ 39{,}34 \\ -9{,}97 \end{pmatrix}$ a vektor pozorování $\begin{pmatrix} 50 \\ 60 \\ 60 \\ 120 \\ 230 \\ 320 \end{pmatrix}$.
Tedy dostáváme MAPE =
$\frac{1}{6}\left(\left|\frac{23{,}98}{50}\right| + \left|\frac{-16{,}68}{60}\right| + \left|\frac{-29{,}34}{60}\right| + \left|\frac{-7{,}34}{120}\right| + \left|\frac{39{,}34}{230}\right| + \left|\frac{-9{,}97}{320}\right|\right) = 0{,}2517$.

Regresní odhad pro x = 110 dostaneme pouhým dosazením do rovnice regresní přímky:
$\hat{y} = 0{,}6868 + 1{,}2665 \cdot 110 = 140$.
Při loňské poptávce 110 kusů by odhad letošní poptávky činil 140 kusů zboží.

Výpočet pomocí systému STATISTICA:
Index determinace je uveden v záhlaví původní výstupní tabulky pod označením R2:

N=6	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					
	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

V našem případě $ID^2 = 0,9447$, tedy variabilita letošní poptávky je z 94,5% vysvětlena regresní přímkou.

Abychom vypočetli MAPE, tak ve výsledcích Vícenásobné regrese zvolíme záložku Rezidua / předpoklady / předpovědi – Reziduální analýza – Uložit – Uložit rezidua a předpovědi – Vybrat vše – OK. Ve vzniklé tabulce přidáme proměnnou chyby a do jejího Dlouhého jména napíšeme
=100*abs(v4/v2)
Pak spočteme průměr této proměnné a zjistíme, že MAPE = 25,17%.

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

Proměnná	Předpovězené hodnoty (Tabulka1 proměnné: Y		
	B-váž.	Hodnota	B-váž. * Hodnot
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.

Při analýze reziduí nejprve posoudíme nezávislost reziduí pomocí Durbinova – Watsonovy statistiky: Na záložce Rezidua/předpoklady/předpovědi zvolíme Reziduální analýza - Pokročilá – Durbinova – Watsonova statistika.

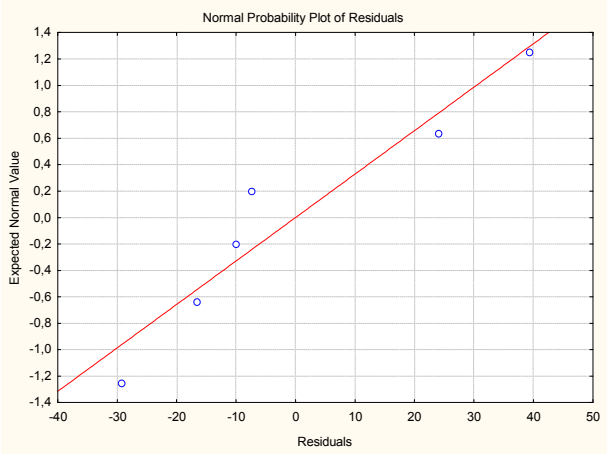
	Durbin-Watsonovo d (poptavka.sta a sériové korelace reziduí	
	Durbin-Watson.d	Sériové korelace
Odhad	2,022847	-0,113505

Tato statistika je blízká číslu 2, tedy rezidua můžeme považovat za nezávislá.

Normalitu reziduí posoudíme Lilieforsovou variantou K-S testu a S-W testem:

Proměnná	Testy normality (Tabulka6)				
	N	max D	Lilliefors p	W	p
Rezidua	6	0,277184	p < ,15	0,911935	0,449251

Ani jeden z testů nezamítá hypotézu o normalitě reziduí na hladině významnosti 0,05. Graficky posoudíme normalitu N-P plotem:



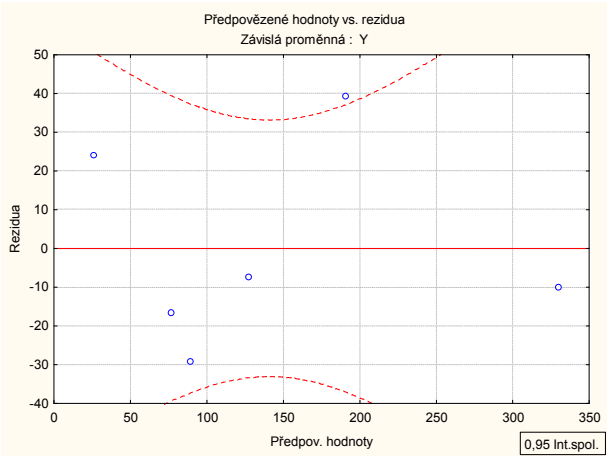
Vidíme, že rezidua se od ideální přímky neodchylují příliš výrazně.

Nulovost střední hodnoty reziduí ověříme jednovýběrovým t-testem:

Proměnná	Test průměrů vůči referenční konstantě (hodnotě) (Tabulka6)							
	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Rezidua	-0,000003	26,13469	6	10,66944	0,00	-0,000000	5	1,000000

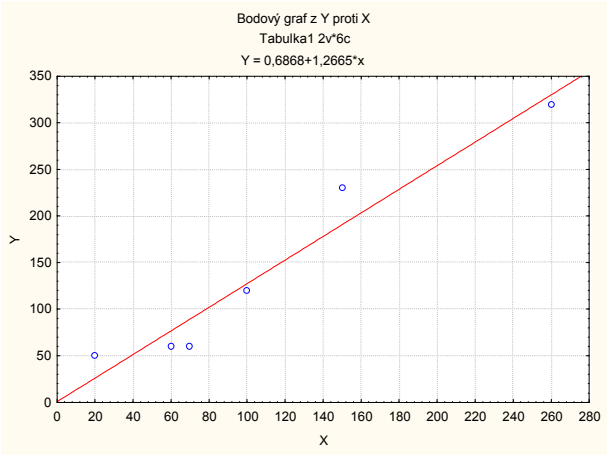
Vidíme, že p-hodnota je 1, tudíž na hladině významnosti 0,05 nezamítáme hypotézu, že rezidua mají nulovou střední hodnotu.

Homoskedasticitu reziduí posoudíme pomocí grafu závislosti reziduí na predikovaných hodnotách veličiny Y: Na záložce Rezidua/předpoklady/předpovědi zvolíme Reziduální analýza – Bodové grafy – Předpovědi vs. Rezidua



Rezidua nevykazují žádnou závislost na predikovaných hodnotách.

Nakonec do dvourozměrného tečkového diagramu nakreslíme regresní přímku. V menu 2D Bodové grafy zvolíme Typ proložení: Lineární, OK.



Vzhled grafu naznačuje, že přímka je vhodným modelem závislosti letošní poptávky na loňské poptávce.

Shrnutí

Jednoduchá lineární regresní analýza slouží k tomu, aby popsala závislost náhodné veličiny Y na veličině X pomocí regresní funkce (tj. podmíněné střední hodnoty $\mathrm{E}(\mathrm{Y}|\mathrm{x})$), která je lineární v parametrech. Důležitým úkolem regresní analýzy je nalezení vhodného typu regresní funkce a odhad jejích parametrů. V případě lineárních regresních modelů odhadujeme neznámé parametry *metodou nejmenších čtverců* na základě znalosti dvourozměrného datového souboru n hodnot veličin X a Y.

Pokud se náhodné odchylky regresního modelu od skutečnosti řídí normálním rozložením s nulovou střední hodnotou a konstantním rozptylem a přitom jsou nezávislé, pak můžeme konstruovat *intervaly spolehlivosti pro regresní parametry*, pomocí *F-testu* ověřovat významnost modelu jako celku a pomocí *dílčích t-testů* ověřovat významnost jednotlivých regresních parametrů.

Vhodnost zvoleného regresního modelu posuzujeme pomocí různých kritérií, např. pomocí *indexu determinace*, pomocí *odhadu reziduálního rozptylu* nebo pomocí *střední absolutní procentuální chyby predikce* (MAPE).

Důležitá je rovněž analýza reziduí, v jejímž průběhu ověřujeme *nezávislost reziduí*, jejich *normalitu*, *nulovost střední hodnoty* a *homoskedasticitu rozptylu*.

Kontrolní otázky

1. Jak je definována podmíněná střední hodnota $\mathrm{E}(\mathrm{Y}|\mathrm{x})$ a k čemu slouží?
2. Popište princip metody nejmenších čtverců.
3. Jak je definován reziduální, regresní a celkový součet čtverců a jaký je mezi nimi vztah?
4. Co platí pro hodnost regresní matice?
5. Jaké vlastnosti má odhad vektoru regresních parametrů získaný metodou nejmenších čtverců?
6. Jak získáme relativní chyby odhadů regresních parametrů?

7. K čemu slouží celkový F-test a dílčí t-testy?
8. Jaká kritéria používáme pro hodnocení kvality regresního modelu?

Autokorekční test

1. Který z následujících předpokladů klasického lineárního regresního modelu $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ je chybný?
a) $h(\mathbf{X}) = p+1 \geq n$
b) $var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$
c) $E(\boldsymbol{\varepsilon}) = \mathbf{0}$

2. Jedno z následujících tvrzení o vektoru **b**, který je získán metodou nejmenších čtverců jako odhad vektoru regresních parametrů **β**, je pravdivé. Které to je?
a) var $\mathbf{b} = \sigma^2(\mathbf{XX}')^{-1}$
b) odhad **b** je vychýlený odhad vektoru **β**
c) odhad **b** je lineární

3. Máte k dispozici výstupní tabulku pro model regresní přímky:

N=19	b*	Std.Err. of b*	b	Std.Err. of b	t(17)	p-value
Intercept			238,4631	22,62066	10,54183	0,000000
Var1	0,964280	0,064244	29,7785	1,98396	15,00958	0,000000

Pokud se hodnota nezávisle proměnné veličiny X zvýší o 5 jednotek, pak regresní odhad hodnoty závisle proměnné veličiny Y se zvýší o:
a) 1192,3 jednotek
b) 148,9 jednotek
c) 29,8 jednotek

4. Máte k dispozici neúplnou tabulku ANOVA pro model regresní přímky:

Effect	Sums of Squares	df	Mean Squares	F	p-value
Regress.	505451,3	1	505451,3	225,2875	0,000000
Residual	38140,9	17			
Total	543592,2				

Odhad rozptylu je:
a) 2243,6
b) 29732,4
c) 28610,1

5. Při výpočtu adjustovaného indexu determinace nepotřebujeme znát:
a) počet pozorování
b) hodnost regresní matice
c) počet parametrů v regresním modelu

6. Výběrový koeficient korelace vypočtený na základě náhodného výběru z dvourozměrného normálního rozložení nabyl hodnoty -0,94. Pokud bychom modelovali závislost veličiny Y na

veličině X pomocí regresní přímky, jakou část variability hodnot veličiny Y by nevysvětlovala regresní přímka?
a) 88,36%
b) 11,64%
c) 6%

Správné odpovědi: 1a), 2c), 3b), 4a), 5b), 6b)

Příklady

1. U osmi náhodně vybraných firem poskytujících odborné konzultace v oblasti jakosti výroby byly v roce 2008 zjištěny počty zaměstnanců (náhodná veličina X) a roční obraty (náhodná veličina Y, v miliónech Kč), jak je uvedeno v tabulce:

Číslo firmy	1	2	3	4	5	6	7	8
X	3	5	5	8	9	11	12	15
Y	0,8	1,2	1,5	1,9	1,8	2,4	2,5	3,1

Předpokládáme, že závislost ročního obratu na počtu zaměstnanců lze popsat regresní přímkou. K dispozici jsou částečné výstupy regresní analýzy ze systému STATISTICA:

N=8	Beta	Sm.chyba beta	B	Sm.chyba B
Abs.člen			0,361207	0,121417
X	0,984798	0,070914	0,181034	0,013036

Efekt	Součet čtverců	sv	Průměr čtverců	F	Úroveň p
Regres.	3,801724	1	3,801724	192,8571	0,000009
Rezid.	0,118276	6	0,019713		
Celk.	3,920000				

a) Napište rovnici regresní přímky vyjadřující závislost Y na X. Interpretujte úsek a směrnici regresní přímky.
b) Najděte 95% intervaly spolehlivosti pro parametry regresní přímky a s jejich pomocí testujte na hladině významnosti hypotézy o nevýznamnosti úseku a směrnice regresní přímky.
c) Vypočtěte index determinace a interpretujte ho.

Výsledek:
ad a) $y = 0{,}361207 + 0{,}181034x$
Pokud firma nebude mít žádné zaměstnance (tzn., že pracují pouze majitelé), bude roční obrat asi 361 000 Kč.
Pokud se zvýší počet zaměstnanců o jednoho, vzroste roční obrat asi o 181 000 Kč.

ad b)
95% interval spolehlivosti pro β_0:

$$d = b_0 - t_{0,975}(6)s_{b_0} = 0{,}361207 - 2{,}4469 \cdot 0{,}121417 = 0{,}064111$$

$$h = b_0 + t_{0,975}(6)s_{b_0} = 0{,}361207 + 2{,}4469 \cdot 0{,}121417 = 0{,}658303$$

Znamená to, že $0{,}06411 < \beta_0 < 0{,}65303$ s pravděpodobností aspoň 0,95.

Protože tento interval neobsahuje číslo 0, na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti úseku regresní přímky.
95% interval spolehlivosti pro β_1:

$$d = b_1 - t_{0,975}(6)s_{b_1} = 0{,}181034 - 2{,}7764 \cdot 0{,}013036 = 0{,}149137$$

$$h = b_1 + t_{0,975}(6)s_{b_1} = 0{,}181034 + 2{,}7764 \cdot 0{,}013036 = 0{,}212932$$

Znamená to, že $0{,}149137 < \beta_1 < 0{,}212932$ s pravděpodobností aspoň 0,95.
Protože tento interval neobsahuje číslo 0, na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti směrnice regresní přímky.
(Tento interval spolehlivosti nám vlastně udává, že při zvýšení počtu zaměstnanců o jednoho se přírůstek ročního obratu firmy bude s pravděpodobností aspoň 0,95 pohybovat v intervalu 149 000 Kč až 213 000 Kč.)

ad c) $ID^2 = \frac{S_R}{S_T} = \frac{3{,}801724}{3{,}92} = 0{,}9698$

Znamená to, že variabilita ročního obratu je z téměř 97 % vysvětlena regresní přímkou.

2. V modelu regresní přímky je index determinace roven 0,8 a reziduální rozptyl je 100. Jaký je rozptyl hodnot závisle proměnné veličiny?
Výsledek:
$s_2^2 = 500$

3. Určitý lék je přepravován v ampulkách, které jsou baleny po 1000 kusech v jednom kartonu. U 10 náhodně vybraných kartonů bylo zjištěno, kolikrát byl karton překládán (veličina X) a počet poškozených ampulek při převzetí zásilky (veličina Y).

X	1	0	2	0	3	1	0	1	2	0
Y	16	9	17	12	22	13	8	15	19	11

Na základě těchto údajů, které považujeme za realizace náhodného výběru z dvourozměrného normálního rozložení, byly vypočteny parametry regresní přímky, která vystihuje závislost počtu poškozených ampulek na počtu překládání: $b_0 = 10{,}2$, $b_1 = 4$. Směrodatné chyby odhadů regresních parametrů jsou: $s_{b_0} = 0{,}663325, s_{b_1} = 0{,}469042$. Na hladině významnosti 0,05 testujte hypotézy o nevýznamnosti parametrů β_0 a β_1. V obou případech vypočtěte hodnotu testové statistiky, najděte kritický obor a napište rozhodnutí o nulové hypotéze.
Výsledek:
Na hladině významnosti $\alpha = 0{,}05$ testujeme H_0: $\beta_0 = 0$ proti H_1: $\beta_0 \neq 0$.

Testová statistika: $t_0 = \frac{b_0}{s_{b_0}} = \frac{10{,}2}{0{,}663325} = 15{,}3771$,

kritický obor:

$$W = \left(-\infty, -t_{1-\alpha/2}(n-p-1)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-p-1), \infty\right) = \left(-\infty, -t_{0,975}(8)\right\rangle \cup \left\langle t_{0,975}(8), \infty\right) =$$

$$= \left(-\infty, -2{,}306\right\rangle \cup \left\langle 2{,}306, \infty\right)$$

Protože se testová statistika realizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_0 (tj. posunutí regresní přímky) zamítáme 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: $t_1 = \frac{b_1}{s_{b_1}} = \frac{4}{0{,}469042} = 8{,}528$,

kritický obor:

$$W = \left(-\infty, -t_{1-\alpha/2}(n-p-1)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-p-1), \infty\right) = \left(-\infty, -t_{0,975}(8)\right\rangle \cup \left\langle t_{0,975}(8), \infty\right) =$$

$$= \left(-\infty, -2{,}306\right\rangle \cup \left\langle 2{,}306, \infty\right)$$

Protože se testová statistika realizuje v kritickém oboru, hypotézu o nevýznamnosti regresního parametru β_1 (tj. směrnice regresní přímky) zamítáme na hladině významnosti 0,05.