Příklady na cvičení Výpočetní statistika, téma **Jednoduchá lineární regrese**

Příklad 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:

$$d = b_1 - t_{0,975}(6)s_{b_1} = 0{,}181034 - 2{,}4469 \cdot 0{,}013036 = 0{,}149137$$

$$h = b_1 + t_{0,975}(6)s_{b_1} = 0{,}181034 + 2{,}4469 \cdot 0{,}013036 = 0{,}212932$$

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.

Příklad 2.: V modelu regresní přímky je index determinace roven 0,8 a reziduální rozptyl je 100. Jaký je rozptyl hodnot závisle proměnné veličiny?

$$ID^2 = 1 - \frac{S_E}{S_T} \Rightarrow S_T = \frac{S_E}{1 - ID^2} = \frac{(n-p)s^2}{1 - ID^2}, s_Y{}^2 = \frac{S_T}{n-1} = \frac{(n-p)s^2}{(n-1)(1 - ID^2)} = |p = 1| =$$

$$= \frac{s^2}{1 - ID^2} = \frac{100}{1 - 0{,}8} = 500$$

Příklad 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:

$$W = (-\infty, -t_{1-\alpha/2}(n-p-1)\rangle \cup \langle t_{1-\alpha/2}(n-p-1), \infty) = (-\infty, -t_{0{,}975}(8)\rangle \cup \langle t_{0{,}975}(8), \infty) =$$
$$= (-\infty, -2{,}306\rangle \cup \langle 2{,}306, \infty).$$

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:

$$W = (-\infty, -t_{1-\alpha/2}(n-p-1)\rangle \cup \langle t_{1-\alpha/2}(n-p-1), \infty) = (-\infty, -t_{0{,}975}(8)\rangle \cup \langle t_{0{,}975}(8), \infty) =$$
$$= (-\infty, -2{,}306\rangle \cup \langle 2{,}306, \infty).$$

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.

Příklad 4.: 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, jak se změní regresní odhad hodnoty závisle proměnné veličiny Y?
Výsledek: Zvýší se o 29,775*5 = 148,9 jednotek.

Příklad 5.: 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		17			
Total	543592.2				

Najděte odhad reziduálního rozptylu.

Výsledek: $s^2 = \frac{543592{,}2 - 505451{,}3}{17} = 2243{,}6$