ANALÝZA DAT V R

Jednoduchá lineární regrese

– máme k dispozici výběr párových hodnot $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ pro proměnné X (nezávisle proměnná, regresor, prediktor) a Y (závisle proměnná, regresand, predikant)

– předpokládáme, že pro naměřené hodnoty platí

$$y = a + bx + e,$$

kde a, b jsou parametry modelu a e je chyba

– parametry a, b se odhadují metodou nejmenších čtverců

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \qquad \hat{a} = \bar{y} - \hat{b}\bar{x}$$

• v R existuje pro odhad parametrů a, b implementovaná funkce `lm(y~x)`

– intervaly o spolehlivosti $(1-\alpha)\cdot 100\%$ pro parametry a, b jsou ve tvaru

$$(\hat{a} \mp t_{n-2}\left(1 - \frac{\alpha}{2}\right) \cdot SE_a)$$

a

$$(\hat{b} \mp t_{n-2}\left(1 - \frac{\alpha}{2}\right) \cdot SE_b),$$

kde

$$SE_a = \sqrt{\frac{s_{y.x}^2 \sum_{i=1}^{n} x_i^2}{s_x^2 (n-1)n}}, \qquad SE_b = \sqrt{\frac{s_{y.x}^2}{s_x^2 (n-1)}}, \qquad s_{y.x}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

– p-hodnoty a testové statistiky testů o parametrech modelu a další získáme pomocí funkce `summary(lm(y~x))`

Příklad

U náhodně vybraných studentů jsme naměřili jejich výšky a váhy:

výška	187	170	180	184	178	180	172	176	186	177
váha	72	60	73	74	72	70	62	70	80	67

a) Zjistěte, zde mezi výškou a váhou existuje lineární vztah.

b) Odhadněte lineární regresní přímku závislosti hmotnosti na výšce.

c) Odhadněte váhu studenta o výšce 179 cm.

d) Testujte významnost parametru b.

Řešení

a) Zda mezi proměnnými existuje lineární vztah, můžeme prověřit sestrojením bodového grafu.

```
x <- c(187, 170, 180, 184, 178, 180, 172, 176, 186, 177)
y <- c(72, 60, 73, 74, 72, 70, 62, 70, 80, 67)
plot(x, y, xlab="vyska [cm]", ylab="vaha [kg]")
```

Graf vykazuje lineární vztah a můžeme body proložit přímkou.

b) Parametry regresní přímky $y = a + bx$ získáme pomocí funkce `lm()`.

```
lm(y~x)
```

Výstupem jsou koeficienty, kde Intercept je odhad parametru a a y je odhad parametru b. Rovnice proložené přímky je tedy $y = -93,24 + 0,912x$.

c) Váhu studenta o výšce 179 cm odhadneme z regresní přímky $y = -93,24 + 0,912x$ dosazením hodnoty 179 za x, tj. $y = 70,008$.

d) Testujeme $H_0 : b = 0$ proti $H_1 : b \neq 0$ a hladinu významnosti zvolíme $\alpha = 0,05$. Vyžijeme funkce `summary(lm())`.

```
summary(lm(y~x))
```

Můžeme buď zkonstruovat 95% oboustranný interval spolehlivosti pro parametr b a zjistit, zda testovaná hodnota, tj. 0, patří do tohoto intervalu. Tedy dosadit do výše uvedeného intervalu, kde SE_b je Std. Error v řádku x u výstupu funkce `summary(lm(y~x))`. Tedy $SE_b = 0,1753$.

```
0.9120-qt(0.975,8)*0.1753
0.9120+qt(0.975,8)*0.1753
```

95% interval spolehlivosti pro parametr b je $(0,5078; 1,3162)$. Hodnota 0 do intervalu nepatří, proto zamítáme nulovou hypotézu, že y na x nezávisí.

Nebo můžeme použít p-hodnotu v tomto řádku, která je označena jako $\Pr(>|t|)$. Z této p-hodnoty, která je rovna $0,000819$ a je menší než stanovená hladina významnosti α, plyne stejný závěr.