# Zápočtová práce z předmětu Statistika II – regresní analýza

Vypracoval: Jiří Neubauer

**Zadání**: Úkolem zápočtové práce je zjištění vztahu mezi výškou a hmotností dětí ve věku od 6,5 do 7 let pomocí regresní analýzy. Pro získání potřebných informací byl pořízen náhodný výběr 32 dětí z prvních tříd na okrese Brno venkov v roce 2010. (V celé práci budeme používat $\alpha = 0{,}05$.)

| Výška | Hmotnost | Výška | Hmotnost | Výška | Hmotnost |
|---|---|---|---|---|---|
| 108 | 19 | 119 | 21 | 121 | 22 |
| 108 | 19 | 119 | 23 | 122 | 22 |
| 110 | 20 | 120 | 21 | 122 | 23 |
| 110 | 20 | 120 | 22 | 123 | 25 |
| 112 | 21 | 120 | 23 | 126 | 24 |
| 114 | 20 | 120 | 23 | 126 | 25 |
| 114 | 21 | 120 | 23 | 126 | 26 |
| 114 | 21 | 121 | 24 | 127 | 26 |
| 118 | 22 | 121 | 22 | 127 | 28 |
| 118 | 21 | 121 | 22 | 128 | 28 |

## Model 1 – přímková regrese

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i, \quad i = 1, \dots n$$

| | $\hat{\beta}_j$ | $s(\hat{\beta}_j)$ | statistika $t$ | $t_{1-\alpha/2}(n-c)$ | $p$-hodnota |
|---|---|---|---|---|---|
| $\hat{\beta}_1$ | −21,18341 | 4,32782 | −4,895 | 2,048 | $3{,}70 \cdot 10^{-5}$ |
| $\hat{\beta}_2$ | 0,36713 | 0,03628 | 10,120 | 2,048 | $7{,}39 \cdot 10^{-11}$ |

*Tab. 1: Tabulka obsahuje odhady regresních parametrů $\hat{\beta}_j$, směrodatné chyby těchto odhadů, $s(\hat{\beta}_j)$, hodnoty testových statistik významnosti regresních parametrů, kritické hodnoty těchto testů a odpovídající p-hodnoty*

Reziduální rozptyl pro daný model je $s_e^2 = \frac{1}{28}\sum_{i=1}^{30}(y_i - \hat{y}_i)^2 = 1{,}237$.

### Intervaly spolehlivosti pro odhady parametrů $\beta_j$

$$\hat{\beta}_j - t_{1-\alpha/2}(n-k) \cdot s(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t_{1-\alpha/2}(n-k) \cdot s(\hat{\beta}_j)$$

$$\begin{aligned} -30{,}049 &< \beta_1 < -12{,}318 \\ 0{,}293 &< \beta_2 < 0{,}441 \end{aligned}$$

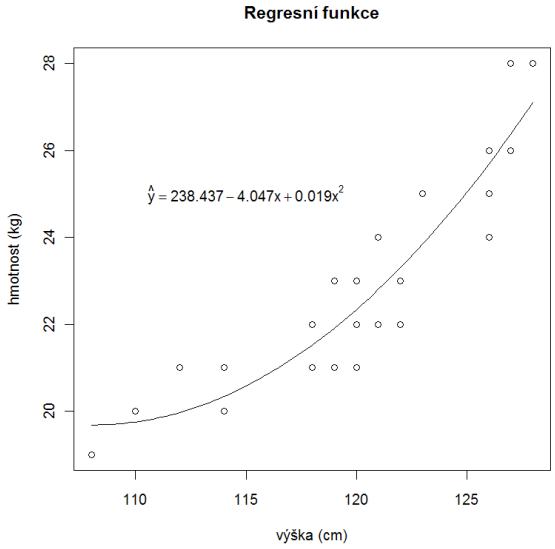
Na základě intervalů spolehlivosti a hodnot jednotlivých $t$-testů můžeme považovat koeficienty $\beta_1$ a $\beta_2$ za statisticky významné.

### Celkový $F$-test modelu

$H : \beta_2 = 0 \to A : \beta_2 \neq 0$
Testové kritérium má hodnotu

$$F = \frac{\frac{S_T(y)}{k-1}}{\frac{S_e(y)}{n-k}} = 102{,}4,$$

odpovídající kritická hodnota $F_{1-\alpha}(k-1, n-k) = F_{0,95}(1, 28) = 4{,}20$ ($p$-hodnota je $7{,}389 \cdot 10^{-11}$). Na hladině významnosti 0,05 můžeme považovat model za statisticky významný. Index determinace má hodnotu $R^2 = 0{,}785$ ($R^2_{adj} = 0{,}778$).




*Obr. 1: Regresní funkce popisující závislost hmotnosti na výšce dětí*

## Model 2 – parabolická regrese

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \epsilon_i, \quad i = 1, \dots n$$

| | $\hat{\beta}_j$ | $s(\hat{\beta}_j)$ | statistika $t$ | $t_{1-\alpha/2}(n-c)$ | $p$-hodnota |
|---|---|---|---|---|---|
| $\hat{\beta}_1$ | 238,436921 | 67,439204 | 3,536 | 2,052 | 0,001490 |
| $\hat{\beta}_2$ | −4,047231 | 1,145483 | −3,533 | 2,052 | 0,001500 |
| $\hat{\beta}_3$ | 0,018720 | 0,004856 | 3,855 | 2,052 | 0.000649 |

*Tab. 2: Tabulka obsahuje odhady regresních parametrů $\hat{\beta}_j$, směrodatné chyby těchto odhadů, $s(\hat{\beta}_j)$, hodnoty testových statistik významnosti regresních parametrů, kritické hodnoty těchto testů a odpovídající p-hodnoty*

Reziduální rozptyl daného modelu je $s_e^2 = \frac{1}{27}\sum_{i=1}^{30}(y_i - \hat{y}_i)^2 = 0{,}828$.

### Intervaly spolehlivosti pro odhady parametrů $\beta_j$

$$\begin{aligned} 100{,}063 &< \beta_1 < 376{,}811 \\ -6{,}398 &< \beta_2 < -1{,}697 \\ 0{,}009 &< \beta_3 < 0{,}029 \end{aligned}$$

Na základě intervalů spolehlivosti hodnot jednotlivých $t$-testů můžeme považovat koeficienty $\beta_1$, $\beta_2$ a $\beta_3$ za statisticky významné.

### Celkový $F$-test modelu

$H : \beta_2 = \beta_3 = 0 \rightarrow A : \beta_2 \neq 0$ nebo $\beta_3 \neq 0$
Testové kritérium má hodnotu

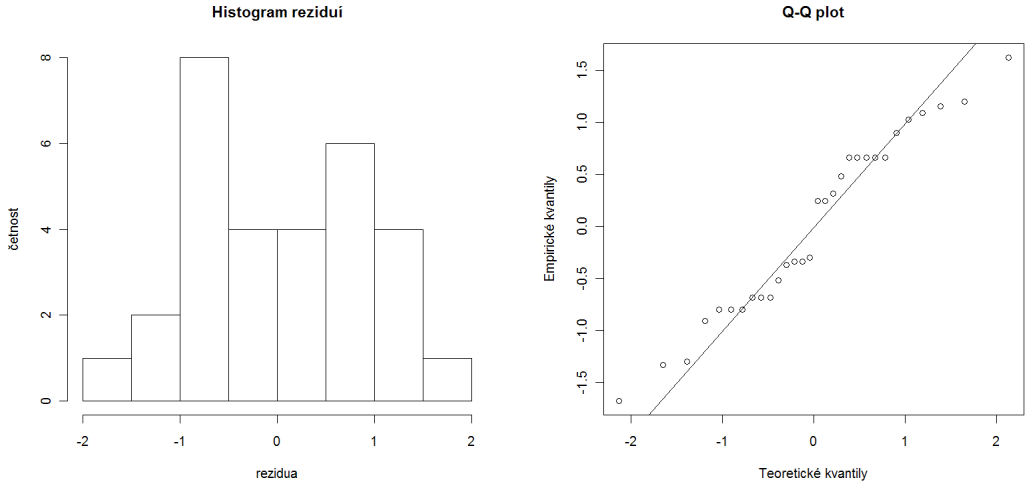
$$F = \frac{\frac{S_T(y)}{k-1}}{\frac{S_e(y)}{n-k}} = 83{,}99,$$

odpovídající kritická hodnota $F_{1-\alpha}(k-1, n-k) = F_{0,95}(2,27) = 3{,}35$ ($p$-hodnota je $2{,}562 \cdot 10^{-12}$). Na hladině významnosti 0,05 můžeme považovat model za statisticky významný. Index determinace má hodnotu $R^2 = 0{,}862$ ($R^2_{adj} = 0{,}851$).

Můžeme konstatovat, že parabolická regresní funkce je pro popis závislosti vhodnější než funkce přímková (menší reziduální rozptyl a větší hodnota indexu determinace). Přidáním dalšího členu do regresní funkce již vede k nevýznamným odhadů regresních parametrů.

### Ověření modelu

Pro ověření vhodnosti modelu zkonstruujeme histogram reziduí a tzv. Q–Q plot reziduí. Normalitu reziduí otestujeme některým z testů normality. Shapiro-Wilkův test (testové kritérium 0,956, $p$-hodnota 0,238) a Lillieforsův test (testové kritérium 0,141, $p$-hodnota 0,133) normalitu reziduí nezamítly. Můžeme tedy říci, že zvolený model je vhodný pro popis studované závislosti.




*Obr. 2: Histogram a Q–Q plot reziduí*

Studovanou závislost hmotnosti na výšce budeme tedy popisovat pomocí parabolické regresní funkce. Zkonstruujeme odhady na základě zvoleného modelu. Jaká je odhadnutá hodnota hmotnosti pro výšku 120 cm? Z odhadnuté regresní funkce dostáváme

$$\hat{y}(\mathbf{x}_0) = 238{,}43692 - 4{,}04723 \cdot 120 + 0{,}01872 \cdot 120^2 = 22{,}338,$$

očekávaná hmotnost dítěte s výškou 120 cm je 22,338 kg. Interval spolehlivosti pro regresní přímku je dán vztahem

$$\widehat{y}(\mathbf{x}_0) - t_{1-\alpha/2}(n-k) \cdot s(\widehat{y}(\mathbf{x}_0)) < y(\mathbf{x}_0) < \widehat{y}(\mathbf{x}_0) + t_{1-\alpha/2}(n-k) \cdot s(\widehat{y}(\mathbf{x}_0)),$$

kde $s(\widehat{y}(\mathbf{x}_0))$ je směrodatná chyba odhadu $\widehat{y}(\mathbf{x}_0)$. Pro výšku 120 cm je interval roven

$$21{,}892 < y(\mathbf{x}_0) < 22{,}785,$$

což znamená, že střední hodnota hmotnosti dětí s výškou 120 cm se bude s pravděpodobností 95 % pohybovat v intervalu $(21{,}892; 22{,}785)$ kg. Interval spolehlivosti pro individuální předpověď určíme ze vztahu

$$\widehat{y}(\mathbf{x}_0) - t_{1-\alpha/2}(n-k) \cdot s_0 < Y_0 < \widehat{y}(\mathbf{x}_0) + t_{1-\alpha/2}(n-k) \cdot s_0,$$

kde $s_0$ je směrodatná chyba odhadu $Y_0$. Po dosazení dostáváme

$$20{,}419 < Y_0 < 24{,}258.$$

Vypočítaný interval říká, že hmotnost dítěte s výškou 120 cm se bude s pravděpodobností 95 % nacházet v intervalu $(20{,}419; 24{,}258)$ kg.