

Jednoduchá lineární regrese – vzorový příklad

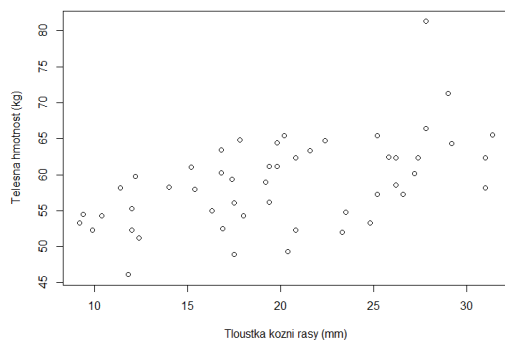
U 51 mladých zdravých žen (převážně studentek z Brna) byla – mimo jiné – zjištěna tělesná hmotnost (proměnná `body.W`, v kg) a tloušťka kožní řasy na boku (proměnná `hip.F`, v mm). Data jsou uložena v souboru `fat.txt`. Úkolem bude modelování závislosti tělesné hmotnosti na tloušťce kožní řasy na boku.

Načtení dat a výpočet číselných charakteristik:

```
fat <- read.table("DATA/fat.txt",header=T)
summary(fat)
body.W      hip.F
Min. :46.10  Min. : 9.20
1st Qu.:54.40 1st Qu.:15.85
Median :58.60 Median :19.80
Mean :58.90  Mean :20.05
3rd Qu.:62.35 3rd Qu.:25.50
Max. :81.30  Max. :31.40
```

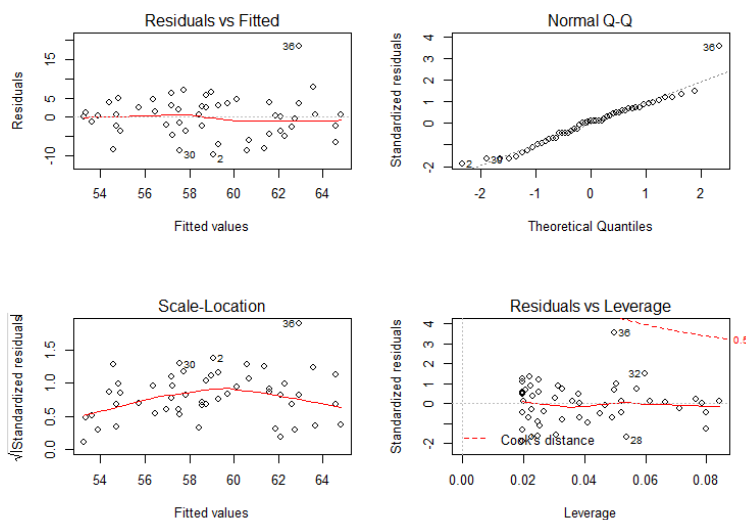
Dvourozměrný tečkový diagram závislosti hmotnosti na tloušťce kožní řasy na boku:

```
plot(fat$hip.F, fat$body.W, xlab='Tloustka kozni rasy (mm)',
     ylab='Telesna hmotnost (kg)')
```



Sestavíme model regrese přímky a pomocí analýzy reziduí ověříme předpoklady modelu:

```
m.weight <- lm(body.W ~ hip.F, data=fat)
par(mfrow=c(2,2))
plot(m.weight)
```



První graf ukazuje střední hodnotu reziduí - pokud je náš model pro data vhodný, bude na prvním grafu červená čára (přibližně) vodorovná kolem 0. Druhý graf je kvantil-kvantilový graf reziduí, pomocí nějž zhodnotíme předpoklad normality. Třetím grafem hodnotíme rozptyl reziduí, pokud je křivka přibližně horizontální a rezidua jsou kolem rozmístěna rovnoměrně, považujeme předpoklad za splněný. Čtvrtý graf slouží k detekci vlivných pozorování.

Předpoklad normality reziduí můžeme dále posoudit Shapirovým-Wilkovým testem, nulovost střední hodnoty pomocí t-testu a nezávislost reziduí pomocí Durbinova-Watsonova testu (v R je třeba načíst knihovnu car).

```
shapiro.test(m.weight$residuals)
```

```
Shapiro-Wilk normality test
```

```
data: m.weight$residuals
```

```
W = 0.95788, p-value = 0.06777
```

Shapiro-Wilkův test nabývá hodnoty s p-hodnotou, v kvantil-kvantilovém grafu jsou rezidua, předpoklad normality tedy považujeme za

```
t.test(m.weight$residuals)
```

```
One Sample t-test
```

```
data: m.weight$residuals
```

```
t = 2.672e-16, df = 50, p-value = 1
```

```
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
```

```
-1.472026 1.472026
```

```
sample estimates:
```

```
mean of x
```

```
1.958239e-16
```

Hypotézu o nulové střední hodnotě reziduí, protože t-test nabývá hodnoty s p-hodnotou, z grafického posouzení také nevidíme problém.

```
library(car)
```

```
durbinWatsonTest(m.weight)
```

```
lag Autocorrelation D-W Statistic p-value
```

```
1 0.09560068 1.75842 0.388
```

```
Alternative hypothesis: rho != 0
```

Durbin-Watsonův test nabývá hodnoty s p-hodnotou, tedy nezávislost reziduí.

Předpoklad rovnosti rozptylů se na základe grafického posouzení (viz 3. graf) zdá mírně porušen, nicméně porušení není závažné. Sestavený model tedy budeme považovat za vhodný.

Podívejme se na podrobné informace o modelu.

summary(m.weight)

Call:

lm(formula = body.W ~ hip.F, data = fat)

Residuals:

Min	1Q	Median	3Q	Max
-9.781	-3.518	0.502	3.278	18.359

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	48.4393	2.4867	19.479	< 2e-16 ***
hip.F	0.5217	0.1184	4.406	5.72e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.287 on 49 degrees of freedom
 Multiple R-squared: 0.2838, Adjusted R-squared: 0.2692
 F-statistic: 19.42 on 1 and 49 DF, p-value: 5.717e-05

MNČ odhady koeficientu a jejich interpretace:

$\beta_0 = \dots\dots\dots$

$\beta_1 = \dots\dots\dots$

Odhad rozptylu:

$s^2 = \dots\dots\dots$

Index determinace (někdy nazýván koeficient determinace a značen R^2 místo ID^2) a jeho interpretace:

$ID^2 = \dots\dots\dots$

Celkový F-test na hladině významnosti 0:05:

F = $\dots\dots\dots$

p-hodnota = $\dots\dots\dots$

závěr $\dots\dots\dots$

Dílčí t-testy

β_0

• hodnota testovací statistiky $\dots\dots\dots$

• p-hodnota $\dots\dots\dots$

• závěr $\dots\dots\dots$

β_1

• hodnota testovací statistiky $\dots\dots\dots$

• p-hodnota $\dots\dots\dots$

• závěr $\dots\dots\dots$

Intervaly spolehlivosti pro regresní koeficienty:

confint(m.weight)

	2.5 %	97.5 %
(Intercept)	43.4420015	53.4365956
hip.F	0.2837486	0.7595599

Interval spolehlivosti pro β_0 : $\dots\dots\dots$

Interval spolehlivosti pro β_1 : $\dots\dots\dots$

Výpočet střední absolutní procentuální chyby predikce:

```
100 * mean(abs(m.weight$residuals/fat$body.W))  
[1] 6.851992
```

MAPE =

Vypočtete odhad tělesné hmotnosti jedince, pokud jste mu naměřili kožní rasu tloušťky 20mm.

```
predict(m.weight, newdata=data.frame(hip.F=20))  
58.87238
```

Odhadnutá hodnota

Na závěr vykreslíme regresní přímku společně s pásem spolehlivosti a predikčním pásem.

```
xx <- seq(min(fat$hip.F), max(fat$hip.F), length=300)  
interval.spol <-  
predict(m.weight,newdata=data.frame(hip.F=xx),interval='confidence')  
pred.interval <-  
predict(m.weight,newdata=data.frame(hip.F=xx),interval='predict')  
plot(fat$hip.F, fat$body.W, xlab='Tloustka kozni rasy (mm)', ylab='Telesna  
hmotnost (kg)')  
lines(xx,interval.spol[,1],col='red')  
lines(xx,interval.spol[,2], col='red', lty=2)  
lines(xx,interval.spol[,3], col='red', lty=2)  
lines(xx,pred.interval[,2], col='blue', lty=2)  
lines(xx,pred.interval[,3], col='blue', lty=2)  
legend("topleft",c('model','IS','pred. int.'), lty=c(1,2,2),  
col=c('red', 'red', 'blue'))
```

