3 Jednoduchý lineární regresní model

Příklad 1. V souboru fat.txt jsou antropometrická data mladých zdravých dospělých žen (převážně studentek vysokých škol z Brna). Zajímá nás závislost tělesné hmotnosti body.W (v kg) na tloušťce kožní řasy na boku hip.F (v mm).

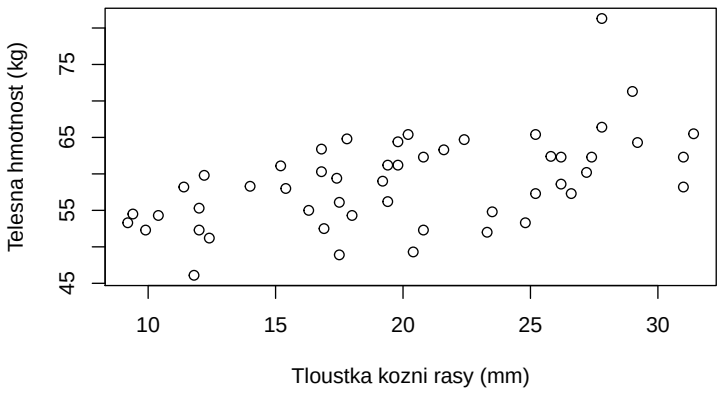
Načteme data a podíváme se na ně. Soubor neobsahuje žádná chybějící pozorování.

```
fat <- read.table("DATA/fat.txt",header=T)
summary(fat)
```

```
##      body.W           BMI            hip.F
##  Min.   :46.10   Min.   :16.63   Min.   : 9.20
##  1st Qu.:54.40   1st Qu.:19.35   1st Qu.:15.85
##  Median :58.60   Median :20.83   Median :19.80
##  Mean   :58.90   Mean   :21.05   Mean   :20.05
##  3rd Qu.:62.35   3rd Qu.:22.18   3rd Qu.:25.50
##  Max.   :81.30   Max.   :28.47   Max.   :31.40
```

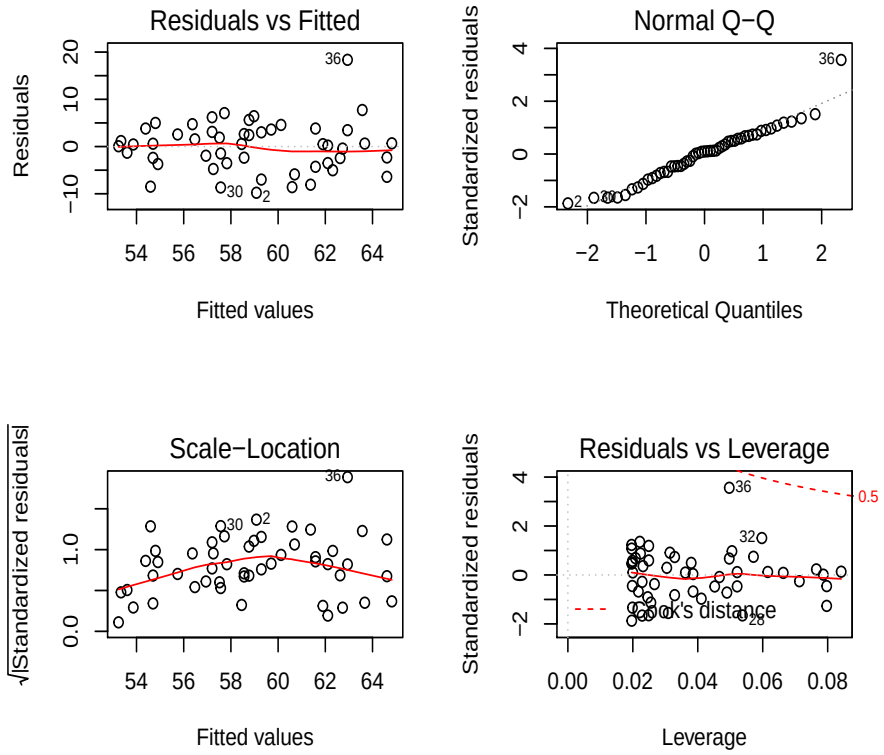
Chceme modelovat závislost tělesné hmotnosti na tloušťce kožní řasy na boku, vykreslíme si tedy hodnoty v bodovém grafu.

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



Sestavíme model regresní 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

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

library(car)
durbinWatsonTest(m.weight)

##  lag Autocorrelation D-W Statistic p-value
##    1      0.09560068       1.75842   0.404
##  Alternative hypothesis: rho != 0
```

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

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

Předpoklad rovnosti rozptylů se na základě grafického posouzení zdá mírně porušen, nicméně porušení není závažné.

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

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 koeficientů a jejich interpretace:
β_0 = ...
β_1 = ...

Odhad rozptylu:
s^2 = ...

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

Celkový F-test na hladině významnosti 0.05:
F = ...
p-hodnota = ...
závěr .....

Dílčí t-testy
β_0

- hodnota testovací statistiky .....
- p-hodnota .....
- závěr .....

β_1

- hodnota testovací statistiky .....
- p-hodnota .....
- závěr .....

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: .....
Interval spolehlivosti pro β_1: .....

Vypočtěte střední absolutní procentuální chybu predikce MAPE: .....

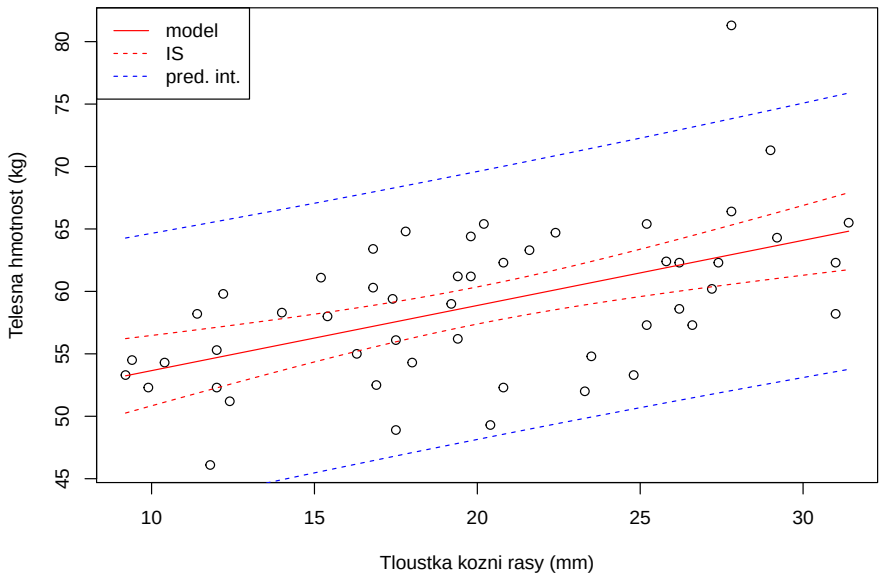
```
100 * mean(abs(m.weight$residuals/fat$body.W))

## [1] 6.851992
```

Vypočtěte odhad tělesné hmotnosti jedince, pokud jste mu naměřili kožní řasu tloušťky $20mm$. .....

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

##        1
## 58.87238
```

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

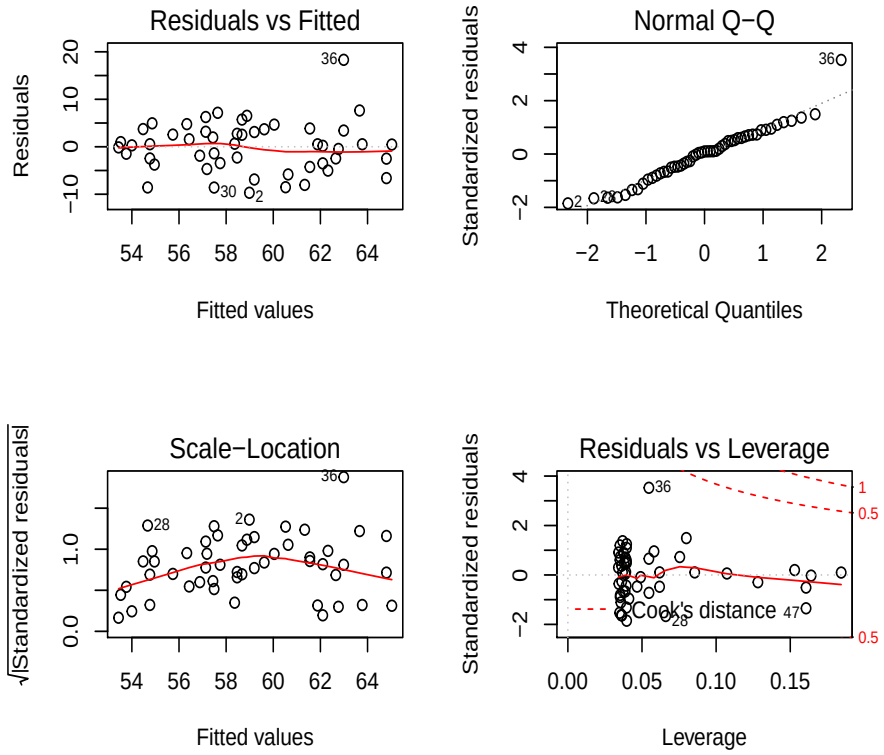
```
xx <- seq(min(fat$hip.F), max(fat$hip.F), length=300)
interval.spol <- predict(m.weight,newdata=data.frame(hip.F=xx),interval='confidence')
pred.interval <- predict(m.weight,newdata=data.frame(hip.F=xx),interval='predict')

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'))
```



Obrázek 1: Výsledný model

Podíváme se nyní, jestli nebude lepší model paraboly $Y_i = \beta_0 + \beta 1 x_i + \beta_2 x_i^2 + \epsilon_i$.

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




```
shapiro.test(m.weight.2$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  m.weight.2$residuals
## W = 0.9582, p-value = 0.06992

t.test(m.weight.2$residuals)

##
##  One Sample t-test
##
## data:  m.weight.2$residuals
## t = 5.3513e-17, df = 50, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.471785  1.471785
## sample estimates:
##    mean of x
## 3.921198e-17

durbinWatsonTest(m.weight.2)

##  lag Autocorrelation D-W Statistic p-value
##    1      0.09448426        1.7613   0.396
##  Alternative hypothesis: rho != 0
```

Z grafického posouzení i z výsledků testování hypotéz budeme považovat předpoklady modelu za splněné. Podívejme se tedy na jeho výsledky.

```
summary(m.weight.2)
```

```
##
## Call:
## lm(formula = body.W ~ hip.F + I(hip.F^2), data = fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6863 -3.4730  0.4686  3.2881 18.3116
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.328893   7.521701   6.558  3.5e-08 ***
## hip.F        0.423936   0.787912   0.538    0.593
## I(hip.F^2)   0.002425   0.019326   0.125    0.901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.341 on 48 degrees of freedom
## Multiple R-squared:  0.284,Adjusted R-squared:  0.2542
## F-statistic: 9.521 on 2 and 48 DF,  p-value: 0.0003292
```

Vidíme, že celkově model vychází významný, ale podle dílčích testů jsou koeficienty β_1 a β_2 nevýznamné.

Pro srovnání modelu paraboly s modelem přímky se podíváme na adjustované indexy determinace (někdy také nazývané adjustované koeficienty determinace):
ID^2_{adj} pro model přímky: .....
ID^2_{adj} pro model paraboly: .....
závěr .....

Modely můžeme srovnat i na základě hodnoty MAPE:
MAPE pro model přímky: .....
MAPE pro model paraboly: .....
závěr .....

```
#parabola
100 * mean(abs(m.weight.2$residuals/fat$body.W))
```

```
## [1] 6.853803
```