

### 3 Dvoufaktorová analýza rozptylu (ANOVA)

**Příklad 1.** Datový soubor `newborns.txt` obsahuje porodní hmotnost novorozenců v gramech (proměnná `weight.C`), jejich pohlaví (proměnná `sex.C`) a vzdělání matky (proměnná `edu.M`). Budeme se zabývat studiem závislosti porodní hmotnosti dítěte na jeho pohlaví a vzdělání jeho matky. Kódování proměnné `edu.M`: 1 - základní, 2 - střední bez maturity, 3 - střední s maturitou, 4 - vysokoškolské.

Načteme data a podíváme se na ně. Soubor obsahuje pozorování s chybějícími hodnotami, ty v analýze použít nemůžete, proto odpovídající řádky vyřadíme pomocí `na.omit()`. Vidíme, že proměnnou `edu.M` R načetlo jako numerickou, my ji potřebujeme kategoriální, ke změně použijeme funkci `factor()`, kde nastavíme odpovídající názvy kategorií.

```
data <- read.table('DATA/newborns.txt', header=TRUE)
summary(data)

##      edu.M      sex.C      weight.C
## Min.   :1.000    f:674    Min.    : 580
## 1st Qu.:1.000    m:729    1st Qu.:2670
## Median :2.000                    Median  :3170
## Mean   :2.129                    Mean    :3072
## 3rd Qu.:3.000                    3rd Qu.:3560
## Max.   :4.000                    Max.    :4970
## NA's   :13                      NA's    :6

data <- na.omit(data)
data$edu.M <- factor(data$edu.M, labels=c('ZS', 'SS', 'SSmat', 'VS'))
```

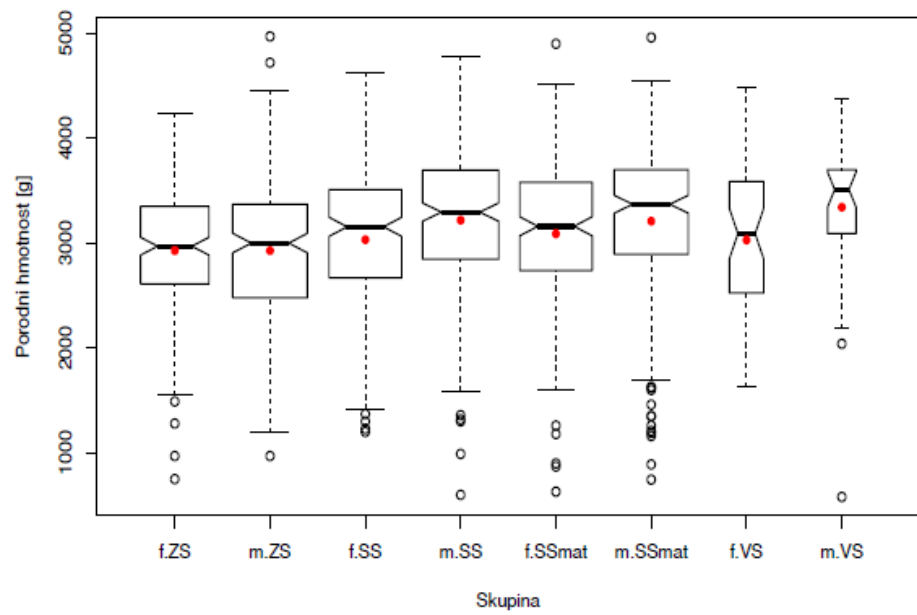
Podíváme se, kolik pozorování máme pro jednotlivé kombinace faktorů.

```
table(data$sex.C, data$edu.M)
```

```
##
##      ZS  SS SSmat  VS
##  f 190 214   213  48
##  m 227 236   223  33
```

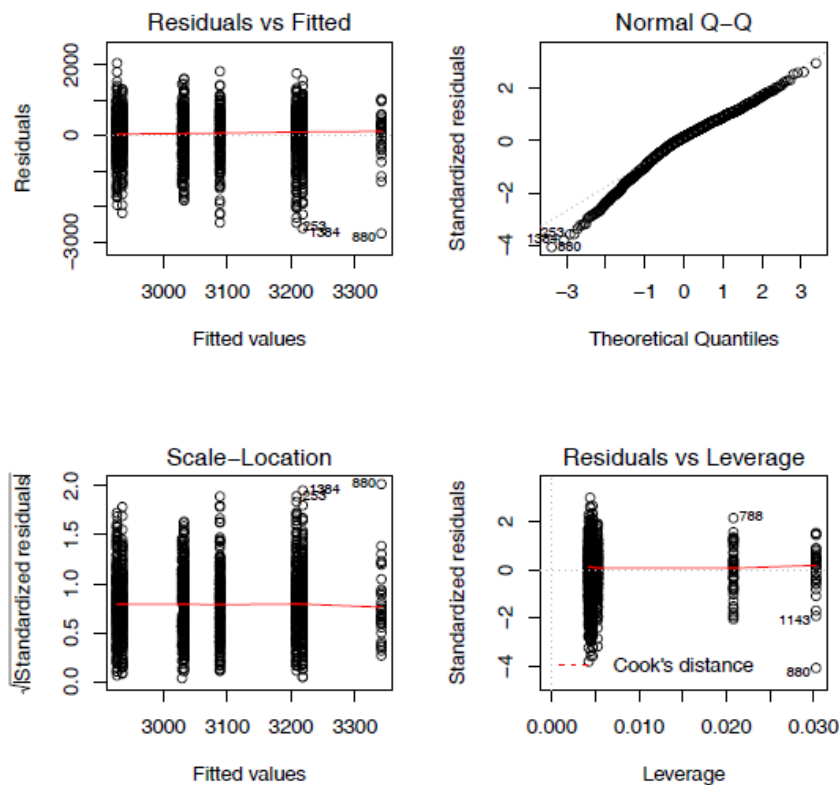
Abychom se lépe seznámili s daty, vykreslíme si krabicové diagramy porodní hmotnosti podle pohlaví dítěte a vzdělání matky. Do grafů zakreslíme červenými body i průměry jednotlivých skupin.

```
boxplot(data$weight.C ~ data$sex.C + data$edu.M, varwidth=T, notch=T,
        xlab='Skupina', ylab='Porodni hmotnost [g]')
mm1 <- mean(data[data$edu.M=='ZS' & data$sex=='f', 'weight.C'])
mm2 <- mean(data[data$edu.M=='ZS' & data$sex=='m', 'weight.C'])
mm3 <- mean(data[data$edu.M=='SS' & data$sex=='f', 'weight.C'])
mm4 <- mean(data[data$edu.M=='SS' & data$sex=='m', 'weight.C'])
mm5 <- mean(data[data$edu.M=='SSmat' & data$sex=='f', 'weight.C'])
mm6 <- mean(data[data$edu.M=='SSmat' & data$sex=='m', 'weight.C'])
mm7 <- mean(data[data$edu.M=='VS' & data$sex=='f', 'weight.C'])
mm8 <- mean(data[data$edu.M=='VS' & data$sex=='m', 'weight.C'])
points(1:8, c(mm2,mm2,mm3,mm4,mm5,mm6,mm7,mm8), col='red', pch=16)
```



Předpoklady modelu ověříme pomocí analýzy reziduí. To znamená, že nejprve sestavíme úplný model - s oběma faktory i interakcí. Pomocí funkce `plot()` použité na sestavený model si zobrazíme diagnostické grafy. První graf ukazuje, zda je model lineární závislosti vhodný (= zobrazená červená křivka je vodorovná kolem 0, pokud by křivka vypadala jinak, značí to, že lineární model není vhodný). Druhý je kvantil-kvantilový graf, pomocí nějž ověřujeme normalitu dat. Třetí graf slouží k ověření předpokladu rovnosti rozptylů: pokud je zobrazená křivka horizontální, svědčí to o homogenitě rozptylů, pokud křivka není horizontální, svědčí to o heterogenitě rozptylů. Čtvrtý graf slouží k detekci vlivných pozorování.

```
model.newborns <- aov(weight.C ~ sex.C*edu.M, data=data)
par(mfrow=c(2,2)) #nastavi zobrazeni 4 grafu najednou (na 2 radky a 2 sloupce)
plot(model.newborns)
```



Pro ověření předpokladu normality ještě aplikujeme Shapirův-Wilkův test na rezidua.

```
shapiro.test(model.newborns$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  model.newborns$residuals
## W = 0.97313, p-value = 2.201e-15
```

ANOVA tabulku vypíšeme pomocí funkce `anova()`.

```
anova(model.newborns)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: weight.C
```

```
##           Df      Sum Sq Mean Sq F value    Pr(>F)
## sex.C      1   3933723 3933723   8.3039 0.004017 **
## edu.M      3  13526215 4508738   9.5178 3.186e-06 ***
## sex.C:edu.M 3   2904470  968157   2.0437 0.105923
## Residuals 1376 651834735  473717
```

```
## ---
```

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

Jednotlivé hypotézy můžeme testovat i pomocí kritického oboru. Vypočítáme si tedy odpovídající kvantily.

```
# pro interakci
qf(0.95, 3, 1376)

## [1] 2.61137

# pro faktor B
qf(0.95, 3, 1376)

## [1] 2.61137

# pro faktor A
qf(0.95, 1, 1376)

## [1] 3.848226
```

## Tukeyova metoda mnohonásobného porovnávání:

```
TukeyHSD(model.newborns, which=c('sex.C', 'edu.M'))

## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = weight.C ~ sex.C * edu.M, data = data)
##
## $sex.C
##          diff          lwr          upr          p adj
## m-f 106.7074  34.0663 179.3486 0.0040173
##
## $edu.M
##          diff          lwr          upr          p adj
## SS-ZS    201.02954    80.69368 321.3654 0.0001090
## SSmat-ZS 222.56046   101.29893 343.8220 0.0000154
## VS-ZS    240.15036    25.18725 455.1135 0.0214337
## SSmat-SS  21.53092   -97.43625 140.4981 0.9665581
## VS-SS    39.12082  -174.55643 252.7981 0.9654441
## VS-SSmat  17.58990  -196.61002 231.7898 0.9966695
```



Hypotézy otestujeme i pomocí Scheffého metody. Ta pro faktor  $A$  zamítne hypotézu  $\alpha_i = \alpha_t$ , když

$$|M_{i..} - M_{t..}| > \sqrt{(a-1) \left( \frac{1}{n_{i.}} + \frac{1}{n_{t.}} \right) \frac{S_E}{n-ab} F_{1-\alpha}(a-1, n-ab)}$$

a pro faktor  $B$  zamítne hypotézu  $\beta_j = \beta_u$ , když

$$|M_{.j.} - M_{.u.}| > \sqrt{(b-1) \left( \frac{1}{n_{.j}} + \frac{1}{n_{.u}} \right) \frac{S_E}{n-ab} F_{1-\alpha}(b-1, n-ab)}$$

Scheffého metoda není v R implementovaná, vypočítáme ji ručně. Nejprve si uložíme všechny potřebné členy. Hodnotu zlomku  $\frac{SE}{n - ab}$  získáme z ANOVA tabulky modelu.

```
a <- 2
b <- 4
n <- nrow(data)
m.boys <- mean(data[data$sex.C == 'm', 'weight.C'])
m.girls <- mean(data[data$sex.C == 'f', 'weight.C'])
n.sex <- as.numeric(table(data$sex.C))

m.zs <- mean(data[data$edu.M == 'ZS', 'weight.C'])
m.ss <- mean(data[data$edu.M == 'SS', 'weight.C'])
m.ssmat <- mean(data[data$edu.M == 'SSmat', 'weight.C'])
m.vs <- mean(data[data$edu.M == 'VS', 'weight.C'])
n.edu <- as.numeric(table(data$edu.M))

res.mean.sq <- anova(model.newborns)['Residuals', 'Mean Sq']
```

Pro každou dvojici si nejprve vypočítáme levou stranu vztahu a pravou stranu vztahu.

```
# chlapci-devcata
abs(m.boys - m.girls)

## [1] 106.7074

sqrt((a-1) * (1/n.sex[1] + 1/n.sex[2]) * res.mean.sq * qf(0.95, a-1, n-a*b))

## [1] 72.64113
```

```
# ZS-SS
abs(m.zs - m.ss)

## [1] 198.9039

sqrt((b-1) * (1/n.edu[1] + 1/n.edu[2]) * res.mean.sq * qf(0.95, b-1, n-a*b))

## [1] 130.9452

# ZS-SSmat
abs(m.zs - m.ssmat)

## [1] 219.0501

sqrt((b-1) * (1/n.edu[1] + 1/n.edu[3]) * res.mean.sq * qf(0.95, b-1, n-a*b))

## [1] 131.9525

# ZS-VS
abs(m.zs - m.vs)

## [1] 225.536

sqrt((b-1) * (1/n.edu[1] + 1/n.edu[4]) * res.mean.sq * qf(0.95, b-1, n-a*b))

## [1] 233.9152
```

```

# SS-SSmat
abs(m.ss - m.ssmat)

## [1] 20.14623

sqrt((b-1) * (1/n.edu[2] + 1/n.edu[3]) * res.mean.sq * qf(0.95, b-1, n-a*b))

## [1] 129.4559

# SS-VS
abs(m.ss - m.vs)

## [1] 26.6321

sqrt((b-1) * (1/n.edu[2] + 1/n.edu[4]) * res.mean.sq * qf(0.95, b-1, n-a*b))

## [1] 232.516

# SSmat-VS
abs(m.ssmat - m.vs)

## [1] 6.48587

sqrt((b-1) * (1/n.edu[3] + 1/n.edu[4]) * res.mean.sq * qf(0.95, b-1, n-a*b))

## [1] 233.0847

```

(Protože proměnná pohlaví má jen dvě kategorie, není pro ni nutné post-hoc testy provádět. Z testování modelu vyšel tento faktor jako významný, tudíž víme, že se chlapci a děvčata liší.)

Nakonec si výsledný model graficky zobrazíme.

```
coefs <- aov(weight.C ~ sex.C + edu.M, data=data)$coefficients
girls <- coefs[1] + c(0, coefs[3], coefs[4], coefs[5])
boys <- coefs[2] + girls
boxplot(data$weight.C~data$edu.M, ylim=c(2800, 3300), border='white',
        xlab='Vzdelani matky', ylab='Porodni hmotnost')
points(1:4,girls, col='red', pch=16)
lines(1:4,girls, col='red', lty=2)
points(1:4, boys, col='blue', pch=17)
lines(1:4, boys, col='blue', lty=2)
legend('topleft', legend=c('divky', 'chlapci'), col=c('red', 'blue'), pch=c(16,17))
```

