

Vzorový příklad na analýzu kovariance

V souboru lrm-foot.txt máme k dispozici antropometrické údaje mladých dospělých lidí (převážně studentů vysokých škol z Brna a Ostravy) – viz př. 1 ze vzorových příkladů na korelační analýzu. Chceme zjistit, zda se muži a ženy liší v tělesné výšce (proměnná body.H, v mm), pokud eliminujeme vliv délky chodidla (proměnná foot.L, v mm).

Načteme datový soubor a podíváme se na jeho číselné charakteristiky:

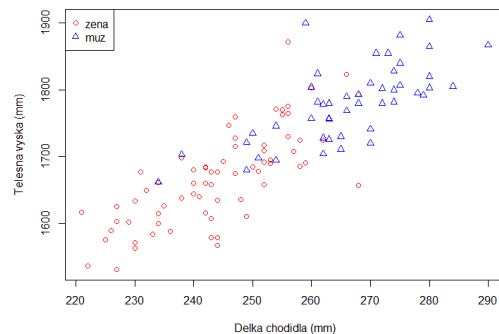
```
foot <- read.table("lrm-foot.txt", header=T)
> summary(foot)
sex      foot.L      body.H
f:70    Min.   :221.0    Min.   :1531
m:47    1st Qu.:242.0    1st Qu.:1658
        Median :253.0    Median :1711
        Mean   :253.1    Mean   :1715
        3rd Qu.:265.0    3rd Qu.:1780
        Max.   :290.0    Max.   :1905
```

Zkontrolujeme, že R pracuje s proměnnou pohlaví jako s faktorem. Pokud by byla v datovém souboru kódována například pomocí 0 a 1, tak by s ní R pracovalo jako s numerickou proměnnou, nikoli kategoriální. V takovém případě bychom ji museli změnit na kategoriální pomocí funkce factor().

```
is.factor(foot$sex)
[1] TRUE
```

Vykreslíme si bodový diagram, ve kterém odlišíme muže a ženy.

```
plot(foot$foot.L, foot$body.H, xlab='Delka chodidla (mm)', ylab='Telesna
vyska (mm)', type='n')
points(foot$foot.L[foot$sex=='f'], foot$body.H[foot$sex=='f'], pch=1,
col='red')
points(foot$foot.L[foot$sex=='m'], foot$body.H[foot$sex=='m'], pch=2,
col='blue')
legend("topleft", c("zena", "muz"), pch = c(1,2), col=c('red', 'blue'))
```



Vypočítáme si rozsahy, výběrové průměry a výběrové směrodatné odchylky tělesné výšky a délky chodidla pro každé pohlaví zvlášť i pro celý datový soubor.

```
table(foot$sex)
```

```
f m
70 47
mean(foot$body.H[foot$sex=='f']) [1] 1670.771
sd(foot$body.H[foot$sex=='f']) [1] 71.26927

mean(foot$body.H[foot$sex=='m']) [1] 1780.064
sd(foot$body.H[foot$sex=='m']) [1] 58.38456

mean(foot$body.H) [1] 1714.675
sd(foot$body.H) [1] 85.25619

mean(foot$foot.L[foot$sex=='f']) [1] 244.3714
sd(foot$foot.L[foot$sex=='f']) [1] 11.43497

mean(foot$foot.L[foot$sex=='m']) [1] 266.2128
sd(foot$foot.L[foot$sex=='m']) [1] 11.46248

mean(foot$foot.L) [1] 253.1453
sd(foot$foot.L) [1] 15.66914
```

	Tělesná výška (mm)			Délka chodidla (mm)	
	rozsah	průměr	sm. odchylka	průměr	sm. odchylka
Ženy					
Muži					
Celkový					

Důležitým předpokladem analýzy kovariance je rovnoběžnost přímk pro každou kategorii faktoru, v našem případě tedy musíme ověřit předpoklad, že regresní přímk modelující závislost tělesné výšky na délce chodidla pro ženy je rovnoběžná s regresní přímkou modelující závislost tělesné výšky na délce chodidla pro muže. Pro ověření rovnoběžnosti sestavíme model s interakcí (tj. model různoběžných přímk) a bez interakce (tj. model rovnoběžných přímk) a otestujeme, zda je model bez interakce dostačující.

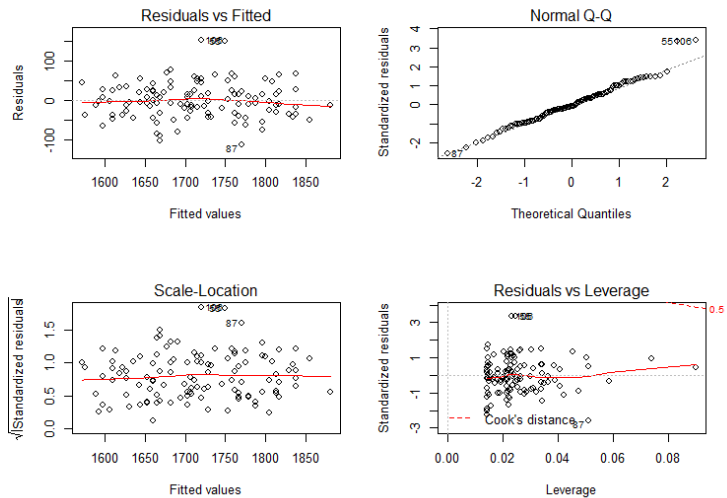
```
model.interakce <- lm(body.H ~ sex * foot.L, data=foot)
model.bez.int <- lm(body.H ~ sex + foot.L, data=foot)
anova(model.bez.int, model.interakce)
Analysis of variance Table
```

```
Model 1: body.H ~ sex + foot.L
Model 2: body.H ~ sex * foot.L
Res.Df  RSS Df Sum of Sq  F Pr(>F)
1     114 236145
2     113 231676 1     4468.8 2.1797 0.1426
```

Hodnota testovací statistiky Fobs = p-hodnota Závěr

Model bez interakce vychází jako dostatečný, tedy předpoklad o rovnoběžnosti regresních přímk budeme považovat za splněný.

Protože se jedná o speciální lineární regresní model, zbývající předpoklady analýzy kovariance můžeme ověřit pomocí analýzy reziduí.



```
shapiro.test(model.bez.int$residuals)
Shapiro-Wilk normality test
```

```
data: model.bez.int$residuals
W = 0.97883, p-value = 0.06131
```

```
t.test(model.bez.int$residuals)
One sample t-test
```

```
data: model.bez.int$residuals
t = 3.7196e-16, df = 116, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -8.261704  8.261704
sample estimates:
 mean of x
1.551554e-15
```

```
library(car)
durbinwatsonTest(model.bez.int)
lag Autocorrelation D-W Statistic p-value
1 -0.03256937 2.061963 0.708
Alternative hypothesis: rho != 0
```

Hypotézu o nulové střední hodnotě reziduí protože t-test nabývá hodnoty s p-hodnotou, z grafického posouzení také nevidíme problém. Shapirův-Wilkův test nabývá hodnoty s p-hodnotou, v kvantil-kvantilovém grafu jsou rezidua, předpoklad normality tedy považujeme za

Předpoklad rovnosti rozptylů se na základě grafického posouzení zdá

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

Vypíšeme si detaily modelu.

```
summary(model.bez.int)
```

```
Call:
lm(formula = body.H ~ sex + foot.L, data = foot)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-114.008  -33.255   -3.586   24.289  151.898
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  634.1086   90.7750   6.985 2.02e-10 ***
sexm          16.6379   11.8006   1.410  0.161
foot.L         4.2422    0.3708  11.441 < 2e-16 ***
```

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

```
Residual standard error: 45.51 on 114 degrees of freedom
Multiple R-squared:  0.7199, Adjusted R-squared:  0.715
F-statistic: 146.5 on 2 and 114 DF, p-value: < 2.2e-16
```

MNC odhady koeficientu a jejich interpretace:

```
β0 = .....
β1 = .....
β2 = .....
Celkový F-test na hladině významnosti 0,05:
F = .....
p-hodnota = .....
závěr .....
```

Z dílčích t-testů se zdá, že faktor pohlaví není významný. Sestavíme proto model, který ho neobsahuje, a otestujeme, zda je tento model dostatečný. Testujeme tedy shodnost regresních přímek.

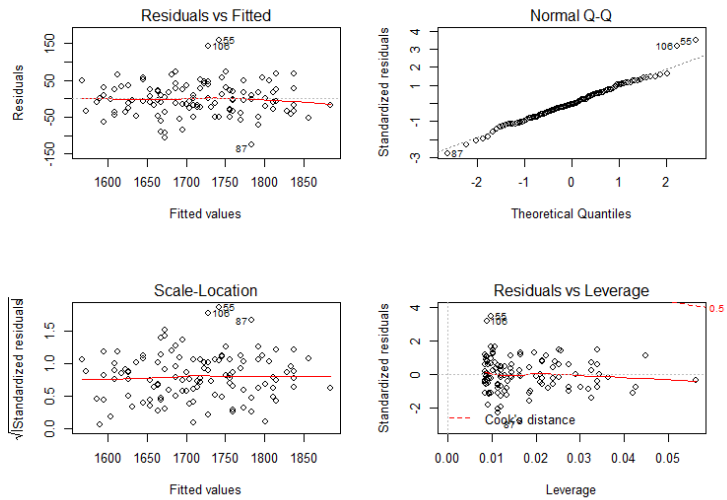
```
model.bez.int.bez.faktoru <- lm(body.H ~ foot.L, data=foot)
anova(model.bez.int.bez.faktoru, model.bez.int)
Analysis of variance Table
```

```
Model 1: body.H ~ foot.L
Model 2: body.H ~ sex + foot.L
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     115 240262
2     114 236145  1    4117.8 1.9879 0.1613
```

Hodnota testovací statistiky Fobs = p-hodnota Závěr

Pro pořádek si ještě ověříme, že výsledný model splňuje předpoklady regresního modelu.

```
par(mfrow=c(2,2))
> plot(model.bez.int.bez.faktoru)
```



`summary(model.bez.int.bez.faktoru)`

Call:
lm(formula = body.H ~ foot.L, data = foot)

Residuals:
Min 1Q Median 3Q Max
-126.021 -28.627 -3.021 28.599 158.388

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 549.9661 68.6931 8.006 1.06e-12 ***
foot.L 4.6010 0.2708 16.987 < 2e-16 ***

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

Residual standard error: 45.71 on 115 degrees of freedom
Multiple R-squared: 0.715, Adjusted R-squared: 0.7126
F-statistic: 288.6 on 1 and 115 DF, p-value: < 2.2e-16

Index determinace a jeho interpretace:

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

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

Výsledný model zakreslíme společně s pásem spolehlivosti.

```
xx <- seq(min(foot$foot.L), max(foot$foot.L), length=300)
> interval.spol <-
predict(model.bez.int.bez.faktoru,newdata=data.frame(foot.L=xx),
+ interval='confidence')
> plot(foot$foot.L, foot$body.H, xlab='Delka chodidla (mm)', ylab='Telesna
vyska (mm)', type='n')
> points(foot$foot.L[foot$sex=='f'], foot$body.H[foot$sex=='f'], pch=1,
col='red')
> points(foot$foot.L[foot$sex=='m'], foot$body.H[foot$sex=='m'], pch=2,
col='blue')
> lines(xx,interval.spol[,1])
> lines(xx,interval.spol[,2], lty=2)
> lines(xx,interval.spol[,3], lty=2)
> legend("topleft", c("zena", "muz"), pch = c(1,2), col=c('red', 'blue'))
```

