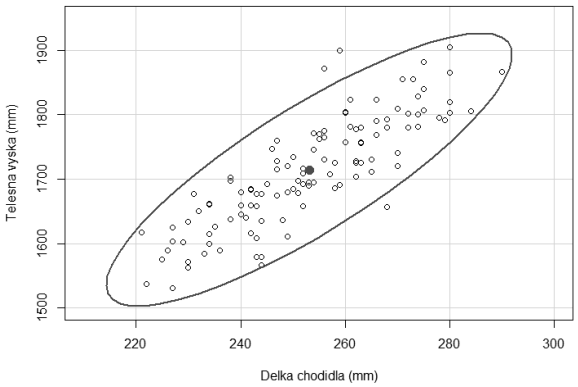
Vzorové příklady na korelační analýzu

Příklad 1.: V souboru lrm-foot.txt máme k dispozici antropometrické údaje 117 mladých dospělých lidí (převážně studentů vysokých škol z Brna a Ostravy). Vypočítejte hodnotu korelačního koeficientu tělesné výšky (proměnná body.H, v mm) a délky chodidla (proměnná foot.L, v mm). Na hladině významnosti 0,05 otestujte hypotézu o nekorelovanosti těchto veličin.
Načteme data: `foot<-read.table("lrm-foot.txt")`
Nejprve orientačně ověříme dvourozměrnou normalitu pomocí tečkového diagramu s 95% elipsou spolehlivosti.

```
library(car)
dataEllipse(foot$foot.L, foot$body.H, level=0.95, xlab='Delka chodidla (mm)',
            ylab='Telesna vyska (mm)', xlim=c(210,300), ylim=c(1500, 1950))
```



Na hladině významnosti 0,05 otestujeme hypotézu, že délka chodidla a tělesná výška jsou nekorelované.

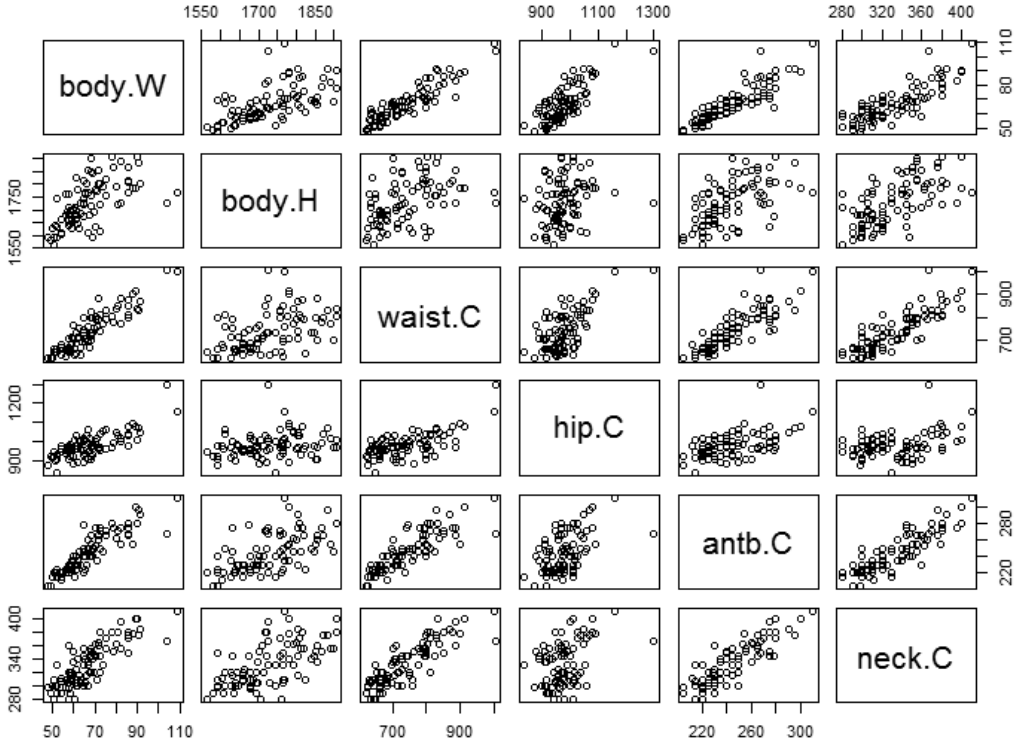
```
cor.test(foot$foot.L, foot$body.H, method='pearson')

        Pearson's product-moment correlation

data:  foot$foot.L and foot$body.H
t = 16.987, df = 115, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7844941 0.8904482
sample estimates:
      cor
0.8456034
```

Realizace výběrového korelačního koeficientu: r = ... Hodnota testovací statistiky: t = ... p-hodnota = ... Interval spolehlivosti: ..... Závěr: .....

Příklad 2.: V souboru cneck.txt máme k dispozici antropometrická data 87 mladých dospělých lidí (převážně studentů vysokých škol z Brna a Ostravy). V souboru máme tyto naměřené veličiny: obvodu krku (proměnná neck.C), tělesná hmotnost (proměnná body.W), tělesná výška (proměnná body.H), obvodu pasu (proměnná waist.C), obvodu boků (proměnná hip.C) a obvodu předloktí (proměnná antb.C). Hmotnost byla měřena v kilogramech, délkové míry v milimetrech. Předpokládáme, že se jedná o náhodný výběr z šestirozměrného normálního rozdělení.

Načteme data: `neck <- read.table("cneck.txt",header=T)`

Vykreslíme si bodové diagramy pro všechny dvojice proměnných. Sloupce id a sex nás nyní nezajímají: `plot(neck[,3:8])`



Vypočítáme realizace výběrové kovarianční a výběrové korelační matice.

```
cov(neck[,3:8], method='pearson')
          body.W    body.H   waist.C     hip.C    antb.C    neck.C
body.W  152.6496  657.5217  965.5304  607.5613  253.5553  322.9343
body.H  657.5217 7645.5031 3468.0445 1302.6721 1191.6834 1726.9988
waist.C 965.5304 3468.0445 7461.3729 3652.7598 1714.3577 2328.7354
hip.C   607.5613 1302.6721 3652.7598 4182.0588  791.0829  813.5605
antb.C  253.5553 1191.6834 1714.3577  791.0829  542.5311  635.5774
neck.C  322.9343 1726.9988 2328.7354  813.5605  635.5774 1007.3063

cor(neck[,3:8], method='pearson')
           body.W    body.H   waist.C     hip.C    antb.C    neck.C
body.W  1.0000000 0.6086383 0.9047087 0.7604090 0.8810742 0.8235417
body.H  0.6086383 1.0000000 0.4591687 0.2303759 0.5851208 0.6223121
waist.C 0.9047087 0.4591687 1.0000000 0.6539080 0.8520787 0.8494347
hip.C   0.7604090 0.2303759 0.6539080 1.0000000 0.5251877 0.3963821
antb.C  0.8810742 0.5851208 0.8520787 0.5251877 1.0000000 0.8597562
neck.C  0.8235417 0.6223121 0.8494347 0.3963821 0.8597562 1.0000000
```

Pokud chceme pro jednotlivé korelační koeficienty otestovat hypotézu o jejich nulovosti, místo toho, abychom pro každou dvojici jednotlivě použili cor.test(), lze použít funkci rcorr() z knihovny Hmisc. Do ní musí vstupovat matice, ale náš soubor je data frame, proto musíme použít funkci as.matrix(), která mění data frame na matici.

```
library(Hmisc)
rcorr(as.matrix(neck[,3:8]), type='pearson')
```

```
        body.W body.H waist.C hip.C antb.C neck.C
body.W    1.00   0.61    0.90  0.76   0.88   0.82
body.H    0.61   1.00    0.46  0.23   0.59   0.62
waist.C   0.90   0.46    1.00  0.65   0.85   0.85
hip.C     0.76   0.23    0.65  1.00   0.53   0.40
antb.C    0.88   0.59    0.85  0.53   1.00   0.86
neck.C    0.82   0.62    0.85  0.40   0.86   1.00

n= 87
P
        body.W body.H waist.C hip.C  antb.C neck.C
body.W          0.0000 0.0000  0.0000 0.0000 0.0000
body.H  0.0000         0.0000  0.0318 0.0000 0.0000
waist.C 0.0000 0.0000          0.0000 0.0000 0.0000
hip.C   0.0000 0.0318 0.0000          0.0000 0.0001
antb.C  0.0000 0.0000 0.0000  0.0000         0.0000
neck.C  0.0000 0.0000 0.0000  0.0001 0.0000
```

Hodnota korelačního koeficientu obvodu krku a tělesné váhy ..... p-hodnota .....
Závěr .....
Hodnota korelačního koeficientu obvodu krku a tělesné výšky ..... p-hodnota .....
Závěr .....
Hodnota korelačního koeficientu obvodu krku a obvodu pasu ..... p-hodnota .....
Závěr .....
Hodnota korelačního koeficientu obvodu krku a obvodu boků ..... p-hodnota .....
Závěr .....
Hodnota korelačního koeficientu obvodu krku a obvodu zápěstí ..... p-hodnota.....
Závěr .....

Koeficient parciální korelace $R_{Y,Z.X}$ měří sílu lineárního vztahu mezi náhodnými veličinami Y a Z při kontrolování efektu náhodného vektoru **X**. Jde tedy o korelační koeficient mezi rezidui $Y - \hat{Y}$ a $Z - \hat{Z}$, přičemž obě veličiny modelujeme pomocí **X**. Podívejme se na na parciální korelační koeficient mezi obvodem krku a tělesnou váhou při eliminaci vlivu tělesné výšky, obvodu pasu, obvodu boků a obvodu předloktí.

```
m1 <- lm(neck.C ~ body.H + waist.C + hip.C + antb.C, data=neck)
y.res <- m1$residuals
m2 <- lm(body.W ~ body.H + waist.C + hip.C + antb.C, data=neck)
z.res <- m2$residuals
r_yz.x <- cor(y.res,z.res)
r_yz.x
[1] 0.2363391
n <- nrow(neck)
> k <- 4
> t.obs <- r_yz.x*sqrt(n-k-2)/sqrt(1-r_yz.x^2)
> t.obs
[1] 2.189067
qt(0.975,n-k-2)
[1] 1.989686
2*(1-pt(abs(t.obs), n-k-2))
[1] 0.0314706
```

Hodnota parciálního korelačního koeficientu obvodu pasu a tělesné váhy ..... Hodnota testovací statistiky t = ... Kritický obor W ..... p-hodnota ..... Závěr .....

Totéž bychom mohli opakovat i pro další dvojice. K výpočtu parciálních korelačních koeficientů lze ale také využít funkci pcor z balíčku ppcor. Ta nám spočítá hodnoty parciálních korelačních koeficientů pro všechny dvojice proměnných s eliminací vlivu zbývajících proměnných. Ne vždy nás zajímají všechny dvojice, proto je potřeba se ve výstupu umět orientovat. Nás zajímají jen parciální korelační koeficienty pro obvod krku a jednotlivé další proměnné.

```
library(ppcor)
pcor(neck[,3:8], method='pearson')
$estimate
          body.W      body.H     waist.C       hip.C      antb.C     neck.C
body.W  1.0000000  0.50401882  0.38150690  0.74878319  0.44531275  0.2363391
body.H  0.5040188  1.00000000 -0.39710449 -0.32733792 -0.05768707  0.1903690
waist.C 0.3815069 -0.39710449  1.00000000  0.07118414  0.09435564  0.4501118
hip.C   0.7487832 -0.32733792  0.07118414  1.00000000 -0.26122384 -0.4442683
antb.C  0.4453127 -0.05768707  0.09435564 -0.26122384  1.00000000  0.2229913
neck.C  0.2363391  0.19036904  0.45011182 -0.44426825  0.22299134  1.0000000

$p.value
              body.W       body.H      waist.C        hip.C       antb.C
neck.C
body.W  0.000000e+00 1.187915e-06 3.728691e-04 3.988139e-16 2.460421e-05 3.1470
60e-02
body.H  1.187915e-06 0.000000e+00 2.016058e-04 2.522330e-03 6.044474e-01 8.4735
92e-02
waist.C 3.728691e-04 2.016058e-04 0.000000e+00 5.225008e-01 3.961711e-01 1.9601
17e-05
hip.C   3.988139e-16 2.522330e-03 5.225008e-01 0.000000e+00 1.706305e-02 2.5840
51e-05
antb.C  2.460421e-05 6.044474e-01 3.961711e-01 1.706305e-02 0.000000e+00 4.2731
06e-02
neck.C  3.147060e-02 8.473592e-02 1.960117e-05 2.584051e-05 4.273106e-02 0.0000
00e+00

$statistic
           body.W     body.H    waist.C       hip.C     antb.C    neck.C
body.W   0.000000  5.2520643  3.7145050 10.1673146  4.4761294  2.189067
body.H   5.252064  0.0000000 -3.8941424 -3.1178097 -0.5200496  1.745237
waist.C  3.714505 -3.8941424  0.0000000  0.6422866  0.8530064  4.536543
hip.C   10.167315 -3.1178097  0.6422866  0.0000000 -2.4355823 -4.463045
antb.C   4.476129 -0.5200496  0.8530064 -2.4355823  0.0000000  2.058761
neck.C   2.189067  1.7452372  4.5365425 -4.4630449  2.0587608  0.000000

$n
[1] 87
$gp
[1] 4
$method
[1] "pearson"
```

Hodnota parciálního korelačního koeficientu obvodu krku a tělesné váhy ..... Hodnota testovací statistiky t = ... p-hodnota ..... Závěr .....
Hodnota parciálního korelačního koeficientu obvodu krku a tělesné výšky ..... Hodnota testovací statistiky t = ... p-hodnota ..... Závěr .....
Hodnota parciálního korelačního koeficientu obvodu krku a obvodu pasu ..... Hodnota testovací statistiky t = ... p-hodnota ..... Závěr .....
Hodnota parciálního korelačního koeficientu obvodu krku a obvodu boků ..... Hodnota testovací statistiky t = ... p-hodnota ..... Závěr .....
Hodnota parciálního korelačního koeficientu obvodu krku a obvodu zápěstí ..... Hodnota testovací statistiky t = ... p-hodnota ..... Závěr .....

Dále nás zajímá hodnota koeficientu vícenásobné korelace obvodu krku s ostatními spojitými proměnnými. Nejprve vypočítáme korelace obvodu krku s vysvětlujícími proměnnými.

```
cor.yx <- cor(neck$neck.C, neck[,3:7], method='pearson')
cor.yx
        body.W    body.H   waist.C     hip.C    antb.C
[1,] 0.8235417 0.6223121 0.8494347 0.3963821 0.8597562
```

Poté i korelační matici vysvětlujících proměnných.

```
cor.xx <- cor(neck[,3:7], method='pearson')
cor.xx
           body.W    body.H   waist.C     hip.C    antb.C
body.W  1.0000000 0.6086383 0.9047087 0.7604090 0.8810742
body.H  0.6086383 1.0000000 0.4591687 0.2303759 0.5851208
waist.C 0.9047087 0.4591687 1.0000000 0.6539080 0.8520787
hip.C   0.7604090 0.2303759 0.6539080 1.0000000 0.5251877
antb.C  0.8810742 0.5851208 0.8520787 0.5251877 1.0000000
```

Pomocí těchto mezivýpočtů můžeme vypočítat hodnotu koeficientu vícenásobné korelace.

```
r.yx <- sqrt(cor.yx %*% solve(cor.xx) %*% t(cor.yx))
> r.yx
          [,1]
[1,] 0.9259102
```

Uvědomíme-li si, že koeficient vícenásobné korelace nám říká, jak dobře lze náhodnou veličinu Y predikovat pomocí náhodného vektoru **X** v rámci lineárního regresního modelu, lze ho vypočítat jako korelační koeficient mezi náhodnou veličinou Y a odhadnutými hodnotami $\hat{Y}$.

```
model <- lm(neck.C ~ body.W + body.H + waist.C + hip.C + antb.C, data=neck)
y.odhad <- model$fitted.values
r.yx <- cor(neck$neck.C, y.odhad, method='pearson')
r.yx
[1] 0.9259102
```

Všimneme si i souvislosti s indexem determinace lineárního regresního modelu.

```
ID <- summary(model)$r.squared
> sqrt(ID)
[1] 0.9259102
```

Nyní chceme na hladině 0,05 testovat hypotézu, že tento koeficient vícenásobné korelace je roven 0. Vypočítáme si tedy hodnotu testovací statistiky a příslušný kritický obor.

```
n <- nrow(neck)
> k <- 5
> F.obs <- (n-k-1)/k * r.yx^2/(1-r.yx^2)
> F.obs
[1] 97.33264
qf(0.95,df1=k, df2=n-k-1)
[1] 2.327269
```

Hodnota testovací statistiky Fobs = ...
Kritický obor W = ...
Závěr .....

Hodnotu testovací statistiky a příslušnou p-hodnotu zjistíme i z výpisu modelu.

```
summary(model)

Call:
lm(formula = neck.C ~ body.W + body.H + waist.C + hip.C + antb.C,
    data = neck)

Residuals:
    Min      1Q  Median      3Q     Max
-32.266  -8.030   1.169   8.493  33.577

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 165.63910   64.79600   2.556   0.0124 *
body.W        1.02594    0.46867   2.189   0.0315 *
body.H        0.04039    0.02314   1.745   0.0847 .
waist.C       0.18260    0.04025   4.537 1.96e-05 ***
hip.C        -0.18166    0.04070  -4.463 2.58e-05 ***
antb.C        0.29120    0.14144   2.059   0.0427 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.35 on 81 degrees of freedom
Multiple R-squared:  0.8573,   Adjusted R-squared:  0.8485
F-statistic: 97.33 on 5 and 81 DF,  p-value: < 2.2e-16
```

Hodnota testovací statistiky Fobs = ...
p-hodnota .....
Závěr .....