cv9.R

Marek

2023-04-13

```r
library(latexpdf)
```

Dvojrozmerny datovy subor, korelacna analyza

Analyzujeme dvojrozmerny datovy subor. Ratame charakteristiky pre jednorozmerne subory, pribudnu charakteristiky pre dvojrozmerny subor. Priklad. Tabulka uvadza cenu pevnych diskov (y) ich kapacitu (x) od vyrobcu ABC. Zratajte vektor strednych hodnot a disperzii.

```r
x <- c(160, 250, 320, 500, 750, 900, 1000, 1500, 2000)
y <- c(789, 800, 851, 874, 1193, 1200, 1335, 1704, 2073)
c(mean(x), mean(y))
```

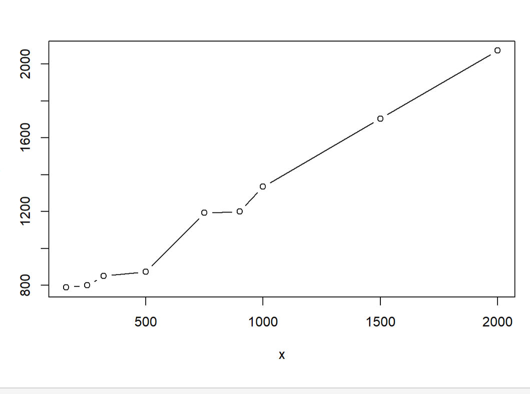
```
## [1]  820.000 1202.111
```

```r
c(var(x), var(y))
```

```
## [1] 376425.0 198922.1
```

Dalej ratame charakteristiku dvojrozmerneho datoveho suboru kovarianciu (Personova, Kendallova, Spearmanova). Je to cislo, ktore nam charakterizuje, ci je medzi zlozkami linearna zavislost alebo nie, ci koreluju. Kym neznormujeme toto cislo, tak mozeme povedat iba jedine, cov = 0 zlozky su nekorelovanea ako cov != 0 su korelovane.

```r
plot(x, y, type = 'b')
```



```r
cov(x, y) #korelovane
```

```
## [1] 271966.2
```

```r
cov(x, y, method = 'kendall')
```

```
## [1] 72
```

```r
cov(x, y, method = 'spearman')
```

```
## [1] 7.5
```

Aby sme zistili, ako silno sa ovplyvnuju, je potrebne nromovat kovarianciu, dostaneme korelacny koeficient, co je cislo medzi -1 a 1. Blizke -1, 1 silna, silna korelacia, silna linearna zavislost. Blizke 0, nekorelovane a linearne nezavisle iba pre normalne rozdelene data. Zlozitejsie zavislosti sa nemusia odhalit. Pearsonov korelacny koeficient je citlivy na vybocujuce hodnoty, vsetky naraz ratat.Pripravime funkciu korelacie.

```r
cor(x, y)
```

```
## [1] 0.9938807
```

```r
korelacie <- function(x, y){c('Pearson' = cor(x, y),
                              'Kendall' = cor(x, y, method = 'kendall'),
                              'Spearman' = cor(x, y, method = 'spearman'))}
korelacie(x, y)
```

```
##   Pearson   Kendall  Spearman
## 0.9938807 1.0000000 1.0000000
```

Testovat korelacny koeficient mozeme, dvojrozmerne data ale musia byt normalne rozdelene, testujeme hypotezu o nulovosti korelacneho koeficientu.

$$H_0 \quad \rho = 0 \quad H_1 \quad \rho \neq 0$$

kniznice pre test normality

```r
library(nortest)
library(mvnormtest)
```

uprava dat, dvojrozmerny subor

```r
data <- rbind(x, y)
mshapiro.test(data)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  Z
## W = 0.86547, p-value = 0.1099
```

P hodnota > 0.05 nezamietam hypotezu o normalite dat mozeme testovat nulovost korelacneho koeficienta

```r
cor.test(x, y) #zamietam H0
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  x and y
## t = 23.806, df = 7, p-value = 5.867e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9700441 0.9987619
## sample estimates:
##       cor 
## 0.9938807
```

```r
cor.test(x, y, method = 'kendall') #zamietam H0
```

```
## 
##  Kendall's rank correlation tau
## 
## data:  x and y
## T = 36, p-value = 5.511e-06
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
## tau 
##   1
```

```r
cor.test(x, y, method = 'spearman') #zamietam H0
```

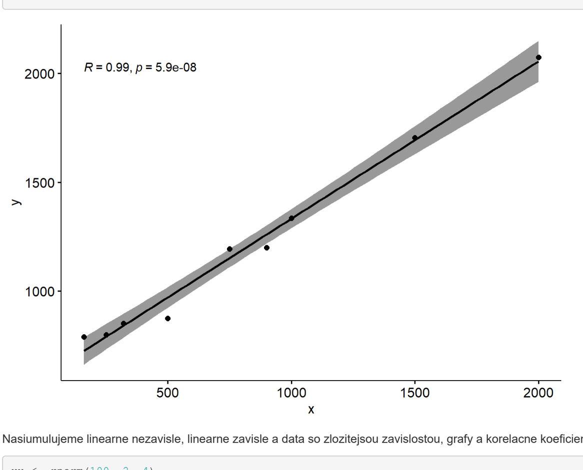
```
## 
##  Spearman's rank correlation rho
## 
## data:  x and y
## S = 0, p-value = 5.511e-06
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho 
##   1
```

graf s viac informaciami, nacitame kniznice

```r
library(ggplot2)
library(ggpubr)
```

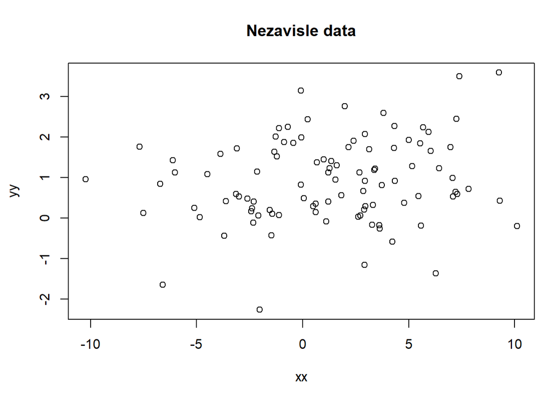
```
## Warning: package 'ggpubr' was built under R version 4.2.3
```

```r
data1 <- data.frame(x, y)
ggscatter(data1, 'x', 'y',
          add = 'reg.line',
          conf.int = T,
          cor.coef = T)
```



Nasiumulujeme linearne nezavisle, linearne zavisle a data so zlozitejsou zavislostou, grafy a korelacne koeficienty

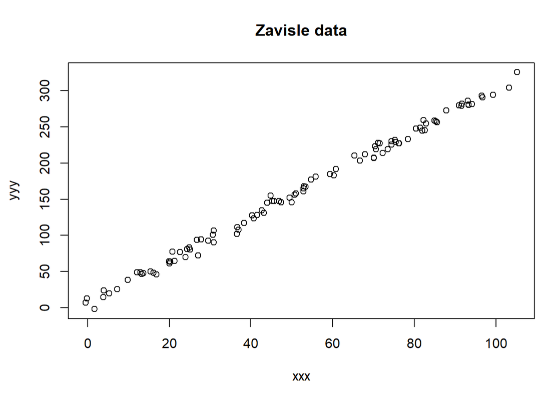
```r
xx <- rnorm(100, 2, 4)
yy <- rnorm(100, 1, 1)
plot(xx, yy, main = 'Nezavisle data')
```




```r
korelacie(xx, yy)
```

```
##   Pearson   Kendall  Spearman
## 0.1788016 0.1034343 0.1647165
```

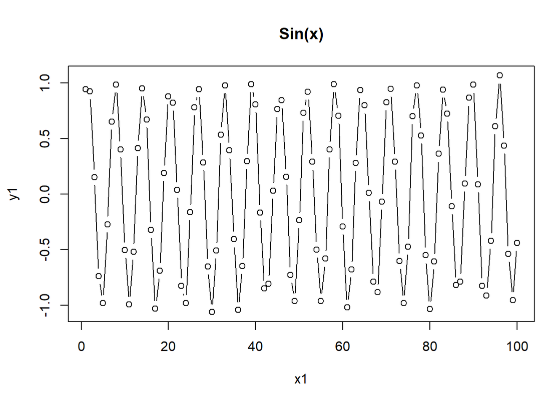
```r
xxx <- 1:100+xx
yyy <- 2+3*xxx+rnorm(100, 2, 6)
plot(xxx, yyy, main = 'Zavisle data')
```




```r
korelacie(xxx, yyy)
```

```
##   Pearson   Kendall  Spearman
## 0.9977918 0.9567677 0.9966637
```

```r
x1 <- 1:100
y1 <- sin(x1) + rnorm(100, 0, 0.05)
plot(x1, y1, main='Sin(x)', type = 'b')
```




```r
korelacie(x1, y1)
```

```
##     Pearson     Kendall    Spearman
## -0.03782328 -0.02101010 -0.02945095
```

Korelacna analyza pre viacrozmerny vektor

Castejsie mame viacrozmerne merania, vtedy ratame korelacne koeficienty vzdy medzi dvomi zlozkami, zapisujeme do matice, ktora sa nazyva korelacna matica. Symetricka matica, na diagonale su jednotky. Vyukovy dataset mtcars, carData, car

```r
library(carData)
library(car)
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```r
View(cor(mtcars))
```

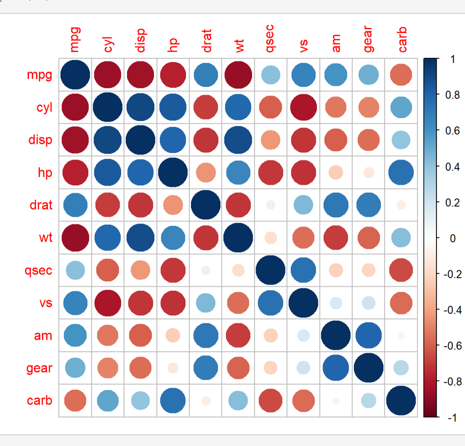
lepsie je to aj vizualizovat, kniznica corrplot

```r
library(corrplot)
```

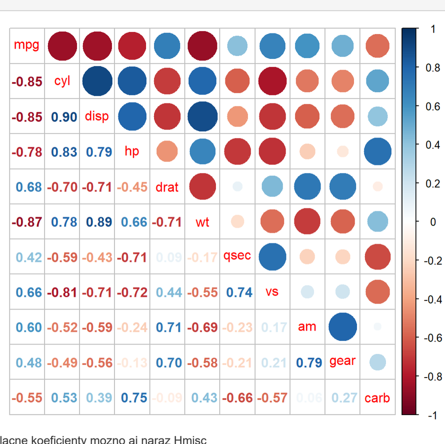
```
## Warning: package 'corrplot' was built under R version 4.2.3
```

```
## corrplot 0.92 loaded
```

```r
km <- cor(mtcars)
corrplot(km)
```



```r
corrplot.mixed(km)
```



Testovat korelacne koeficienty mozno aj naraz Hmisc

```r
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## 
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
## 
##     format.pval, units
```

```r
rcorr(as.matrix(mtcars))
```

```
##        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
## qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
## vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
## am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
## gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
## carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
## 
## n= 32 
## 
## 
## P
##      mpg    cyl    disp   hp     drat   wt     qsec   vs     am     gear  
## mpg         0.0000 0.0000 0.0000 0.0000 0.0000 0.0171 0.0000 0.0003 0.0054
## cyl  0.0000        0.0000 0.0000 0.0000 0.0000 0.0004 0.0000 0.0022 0.0042
## disp 0.0000 0.0000        0.0000 0.0000 0.0000 0.0131 0.0000 0.0004 0.0010
## hp   0.0000 0.0000 0.0000        0.0100 0.0000 0.0000 0.0000 0.1798 0.4930
## drat 0.0000 0.0000 0.0000 0.0100        0.0000 0.6196 0.0117 0.0000 0.0000
## wt   0.0000 0.0000 0.0000 0.0000 0.0000        0.3389 0.0010 0.0000 0.0005
## qsec 0.0171 0.0004 0.0131 0.0000 0.6196 0.3389        0.0000 0.2057 0.2425
## vs   0.0000 0.0000 0.0000 0.0000 0.0117 0.0010 0.0000        0.3570 0.2579
## am   0.0003 0.0022 0.0004 0.1798 0.0000 0.0000 0.2057 0.3570        0.0000
## gear 0.0054 0.0042 0.0010 0.4930 0.0000 0.0005 0.2425 0.2579 0.0000       
## carb 0.0011 0.0019 0.0253 0.0000 0.6212 0.0146 0.0000 0.0007 0.7545 0.1290
##      carb  
## mpg  0.0011
## cyl  0.0019
## disp 0.0253
## hp   0.0000
## drat 0.6212
## wt   0.0146
## qsec 0.0000
## vs   0.0007
## am   0.7545
## gear 0.1290
## carb
```

Este jedna vhodna kniznica

```r
library(correlation)
```

```
## Warning: package 'correlation' was built under R version 4.2.3
```

```r
correlation(data1)
```

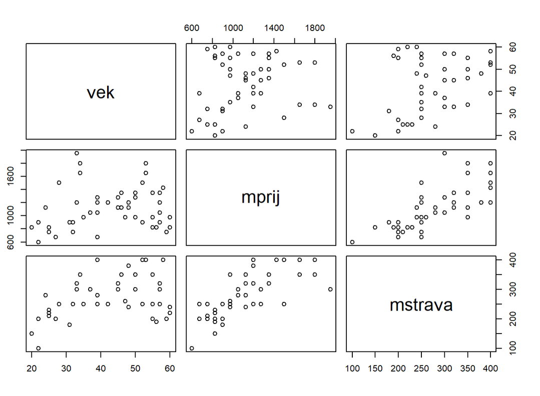
```
## # Correlation Matrix (pearson-method)
## 
## Parameter1 | Parameter2 |    r |       95% CI |  t(7) |         p
## -----------------------------------------------------------------
## x          |          y | 0.99 | [0.97, 1.00] | 23.81 | < .001***
## 
## p-value adjustment method: Holm (1979)
## Observations: 9
```

Grafy, ktore nam umoznia zistit strukturu dat, zavislosti v datach Nase data dataE.xlsx

```r
library(readxl)
data <- read_xlsx('dataE.xlsx')
head(data)
```

```
## # A tibble: 6 × 13
##   respon…¹ pohla…²   vek vzdel…³  stav fajciar praca  clen mprij mstrava rdovo…⁴
##      <dbl> <chr>   <dbl>   <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl>   <dbl>   <dbl>
## 1        1 m          28       3     1       0     2     5  1500     250   10000
## 2        2 m          22       2     1       0     1     4   900     200     350
## 3        3 m          50       1     2       0     2     4   975     350     300
## 4        4 z          39       2     2       0     3     5  1200     400     500
## 5        5 z          53       3     2       0     1     5  1800     400   15000
## 6        6 z          55       1     3       0     3     2   825     200     200
## # … with 2 more variables: pocit <dbl>, invest <dbl>, and abbreviated variable
## #   names ¹respondent, ²pohlavie, ³vzdelanie, ⁴rdovolenka
```

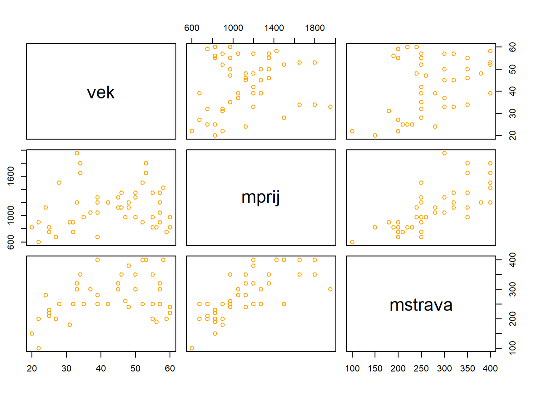
```r
pairs(data[,c('vek', 'mprij', 'mstrava')])
```



```r
korelacie(data$mprij, data$mstrava)
```

```
##   Pearson   Kendall  Spearman
## 0.7276575 0.6340417 0.7882542
```

```r
pairs(data[,c('vek', 'mprij', 'mstrava')], col='orange')
```



```r
library(GGally)
```

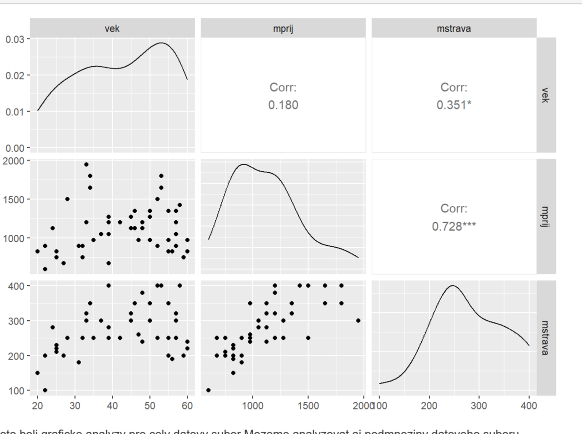
```
## Warning: package 'GGally' was built under R version 4.2.3
```

```
## Registered S3 method overwritten by 'GGally':
##   method from   
##   +.gg   ggplot2
```

```
## 
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:latexpdf':
## 
##     wrap
```

```r
ggpairs(data[,c('vek', 'mprij', 'mstrava')])
```



toto boli graficke analyzy pre cely datovy subor Mozeme analyzovat aj podmnoziny datoveho suboru

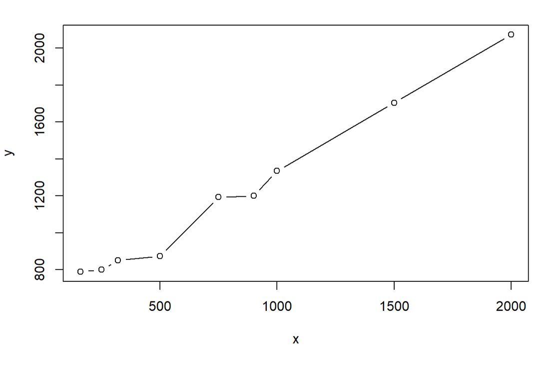
```r
ggpairs(data[,c('pohlavie', 'vek', 'mprij', 'mstrava')],
        aes(colour=pohlavie, alpha = 0.3))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Regresna analyza

Ked korelacne koeficienty su vysoke, teda data su linearne zavisle, tak nasleduje regresna analyza. Snazime sa zavislost popisat vhodnou funkciou - regresna funkcia, priklad zo zaciatku, prelozime priamku y=a+bx

```r
plot(x, y, type = 'b')
```



```r
lr1 <- lm(y~x)
lr1
```

```
## 
## Call:
## lm(formula = y ~ x)
## 
## Coefficients:
## (Intercept)            x  
##    609.6629       0.7225
```

```r
summary(lr1)
```

```
## 
## Call:
## lm(formula = y ~ x)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -96.912   2.839  10.138  18.341  63.737 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 609.66288   30.45571   20.02 1.94e-07 ***
## x             0.72250    0.03035   23.81 5.87e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 52.67 on 7 degrees of freedom
## Multiple R-squared:  0.9878, Adjusted R-squared:  0.9861 
## F-statistic: 566.7 on 1 and 7 DF,  p-value: 5.867e-08
```

Multiple R-squared -> cim blizsie k 1, tym lepsi model -> 98% dat je vysvetlenych nasou priamkou

```r
#abline(lr1, col = 'red')
```

Ak konstantu neuvazujeme

```r
lr2 <- lm(y~0+x)
summary(lr2)
```

```
## 
## Call:
## lm(formula = y ~ 0 + x)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -364.9  103.0  264.5  460.9  594.0 
## 
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)    
## x   1.2189     0.1249    9.76 1.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 376 on 8 degrees of freedom
## Multiple R-squared:  0.9225, Adjusted R-squared:  0.9128 
## F-statistic: 95.26 on 1 and 8 DF,  p-value: 1.017e-05
```

```r
#abline(lr2, col = 'green')
```

prelozime kvadraticku funckiu

```r
lr3 <- lm(y~poly(x, 2))
summary(lr3)
```

```
## 
## Call:
## lm(formula = y ~ poly(x, 2))
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -91.634 -12.870   2.168  25.098  58.875 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1202.11      17.36  69.241 6.11e-10 ***
## poly(x, 2)1  1253.78      52.08  24.072 3.38e-07 ***
## poly(x, 2)2    56.04      52.08   1.076    0.323    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 52.08 on 6 degrees of freedom
## Multiple R-squared:  0.9898, Adjusted R-squared:  0.9864 
## F-statistic: 290.3 on 2 and 6 DF,  p-value: 1.07e-06
```

```r
#lines(x, fitted(lr3), col = 'blue')
```