Ekonometrija 1

Prvi seminar:
Uvod v programski paket R/RStudio.

Na prvem seminarju bomo najprej spoznali zasnovo in osnove dela s programskim paketom R. Srečali se bomo z dvema pristopoma k analizi podatkov. Na konkretnem primeru si bomo pogledali pregled in urejanje podatkov, kreiranje različnih diagramov, delo s skalarji in matrikami, transformiranje spremenljivk, uporabo statističnih porazdelitev in testiranje ničelnih hipotez. Nato si bomo pogledali še primer časovno serije, kjer se bomo osredotočili na opredelitev časovne dimenzije, kreiranje periodičnih komponent, uporabo nepravih spremenljivk in uporabo odlogov.

♣-♣-♣

Primer 1: Na voljo imamo podatke za spremenljivke y, x_1, x_2 in x_3. Za navedene spremenljivke imamo po 8 opazovanj, kot je prikazano v tabeli. Podatki se že nahajajo v podatkovni datoteki `osnove_R.rds`, programska koda, ki jo boste potrebovali, pa v datoteki `osnove_R-ukazi.R`.

i	1	2	3	4	5	6	7	8
y_i	2	2	1	5	–4	1	4	1
x_{1i}	1	1	1	1	1	1	1	1
x_{2i}	1	2	0	–1	1	–1	–2	0
x_{3i}	–1	–1	2	–4	3	0	2	–1

a) Proučite podatke s pomočjo različnih ukazov za pregled podatkov. Kako bi najlažje ročno uredili podatke za konkretne spremenljivke in konkretna opazovanja v vaši bazi?
b) Proučite podatke še grafično s pomočjo različnih diagramov. Uporabite razsevni diagram, linijski diagram in histogram.
c) Na osnovi obstoječih spremenljivk iz podatkovne baze z različnimi transformacijami generirajte nekaj novih spremenljivk. Uporabite množenje, absolutne vrednosti, logaritmiranje, antilogaritmiranje ter standardiziranje.
d) Prikličite iz okolja R rezultate izvedbe enostavnejšega ukaza `psych::describe` ter kompleksnejšega ukaza `lm`. Kako bi jih shranili za kasnejšo uporabo?
e) Prikažite kovariančno in korelacijsko matriko spremenljivk y, x_2 in x_3. Ugotovite tudi statistično značilnost izračunanih korelacijskih koeficientov.

Izpis rezultatov obdelav v programskem paketu R:

a) Pregled podatkov

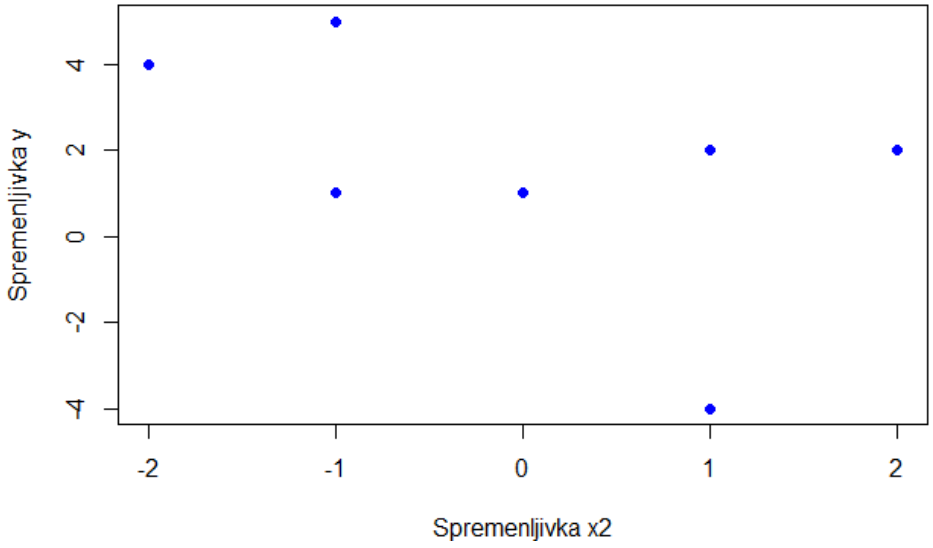
```
> sapply(osnove_R,class)
      obs         y          x1         x2         x3
[1,] "labelled" "labelled" "labelled" "labelled" "labelled"
[2,] "integer"  "integer"  "integer"  "integer"  "integer"
```

```
> psych::describe(osnove_R[,-c(1)], type=1)
   vars n mean   sd median trimmed  mad min max range  skew kurtosis   se
y     1 8  1.5 2.67    1.5     1.5 0.74  -4   5     9 -0.86     0.53 0.94
x1    2 8  1.0 0.00    1.0     1.0 0.00   1   1     0   NaN      NaN 0.00
x2    3 8  0.0 1.31    0.0     0.0 1.48  -2   2     4  0.00    -1.00 0.46
x3    4 8  0.0 2.27   -0.5     0.0 2.22  -4   3     7 -0.31    -0.70 0.80

> psych::describe(osnove_R$y, type=1)
   vars n mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 8  1.5 2.67    1.5     1.5 0.74  -4   5     9 -0.86     0.53 0.94

> Hmisc::describe(osnove_R$y)
osnove_R$y : Spremenljivka y
       n  missing distinct     Info     Mean      Gmd
       8        0        5     0.94      1.5    2.929

lowest : -4  1  2  4  5, highest: -4  1  2  4  5

Value         -4     1     2     4     5
Frequency      1     3     2     1     1
Proportion 0.125 0.375 0.250 0.125 0.125

> quantile(osnove_R$y, c(.01, .05, .1, .25, .5, .75, .9, .95, .99))
Spremenljivka y
 1%  5% 10% 25% 50% 75% 90% 95% 99%
 -3  -2   0   1   1   2   4   4   4

> desc=stat.desc(osnove_R[, -c(1)])
> round(desc, 2)
                   y x1    x2    x3
nbr.val         8.00  8  8.00  8.00
nbr.null        0.00  0  2.00  1.00
nbr.na          0.00  0  0.00  0.00
min            -4.00  1 -2.00 -4.00
max             5.00  1  2.00  3.00
range           9.00  0  4.00  7.00
sum            12.00  8  0.00  0.00
median          1.50  1  0.00 -0.50
mean            1.50  1  0.00  0.00
SE.mean         0.94  0  0.46  0.80
CI.mean.0.95    2.23  0  1.09  1.90
var             7.14  0  1.71  5.14
std.dev         2.67  0  1.31  2.27
coef.var        1.78  0   Inf   Inf

> freq=table(osnove_R$y, exclude=NULL)
> percent=prop.table(freq)*100
> cum=cumsum(percent)

> cbind(freq,percent,cum)
   freq percent   cum
-4    1    12.5  12.5
1     3    37.5  50.0
2     2    25.0  75.0
4     1    12.5  87.5
5     1    12.5 100.0

> osnove_R[,2:5]
   y x1 x2 x3
1  2  1  1 -1
2  2  1  2 -1
3  1  1  0  2
4  5  1 -1 -4
5 -4  1  1  3
6  1  1 -1  0
7  4  1 -2  2
8  1  1  0 -1
```

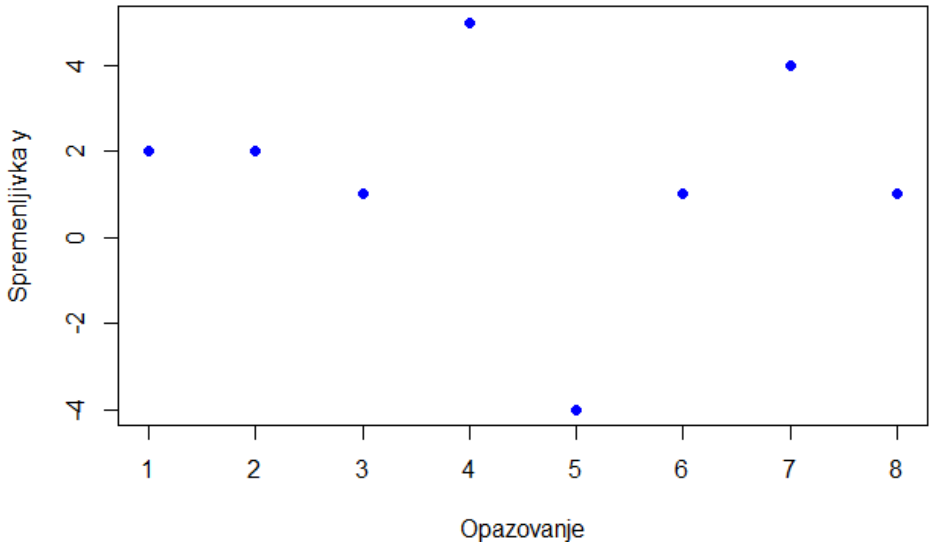
```
> osnove_R[4:8,2:5]
   y x1 x2 x3
4  5  1 -1 -4
5 -4  1  1  3
6  1  1 -1  0
7  4  1 -2  2
8  1  1  0 -1

> osnove_R[osnove_R$x2>=0,3:4]
  x1 x2
1  1  1
2  1  2
3  1  0
5  1  1
8  1  0
```

b) Diagrami v R

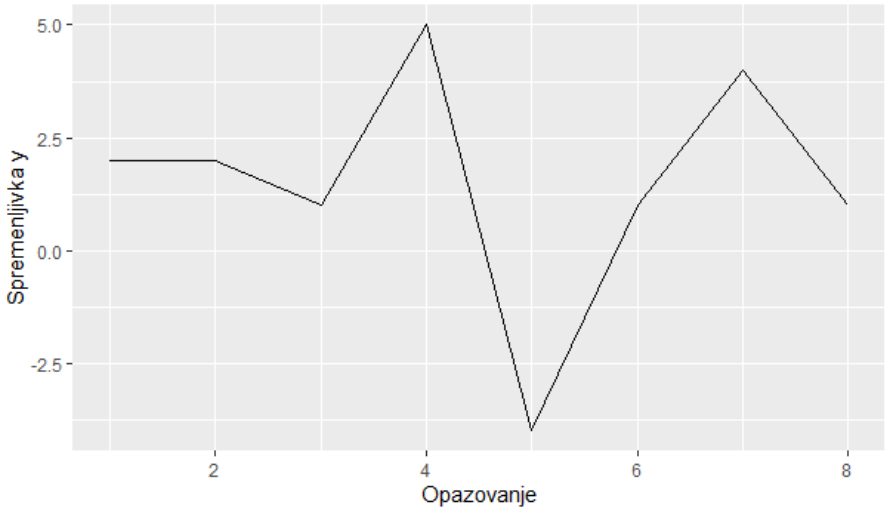
```
> plot(osnove_R$x2, osnove_R$y, col="blue", xlab=label(osnove_R$x2),
  ylab=label(osnove_R$y), pch=16)
```



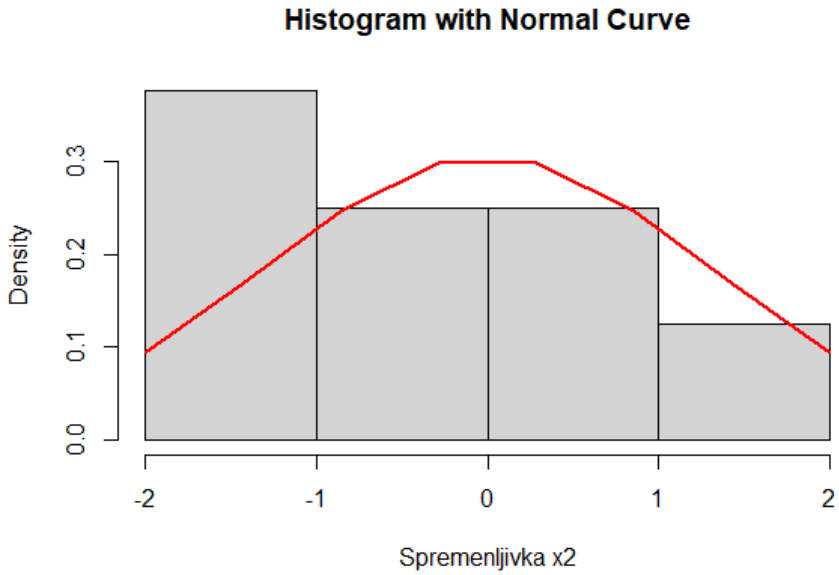

```
> plot(osnove_R$obs, osnove_R$y, col="blue", xlab=label(osnove_R$obs),
  ylab=label(osnove_R$y), pch=16)
```



```
> ts_plot = ggplot(osnove_R, aes(x = .data$obs, y = .data$y)) + geom_line() +
  labs(x=label(osnove_R$obs), y=label(osnove_R$y))
> plot(ts_plot)
```




```
> x = osnove_R$x2
> h = hist(x, breaks=4, col="light grey", xlab=label(osnove_R$x2),
  main="Histogram with Normal Curve", freq=FALSE)
> xfit = seq(min(x), max(x), length=length(x))
> yfit = dnorm(xfit, mean=mean(x), sd=sd(x))
> lines(xfit, yfit, col="red", lwd=2)
```



c) Generiranje novih spremenljivk

```
> osnove_R$yx2=100*osnove_R$y*osnove_R$x2

> osnove_R$x2sq=osnove_R$x2^2

> osnove_R$x2a=abs(osnove_R$x2)

> osnove_R$lx2=log(osnove_R$x2)
Warning message:
In log(osnove_R$x2) : NaNs produced
```

```
> osnove_R$ex2=exp(osnove_R$lx2)

> osnove_R[, -c(1)]
   y x1 x2 x3  yx2 x2sq x2a       lx2 ex2
1  2  1  1 -1  200    1   1 0.0000000   1
2  2  1  2 -1  400    4   2 0.6931472   2
3  1  1  0  2    0    0   0      -Inf   0
4  5  1 -1 -4 -500    1   1       NaN NaN
5 -4  1  1  3 -400    1   1 0.0000000   1
6  1  1 -1  0 -100    1   1       NaN NaN
7  4  1 -2  2 -800    4   2       NaN NaN
8  1  1  0 -1    0    0   0      -Inf   0

> osnove_R$lx2[is.na(osnove_R$lx2)] = 0
> osnove_R$lx2[which(osnove_R$lx2==-Inf)] = 0

> osnove_R$x2s=zscore(osnove_R$x2)

> osnove_R[, -c(1)]
   y x1 x2 x3  yx2 x2sq x2a       lx2 ex2        x2s
1  2  1  1 -1  200    1   1 0.0000000   1  0.7637626
2  2  1  2 -1  400    4   2 0.6931472   2  1.5275252
3  1  1  0  2    0    0   0 0.0000000   0  0.0000000
4  5  1 -1 -4 -500    1   1 0.0000000 NaN -0.7637626
5 -4  1  1  3 -400    1   1 0.0000000   1  0.7637626
6  1  1 -1  0 -100    1   1 0.0000000 NaN -0.7637626
7  4  1 -2  2 -800    4   2 0.0000000 NaN -1.5275252
8  1  1  0 -1    0    0   0 0.0000000   0  0.0000000

> osnove_R$yx2 = NULL
> osnove_R$x2sq = NULL
> osnove_R$x2a = NULL
> osnove_R$lx2 = NULL
> osnove_R$ex2 = NULL
> osnove_R$x2s = NULL
```

d) Priklic podatkov iz okolja R

```
> psych::describe(osnove_R[,-c(1)], type=1)
   vars n mean   sd median trimmed  mad min max range  skew kurtosis   se
y     1 8  1.5 2.67    1.5     1.5 0.74  -4   5     9 -0.86     0.53 0.94
x1    2 8  1.0 0.00    1.0     1.0 0.00   1   1     0   NaN      NaN 0.00
x2    3 8  0.0 1.31    0.0     0.0 1.48  -2   2     4  0.00    -1.00 0.46
x3    4 8  0.0 2.27   -0.5     0.0 2.22  -4   3     7 -0.31    -0.70 0.80

> tab_describe = psych::describe(osnove_R[,-c(1)], type=1)
> str(tab_describe)
Classes ‘psych’, ‘describe’ and 'data.frame':  4 obs. of  13 variables:
 $ vars    : int  1 2 3 4
 $ n       : num  8 8 8 8
 $ mean    : num  1.5 1 0 0
 $ sd      : num  2.67 0 1.31 2.27
 $ median  : num  1.5 1 0 -0.5
 $ trimmed : num  1.5 1 0 0
 $ mad     : num  0.741 0 1.483 2.224
 $ min     : num  -4 1 -2 -4
 $ max     : num  5 1 2 3
 $ range   : num  9 0 4 7
 $ skew    : num  -0.864 NaN 0 -0.314
 $ kurtosis: num  0.534 NaN -1 -0.704
 $ se      : num  0.945 0 0.463 0.802

> mean_y = tab_describe$mean[1]
> median_x3 = tab_describe$median[4]
```

```
> mean_y; median_x3
[1] 1.5
[1] -0.5

> regression = lm(y ~ x2 + x3, data = osnove_R)
> summary(regression)

Call:
lm(formula = y ~ x2 + x3, data = osnove_R)

Residuals:
Spremenljivka y
 1  2  3  4  5  6  7  8
 0  1  0  0 -2 -1  2 -1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5000     0.6661   2.252   0.0741 .
x2           -1.0000     0.5439  -1.839   0.1254
x3           -0.7500     0.3140  -2.388   0.0625 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.884 on 5 degrees of freedom
Multiple R-squared:  0.645,     Adjusted R-squared:  0.503
F-statistic: 4.542 on 2 and 5 DF,  p-value: 0.07509

> summary_regression = summary(regression)
> str(summary_regression)
List of 11
 $ call         : language lm(formula = y ~ x2 + x3, data = osnove_R)
 $ terms        :Classes 'terms', 'formula'  language y ~ x2 + x3
  .. ..- attr(*, "variables")= language list(y, x2, x3)
  .. ..- attr(*, "factors")= int [1:3, 1:2] 0 1 0 0 0 1
  .. .. ..- attr(*, "dimnames")=List of 2
  .. .. .. ..$ : chr [1:3] "y" "x2" "x3"
  .. .. .. ..$ : chr [1:2] "x2" "x3"
  .. ..- attr(*, "term.labels")= chr [1:2] "x2" "x3"
  .. ..- attr(*, "order")= int [1:2] 1 1
  .. ..- attr(*, "intercept")= int 1
  .. ..- attr(*, "response")= int 1
  .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
  .. ..- attr(*, "predvars")= language list(y, x2, x3)
  .. ..- attr(*, "dataClasses")= Named chr [1:3] "numeric" "numeric" "numeric"
  .. .. ..- attr(*, "names")= chr [1:3] "y" "x2" "x3"
 $ residuals    : 'labelled' Named num [1:8] 0.75 1.75 1 -0.5 -2.25 -1.5 2 -1.25
  ..- attr(*, "label")= chr "Spremenljivka y"
  ..- attr(*, "names")= chr [1:8] "1" "2" "3" "4" ...
 $ coefficients : num [1:3, 1:4] 1.5 -1 -0.75 0.666 0.544 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3] "(Intercept)" "x2" "x3"
  .. ..$ : chr [1:4] "Estimate" "Std. Error" "t value" "Pr(>|t|)"
 $ aliased      : Named logi [1:3] FALSE FALSE FALSE
  ..- attr(*, "names")= chr [1:3] "(Intercept)" "x2" "x3"
 $ sigma        : num 1.88
 $ df           : int [1:3] 3 5 3
 $ r.squared    : num 0.645
 $ adj.r.squared: num 0.503
 $ fstatistic   : Named num [1:3] 4.54 2 5
  ..- attr(*, "names")= chr [1:3] "value" "numdf" "dendf"
 $ cov.unscaled : num [1:3, 1:3] 1.25e-01 9.81e-18 -1.05e-34 9.81e-18 8.33e-02 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3] "(Intercept)" "x2" "x3"
  .. ..$ : chr [1:3] "(Intercept)" "x2" "x3"
 - attr(*, "class")= chr "summary.lm"
```

```
> regression_r2 = summary_regression$r.squared
> regression_r2
[1] 0.645

> regression_varcov = tril(vcov(regression))
> regression_varcov
3 x 3 Matrix of class "dtrMatrix"
             (Intercept)           x2           x3
(Intercept)  4.437500e-01            .            .
x2           3.483643e-17  2.958333e-01            .
x3          -3.721619e-34 -3.160425e-18  9.861111e-02

> regression_varcov33 = regression_varcov[3,3]
> regression_varcov33
[1] 0.09861111
```

e) Kovariance in korelacija

```
> df = data.frame(osnove_R)[, -c(1,3)]

> cov(df)
           y        x2        x3
y   7.142857 -1.714286 -3.857143
x2 -1.714286  1.714286  0.000000
x3 -3.857143  0.000000  5.142857

> cor(df)
            y         x2         x3
y   1.0000000 -0.4898979 -0.6363961
x2 -0.4898979  1.0000000  0.0000000
x3 -0.6363961  0.0000000  1.0000000

> rcorr(as.matrix(df))
       y    x2    x3
y   1.00 -0.49 -0.64
x2 -0.49  1.00  0.00
x3 -0.64  0.00  1.00

n= 8

P
   y      x2     x3
y         0.2178 0.0898
x2 0.2178        1.0000
x3 0.0898 1.0000
```

■

Primer 2: V priloženi podatkovni datoteki `osnove_casovnih_vrst.rds` se nahaja časovna vrsta z začetkom v letu 1950. Programska koda, ki jo boste potrebovali, se nahaja v datoteki `osnove_casovnih_vrst-ukazi.R`.

a) Odprite podatkovno datoteko v programskem paketu R. Proučite podatke s pomočjo različnih ukazov za pregled podatkov.
b) Opredelite časovno dimenzijo podatkov. Nato sortirajte podatkovno bazo po časovni dimenziji ter zamenjajte vrstni red spremenljivk. Generirajte trend, neprave spremenljivke za četrtletja in ciklično komponento.

c) Generirajte nepravo spremenljivko, ki zavzame vrednost 1, če ima naša spremenljivka vrednost, ki je večja ali enaka 80 % njene mediane ali pa manjša od dveh tretjin njene aritmetične sredine, v ostalih primerih pa zavzame vrednost 0.
d) Generirajte prve in četrte odloge naše spremenljivke ter druge vodeče odloge. Generirajte še prve diference naše spremenljivke.

Izpis rezultatov obdelav v programskem paketu R:

a) Pregled podatkov

```
> sapply(osnove_casovnih_vrst,class)
     kvartal    spr
[1,] "labelled" "labelled"
[2,] "Date"     "numeric"

> psych::describe(osnove_casovnih_vrst$spr, type=1)
   vars   n    mean      sd median trimmed     mad    min    max  range skew
X1    1 204 4562.65 2113.96 4142.2 4409.99 2601.81 1610.5 9303.9 7693.4 0.47
   kurtosis     se
X1    -0.87 148.01

> Hmisc::describe(osnove_casovnih_vrst$spr)
osnove_casovnih_vrst$spr : Casovna spremenljivka (v enotah mere)
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
     204        0      203        1     4563     2409     1887     2068     2602
     .50      .75      .90      .95
    4142     6294     7604     8436

lowest : 1610.5 1658.8 1723.0 1753.9 1773.5,
highest: 9049.9 9102.5 9229.4 9260.1 9303.9


b) Opredelitev casovne dimenzije in generiranje periodicnih komponent

> osnove_casovnih_vrst$kvartal = seq(as.Date("1950/1/1"), as.Date("2000/12/1"),
  by="quarter")

> osnove_casovnih_vrst = osnove_casovnih_vrst[order(osnove_casovnih_vrst$kvartal),]

> osnove_casovnih_vrst = osnove_casovnih_vrst[, c(2,1)]
> colnames(osnove_casovnih_vrst)
[1] "spr"     "kvartal"

> osnove_casovnih_vrst = osnove_casovnih_vrst[, c(2,1)]
> colnames(osnove_casovnih_vrst)
[1] "kvartal" "spr"

> osnove_casovnih_vrst$t = seq_along(osnove_casovnih_vrst$spr)

> osnove_casovnih_vrst$q = get_quarter(osnove_casovnih_vrst$kvartal)
> osnove_casovnih_vrst$d = dummy_cols(osnove_casovnih_vrst$q)

> osnove_casovnih_vrst$q = NULL

> osnove_casovnih_vrst$t2 = osnove_casovnih_vrst$t^2
> osnove_casovnih_vrst$t3 = osnove_casovnih_vrst$t^3

> osnove_casovnih_vrst[1:12,]
      kvartal    spr  t d..data d..data_1 d..data_2 d..data_3 d..data_4  t2   t3
1  1950-01-01 1610.5  1       1         1         0         0         0   1    1
2  1950-04-01 1658.8  2       2         0         1         0         0   4    8
```

```
3  1950-07-01 1723.0  3       3         0         0         1         0   9   27
4  1950-10-01 1753.9  4       4         0         0         0         1  16   64
5  1951-01-01 1773.5  5       1         1         0         0         0  25  125
6  1951-04-01 1803.7  6       2         0         1         0         0  36  216
7  1951-07-01 1839.8  7       3         0         0         1         0  49  343
8  1951-10-01 1843.3  8       4         0         0         0         1  64  512
9  1952-01-01 1864.7  9       1         1         0         0         0  81  729
10 1952-04-01 1866.2 10       2         0         1         0         0 100 1000
11 1952-07-01 1878.0 11       3         0         0         1         0 121 1331
12 1952-10-01 1940.2 12       4         0         0         0         1 144 1728

> osnove_casovnih_vrst = osnove_casovnih_vrst[, c(1,2)]
> colnames(osnove_casovnih_vrst)
[1] "kvartal" "spr"
```

c) Generiranje neprave spremenljivke

```
> psych::describe(osnove_casovnih_vrst$spr, type=1)
   vars   n    mean      sd median trimmed     mad    min    max  range skew
X1    1 204 4562.65 2113.96 4142.2 4409.99 2601.81 1610.5 9303.9 7693.4 0.47
   kurtosis     se
X1    -0.87 148.01

> mean_spr = psych::describe(osnove_casovnih_vrst$spr, type=1)$mean
> median_spr = psych::describe(osnove_casovnih_vrst$spr, type=1)$median

> mean_spr; median_spr
[1] 4562.646
[1] 4142.2

> osnove_casovnih_vrst$d = 0
> osnove_casovnih_vrst$d[which(osnove_casovnih_vrst$spr>=0.8*median_spr |
  osnove_casovnih_vrst$spr<(2/3)*mean_spr)] = 1

> freq=table(osnove_casovnih_vrst$d, exclude=NULL)
> percent=prop.table(freq)*100
> cum=cumsum(percent)

> cbind(freq,percent,cum)
  freq   percent        cum
0    9  4.411765   4.411765
1  195 95.588235 100.000000

> osnove_casovnih_vrst$d = NULL
```

d) Generiranje odlozenih in vodecih spremenljivk

```
> osnove_casovnih_vrst$spr_lag1 = shift(osnove_casovnih_vrst$spr, n=1,
  type=c("lag"))
> osnove_casovnih_vrst$spr_lag4 = shift(osnove_casovnih_vrst$spr, n=4,
  type=c("lag"))
> osnove_casovnih_vrst$spr_lead2 = shift(osnove_casovnih_vrst$spr, n=2,
  type=c("lead"))

> osnove_casovnih_vrst$spr_diff1 = osnove_casovnih_vrst$spr –
  osnove_casovnih_vrst$spr_lag1

> osnove_casovnih_vrst[1:10,]
      kvartal    spr spr_lag1 spr_lag4 spr_lead2 spr_diff1
1  1950-01-01 1610.5       NA       NA    1723.0        NA
2  1950-04-01 1658.8   1610.5       NA    1753.9  48.30005
3  1950-07-01 1723.0   1658.8       NA    1773.5  64.19995
4  1950-10-01 1753.9   1723.0       NA    1803.7  30.90002
```

```
5  1951-01-01 1773.5   1753.9   1610.5    1839.8  19.59998
6  1951-04-01 1803.7   1773.5   1658.8    1843.3  30.19995
7  1951-07-01 1839.8   1803.7   1723.0    1864.7  36.10010
8  1951-10-01 1843.3   1839.8   1753.9    1866.2   3.50000
9  1952-01-01 1864.7   1843.3   1773.5    1878.0  21.39990
10 1952-04-01 1866.2   1864.7   1803.7    1940.2   1.50000
```

```
> osnove_casovnih_vrst[195:204,]
       kvartal    spr spr_lag1 spr_lag4 spr_lead2 spr_diff1
195 1998-07-01 8528.5   8442.9   8216.6    8733.5  85.59961
196 1998-10-01 8667.9   8528.5   8272.9    8771.2 139.40039
197 1999-01-01 8733.5   8667.9   8396.3    8871.5  65.59961
198 1999-04-01 8771.2   8733.5   8442.9    9049.9  37.70020
199 1999-07-01 8871.5   8771.2   8528.5    9102.5 100.29980
200 1999-10-01 9049.9   8871.5   8667.9    9229.4 178.40039
201 2000-01-01 9102.5   9049.9   8733.5    9260.1  52.59961
202 2000-04-01 9229.4   9102.5   8771.2    9303.9 126.90039
203 2000-07-01 9260.1   9229.4   8871.5        NA  30.69922
204 2000-10-01 9303.9   9260.1   9049.9        NA  43.80078
```

■