Мастер академске студије
Информациони инжењеринг

Статистика у информационом инжењерингу

Основне анализе података помоћу језика R

(материјали за предавања)

Садржај

1. **Основне статистичке анализе**
2. Дескриптивне статистике
3. Статистички тестови
4. Анализа варијансе
5. Регресиона анализа
6. Додатне напомене
7. Извори и литература

Основне статистичке анализе

- израчунавање дескриптивних статистика
- основно статистичко тестирање
- анализа варијансе
- регресиона анализа

Скупови података коришћени у примерима

- скупови података **`mostovi.svi`** и **`mostovi`**
  - скуп података *Pittsburgh Bridges*
    - аутори *Yoram Reich* и *Steven Fenves*
    - електронска локација (доступни скуп података и пратеће информације)
      - http://archive.ics.uci.edu/dataset/18/pittsburgh+bridges
      - https://doi.org/10.24432/C5RP5H
      - репозиторијум *UC Irvine Machine Learning Repository* на електронској локацији http://archive.ics.uci.edu/
        - скуп података дониран за репозиторијум 31. 7. 1990.
    - датотека *pittsburgh+bridges.zip* (преузето 13. 3. 2024)
      - електронска локација
        - http://archive.ics.uci.edu/static/public/18/pittsburgh+bridges.zip
      - подаци у датотекама *bridges.data.version1* и *bridges.data.version2*
    - лиценца *Creative Commons Attribution 4.0 International (CC BY 4.0)*
      - електронска локација
        - https://creativecommons.org/licenses/by/4.0/legalcode
  - скуп података *Pittsburgh Bridges* читан, обрађиван и анализиран језиком *R*
    - активности над скупом података и резултати представљени у наставку
    - датотеке визуализација генерисане за формат *TIFF* (15,5 x 15,5 cm, 300 *DPI*)

Основне статистичке анализе

Скупови података коришћени у примерима

- скуп података **`mostovi.svi`**
  - подаци о мостовима Питсбурга (САД)
  - 108 записа
  - 16 обележја укупно (нека обележја дата и у номиналној и у континуалној верзији)
    - намена, тип, распон, време изградње, дужина, број трака...
  - коришћен скуп података *Pittsburgh Bridges*
    - подаци из датотека *bridges.data.version1* и *bridges.data.version2* након учитавања су спојени у један скуп података, који је даље обрађиван и анализиран

Основне статистичке анализе

Скупови података коришћени у примерима

- скуп података **`mostovi`**
  - записи из скупа података **`mostovi.svi`** који немају недостајућих вредности
  - 70 записа
  - 16 обележја

Основне статистичке анализе

Скупови података коришћени у примерима

скупови података **mostovi.svi** и **mostovi** – припрема

```
mostovi.v1 <- read.csv("bridges.data.version1",
                       header=F, na.strings = "?",
                       stringsAsFactors=T)
mostovi.v2 <- read.csv("bridges.data.version2",
                       header=F, na.strings = "?",
                       stringsAsFactors=T)
zaglav.v1 <- c("IDENTIF", "RIVER", "LOCATION",
               "ERECTED_C", "PURPOSE", "LENGTH_C",
               "LANES_C", "CLEARG", "TORD", "MATERIAL",
               "SPAN", "RELL", "TYPE")
zaglav.v2 <- c("IDENTIF", "RIVER", "LOCATION",
               "ERECTED_N", "PURPOSE", "LENGTH_N",
               "LANES_N", "CLEARG", "TORD", "MATERIAL",
               "SPAN", "RELL", "TYPE")
colnames(mostovi.v1) <- zaglav.v1
colnames(mostovi.v2) <- zaglav.v2
mostovi.svi <- merge(mostovi.v1, mostovi.v2)
mostovi <- na.omit(mostovi.svi)
```

УЛАЗ

Основне статистичке анализе

Скупови података коришћени у примерима

- скупови података **djaci.mat** и **djaci.mat.GP**
  - скуп података *Student Performance*
    - аутор *Paulo Cortez*
    - електронска локација (доступни скуп података и пратеће информације)
      - http://archive.ics.uci.edu/dataset/320/student+performance
      - https://doi.org/10.24432/C5TG7T
      - репозиторијум *UC Irvine Machine Learning Repository* на електронској локацији http://archive.ics.uci.edu/
        - скуп података дониран за репозиторијум 26. 11. 2014.
    - датотека *student+performance.zip* (преузето 13. 3. 2024)
      - електронска локација
        - http://archive.ics.uci.edu/static/public/320/student+performance.zip
      - подаци у датотекама *student-mat.csv* и *student-por.csv* које из датотеке *student.zip*
    - лиценца *Creative Commons Attribution 4.0 International (CC BY 4.0)*
      - електронска локација
        - https://creativecommons.org/licenses/by/4.0/legalcode
  - скуп података *Student Performance* читан, обрађиван и анализиран језиком *R*
    - активности над скупом података и резултати представљени у наставку
    - датотеке визуализација генерисане за формат *TIFF* (15,5 x 15,5 cm, 300 *DPI*)

Основне статистичке анализе

Скупови података коришћени у примерима

- скупови података **`djaci.mat`** и **`djaci.mat.GP`**
  - додатне информације у вези с подацима из скупа података *Student Performance*
    - P. Cortez & A. Silva. Using data mining to predict secondary school student performance. In A. E. S. Carvalho Brito & J. Manuel Feliz-Teixeira Eds., Proceedings of 5th Future Business Technology Conference (FUBUTEC'2008) pp. 5-12, Porto, Portugal, April 9-11, 2008, EUROSIS-ETI, ISBN 978-9077381-39-7.

Основне статистичке анализе

Скупови података коришћени у примерима

- скуп података **`djaci.mat`**
  - подаци о оценама из математике за ђаке из две средње школе (Португал)
  - 395 записа
  - 33 обележја
    - ознака школе, лични подаци, личне навике, изостанци, резултати из математике...
  - коришћен скуп података *Student Performance*
    - подаци из датотеке *student-mat.csv* из датотеке *student.zip* након учитавања су даље обрађивани и анализирани

Основне статистичке анализе

Скупови података коришћени у примерима

- скуп података **`djaci.mat.GP`**
  - записи из скупа података **`djaci.mat`** који одговарају ђацима из школе *Gabriel Pereira* (*GP*)
  - 349 записа
  - 33 обележја

Основне статистичке анализе

Скупови података коришћени у примерима

скупови података **djaci.mat** и **djaci.mat.GP** – припрема

```
djaci.mat <- read.csv("student-mat.csv",
                      header=T, sep=";",
                      stringsAsFactors=T)

djaci.mat.GP <- djaci.mat[djaci.mat$school == "GP", ]
```

УЛАЗ

Основне статистичке анализе

Скупови података коришћени у примерима

- скуп података **`nekretnine`**
  - скуп података *Real Estate Valuation*
    - аутор *I-Cheng Yeh*
    - електронска локација (доступни скуп података и пратеће информације)
      - https://archive.ics.uci.edu/dataset/477/real+estate+valuation+data+set
      - https://doi.org/10.24432/C5J30W
      - репозиторијум *UC Irvine Machine Learning Repository* на електронској локацији http://archive.ics.uci.edu/
        - скуп података дониран за репозиторијум 17. 8. 2018.
    - датотека *real+estate+valuation+data+set.zip* (преузето 13. 3. 2024)
      - електронска локација
        - https://archive.ics.uci.edu/static/public/477/real+estate+valuation+data+set.zip
      - подаци у датотеци *Real estate valuation data set.xlsx*
    - лиценца *Creative Commons Attribution 4.0 International (CC BY 4.0)*
      - електронска локација
        - https://creativecommons.org/licenses/by/4.0/legalcode
  - скуп података *Real Estate Valuation* читан, обрађиван и анализиран језиком *R*
    - активности над скупом података и резултати представљени у наставку
    - датотеке визуализација генерисане за формат *TIFF* (15,5 x 15,5 cm, 300 *DPI*)

Основне статистичке анализе

Скупови података коришћени у примерима

- скуп података **`nekretnine`**
  - додатне информације у вези с подацима из скупа података *Real Estate Valuation*
    - I-C. Yeh & T.-K. Hsu (2018). Building real estate valuation models with comparative approach through case-based reasoning. Applied Soft Computing, 65, 260-271.

Основне статистичке анализе

Скупови података коришћени у примерима

- скуп података **`nekretnine`**
  - подаци о тржишту некретнина на Тајвану
  - 414 записа
  - 8 обележја
    - идентификатор, датум трансакције, старост некретнине, удаљеност од најближе метро станице, број продавница у близини, географска ширина, географска дужина и цена по јединици површине
  - коришћен скуп података *Real Estate Valuation*
    - подаци из датотеке *Real estate valuation data set.xlsx* након учитавања су даље обрађивани и анализирани

Основне статистичке анализе

Скупови података коришћени у примерима

скуп података **nekretnine** – припрема

```
library(readxl)  # potrebno instalirati paket

nekretnine <- read_excel(
  "Real estate valuation data set.xlsx", trim_ws=T)

names(nekretnine) <- c("id","date","age","distanceMRT",
                       "storesCount", "lat","long","priceUnit")
```

УЛАЗ

Садржај

1. Основне статистичке анализе
2. **Дескриптивне статистике**
3. Статистички тестови
4. Анализа варијансе
5. Регресиона анализа
6. Додатне напомене
7. Извори и литература

Дескриптивне статистике

Основне дескриптивне статистике

аритметичка средина (енгл. *mean*)
медијана (енгл. *median*)
модус (енгл. *mode*)
варијанса (енгл. *variance*)
стандардна девијација (енгл. *standard deviation*)
распон (енгл. *range*)
квантили (енгл. *quantiles*)
коефицијент асиметричности (енгл. *skewness*)
коефицијент спљоштености (енгл. *kurtosis*)

Основне дескриптивне статистике

- аритметичка средина
  - функција **`mean(...)`**
- медијана
  - функција **`median(...)`**
- модус
  - не постоји одговарајућа уграђена функција
    - може једноставно имплементирати
  - одређивање модуса
    - функција **`Mode(...)`** у додатном пакету **`pracma`**

Дескриптивне статистике

Основне дескриптивне статистике

- варијанса
  - функција **`var(...)`**
- стандардна девијација
  - функција **`sd(...)`**
- распон
  - функција **`range(...)`**
- квантили
  - функција **`quantile(...)`**
    - могућност задавања реда квантила преко аргумента **`probs`**

Дескриптивне статистике

Основне дескриптивне статистике

- коефицијент асиметричности
  - функција **`skewness(...)`** у додатном пакету **`moments`**
- коефицијент спљоштености
  - функција **`kurtosis(...)`** у додатном пакету **`moments`**

Дескриптивне статистике

Основне дескриптивне статистике – Пример 1

```
> mean(mostovi$LENGTH_C)
[1] 1597.657
> median(mostovi$LENGTH_C)
[1] 1325
> library(pracma)  # potrebno instalirati paket
> Mode(mostovi$LENGTH_C)
[1] 1000
>
```

КОНЗОЛА

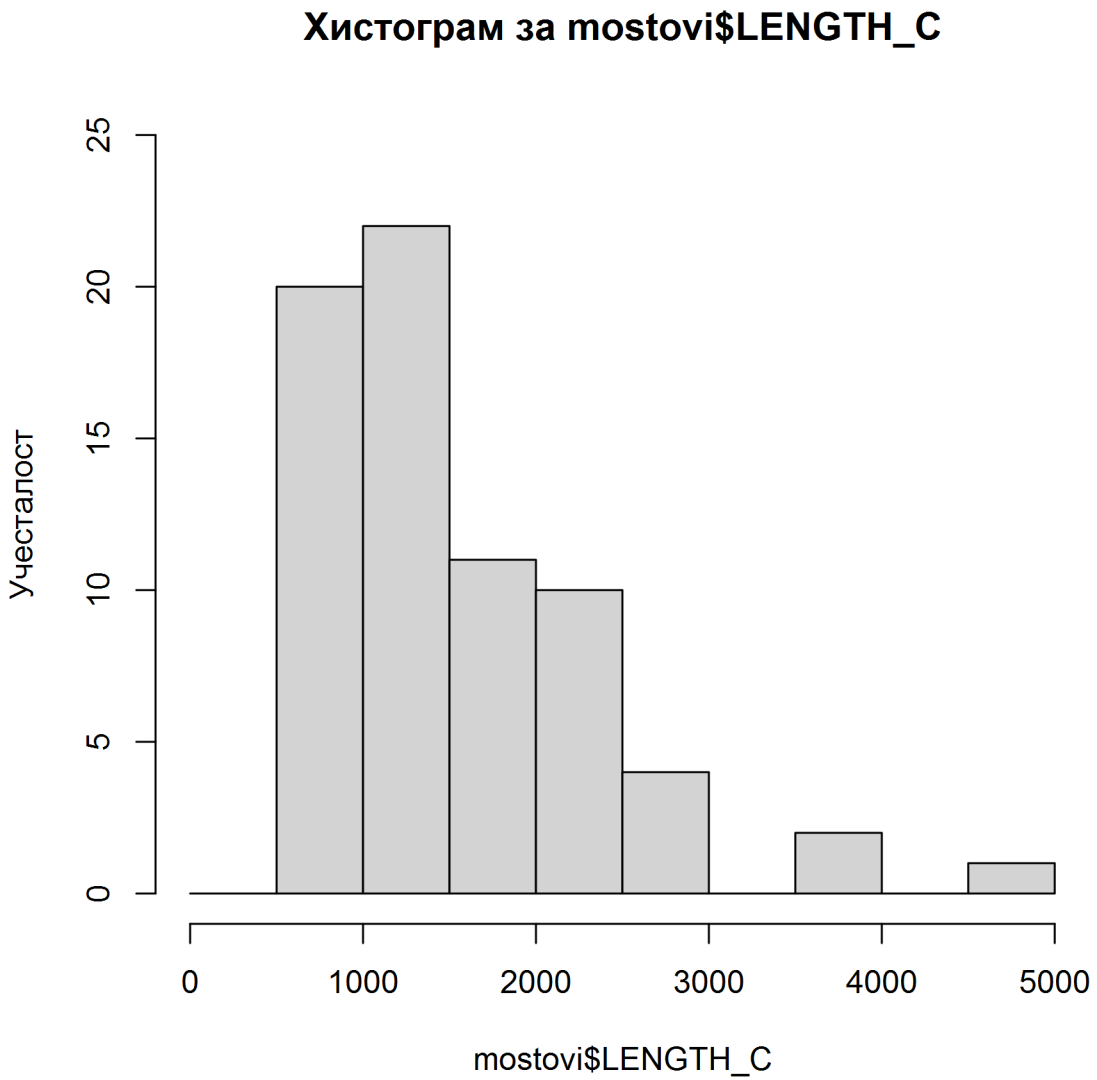
Дескриптивне статистике

Основне дескриптивне статистике – Пример 1

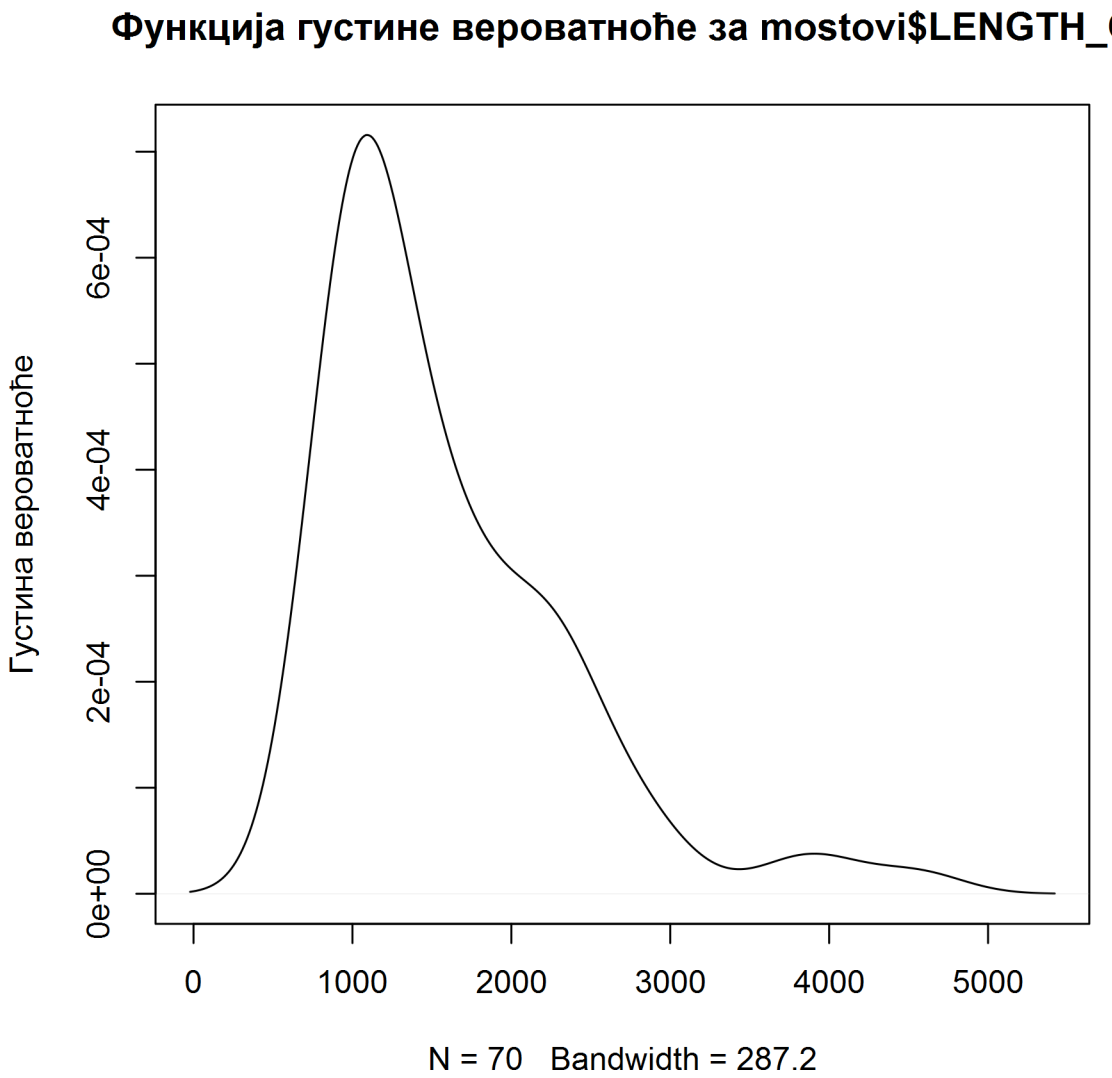
```
> var(mostovi$LENGTH_C)
[1] 608770.8
> sd(mostovi$LENGTH_C)
[1] 780.2377
> range(mostovi$LENGTH_C)
[1]  840 4558
> quantile(mostovi$LENGTH_C)
   0%   25%   50%   75%  100%
 840  1000  1325  2000  4558
> library(moments)  # potrebno instalirati paket
> skewness(mostovi$LENGTH_C)
[1] 1.633127
> kurtosis(mostovi$LENGTH_C)
[1] 5.849141
>
```

КОНЗОЛА

Основне дескриптивне статистике – Пример 1



Основне дескриптивне статистике – Пример 1



Садржај

1. Основне статистичке анализе
2. Дескриптивне статистике
3. **Статистички тестови**
4. Анализа варијансе
5. Регресиона анализа
6. Додатне напомене
7. Извори и литература

Статистички тестови

Основни статистички тестови

- испитивање нормалности расподеле
- поређење два независна узорка
- поређење два зависна узорка
- испитивање независности два обележја

Основни статистички тестови

- испитивање нормалности расподеле
  - тест Шапиро-Вилка
    - функција **`shapiro.test(...)`**
  - додатни тестови
    - у додатном пакету **`nortest`**

Основни статистички тестови

- поређење два независна узорка по нумеричком обележју
  - t-тест за два независна узорка
    - функција **`t.test(...)`**
  - U-тест Ман-Витнија / Вилкоксонов тест суме рангова
    - функција **`wilcox.test(...)`**
- поређење два зависна узорка по нумеричком обележју
  - t-тест за два зависна узорка
    - функција **`t.test(..., paired=T)`**
  - Вилкоксонов тест означених рангова (еквивалентних парова)
    - функција **`wilcox.test(..., paired=T)`**

Основни статистички тестови

- испитивање независности два категоријска обележја
  - χ^2 тест независности
    - функција **`chisq.test(...)`**
  - Фишеров тест тачне вероватноће
    - функција **`fisher.test(...)`**

Основни статистички тестови

- корекције при вишеструком поређењу
  - корекција p-вредности
    - функција **`p.adjust(...)`**

Статистички тестови

Испитивање нормалности расподеле – Пример 2

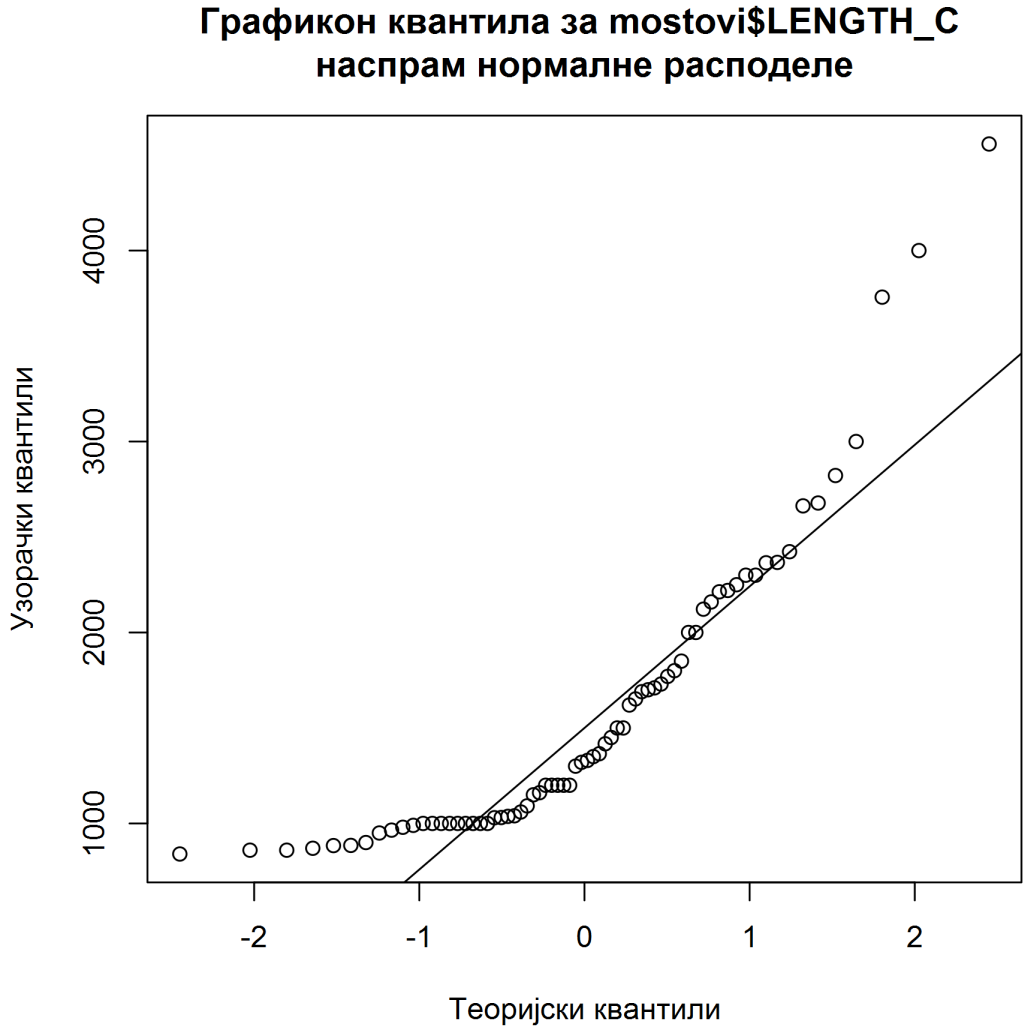
тест Шапиро-Вилка

```
> svt <- shapiro.test(mostovi$LENGTH_C)
> svt

   Shapiro-Wilk normality test

data:  mostovi$LENGTH_C
W = 0.82138, p-value = 9.267e-08

> svt$statistic
        W
0.8213753
> svt$p.value
[1] 9.267153e-08
>
```

КОНЗОЛА

Испитивање нормалности расподеле – Пример 2

графичка провера нормалности расподеле



Поређење два независна узорка – Пример 3А

t-тест с претпоставком једнаких варијанси (дисперзија)

```
> t.test(djaci.mat.GP[djaci.mat.GP$sex == "M", ]$G1,
djaci.mat.GP[djaci.mat.GP$sex == "F", ]$G1, var.equal = T)

    Two Sample t-test

data:  djaci.mat.GP[djaci.mat.GP$sex == "M", ]$G1 and
djaci.mat.GP[djaci.mat.GP$sex == "F", ]$G1
t = 2.1419, df = 347, p-value = 0.03289
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.06198053 1.45424832
sample estimates:
mean of x mean of y
 11.33735  10.57923

>
```

КОНЗОЛА

Поређење два независна узорка – Пример 3А

t-тест с Велчовом апроксимацијом (Велчов *t*-тест)

```
> t.test(djaci.mat.GP[djaci.mat.GP$sex == "M", ]$G1,
djaci.mat.GP[djaci.mat.GP$sex == "F", ]$G1)

    Welch Two Sample t-test

data:  djaci.mat.GP[djaci.mat.GP$sex == "M", ]$G1 and
djaci.mat.GP[djaci.mat.GP$sex == "F", ]$G1
t = 2.1324, df = 335.03, p-value = 0.0337
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.05876594 1.45746291
sample estimates:
mean of x mean of y
 11.33735  10.57923

>
```

КОНЗОЛА

Поређење два независна узорка – Пример 3А

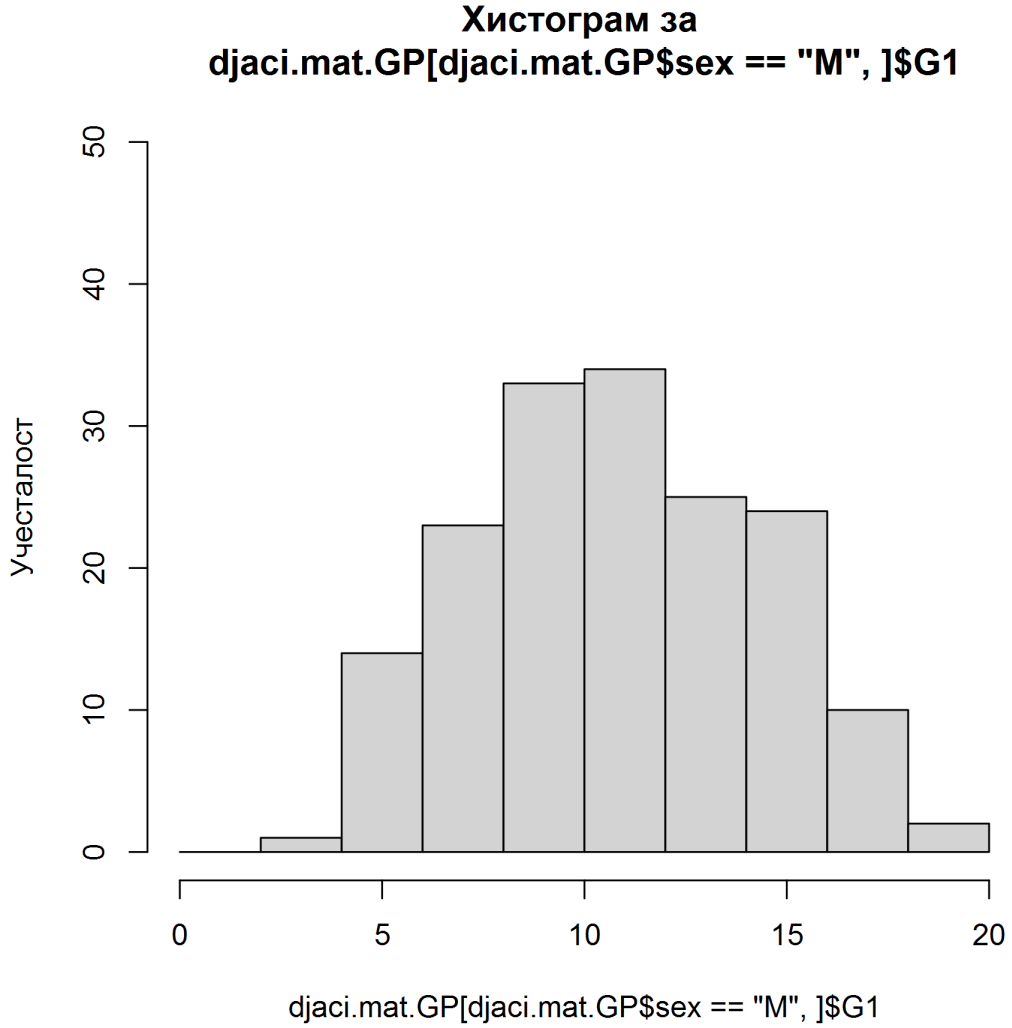
провера величине и варијансе за оба узорка

```
> nrow(djaci.mat.GP[djaci.mat.GP$sex == "M", ])
[1] 166
> nrow(djaci.mat.GP[djaci.mat.GP$sex == "F", ])
[1] 183
> var(djaci.mat.GP[djaci.mat.GP$sex == "M", ]$G1)
[1] 11.95823
> var(djaci.mat.GP[djaci.mat.GP$sex == "F", ]$G1)
[1] 9.948358
>
```

КОНЗОЛА

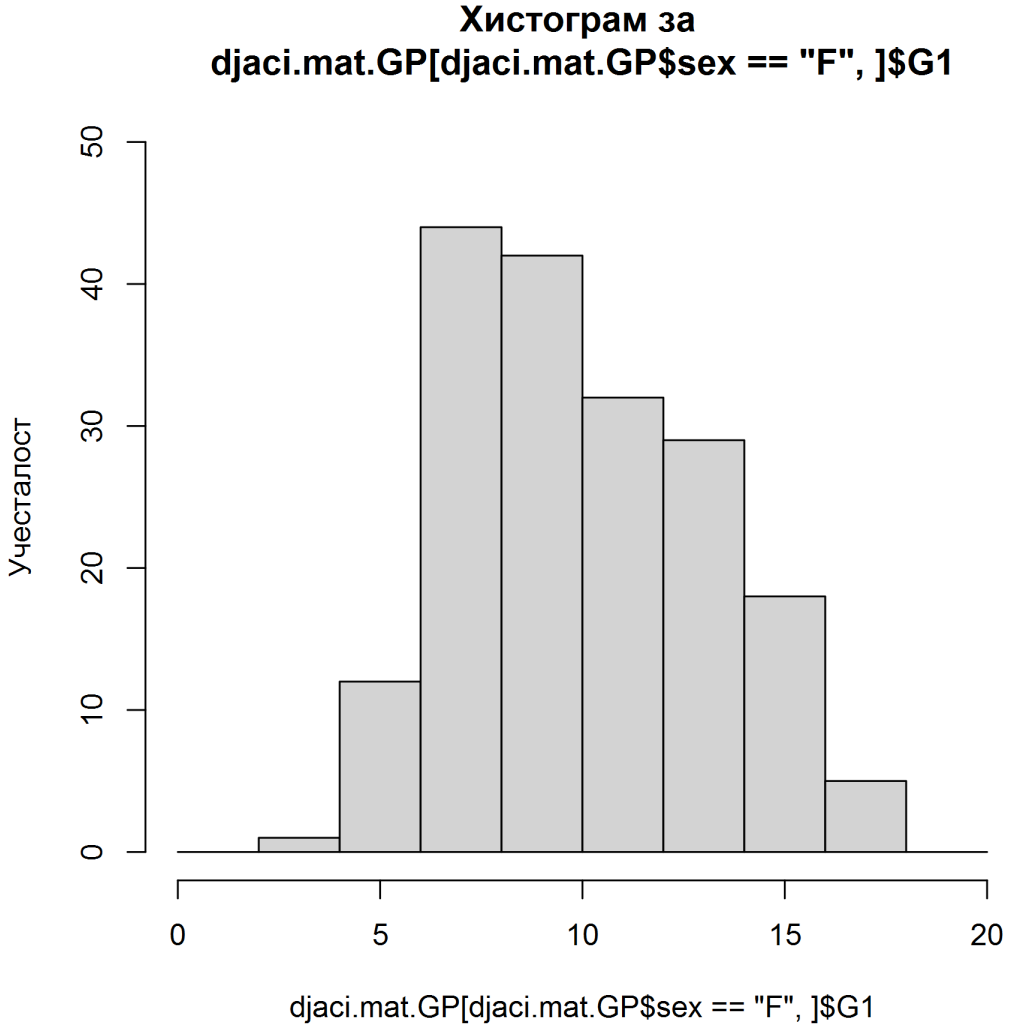
Поређење два независна узорка – Пример 3А

графичка провера расподеле за први узорак



Поређење два независна узорка – Пример 3А

графичка провера расподеле за други узорак



Поређење два независна узорка – Пример 3Б

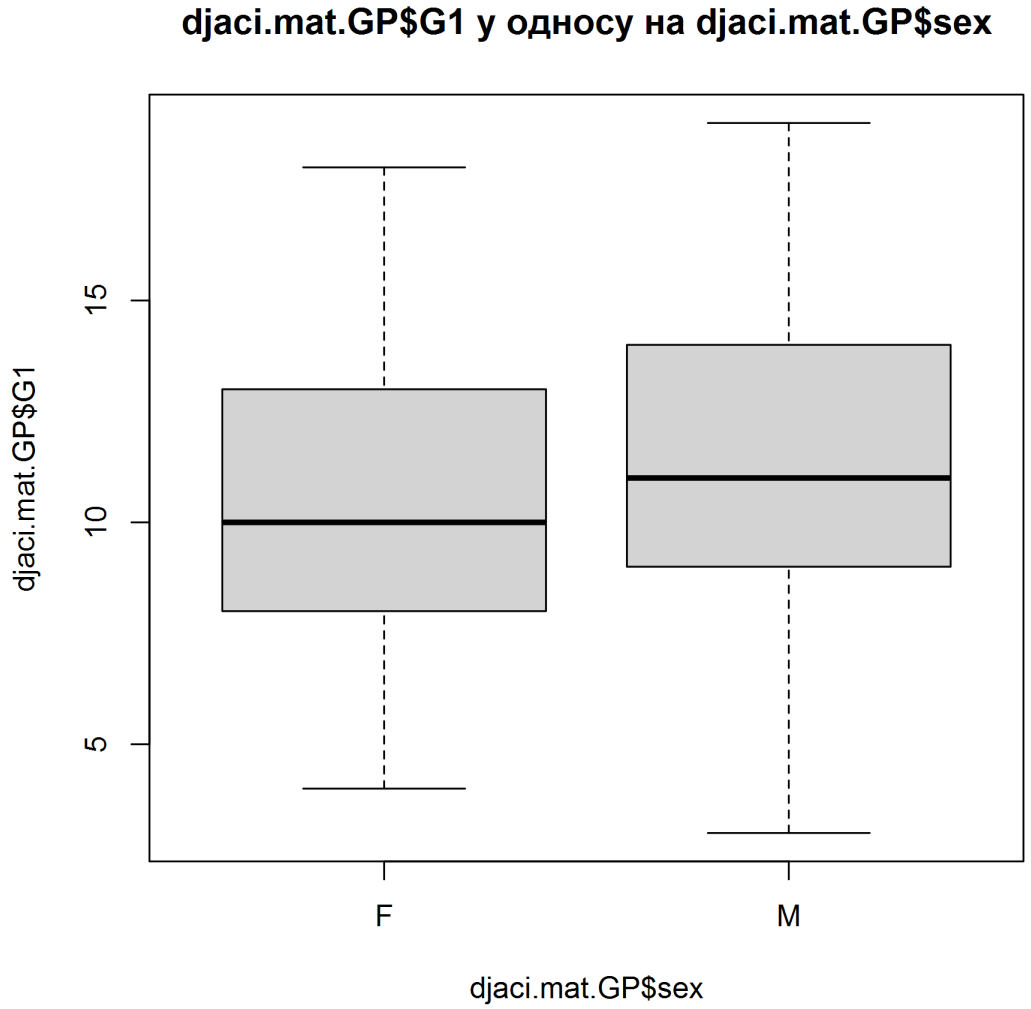
U-тест Ман-Витнија / Вилкоксонов тест суме рангова

```
> wilcox.test(djaci.mat.GP$G1 ~ djaci.mat.GP$sex)

    Wilcoxon rank sum test with continuity correction

data:  djaci.mat.GP$G1 by djaci.mat.GP$sex
W = 13191, p-value = 0.03314
alternative hypothesis: true location shift is not equal to 0

>
```

КОНЗОЛА

Поређење два независна узорка – Пример 3Б

графичка провера расподеле за оба узорка



Статистички тестови

Испитивање независности два обележја – Пример 4А

припрема података

```
> tabela.mostovi <- table(mostovi[, c("LENGTH_N", "MATERIAL")])
> tabela.mostovi
          MATERIAL
LENGTH_N IRON STEEL WOOD
  LONG      0    19    0
  MEDIUM    4    28    8
  SHORT     0    10    1
>
```

КОНЗОЛА

Испитивање независности два обележја – Пример 4А

χ^2 тест независности

```
> chst <- chisq.test(tabela.mostovi)
Warning message:
In chisq.test(tabela.mostovi) : Chi-squared approximation may be
incorrect
> chst

    Pearson's Chi-squared test

data:  tabela.mostovi
X-squared = 8.7193, df = 4, p-value = 0.06851

> chst$expected
         MATERIAL
LENGTH_N      IRON     STEEL     WOOD
  LONG   1.0857143 15.471429 2.442857
  MEDIUM 2.2857143 32.571429 5.142857
  SHORT  0.6285714  8.957143 1.414286
>
```

КОНЗОЛА

Статистички тестови

Испитивање независности два обележја – Пример 4Б

Фишеров тест тачне вероватноће

```
> fisher.test(tabela.mostovi)

    Fisher's Exact Test for Count Data

data:  tabela.mostovi
p-value = 0.07448
alternative hypothesis: two.sided

>
```

КОНЗОЛА

Испитивање независности два обележја – Пример 4B

преуређивање података

```
> duzina.nova <- mostovi$LENGTH_N
> levels(duzina.nova)
[1] "LONG"   "MEDIUM" "SHORT"
> levels(duzina.nova) <- c("LONG", "SHORT_OR_MEDIUM",
"SHORT_OR_MEDIUM")
> mostovi$LENGTH_N_T <- duzina.nova
> tabela.mostovi.nova <- table(mostovi[, c("LENGTH_N_T",
"MATERIAL")])
> tabela.mostovi.nova
                MATERIAL
LENGTH_N_T       IRON STEEL WOOD
  LONG              0    19    0
  SHORT_OR_MEDIUM   4    38    9
>
```

КОНЗОЛА

Статистички тестови

Испитивање независности два обележја – Пример 4B

Фишеров тест тачне вероватноће

```
> fisher.test(tabela.mostovi.nova)

    Fisher's Exact Test for Count Data

data:  tabela.mostovi.nova
p-value = 0.04314
alternative hypothesis: two.sided

>
```

КОНЗОЛА

Садржај

1. Основне статистичке анализе
2. Дескриптивне статистике
3. Статистички тестови
4. **Анализа варијансе**
5. Регресиона анализа
6. Додатне напомене
7. Извори и литература

Анализа варијансе

Основна анализа варијансе

- анализа варијансе (ANOVA)
  - функција **`aov(...)`**
- *post hoc* анализа помоћу Тукијевог теста
  - функција **`TukeyHSD(...)`**
- H-тест Краскал-Волиса
  - функција **`kruskal.test(...)`**

Анализа варијансе – Пример 5A

анализа варијансе – припрема података и спровођење анализе

```
> djaci.mat.GP$studytime_F <- as.factor(djaci.mat.GP$studytime)
> a <- aov(G1 ~ studytime_F, data=djaci.mat.GP)
>
```

КОНЗОЛА

Анализа варијансе – Пример 5A

анализа варијансе – приказ резултата

```
> a
Call:
   aov(formula = G1 ~ studytime_F, data = djaci.mat.GP)

Terms:
                studytime_F Residuals
Sum of Squares      106.400  3727.336
Deg. of Freedom           3       345

Residual standard error: 3.286924
Estimated effects may be unbalanced
> summary(a)
              Df Sum Sq Mean Sq F value Pr(>F)
studytime_F    3    106   35.47   3.283 0.0211 *
Residuals    345   3727   10.80
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

КОНЗОЛА

Анализа варијансе – Пример 5A

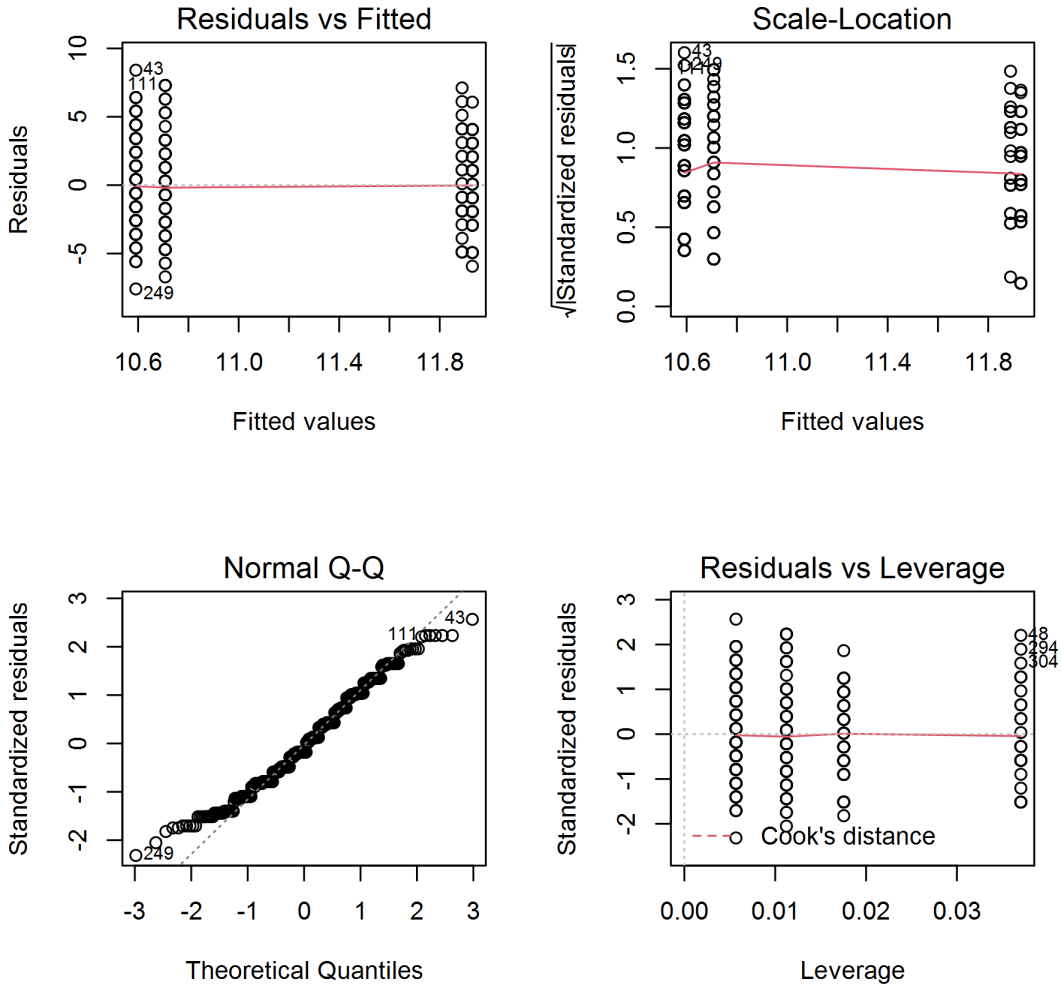
анализа варијансе – графичка провера резидуалâ

```
> layout(matrix(1:4, 2, 2))
> plot(a)
> layout(matrix(1))
>
```

КОНЗОЛА

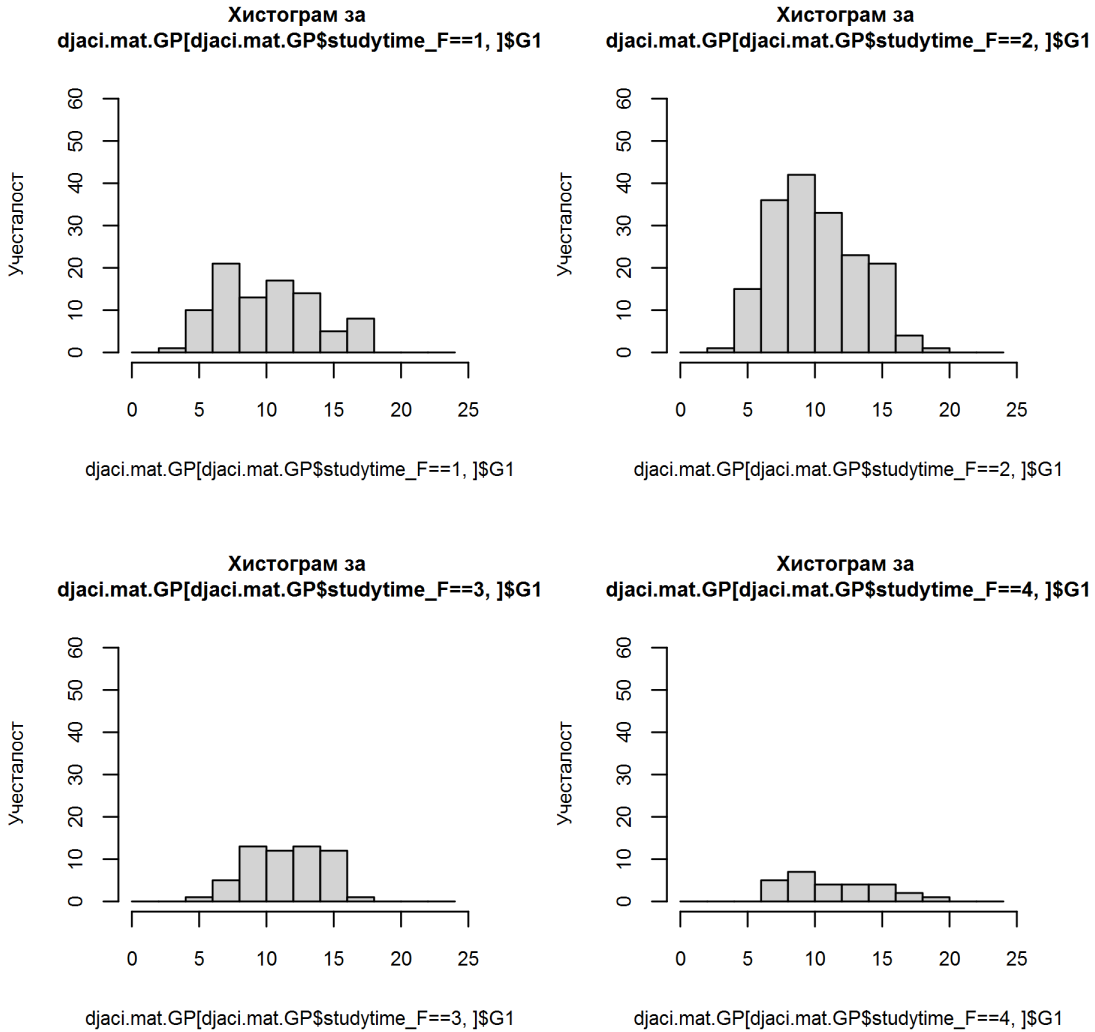
Анализа варијансе – Пример 5А

анализа варијансе – графичка провера резидуалâ



Анализа варијансе – Пример 5А

анализа варијансе – графичка провера расподеле по групама



Анализа варијансе – Пример 5Б

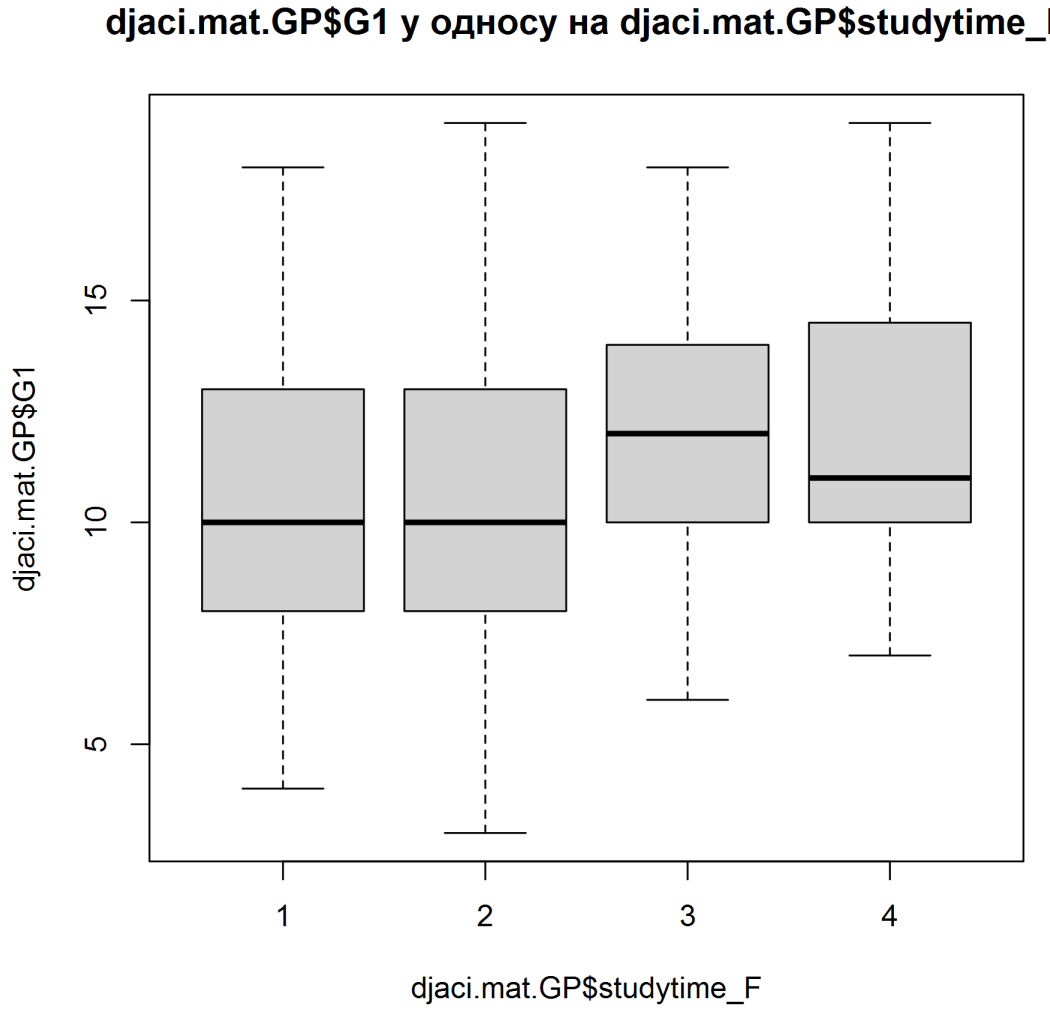
post hoc анализа помоћу Тукијевог теста

```
> TukeyHSD(a)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = G1 ~ studytime_F, data = djaci.mat.GP)

$studytime_F
            diff          lwr       upr     p adj
2-1 -0.11695608 -1.22065233 0.9867402 0.9928357
3-1  1.22195939 -0.21757473 2.6614935 0.1276389
4-1  1.18102372 -0.68333757 3.0453850 0.3601113
3-2  1.33891547  0.04572598 2.6321050 0.0392125
4-2  1.29797980 -0.45585132 3.0518109 0.2254491
4-3 -0.04093567 -2.02336768 1.9414963 0.9999455

>
```

КОНЗОЛА

Анализа варијансе – Пример 5Б

графичка провера расподеле по групама



Анализа варијансе – Пример 5B

H-тест Краскал-Волиса

```
> kruskal.test(G1 ~ studytime_F, data=djaci.mat.GP)

    Kruskal-Wallis rank sum test

data:  G1 by studytime_F
Kruskal-Wallis chi-squared = 10.037, df = 3, p-value = 0.01825

>
```

КОНЗОЛА

Садржај

1. Основне статистичке анализе
2. Дескриптивне статистике
3. Статистички тестови
4. Анализа варијансе
5. **Регресиона анализа**
6. Додатне напомене
7. Извори и литература

Регресиона анализа

Формуле у језику *R*

знак **~**

однос између зависне и независне променљиве

```
y ~ x
```

знак **+**

линеарна веза између променљивих

```
y ~ x1 + x2
```

знак **0**

уклањање слободног члана

```
y ~ 0 + x
```

знак **1**

експлицитно навођење слободног члана

```
y ~ 1 + x
```

Регресиона анализа

Формуле у језику *R*

- функција **`I()`**
  - очување уобичајене аритметичке интерпретације задатог израза

    ```
    y ~ x + I(x^2) + I(x^3)
    ```

- знак **`:`**
  - интеракција између променљивих

    ```
    y ~ x1:x2
    ```

- знак **`*`**
  - променљиве са интеракцијама између променљивих

    ```
    y ~ x1*x2
    ```

    исто као `y ~ x1 + x2 + I(x1 * x2)`

- знак **`^`**
  - променљиве са интеракцијама између променљивих, до одређеног степена

    ```
    y ~ (x1 + x2 + x3)^2
    ```

    исто као `y ~ (x1 + x2 + x3)*(x1 + x2 + x3)`

Регресиона анализа

Формуле у језику *R*

знак **-**

уклањање променљивих

```
y ~ a*b - a:b
```

знак **.**

означава све оне променљиве које нису већ експлицитно наведене

```
y ~ .
```

Регресиона анализа

Регресиони модели

- основни модели
  - функција **`lm(...)`**
- напредни модели
  - функција **`glm(...)`**

Регресиони модели – Пример 6A

формирање регресионог модела

```
1  modelA <- lm(priceUnit ~ age,
2               data=nekretnine)
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
```

УЛАЗ

Регресиони модели – Пример 6А

приказ појединости о регресионом моделу

```
1  summary(modelA)
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
```

УЛАЗ

Регресиони модели – Пример 6A

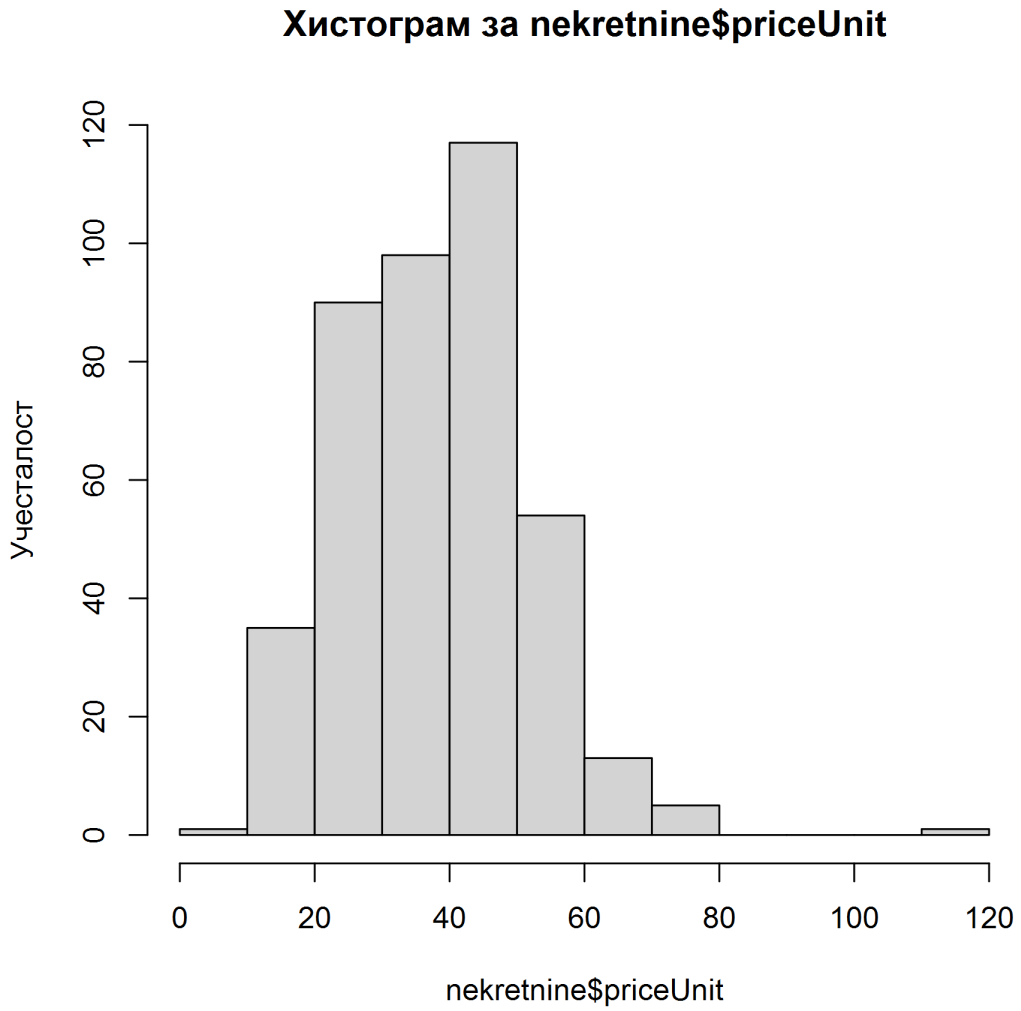
приказ појединости о регресионом моделу

```
 1
 2 Call:
 3 lm(formula = priceUnit ~ age, data = nekretnine)
 4
 5 Residuals:
 6     Min      1Q  Median      3Q     Max
 7 -31.113 -10.738   1.626   8.199  77.781
 8
 9 Coefficients:
10             Estimate Std. Error t value Pr(>|t|)
11 (Intercept) 42.43470    1.21098  35.042  < 2e-16 ***
12 age         -0.25149    0.05752  -4.372 1.56e-05 ***
13 ---
14 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
15
16 Residual standard error: 13.32 on 412 degrees of freedom
17 Multiple R-squared:  0.04434, Adjusted R-squared:  0.04202
18 F-statistic: 19.11 on 1 and 412 DF,  p-value: 1.56e-05
19
```

ИЗЛАЗ

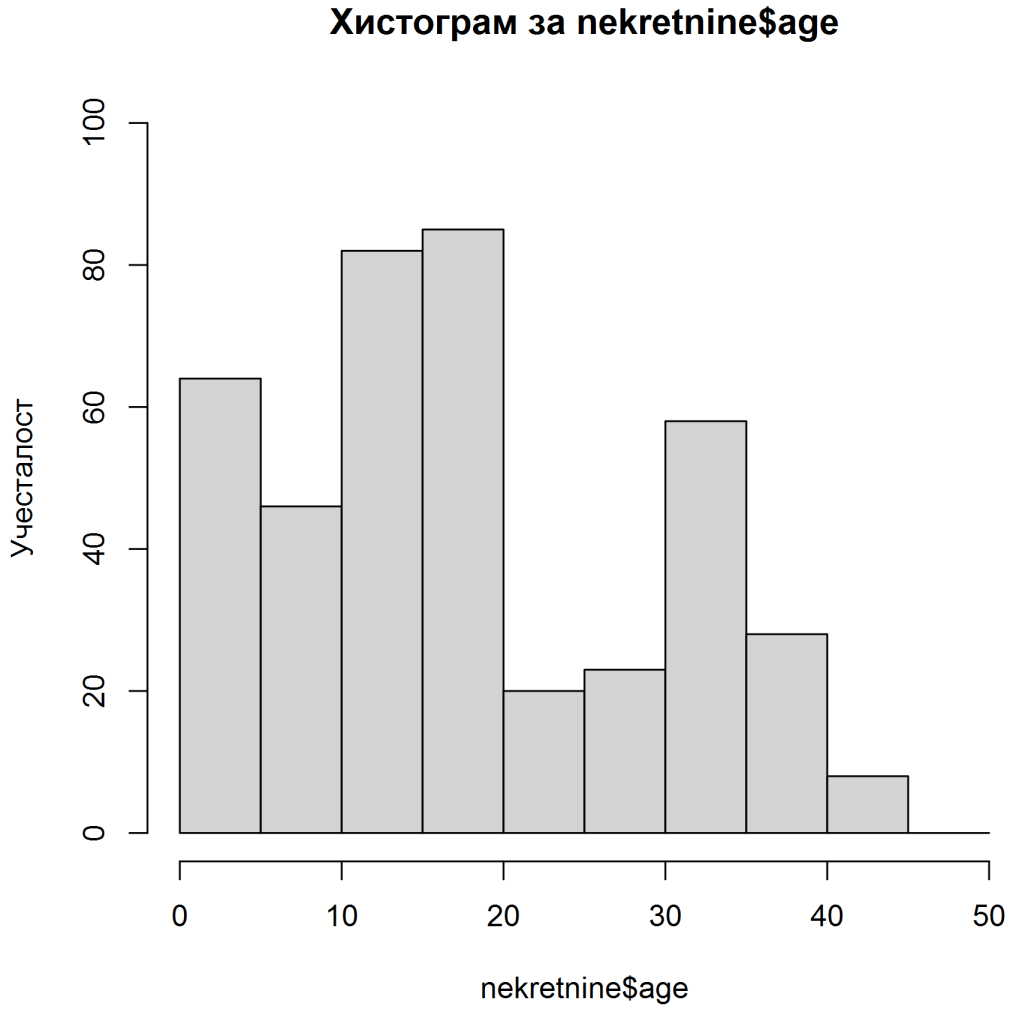
Регресиони модели – Пример 6А

графичка провера расподеле зависног обележја



Регресиони модели – Пример 6А

графичка провера расподеле независног обележја



Регресиони модели – Пример 6A

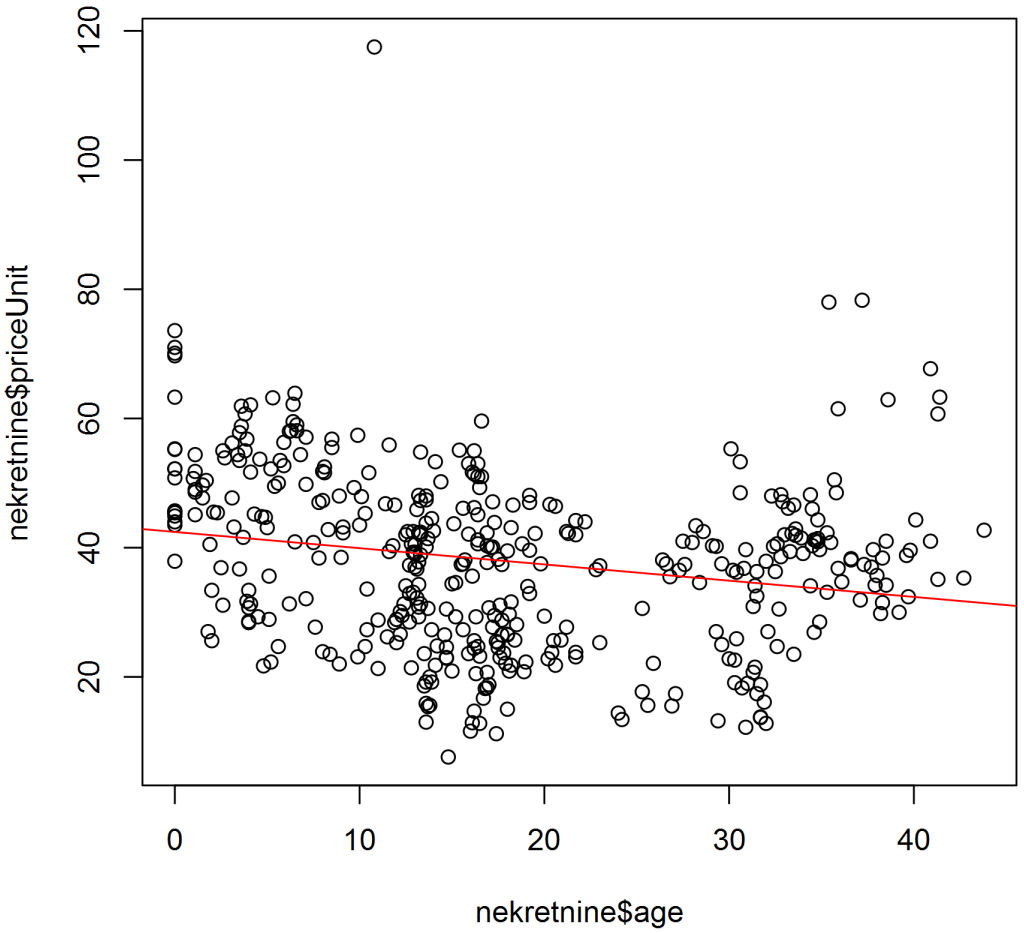
графички приказ података и регресионог модела

```
1  plot(nekretnine$age,
2       nekretnine$priceUnit)
3
4  abline(modelA,
5         col="red")
6
7
8
9
10
11
12
13
14
15
16
17
18
19
```

УЛАЗ

Регресиони модели – Пример 6A

графички приказ података и регресионог модела



Регресиони модели – Пример 6A

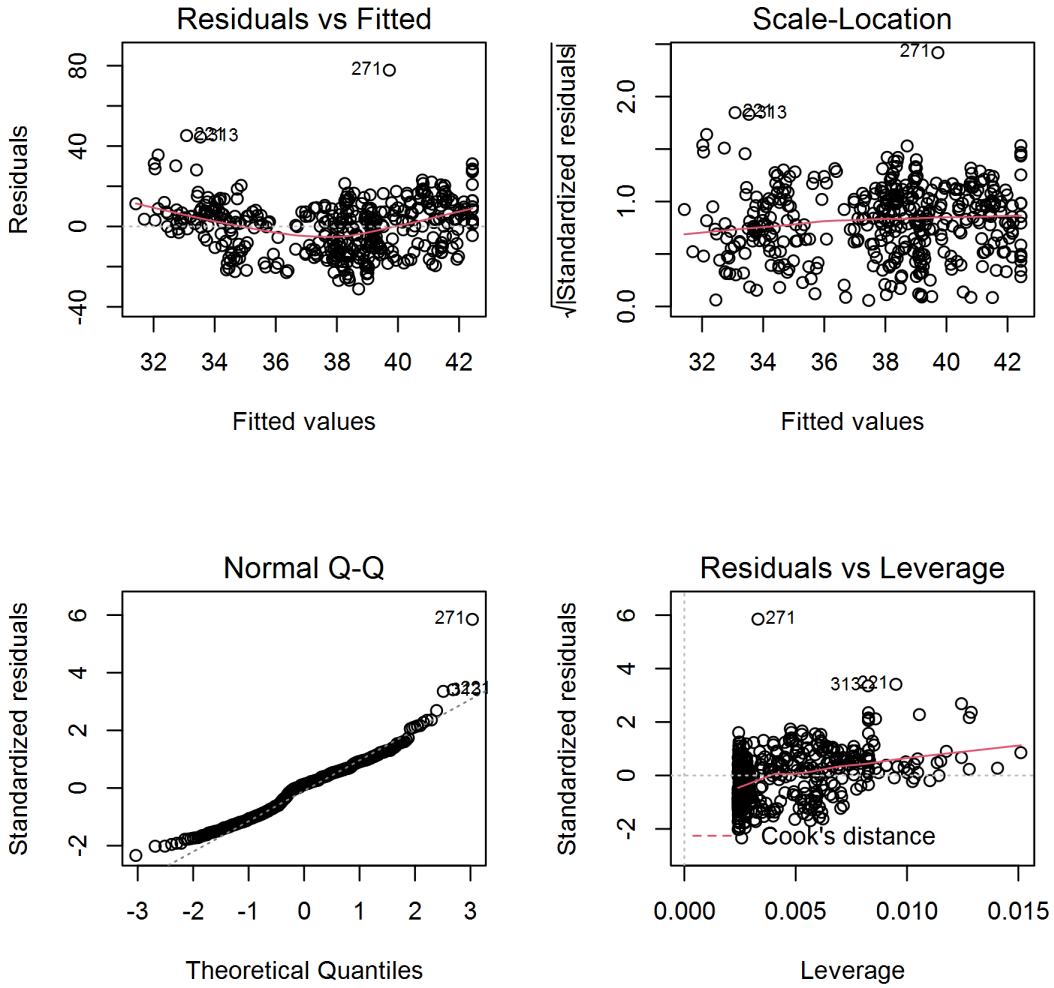
графичка провера резидуалâ

```
1  layout(matrix(1:4, 2, 2))
2
3  plot(modelA)
4
5  layout(matrix(1))
6
7
8
9
10
11
12
13
14
15
16
17
18
19
```

УЛАЗ

Регресиони модели – Пример 6A

графичка провера резидуалâ



Регресиони модели – Пример 6Б

формирање регресионог модела

```
1  modelB <- lm(priceUnit ~ poly(age, 2),
2               data=nekretnine)
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
```

УЛАЗ

Регресиони модели – Пример 6Б

приказ појединости о регресионом моделу

```
1  summary(modelB)
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
```

УЛАЗ

Регресиони модели – Пример 6Б

приказ појединости о регресионом моделу

```
 1
 2 Call:
 3 lm(formula = priceUnit ~ poly(age, 2), data = nekretnine)
 4
 5 Residuals:
 6     Min      1Q  Median      3Q     Max
 7 -26.542  -9.085  -0.445   8.260  79.961
 8
 9 Coefficients:
10               Estimate Std. Error t value Pr(>|t|)
11 (Intercept)     37.980      0.599  63.406  < 2e-16 ***
12 poly(age, 2)1  -58.225     12.188  -4.777 2.48e-06 ***
13 poly(age, 2)2  109.635     12.188   8.995  < 2e-16 ***
14 ---
15 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
16
17 Residual standard error: 12.19 on 411 degrees of freedom
18 Multiple R-squared:  0.2015, Adjusted R-squared:  0.1977
19 F-statistic: 51.87 on 2 and 411 DF,  p-value: < 2.2e-16
```

ИЗЛАЗ

Регресиони модели – Пример 6Б

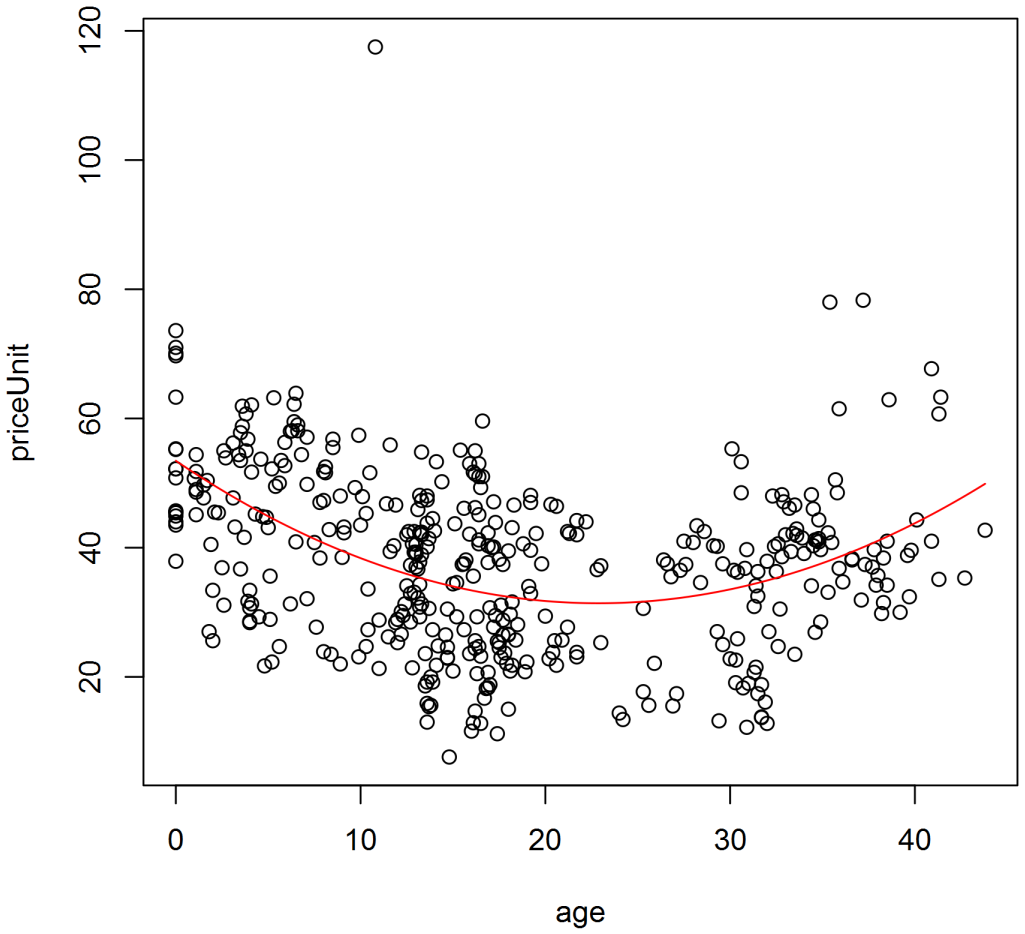
графички приказ података и регресионог модела

```
1  nekretnine.predv <- data.frame(
2    age=seq(min(nekretnine$age), max(nekretnine$age), 0.01))
3
4  nekretnine.predv$priceUnit <- predict(modelB, nekretnine.predv)
5
6  plot(priceUnit ~ age,
7       data=nekretnine)
8
9  lines(nekretnine.predv$age,
10       nekretnine.predv$priceUnit,
11       col="red")
12
13
14
15
16
17
18
19
```

УЛАЗ

Регресиони модели – Пример 6Б

графички приказ података и регресионог модела



Регресиони модели – Пример 6Б

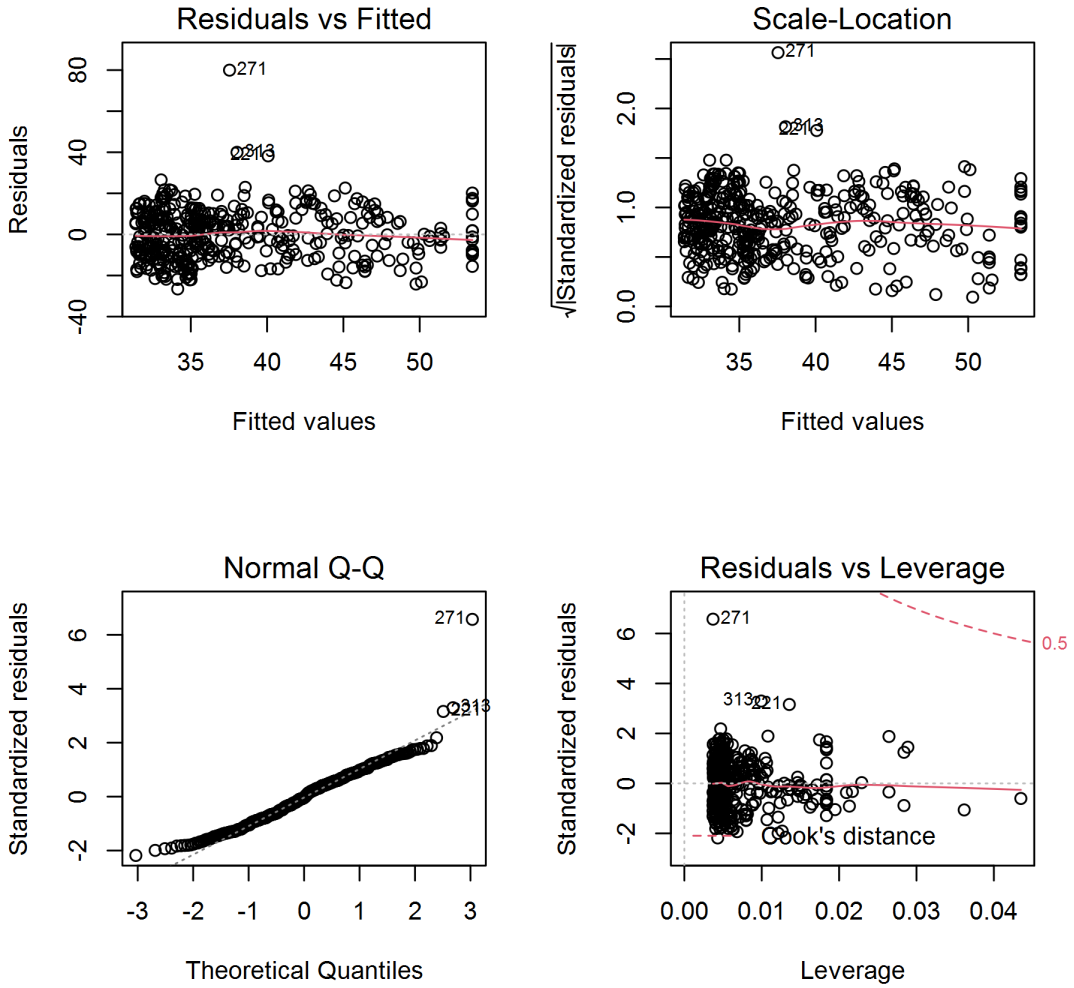
графичка провера резидуалâ

```
1  layout(matrix(1:4, 2, 2))
2
3  plot(modelB)
4
5  layout(matrix(1))
6
7
8
9
10
11
12
13
14
15
16
17
18
19
```

УЛАЗ

Регресиони модели – Пример 6Б

графичка провера резидуалâ



Садржај

1. Основне статистичке анализе
2. Дескриптивне статистике
3. Статистички тестови
4. Анализа варијансе
5. Регресиона анализа
6. **Додатне напомене**
7. Извори и литература

Додатне напомене

Претпоставке у статистичким процедурама

- статичке процедуре углавном подразумевају задовољење одређених претпоставки
  - нарушавање претпоставки може довести у питање остварени закључак
  - понекад нарушавање неке претпоставке не мора суштински утицати на остварени закључак
    - зависно од конкретне статистичке процедуре и њене конкретне претпоставке која је нарушена
      - зависно од степена нарушавања претпоставке
  - поједине статистичке процедуре су робустне
    - релативно „отпорне” на одступања од претпоставки

Додатне напомене

Претпоставке у статистичким процедурама

- приликом примене неке одабране статистичке процедуре потребно испитати задовољење њених претпоставки
  - на основу степена задовољења претпоставки треба одлучити да ли је одабрана статистичка процедура прикладна за примену у конкретном случају
- може се десити да одабрана статистичка процедура није прикладна за примену у конкретном случају
  - може се покушати с трансформацијом података како би нарушена претпоставка постала задовољена и одабрана статистичка процедура постала прикладна
  - може се покушати с неком другом статистичком процедуром

Поређење два независна узорка

t-тест за два независна узорка

основне претпоставке

- независност података
- нормалност расподела
- једнакост варијанси

одступања од претпоставки

- *t*-тест је релативно робустан
- одступање од претпоставки је већи проблем у случају мањих узорака
  - препорука да буде преко 15 појава у свакој групи
- одступање расподеле од нормалне је већи проблем него разлике у варијанси
- када постоје умерене разлике у варијанси а величине узорака су приближне обично се и даље користи *t*-тест
- када се поред умерене разлике у варијанси и величине узорака битно разликују прикладнији је Велчов *t*-тест (*t*-тест неједнаких варијанси)
- када постоје значајне разлике у варијанси користи се модификовани *t*-тест (нпр. Велчов *t*-тест) или одговарајући непараметарски тест (али не *U*-тест Ман-Витнија)

Испитивање независности два обележја

- χ^2 тест независности
  - основне претпоставке
    - очекиване учесталости да нису премале
      - очекиване учесталости за сваку ћелију табеле контингенције најмање 1
      - 80% ћелија табеле контингенције с очекиваним учесталостима преко 5
  - одступања од претпоставки
    - може покушати смањење димензија табеле контингенције кроз спајање ћелија међу којима су и проблематичне ћелије
    - може применити Фишеров тест тачне вероватноће

Додатне напомене

Анализа варијансе

- једнофакторска анализа варијансе
  - основне претпоставке
    - независност података
    - расподела на нивоу сваке групе је нормална
    - варијанса је једнака за све групе
  - одступања од претпоставки
    - када постоје мања одступања од претпоставки обично се и даље користи једнофакторска анализа варијансе
    - знатне разлике у варијанси не морају бити проблем ако су групе приближних величина
    - у случају нарушења претпоставки може се покушати с одговарајућом непараметарском анализом
      - примена H-теста Краскал-Волиса

Додатне напомене

Регресиона анализа

прости линеарни регресиони модели

основне претпоставке

- независно и зависно обележје су у линеарној вези
- вредности независног обележја су тачно измерене
- вредности зависног обележја су међусобно статистички независне
- варијанса зависног обележја је иста за све вредности независног обележја
- расподела зависног обележја је приближно нормална за све вредности независног обележја

Садржај

1. Основне статистичке анализе
2. Дескриптивне статистике
3. Статистички тестови
4. Анализа варијансе
5. Регресиона анализа
6. Додатне напомене
7. **Извори и литература**

Извори и литература

Основни извори и литература

- R Project. R: A language and environment for statistical computing – Reference index – The R Core Team – Version 4.5.1 (2025-06-13). Internet: https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf
- Adler J. R in a nutshell: A desktop quick reference. 2nd edition. O'Reilly; 2012.
- Hoffman JIE. Biostatistics for medical and biomedical practitioners. Academic Press; 2015.
- Riffenburgh RH. Statistics in medicine. 3rd edition. Academic Press; 2012.

Извори и литература

Основни извори података

- скупови података **`mostovi.svi`** и **`mostovi`**
  - скуп података *Pittsburgh Bridges*
    - аутори *Yoram Reich* и *Steven Fenves*
    - електронска локација (доступни скуп података и пратеће информације)
      - http://archive.ics.uci.edu/dataset/18/pittsburgh+bridges
      - https://doi.org/10.24432/C5RP5H
      - репозиторијум *UC Irvine Machine Learning Repository* на електронској локацији http://archive.ics.uci.edu/
        - скуп података дониран за репозиторијум 31. 7. 1990.
    - датотека *pittsburgh+bridges.zip* (преузето 13. 3. 2024)
      - електронска локација
        - http://archive.ics.uci.edu/static/public/18/pittsburgh+bridges.zip
      - подаци у датотекама *bridges.data.version1* и *bridges.data.version2*
    - лиценца *Creative Commons Attribution 4.0 International (CC BY 4.0)*
      - електронска локација
        - https://creativecommons.org/licenses/by/4.0/legalcode
  - скуп података *Pittsburgh Bridges* читан, обрађиван и анализиран језиком *R*

Основни извори података

- скупови података **`djaci.mat`** и **`djaci.mat.GP`**
  - скуп података *Student Performance*
    - аутор *Paulo Cortez*
    - електронска локација (доступни скуп података и пратеће информације)
      - http://archive.ics.uci.edu/dataset/320/student+performance
      - https://doi.org/10.24432/C5TG7T
      - репозиторијум *UC Irvine Machine Learning Repository* на електронској локацији http://archive.ics.uci.edu/
        - скуп података дониран за репозиторијум 26. 11. 2014.
    - датотека *student+performance.zip* (преузето 13. 3. 2024)
      - електронска локација
        - http://archive.ics.uci.edu/static/public/320/student+performance.zip
      - подаци у датотекама *student-mat.csv* и *student-por.csv* које из датотеке *student.zip*
    - лиценца *Creative Commons Attribution 4.0 International (CC BY 4.0)*
      - електронска локација
        - https://creativecommons.org/licenses/by/4.0/legalcode
  - скуп података *Student Performance* читан, обрађиван и анализиран језиком *R*

Основни извори података

- скуп података **`nekretnine`**
  - скуп података *Real Estate Valuation*
    - аутор *I-Cheng Yeh*
    - електронска локација (доступни скуп података и пратеће информације)
      - https://archive.ics.uci.edu/dataset/477/real+estate+valuation+data+set
      - https://doi.org/10.24432/C5J30W
      - репозиторијум *UC Irvine Machine Learning Repository* на електронској локацији http://archive.ics.uci.edu/
        - скуп података дониран за репозиторијум 17. 8. 2018.
    - датотека *real+estate+valuation+data+set.zip* (преузето 13. 3. 2024)
      - електронска локација
        - https://archive.ics.uci.edu/static/public/477/real+estate+valuation+data+set.zip
      - подаци у датотеци *Real estate valuation data set.xlsx*
    - лиценца *Creative Commons Attribution 4.0 International (CC BY 4.0)*
      - електронска локација
        - https://creativecommons.org/licenses/by/4.0/legalcode
  - скуп података *Real Estate Valuation* читан, обрађиван и анализиран језиком *R*

Мастер академске студије
Информациони инжењеринг

Статистика у информационом инжењерингу

Основне анализе података помоћу језика R

(материјали за предавања)