

SAP - Uspjeh učenika u nastavi

Petar Dragojević, Magda Radić, Tomislav Prhat, Matej Ciglencečki

Deskriptivna analiza

Osnove

Učitavamo podatke, provjeravamo kojeg je oblika skup podataka i od kojih se stupaca sastoji.

```
students_org <- readxl::read_excel("student_data.xlsx")
```

```
# 370 rows, 39 columns  
dim(students_org)
```

```
# Show column names  
names(students_org)
```

Provjeravamo prvih par redataka podatkovnog skupa

```
# Show first few rows  
head(students_org)
```

Saznajemo osnovne podatke za svaki stupac

```
# Show details for each column  
summary(students_org)
```

Provjeravamo koji su stupci kojeg tipa: numerički, kategorički...

```
# Check the class of the column. 'numeric', 'character'...  
sapply(students_org, class)
```

Provjeravamo postoje li nevažeći podaci koji prelaze maksimalne vrijednosti specificirane u uputama o podacima. Sve vrijednosti su dozvoljenom intervalu.

```
colMax <- students_org %>%  
  select(where(is.numeric)) %>%  
  sapply(., max, na.rm = TRUE)  
colMax  
# Every column has normal maximum value
```

Vrijednosti svih atributa padaju u dozvoljen interval

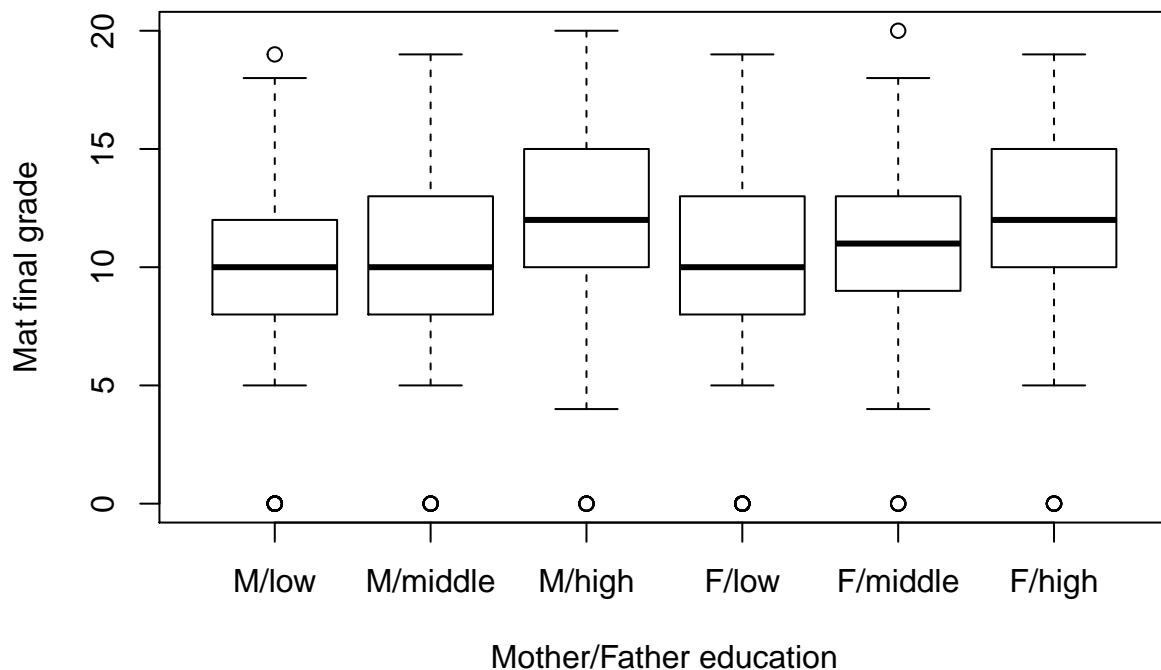
Izbacivanje svih NaN/NA/null vrijednosti iz podatkovnog skupa. Na sreću, takvih vrijednosti nije bilo.

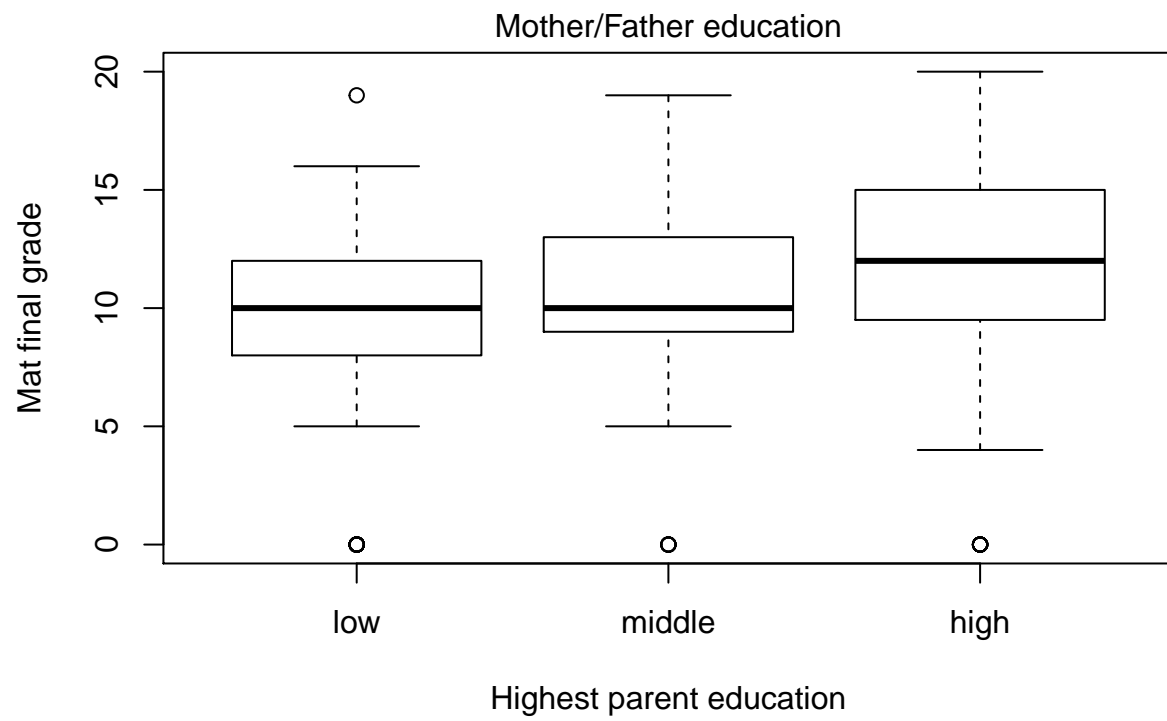
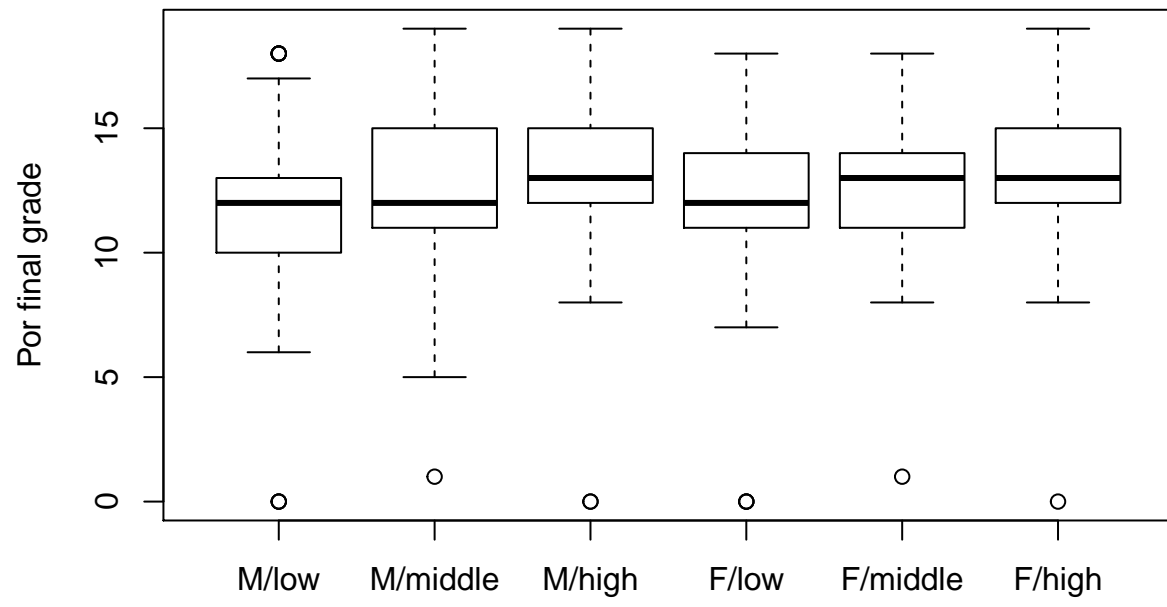
```
# Are there any na values?  
students_org %>%  
  filter(is.na(.))  
sum(apply(students_org, 2, is.nan))  
students_org %>%  
  filter(is.null(.)) %>%  
  summarise(n = n())
```

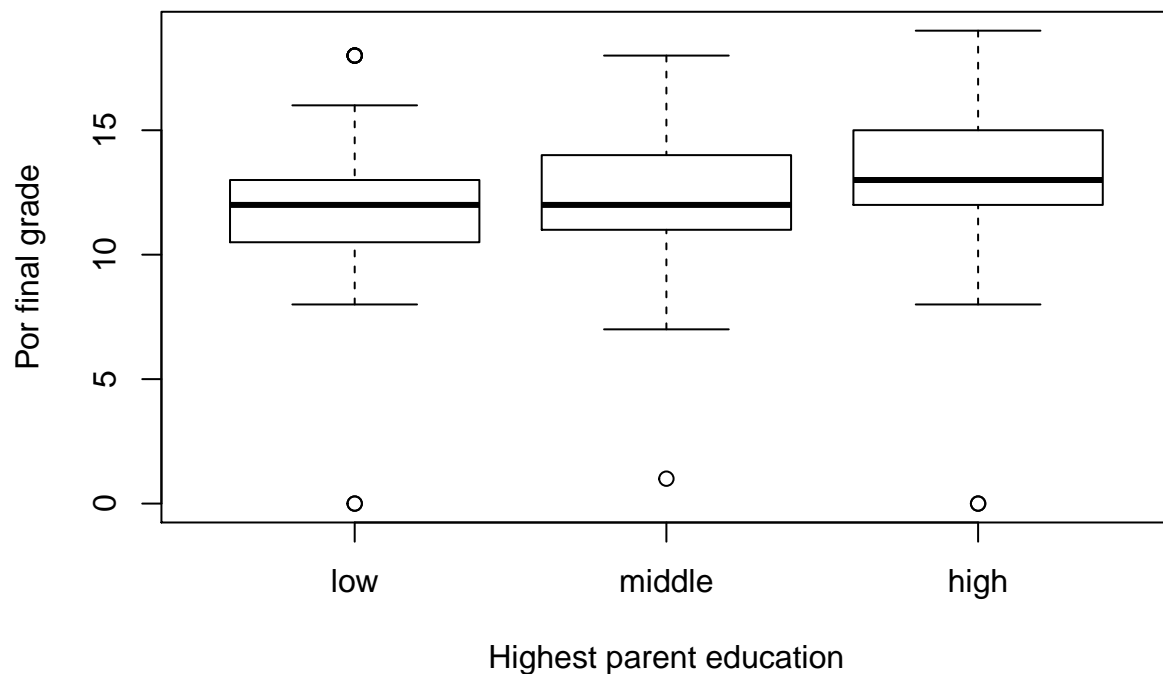
```
# Drop these values just in case they show up with another dataset We will
# continue using 'student' variable
students <- students_org %>%
  filter_all(all_vars(!is.na(.) & !is.nan(.) & !is.null(.)))
students_clean <- students
```

Zavisnost između edukacije roditelja i uspješnosti

```
# Mat_grade i Por_grade prebacivanje ocjena u engleski način ocjenjivanja
students <- students %>%
  mutate(Mat_grade = case_when(G3_mat < 10 ~ "F", G3_mat >= 10 & G3_mat < 14 ~
    "C", G3_mat >= 14 & G3_mat < 16 ~ "B", G3_mat >= 16 ~ "A"))
students <- students %>%
  mutate(Por_grade = case_when(G3_por < 10 ~ "F", G3_por >= 10 & G3_mat < 14 ~
    "C", G3_por >= 14 & G3_mat < 16 ~ "B", G3_por >= 16 ~ "A"))
# MeduMod i FeduMod grupiranje edukacije roditelja u veće podgrupe
students <- students %>%
  mutate(MeduMod = case_when(Medu == "0" | Medu == "1" | Medu == "2" ~ "0", Medu ==
    "3" ~ "1", Medu == "4" ~ "2"))
students <- students %>%
  mutate(FeduMod = case_when(Fedu == "0" | Medu == "1" | Fedu == "2" ~ "0", Fedu ==
    "3" ~ "1", Fedu == "4" ~ "2"))
# za edukaciju roditelja uzimamo onu koja je veća
students$highestparentedu <- pmax(students$MeduMod, students$FeduMod)
```







H0: Ocjena iz matematike i edukacija više educiranog roditelja su nezavisna obilježja H1: Ocjena iz matematike i edukacija više educiranog roditelja su zavisna obilježja

```
tbl = table(students$highestparentedu, students$Mat_grade)
added_margins_tbl = addmargins(tbl)
print(added_margins_tbl)
```

```
##
##      A  B  C  F Sum
##  0    4  9 48 39 100
##  1    7  8 37 25  77
##  2   25 35 54 38 152
## Sum  36 52 139 102 329
```

```
chisq.test(tbl, correct = F)
```

```
##
## Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 25.134, df = 6, p-value = 0.0003224
```

p-value testa iznosi manje od 0.05 stoga odbacujemo hipotezu H0, te zaključujemo da su edukacija više educiranog roditelja i završna ocjena iz matematike zavisne.

H0: Ocjena iz matematike i edukacija majke su nezavisna obilježja H1: Ocjena iz matematike i edukacija majke su zavisna obilježja

```
tbl = table(students$MeduMod, students$Mat_grade)
added_margins_tbl = addmargins(tbl)
print(added_margins_tbl)
```

```
##
##      A  B  C  F Sum
##  0    7 17 68 56 148
```

```
##    1    11  12  37  33  93
##    2    22  30  45  32 129
##   Sum   40  59 150 121 370
```

```
chisq.test(tbl, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 22.482, df = 6, p-value = 0.0009898
```

p-value Testa nezavisnosti iznosi manje od 0.05 stoga odbacujemo hipotezu H0, te zaključujemo da su edukacija majke i završna ocjena iz matematike zavisne.

H0: Ocjena iz matematike i edukacija oca su nezavisna obilježja H1: Ocjena iz matematike i edukacija oca su zavisna obilježja

```
tbl2 = table(students$FeduMod, students$Mat_grade)
added_margins_tbl2 = addmargins(tbl2)
print(added_margins_tbl2)
```

```
##
##          A    B    C    F Sum
##    0      11   19   64   49 143
##    1      11   12   40   30  93
##    2      14   21   35   23  93
##   Sum     36   52  139  102 329
```

```
chisq.test(tbl2, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl2
## X-squared = 9.0666, df = 6, p-value = 0.1699
```

p-value iznosi više od 0.05 pa ne možemo odbaciti hipotezu H0.

H0: Ocjena iz portugala i edukacija više educiranog roditelja su nezavisna obilježja H1: Ocjena iz portugala i edukacija više educiranog roditelja su zavisna obilježja

```
tbl = table(students$highestparentedu, students$Por_grade)
added_margins_tbl = addmargins(tbl)
print(added_margins_tbl)
```

```
##
##          A    B    C    F Sum
##    0       3    3   77   10  93
##    1       3    6   57    5  71
##    2      12   23   85    7 127
##   Sum     18   32  219   22 291
```

```
chisq.test(tbl, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 19.409, df = 6, p-value = 0.003526
```

```
# očekivane frekvencije svih razreda moraju biti veće ili jednake 5
for (col_names in colnames(added_margins_tbl)) {
  for (row_names in rownames(added_margins_tbl)) {
    if (!(row_names == "Sum" | col_names == "Sum")) {
      cat("Očekivane frekvencije za razred ", col_names, "-", row_names, ": ",
          (added_margins_tbl[row_names, "Sum"] * added_margins_tbl["Sum", col_names])/added_margins_tbl["Sum", "Sum"], "\n")
    }
  }
}
```

```
## Očekivane frekvencije za razred A - 0 : 5.752577
## Očekivane frekvencije za razred A - 1 : 4.391753
## Očekivane frekvencije za razred A - 2 : 7.85567
## Očekivane frekvencije za razred B - 0 : 10.2268
## Očekivane frekvencije za razred B - 1 : 7.80756
## Očekivane frekvencije za razred B - 2 : 13.96564
## Očekivane frekvencije za razred C - 0 : 69.98969
## Očekivane frekvencije za razred C - 1 : 53.43299
## Očekivane frekvencije za razred C - 2 : 95.57732
## Očekivane frekvencije za razred F - 0 : 7.030928
## Očekivane frekvencije za razred F - 1 : 5.367698
## Očekivane frekvencije za razred F - 2 : 9.601375
```

```
# Vidimo da postoje očekivane frekvencije manje od 5 pa koristimo fisher.test()
# umjesto chisq.test()
fisher.test(tbl)
```

```
##
## Fisher's Exact Test for Count Data
##
## data: tbl
## p-value = 0.003003
## alternative hypothesis: two.sided
```

chisq.test je nepouzdan pošto su očekivane frekvencije pojedinih razreda manje od 5, radi toga koristimo fisher.test. p-value Fesherovog testa iznosi manje od 0.05 stoga odbacujemo hipotezu H0, te zaključujemo da su edukacija više educiranog roditelja i završna ocjena iz portugala zavisne.

H0: Ocjena iz portugala i edukacija majke su nezavisna obilježja H1: Ocjena iz portugala i edukacija majke su zavisna obilježja

```
tbl = table(students$MeduMod, students$Por_grade)
added_margins_tbl = addmargins(tbl)
print(added_margins_tbl)
```

```
##
##           A    B    C    F Sum
## 0         5    9 108  16 138
## 1         6    9  62   8  85
## 2         9   19  72   5 105
## Sum      20   37 242  29 328
```

```
chisq.test(tbl, correct = F)
```

```
##
## Pearson's Chi-squared test
```

```
##
## data:  tbl
## X-squared = 13.658, df = 6, p-value = 0.0337
```

p-value Testa nezavisnosti iznosi manje od 0.05 stoga odbacujemo hipotezu H0, te zaključujemo da su edukacija majke i završna ocjena iz portugala zavisne.

H0: Ocjena iz portugala i edukacija oca su nezavisna obilježja H1: Ocjena iz portugala i edukacija oca su zavisna obilježja

```
tbl2 = table(students$FeduMod, students$Por_grade)
added_margins_tbl2 = addmargins(tbl2)
print(added_margins_tbl2)
```

```
##
##           A    B    C    F Sum
## 0         8   10   99   14 131
## 1         3    7   67    3  80
## 2         7   15   53    5  80
## Sum      18   32  219   22 291
```

```
chisq.test(tbl2, correct = F)
```

```
##
## Pearson's Chi-squared test
##
## data:  tbl2
## X-squared = 12.75, df = 6, p-value = 0.04719
```

```
# očekivane frekvencije svih razreda moraju biti veće ili jednake 5
for (col_names in colnames(added_margins_tbl2)) {
  for (row_names in rownames(added_margins_tbl2)) {
    if (!(row_names == "Sum" | col_names == "Sum")) {
      cat("Očekivane frekvencije za razred ", col_names, "-", row_names, ": ",
          (added_margins_tbl2[row_names, "Sum"] * added_margins_tbl2["Sum",
            col_names])/added_margins_tbl2["Sum", "Sum"], "\n")
    }
  }
}
```

```
## Očekivane frekvencije za razred A - 0 : 8.103093
## Očekivane frekvencije za razred A - 1 : 4.948454
## Očekivane frekvencije za razred A - 2 : 4.948454
## Očekivane frekvencije za razred B - 0 : 14.4055
## Očekivane frekvencije za razred B - 1 : 8.797251
## Očekivane frekvencije za razred B - 2 : 8.797251
## Očekivane frekvencije za razred C - 0 : 98.58763
## Očekivane frekvencije za razred C - 1 : 60.20619
## Očekivane frekvencije za razred C - 2 : 60.20619
## Očekivane frekvencije za razred F - 0 : 9.90378
## Očekivane frekvencije za razred F - 1 : 6.04811
## Očekivane frekvencije za razred F - 2 : 6.04811
```

```
# Vidimo da postoje očekivane frekvencije manje od 5 pa koristimo fisher.test()
# umjesto chisq.test()
fisher.test(tbl)
```

```
##
```

```
## Fisher's Exact Test for Count Data
##
## data:  tbl
## p-value = 0.03133
## alternative hypothesis: two.sided
```

Kod usporedbe s edukacijom oca koristimo Fisherov test gdje je p-value manji od 0.05 pa odbacujemo H_0 i zaključujemo da su edukacija oca i završna ocjena iz portugala zavisne.

Koja škola je bolja u matematici a koja u portugalskom?

Na ovo pitanje odgovoriti ćemo provedbom t-testa koristeći 4 različita podatkovna skupa. Razdvajanje podatkovnog skupa na dvije škole (GP, MS) te na dva predmeta (matematika i portugalski) dobivamo sljedeće podatkovne skupove: `gp_mat`, `gp_por`, `ms_mat`, `ms_por`

```
# Show average grade for all schools
schools <- students %>%
  select("school") %>%
  distinct(.)
schools # [GP, MS]
subject_final_grade_names <- names(students)[grepl("G3", names(students))]

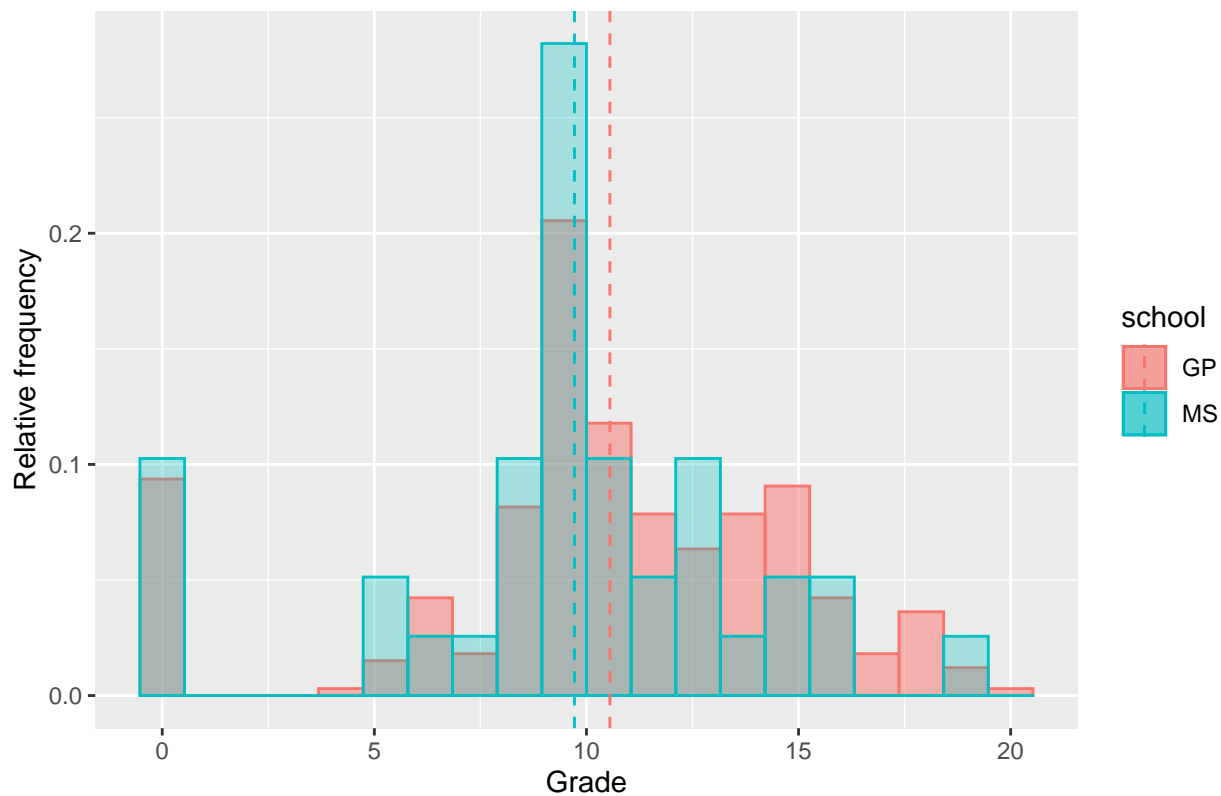
# all_of Note: Using an external vector in selections is ambiguous. Use
# `all_of(vars)` instead of `vars` to silence this message.
students_final_grade <- students %>%
  select("school", all_of(subject_final_grade_names))

# Select only the subject grade and school
gp_mat <- students_final_grade %>%
  filter(school == "GP") %>%
  select(G3_mat, school)
gp_por <- students_final_grade %>%
  filter(school == "GP") %>%
  select(G3_por, school)
ms_mat <- students_final_grade %>%
  filter(school == "MS") %>%
  select(G3_mat, school)
ms_por <- students_final_grade %>%
  filter(school == "MS") %>%
  select(G3_por, school)
```

Prikaz relativnih frekvencija predmeta i škola

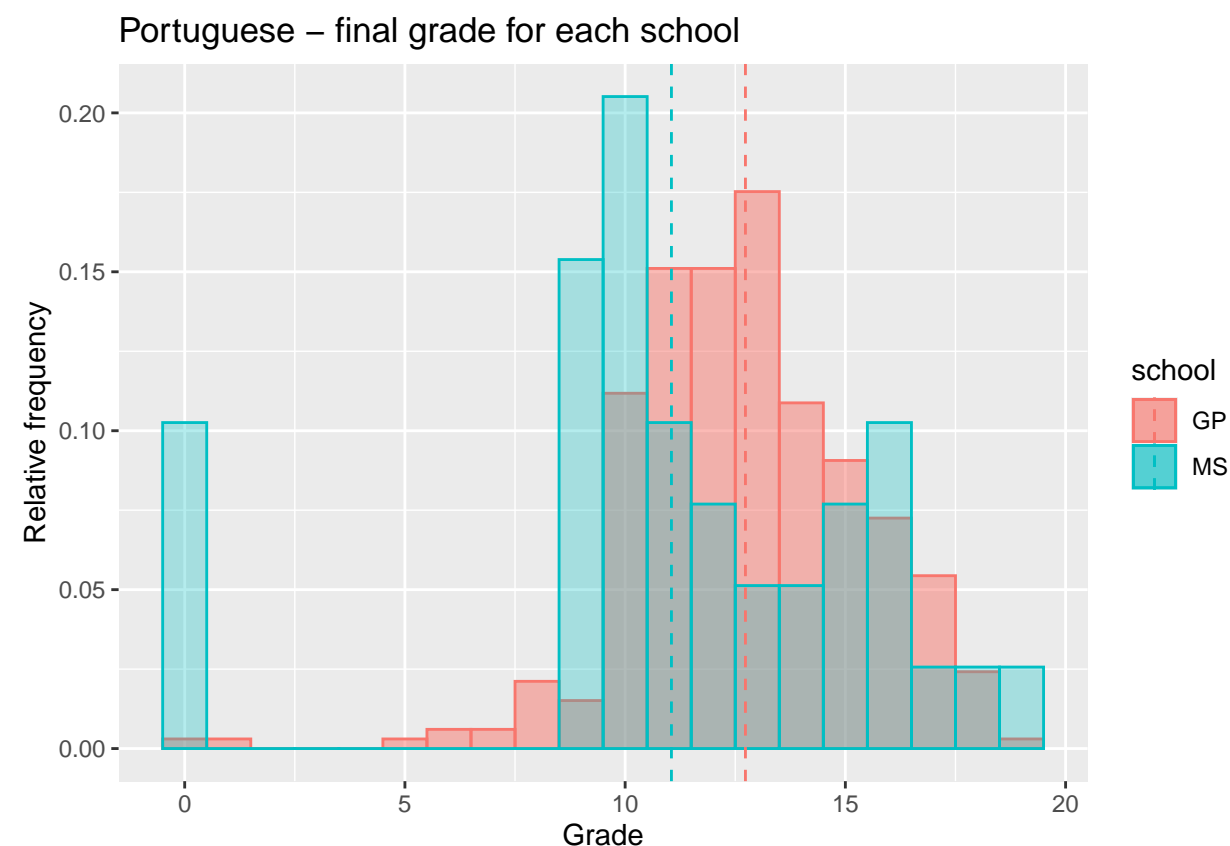
Iz grafa relativne frekvencije možemo usporediti vertikalne crte koje određuju srednju vrijednost ocjene za pojedinu školu i također dobiti osjećaj za normalnost podataka. Konstruirati ćemo jednosmjerni T-test a alternativa će ići u korist škole koja ima veću srednju vrijednost čime ćemo provjeriti je li ta škola statistički značajno bolja u matematici/portugalskom.

Matematika - prikaz relativnih frekvencija i srednjih vrijednosti
 Mathematics – final grade for each school



Na grafu za matematiku vidi se da škola GP ima veću srednju vrijednost od škole MS

Portugalski - prikaz relativnih frekvencija i srednjih vrijednosti



Na grafu za portugalski vidi se da škola GP ima veću srednju vrijednost od škole MS

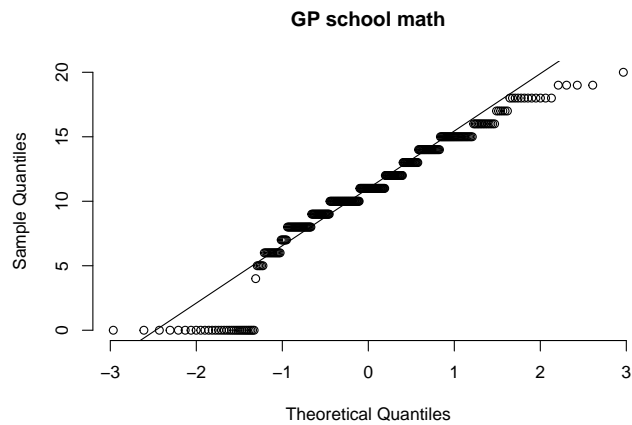
Provjera normalnosti

Normalnost se provjerva na više načina. U sljedećim koracima biti će prikazani `qqnorm` grafovi i provedeni Lilliefors i Kolmogorov-Smirnov testovi na temelju kojih će se pretpostaviti (ne)normalnost.

```
nrow(gp_mat)
nrow(gp_por)
nrow(ms_mat)
nrow(ms_por)
```

n - broj podataka za matematiku je 331 a za portugalski 39

```
qqnorm(gp_mat$grade, pch = 1, frame = FALSE, main = "GP school math")
qqline(gp_mat$grade)
```



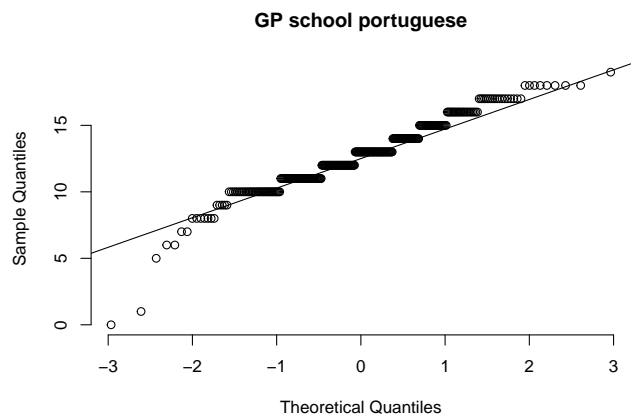
```
lillie.test(gp_mat$grade)["p.value"]
```

```
## $p.value
## [1] 7.814771e-14
```

```
ks.test(gp_mat$grade, "pnorm", mean(gp_mat$grade), sd(gp_mat$grade))["p.value"]
```

```
## $p.value
## [1] 5.330255e-05
```

```
qqnorm(gp_por$grade, pch = 1, frame = FALSE, main = "GP school portuguese")
qqline(gp_por$grade)
```



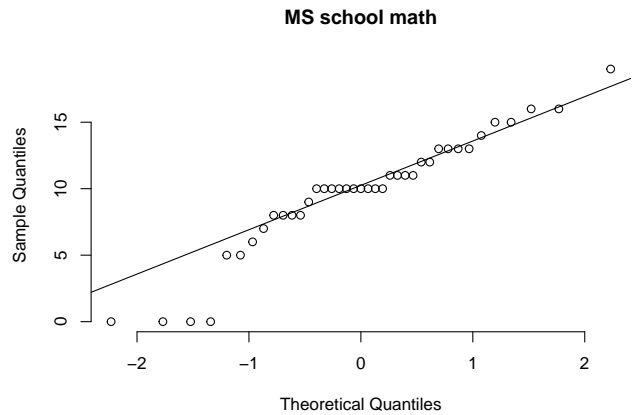
```
lillie.test(gp_por$grade)["p.value"]
```

```
## $p.value
## [1] 1.673428e-09
```

```
ks.test(gp_por$grade, "pnorm", mean(gp_por$grade), sd(gp_por$grade))["p.value"]
```

```
## $p.value
## [1] 0.001247681
```

```
qqnorm(ms_mat$grade, pch = 1, frame = FALSE, main = "MS school math")
qqline(ms_mat$grade)
```



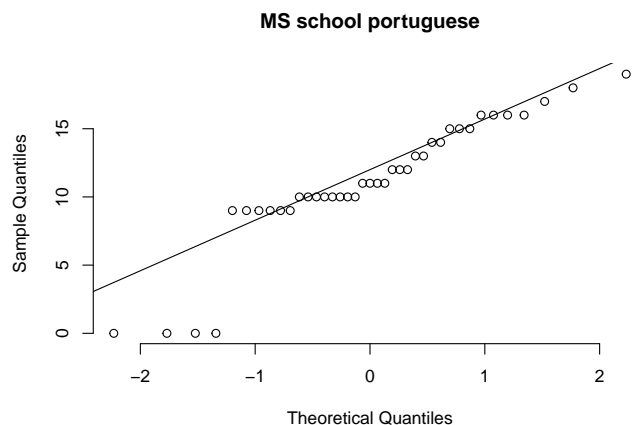
```
lillie.test(ms_mat$grade)["p.value"]
```

```
## $p.value
## [1] 0.0009170632
```

```
ks.test(ms_mat$grade, "pnorm", mean(ms_mat$grade), sd(ms_mat$grade))["p.value"]
```

```
## $p.value
## [1] 0.1131777
```

```
qqnorm(ms_por$grade, pch = 1, frame = FALSE, main = "MS school portuguese")
qqline(ms_por$grade)
```



```
lillie.test(ms_por$grade)["p.value"]
```

```
## $p.value
## [1] 1.951046e-05
```

```
ks.test(ms_por$grade, "pnorm", mean(ms_por$grade), sd(ms_por$grade))["p.value"]
```

```
## $p.value
## [1] 0.03355273
```

Repovi su prisutni na lijevoj strani podataka zbog čega je p vrijednost skoro uvijek manja od 0.05 za Kolmogorov-Smirnov i Lillieforsov test. Grafički, na temelju rezultata određujemo da za sve skupove vrijedi da proizlaze iz normalne distribucije ali s opaskom da postoje stršće vrijednosti na lijevoj strani distribucije.

F-test - test o jednakosti varijanca

Važno je napomenuti da je test o varijanci iznimno osjetljiv na normalnost. Test će biti proveden zbog vježbe ali njegov **rezultat se neće uzeti u obzir** jer podaci nisu normalno distribuirani.

p – vjerojatnost da pod H_0 dobijemo vrijednost koja je jednako ili više ekstremna nego vrijednost koji bi dobili izračunom iz uzorka kojeg imamo

Ako je $p < \alpha$, odbacujemo hipotezu H_0 u korist hipoteze H_1 :

- pada u desni ili lijevi rep => odbacivanje

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \neg H_0$$

Poredak argumenata za `var.test` nije bitna ali generalno vrijedi:

$$\frac{\sigma_1^2}{\sigma_2^2}, \quad \sigma_1^2 > \sigma_2^2$$

```
cat("Mathematics variances", var(gp_mat$grade), var(ms_mat$grade))
```

```
## Mathematics variances 21.38735 19.89204
```

```
cat("Portugeuse variances", var(gp_por$grade), var(ms_por$grade))
```

```
## Portugeuse variances 6.839605 22.1552
```

Na prvi pogled čini se da će H_0 hipoteza za portugalski biti odbačena zbog toga što su varijance značajno drugačije. Potrebno je provesti f-test da se uvjerimo da se radi o statistički značajnoj razlici varijanci.

Konstruirajmo i provedimo testove o varijanci:

```
alpha <- 0.05
```

```
# H0 - Variance of GP_MAT and MS_MAT are equal H1 - not H0
```

```
mat_f_test <- var.test(gp_mat$grade, ms_mat$grade, alternative = "two.sided") # F = 1.0752, p = 0.817
```

```
# H0 - Variance of GP_POR and MS_MAT are equal H1 - not H0
```

```
por_f_test <- var.test(gp_por$grade, ms_por$grade, alternative = "two.sided") # F = 0.30871, p = 1.217
```

```
var_equal_mat <- if (mat_f_test$p.value < alpha) FALSE else TRUE
```

```
cat_reject_h0("Matematika - test o jednakosti varijanca:", !var_equal_mat)
```

```
## Matematika - test o jednakosti varijanca:
```

```
## Ne odbacujemo hipotezu H0
```

```
var_equal_por <- if (por_f_test$p.value < alpha) FALSE else TRUE
```

```
cat_reject_h0("Portugalski - test o jednakosti varijanca:", !var_equal_por)
```

```
## Portugalski - test o jednakosti varijanca:
```

```
## Odbacujemo hipotezu H0 u korist hipoteze H1
```

T-test - testiranje jednakosti srednje vrijednosti ocjena za dvije škole uz nepoznate varijance

Uz to što je n veći od 30 za oba podatkovna skupa i uz činjenicu da je t-test robustan na (ne)normalnost provodimo t-test srednje vrijednosti za oba predmeta.

Zbog prethodno dobivenih srednje vrijednosti o ocjenama (koje idu u korist škole GP) postavljena je jednosmjerna alternativa hipoteza.

Ponovno, zbog toga što test o varijanci nije robustan na nenormalnost pretpostaviti ćemo da varijance uzoraka nisu jednake.

```
# H0 - GP school has equal grades to in mathematics to MS (GP=MS) H1 - GP>MS
mat_t_test <- t.test(gp_mat$grade, ms_mat$grade, alt = "greater", var.equal = FALSE)
is_gp_mat_better <- if (mat_t_test$p.value < alpha) TRUE else FALSE
cat_reject_h0("Matematika - t-test:", is_gp_mat_better)
## Matematika - t-test:
## Ne odbacujemo hipotezu H0

# H0 - GP school has equal grades to in Portuguese to MS (GP=MS) H1 - GP>MS
por_t_test <- t.test(gp_por$grade, ms_por$grade, alt = "greater", var.equal = FALSE)
is_gp_por_better <- if (por_t_test$p.value < alpha) TRUE else FALSE
cat_reject_h0("Portugalski t-test:", is_gp_por_better)
## Portugalski t-test:
## Odbacujemo hipotezu H0 u korist hipoteze H1
```

Za matematiku, nismo odbacili hipotezu H0 i zbog čega ne možemo zaključiti da škola GP ima bolje ocjene iz matematike od škole MS.

Za portugalski, odbacujemo hipotezu H0 u korist hipoteze H1 i zaključujemo da je škola GP ima bolje ocjene iz portugalskog od škole MS.

Jesu li učenici uspješniji u matematici ili glavnom jeziku?

```
students_org %>%
  summarise(Mean.G3_mat = mean(G3_mat), Mean.G3_por = mean(G3_por), ) -> summary.result1
summary.result1

## # A tibble: 1 x 2
##   Mean.G3_mat Mean.G3_por
##   <dbl>      <dbl>
## 1      10.5      12.6

students_org %>%
  summarise(Med.G3_mat = median(G3_mat), Med.G3_por = median(G3_por), ) -> summary.result2
summary.result2

## # A tibble: 1 x 2
##   Med.G3_mat Med.G3_por
##   <dbl>      <dbl>
## 1        11        13

students_org %>%
  summarise(Mean.G3_mat = mean(G3_mat, trim = 0.1), Mean.G3_por = mean(G3_por,
    trim = 0.1), ) -> summary.result3
summary.result3

## # A tibble: 1 x 2
##   Mean.G3_mat Mean.G3_por
##   <dbl>      <dbl>
## 1      10.9      12.6
```

```
(1 - summary.result3/summary.result1) * 100
```

```
## Mean.G3_mat Mean.G3_por
## 1 -4.016012 -0.7265877
```

Kao što je vidljivo iz podataka, učenici su malo uspješniji u glavnom jeziku (portugalskom), ali ako gleda prema samoj ocjeni obje skupine spadaju u ocjenu "C". Čak i ako uzmemo podrezanu srednju vrijednost (10%), rezultat se promijeni za ~1%.

```
students_org %>%
  summarise(IQR.G3_mat = IQR(G3_mat), IQR.G3_por = IQR(G3_por), ) -> summary.result4
summary.result4
```

```
## # A tibble: 1 x 2
## IQR.G3_mat IQR.G3_por
## <dbl> <dbl>
## 1 6 3
```

```
students_org %>%
  summarise(Var.G3_mat = var(G3_mat), Var.G3_por = var(G3_por), ) -> summary.result5
summary.result5
```

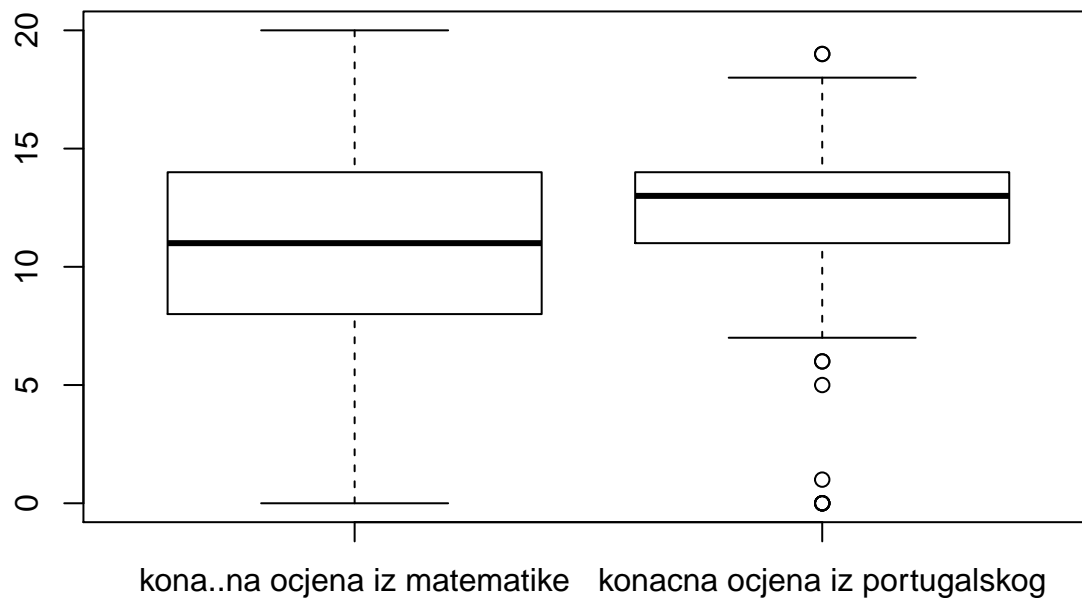
```
## # A tibble: 1 x 2
## Var.G3_mat Var.G3_por
## <dbl> <dbl>
## 1 21.2 8.67
```

```
students_org %>%
  summarise(sd.G3_mat = sd(G3_mat), sd.G3_por = sd(G3_por), ) -> summary.result6
summary.result6
```

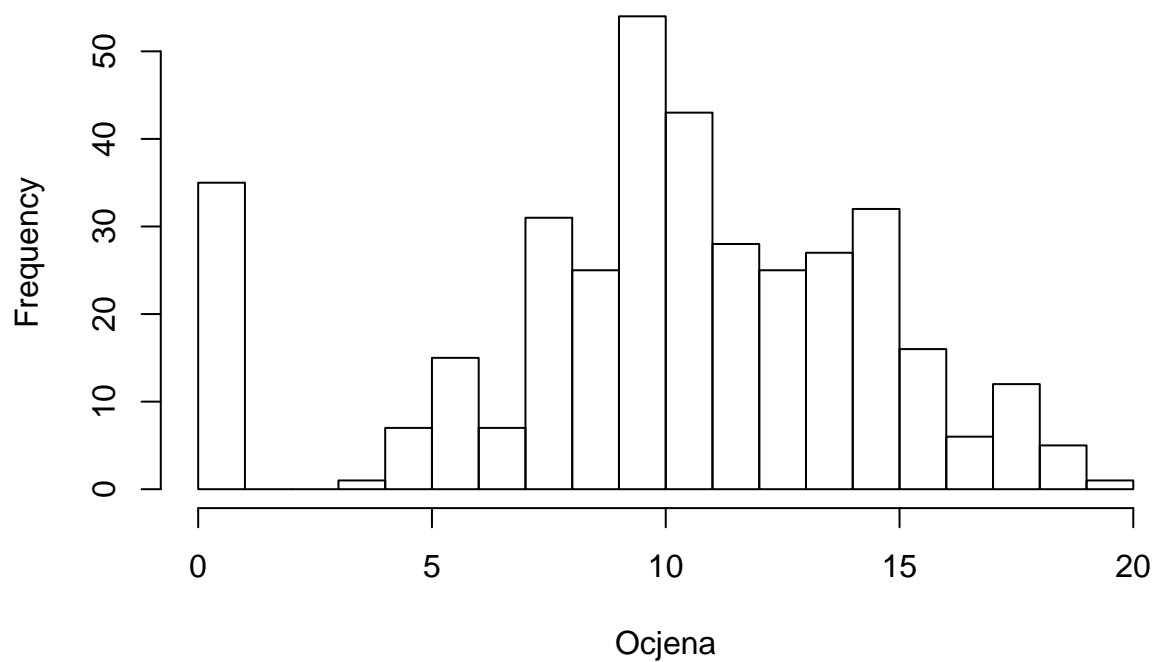
```
## # A tibble: 1 x 2
## sd.G3_mat sd.G3_por
## <dbl> <dbl>
## 1 4.61 2.94
```

Ako gledamo raspršenost varijabli vidimo da ocjene iz portugalskog jezika imaju manje sve tri mjere (IQR, varijanca i standardna devijacija) vidimo da se ocjene iz portugalskog manje manje odmiču od srednje vrijednosti nego ocjene iz matematike.

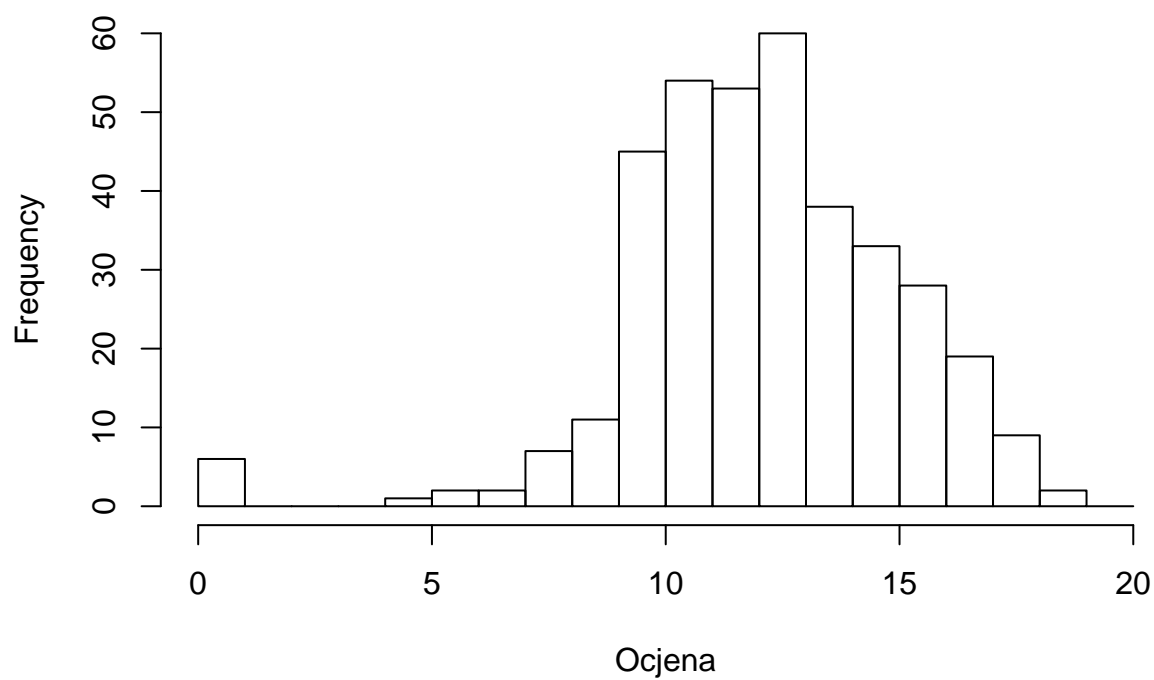
Boxplot konacnih ocjena iz matematike i portugala



Histogram ocjena iz matematike



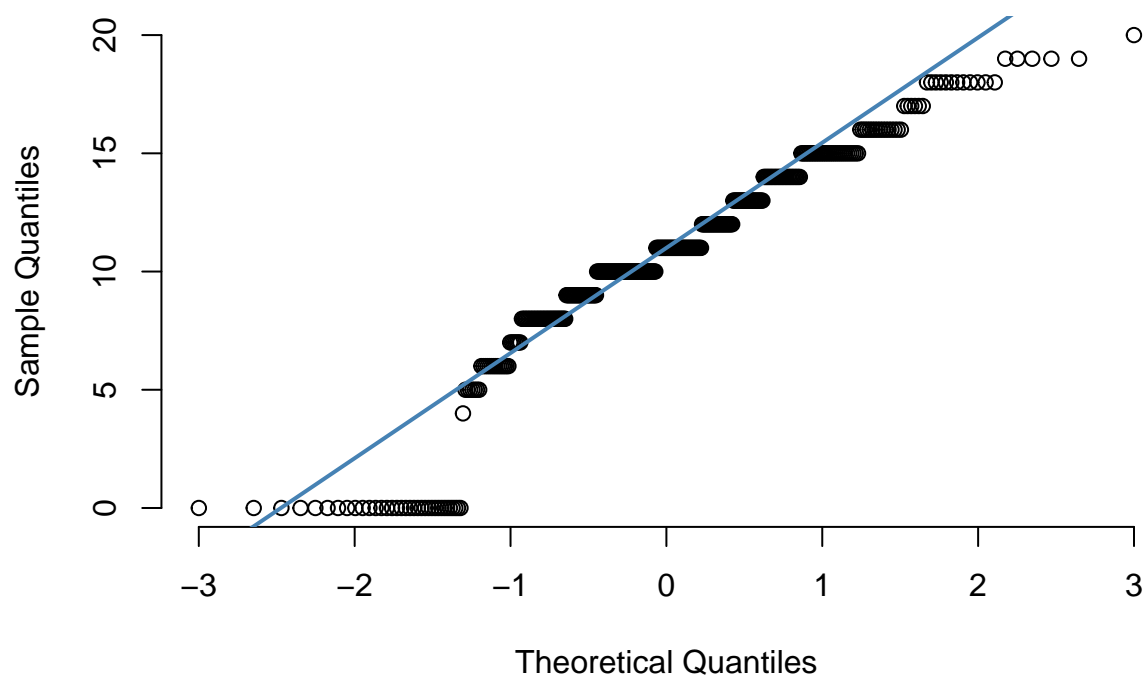
Histogram ocjena iz portugalskog



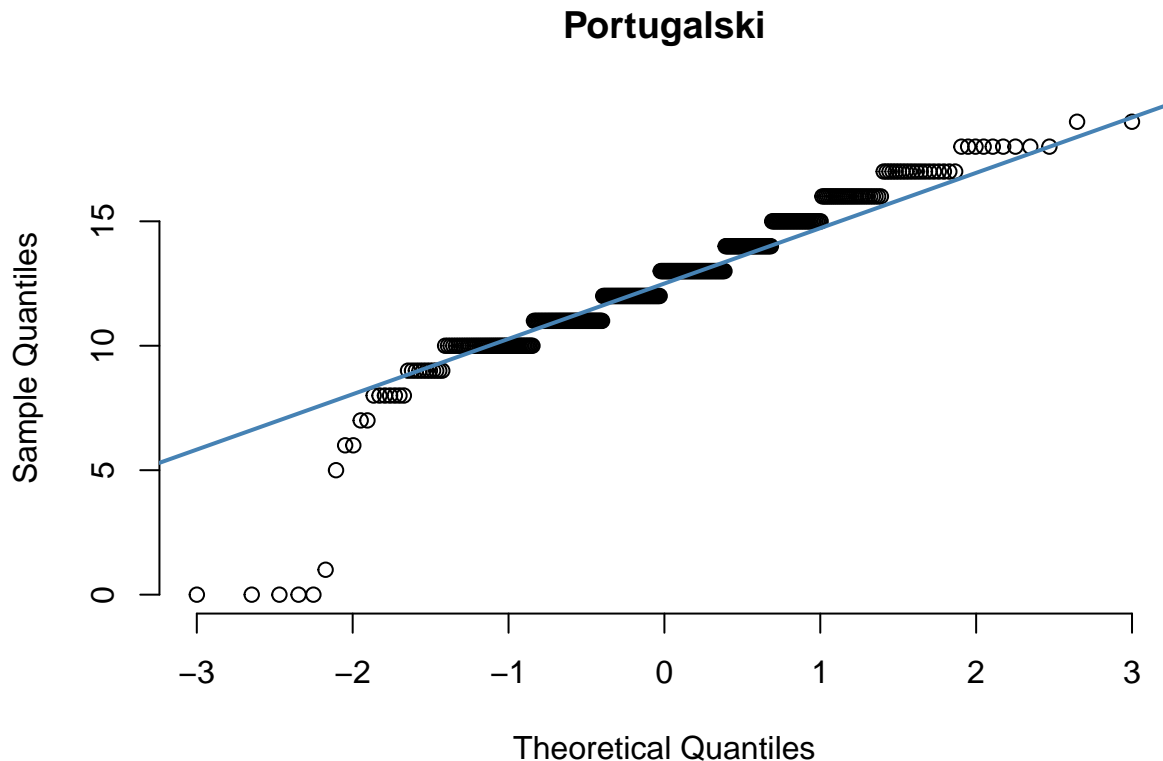
Prije testa varijance radimo provjeru normalnosti podataka pomoću qqplotova te Lilliefors i Kolmogorov-Smirnov testova:

```
qqnorm(students_org$G3_mat, pch = 1, frame = FALSE, main = "Matematika")  
qqline(students_org$G3_mat, col = "steelblue", lwd = 2)
```

Matematika



```
qqnorm(students_org$G3_por, pch = 1, frame = FALSE, main = "Portugalski")
qqline(students_org$G3_por, col = "steelblue", lwd = 2)
```



```
print("Matematika:")
```

```
## [1] "Matematika:"
```

```
lillie.test(students_org$G3_mat)["p.value"]
```

```
## $p.value
```

```
## [1] 2.653956e-17
```

```
ks.test(students_org$G3_mat, "pnorm", mean(students_org$G3_mat), sd(students_org$G3_mat))["p.value"]
```

```
## $p.value
```

```
## [1] 4.296089e-06
```

```
print("Portugalski:")
```

```
## [1] "Portugalski:"
```

```
lillie.test(students_org$G3_por)["p.value"]
```

```
## $p.value
```

```
## [1] 1.236014e-12
```

```
ks.test(students_org$G3_por, "pnorm", mean(students_org$G3_por), sd(students_org$G3_por))["p.value"]
```

```
## $p.value
```

```
## [1] 0.0001241436
```

Jako male p-vrijednosti kod Lilliefors i Kolmogorov-Smirnov testova dolaze zbog repova s lijeve strane. Grafički, na temelju rezultata određujemo da za sve skupove vrijedi da proizlaze iz normalne distribucije ali s opaskom da postoje stršće vrijednosti na lijevoj strani distribucije.

F-test

Važno je napomenuti da je test o varijanci iznimno osjetljiv na normalnost. Test će biti proveden zbog vježbe ali njegov rezultat se neće uzeti u obzir.

```
var(students_org$G3_mat)

## [1] 21.24131
var(students_org$G3_por)

## [1] 8.665092
var.test(students_org$G3_mat, students_org$G3_por)

##
## F test to compare two variances
##
## data: students_org$G3_mat and students_org$G3_por
## F = 2.4514, num df = 369, denom df = 369, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 1.998239 3.007245
## sample estimates:
## ratio of variances
## 2.451366
```

Zbog jako male vrijednosti odbacujemo hipotezu H_0 da su varijance dva uzorka jednake.

T-test

```
t.test(students_org$G3_por, students_org$G3_mat, alternative = "greater", var.equal = FALSE)

##
## Welch Two Sample t-test
##
## data: students_org$G3_por and students_org$G3_mat
## t = 7.3485, df = 627.1, p-value = 3.137e-13
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
## 1.620861 Inf
## sample estimates:
## mean of x mean of y
## 12.55405 10.46486
```

Zbog jako male p-vrijednosti odbacujemo hipotezu H_0 da su prosjeci ocjena jednaki u korist hipoteze H_1 da je prosjek ocjena iz portugalskog značajno veći od prosjeka ocjena iz matematike.

Kako vrijeme putovanja do škole utječe na uspjeh učenika?

Na ovo pitanje odgovirit ćemo ANOVA-om. Pretpostavke ANOVA-e su:

- nezavisnost pojedinih podataka u uzorcima
- normalna razdioba podataka
- homogenost varijanci među populacijama

Postavljamo hipotezu H_0 koja glasi, srednja vrijednost grupa su podjednake.

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1 : \neg H_0$$

S obizrom da se radi o različitim školama i različitim predmetima možemo pretpostaviti nezavisnost ocjena.

Ukoliko nakon provedbe ANOVA-e odbacimo H_0 hipotezu možemo zaključiti da su srednje vrijednosti međusobno različite, tj. da vrijeme putovanje utječe na uspjeh učenika.

Obrada kategoričkih stupaca

Kao grupe koristiti će se vrijednosti iz stupca `traveltime`. Prvo je potrebno pretvoriti stupac `traveltime` u kategoričke podatke (s poretkom). `traveltime` se sastoji od 4 mogućih vrijednosti koje definiraju potrebno vrijeme od škole do doma:

- < 15min
- 15 - 30 min
- 30 - 60 min
- > 60 min

Nadalje, zadnju kategoriju (60min+) spojiti ćemo sa predzadnjom kategorijom (30-60min) zbog toga što se u zadnjoj kategoriji nalaze samo 8 podataka dok se u preostalim kategorijama nalazi puno veći broj podataka.

```
count(students, students$traveltime)
```

```
## # A tibble: 4 x 2
##   `students$traveltime`      n
##             <dbl> <int>
## 1                   1    242
## 2                   2     99
## 3                   3     21
## 4                   4      8

students <- students_clean
students$traveltime <- factor(students$traveltime, ordered = TRUE, labels = c("0 - 15 min",
  "15 - 30 min", "> 30 min", "> 30 min"))
```

Za uspjeh koristiti ćemo zbog varijabli `G[1,2,3]_mat` i `G[1,2,3]_por` koji ćemo spremiti u novu varijablu `G_total`.

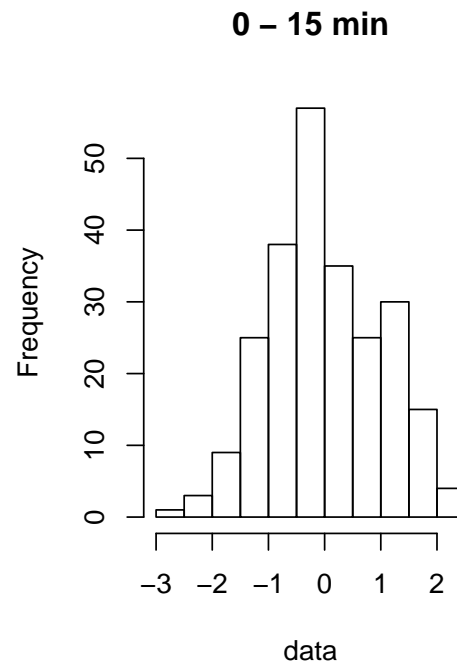
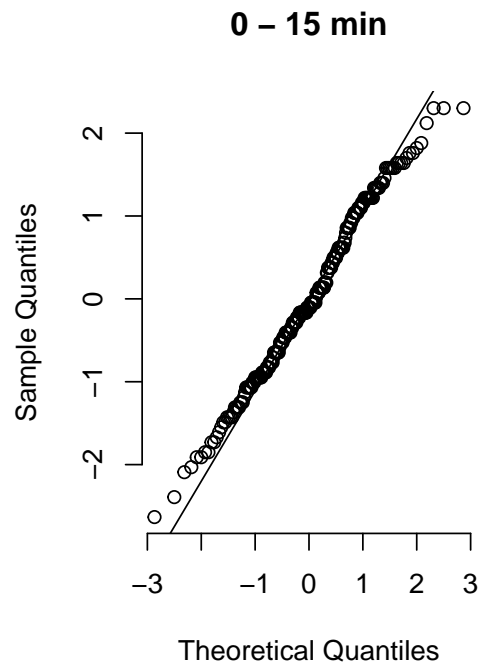
```
students$G3_total <- students$G3_mat + students$G3_por
students$G2_total <- students$G2_mat + students$G2_por
students$G1_total <- students$G1_mat + students$G1_por
students$G_total <- students$G1_total + students$G2_total + students$G3_total
```

ANOVA je robustna na blaga odstupanja što se tiče normalnosti. Svedjedno, testirati ćemo normalnost varijable `G_total` nad cijelim podatkovnim skupom, a zatim nad `G_total` za svaku pojedinu grupu `traveltime`-a.

```
model = lm(students$G_total ~ students$traveltime)

par(mfrow = c(1, 2)) # 2 plots in 1 row

timeperiod = "0 - 15 min"
data <- rstandard(model)[students$traveltime == timeperiod]
qqnorm(data, pch = 1, frame = FALSE, main = timeperiod)
qqline(data)
hist(data, main = timeperiod)
```



```
lillie.test(data)["p.value"]
```

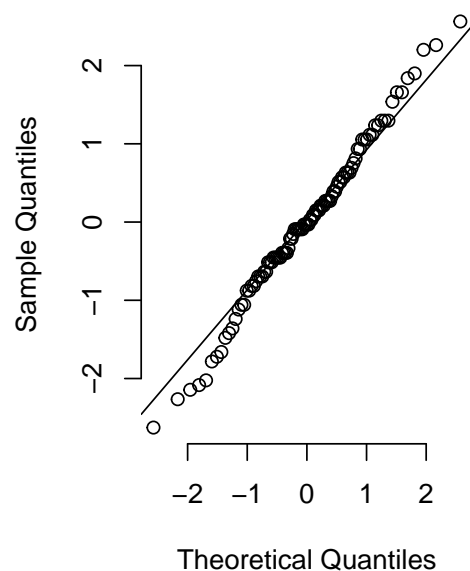
```
## $p.value
## [1] 0.008983716
```

```
ks.test(data, "pnorm", mean = mean(data), sd = sd(data))["p.value"]
```

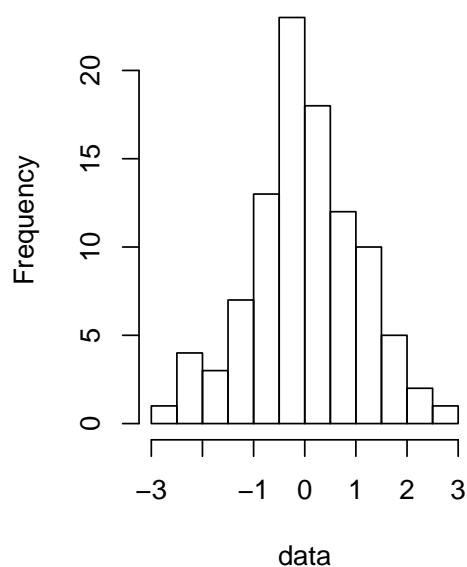
```
## $p.value
## [1] 0.2157153
```

```
timeperiod = "15 - 30 min"
data <- rstandard(model)[students$traveltime == timeperiod]
qqnorm(data, pch = 1, frame = FALSE, main = timeperiod)
qqline(data)
hist(data, main = timeperiod)
```

15 – 30 min



15 – 30 min



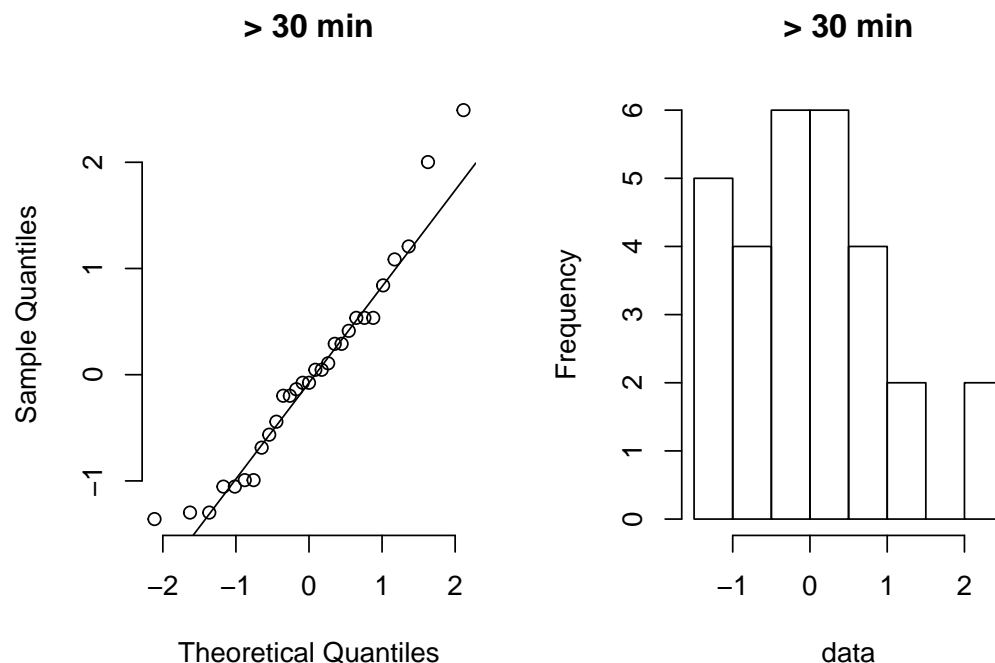
```
lillie.test(data)["p.value"]
```

```
## $p.value
## [1] 0.5782076
```

```
ks.test(data, "pnorm", mean = mean(data), sd = sd(data))["p.value"]
```

```
## $p.value
## [1] 0.897279
```

```
timeperiod = "> 30 min"
data <- rstandard(model)[students$traveltime == timeperiod]
qqnorm(data, pch = 1, frame = FALSE, main = timeperiod)
qqline(data)
hist(data, main = timeperiod)
```



```
lillie.test(data)["p.value"]
```

```
## $p.value
## [1] 0.4329395
```

```
ks.test(data, "pnorm", mean = mean(data), sd = sd(data))["p.value"]
```

```
## $p.value
## [1] 0.8440515
```

Na svakom grafu možemo vidjeti da podaci uglavnom prate normalnu distribuciju uz manji broj stršućih vrijednosti (lijevi rep). Nadalje, p vrijednosti Lillieforsovog testa nisu uvijek iznad 0.05 međutim za sve Kolmogorov–Smirnov testove p vrijednosti su iznad 0.05.

Lilliefors koristimo ako nam nije poznata varijanca i srednja vrijednost populacije, što je s ovim podacima i slučaj. Poznato je da Lilliefors konzervativniji i da odbacuje hipotezu H_0 češće nego Kolmogorov–Smirnov.

S obzirom na manja odstupanja, ne toliko male p vrijednosti i grafički izgled `qqnorm`-a i histograma pretpostaviti ćemo da su podaci uzrokovani iz normalne distribucije.

Homogenost varijanci - Bartlettov test

Prvo je potrebno postaviti hipoteze H_0 i H_1 :

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

$$H_1 : \neg H_0$$

```
var(students$G_total[students$traveltime == "> 30 min"])
```

```
## [1] 241.6897
```

```
var(students$G_total[students$traveltime == "15 - 30 min"])
```

```
## [1] 296.1703
```

```
var(students$G_total[students$traveltime == "> 30 min"])
```

```
## [1] 241.6897
```

```
bartlett.test(students$G_total ~ students$traveltime)
```

```
##
## Bartlett test of homogeneity of variances
##
## data:  students$G_total by students$traveltime
## Bartlett's K-squared = 0.48546, df = 2, p-value = 0.7845
```

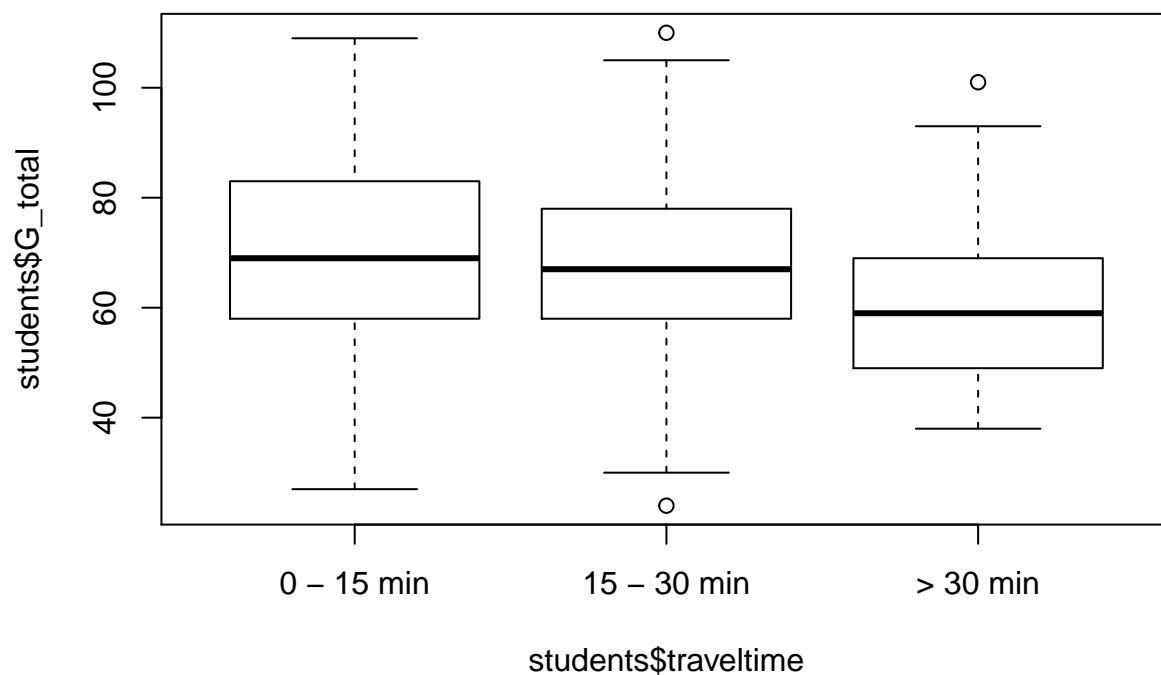
Vidimo da su vrijednosti varijance slične. S obzirom da je p vrijednost testa veća od 0.05 ne odbacujemo H_0 čime zadovoljavamo ANOVA pretpostavku o homogenosti varijanca.

ANOVA - Jesu li srednje vrijednosti za različite grupe drugačije?

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1 : \neg H_0$$

```
boxplot(students$G_total ~ students$traveltime)
```



Grafički možemo pretpostaviti da se vrijeme putovanja utječe na uspjeh učenika. Naravno, ANOVA-om je potrebno provjeriti koliko je ta razlika statistički značajna.

```
model = lm(students$G_total ~ students$traveltime)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: students$G_total
##              Df Sum Sq Mean Sq F value    Pr(>F)
## students$traveltime  2    3185   1592.35    5.7419 0.003504 **
## Residuals          367   101777    277.32
## ---
```



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

ANOVA nam govori da postoji razlika između grupa `traveltime`. Iako nije strogo značajna i dalje se radi o značajnoj p vrijednosti koja se nalazi između 0.001 i 0.01. Možemo zaključiti da za različite grupe vremena putovanja imaju utjecaj na učenikov uspjeh.

Predviđanje uspjeha na kraju školske godine drugim varijablama iz skupa podataka

Transformirajmo kategoričke varijable u dummy varijable.

```
require(fastDummies)
students_dummies = dummy_cols(students, remove_first_dummy = TRUE, remove_selected_columns = TRUE)
students_dummies
```

Sada provodimo individualne jednostavne linearne regresije `G3_mat` i `G3_por` ovisno o svakoj od varijabli iz skupa, te spremamo R^2 vrijednosti i p-vrijednosti F-testova za jednostavnu linearnu regresiju u tablice `modelsMat` i `modelsPor`

```
varName = c()
rSquaredM = c()
pValueofFM = c()
rSquaredP = c()
pValueofFP = c()

for (i in 1:ncol(students_dummies)) {
  if (i != 18 && i != 21) {
    colName = colnames(students_dummies)[i]
    names(students_dummies)[i] = "temp"

    modelMat = lm(formula = G3_mat ~ temp, data = students_dummies)
    modelPor = lm(formula = G3_por ~ temp, data = students_dummies)
    names(students_dummies)[i] = colName
    m = summary(modelMat)
    p = summary(modelPor)
    varName = append(varName, colName)
    rSquaredM = append(rSquaredM, m$r.squared)
    pValueofFM = append(pValueofFM, pf(m$fstatistic[1], m$fstatistic[2], m$fstatistic[3],
    lower.tail = FALSE))
    rSquaredP = append(rSquaredP, p$r.squared)
    pValueofFP = append(pValueofFP, pf(p$fstatistic[1], p$fstatistic[2], p$fstatistic[3],
    lower.tail = FALSE))
  }

  modelsMat = data.frame(varName, rSquaredM, pValueofFM)
  modelsPor = data.frame(varName, rSquaredP, pValueofFP)
}
```

Predviđanje konačne ocjene iz matematike

Pogledajmo koje su se varijable ispostavile najboljim prediktorima za `G3_mat` (poredano po R^2 vrijednostima)

```
modelsMat[order(-modelsMat$rSquaredM), ]
##          varName    rSquaredM    pValueofFM
## 20      G3_total 8.485015e-01 7.068668e-153
```

```

## 17          G2_mat 8.220203e-01 5.360474e-140
## 23          G_total 7.768376e-01 6.599137e-122
## 21          G2_total 7.361574e-01 1.630294e-108
## 16          G1_mat 6.482165e-01 1.686172e-85
## 22          G1_total 5.791317e-01 3.796506e-71
## 19          G2_por 3.036584e-01 8.998815e-31
## 18          G1_por 2.620521e-01 4.197705e-26
## 6      failures_mat 1.392004e-01 1.153185e-13
## 48      higher_yes 5.127693e-02 1.091204e-05
## 2          Medu 4.442857e-02 4.374826e-05
## 1          age 3.341011e-02 4.095595e-04
## 15      absences_por 2.982017e-02 8.514064e-04
## 3          Fedu 2.468901e-02 2.437230e-03
## 50      romantic_yes 2.212492e-02 4.138021e-03
## 4          traveltime 1.763917e-02 3.817115e-02
## 25          sex_M 1.724362e-02 1.146151e-02
## 41      guardian_other 1.572560e-02 1.579910e-02
## 10          goout 1.524576e-02 1.749543e-02
## 29      Mjob_health 1.177072e-02 3.697833e-02
## 39      reason_reputation 1.117669e-02 4.211523e-02
## 7          failures_por 1.110634e-02 4.277112e-02
## 26          address_U 1.093968e-02 4.436825e-02
## 27          famsize_LE3 1.082667e-02 4.548664e-02
## 31      Mjob_services 9.438170e-03 6.192712e-02
## 36      Fjob_teacher 9.040613e-03 6.771572e-02
## 30      Mjob_other 8.462013e-03 7.719322e-02
## 44      paid_mat_yes 8.230946e-03 8.136636e-02
## 42      schoolsup_yes 7.140813e-03 1.046198e-01
## 49      internet_yes 6.747007e-03 1.147224e-01
## 5          studytime 5.772488e-03 1.446729e-01
## 32      Mjob_teacher 3.525534e-03 2.545926e-01
## 13          health 3.454390e-03 2.594503e-01
## 47      nursery_yes 3.256085e-03 2.736086e-01
## 24          school_MS 3.102947e-03 2.852075e-01
## 34      Fjob_other 2.863524e-03 3.046207e-01
## 28      Pstatus_T 2.595037e-03 3.284709e-01
## 33      Fjob_health 2.576309e-03 3.302248e-01
## 43      famsup_yes 2.517837e-03 3.357810e-01
## 8          famrel 1.991185e-03 3.920770e-01
## 38      reason_other 1.856206e-03 4.086276e-01
## 11          Dalc 1.774706e-03 4.191188e-01
## 12          Walc 1.350276e-03 4.810145e-01
## 46      activities_yes 1.260180e-03 4.960353e-01
## 40      guardian_mother 7.775170e-04 5.928930e-01
## 37          reason_home 6.119266e-04 6.352923e-01
## 14      absences_mat 3.779776e-04 7.093439e-01
## 9          freetime 4.639719e-05 8.961068e-01
## 35      Fjob_services 4.089570e-05 9.024280e-01
## 45      paid_por_yes 3.761621e-05 9.064032e-01

```

Razmotrit ćemo prvih 10 najboljih prediktora. Najprije provjerimo jesu li neke od tih varijabli visoko korelirane:

```
cor(cbind(students_dummies$G2_mat, students_dummies$G1_mat, students_dummies$G2_por,
students_dummies$G1_por, students_dummies$failures_mat, students_dummies$higher_yes,
students_dummies$Medu, students_dummies$age, students_dummies$absences_por, students_dummies$Fedu))
##           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,]  1.0000000  0.8567705  0.57804548  0.5382023 -0.3686201  0.2136061
## [2,]  0.8567705  1.0000000  0.59826670  0.5810192 -0.3863053  0.2238001
## [3,]  0.5780455  0.5982667  1.00000000  0.8874806 -0.3508856  0.2985681
## [4,]  0.5382023  0.5810192  0.88748063  1.0000000 -0.2912844  0.2771680
## [5,] -0.3686201 -0.3863053 -0.35088560 -0.2912844  1.0000000 -0.3659806
## [6,]  0.2136061  0.2238001  0.29856805  0.2771680 -0.3659806  1.0000000
## [7,]  0.2129388  0.2220518  0.21506989  0.2009143 -0.2231552  0.1571828
## [8,] -0.1654850 -0.1042716 -0.03497507 -0.1023385  0.1476343 -0.2455488
## [9,] -0.1800028 -0.1502991 -0.18243277 -0.1472927  0.1689433 -0.1154271
## [10,] 0.1681930  0.2053066  0.17844395  0.1487680 -0.2259357  0.1703762
##           [,7]      [,8]      [,9]      [,10]
## [1,]  0.21293884 -0.16548500 -0.18000282  0.16819303
## [2,]  0.22205178 -0.10427157 -0.15029911  0.20530662
## [3,]  0.21506989 -0.03497507 -0.18243277  0.17844395
## [4,]  0.20091426 -0.10233849 -0.14729266  0.14876797
## [5,] -0.22315525  0.14763426  0.16894333 -0.22593571
## [6,]  0.15718279 -0.24554880 -0.11542706  0.17037620
## [7,]  1.00000000 -0.11524237  0.02165622  0.63603508
## [8,] -0.11524237  1.00000000  0.09332766 -0.12381596
## [9,]  0.02165622  0.09332766  1.00000000  0.02517965
## [10,] 0.63603508 -0.12381596  0.02517965  1.00000000
```

Kao i očekivano ocjene G2_mat i G1_mat visoko su korelirane, isto kao i G2_por i G1_por, a značajna je i korelacija između ocjena matematike i portugala. Osim toga uočimo koreliranost razina edukacije majke i oca.

```
cor(students_dummies$Medu, students_dummies$Fedu)
## [1] 0.6360351
```

Zasad nećemo eliminirati nijedan regresor. Izgradimo linearni model od gore izdvojenih varijabli za G3_mat:

```
multiMat = lm(G3_mat ~ G2_mat + G1_mat + G2_por + G1_por + failures_mat + higher_yes +
Medu + age + absences_por + Fedu, data = students_dummies)
summary(multiMat)
##
## Call:
## lm(formula = G3_mat ~ G2_mat + G1_mat + G2_por + G1_por + failures_mat +
##     higher_yes + Medu + age + absences_por + Fedu, data = students_dummies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.3643 -0.3881  0.2843  0.9220  3.3666
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.40447    1.77472   0.228   0.8199
## G2_mat        0.95943    0.05325  18.018 <2e-16 ***
## G1_mat        0.13853    0.06182   2.241  0.0257 *
## G2_por        0.07276    0.09549   0.762  0.4466
## G1_por       -0.04011    0.08842  -0.454  0.6504
## failures_mat -0.17299    0.16603  -1.042  0.2981
```

```
## higher_yes      0.26655      0.55720      0.478      0.6327
## Medu            0.08720      0.12252      0.712      0.4771
## age             -0.14087      0.09154     -1.539      0.1247
## absences_por    -0.00179      0.02163     -0.083      0.9341
## Fedu            -0.10921      0.12210     -0.894      0.3717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.936 on 359 degrees of freedom
## Multiple R-squared:  0.8283, Adjusted R-squared:  0.8235
## F-statistic: 173.1 on 10 and 359 DF,  p-value: < 2.2e-16
```

Pojednostavimo sad model, uzevši 5 varijabli s najnižim p-vrijednostima

```
multiMat2 = lm(data = students_dummies, G3_mat ~ G2_mat + G1_mat + age + failures_mat +
  G2_por)
summary(multiMat2)
##
## Call:
## lm(formula = G3_mat ~ G2_mat + G1_mat + age + failures_mat +
##     G2_por, data = students_dummies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.3354 -0.3486  0.2398  0.9426  3.3867
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.53966    1.53698   0.351  0.7257
## G2_mat         0.96419    0.05249  18.369 <2e-16 ***
## G1_mat         0.13056    0.06045   2.160  0.0314 *
## age           -0.14085    0.08711  -1.617  0.1068
## failures_mat  -0.19456    0.15636  -1.244  0.2142
## G2_por         0.04284    0.05217   0.821  0.4121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.926 on 364 degrees of freedom
## Multiple R-squared:  0.8277, Adjusted R-squared:  0.8253
## F-statistic: 349.6 on 5 and 364 DF,  p-value: < 2.2e-16
```

Nešto nam se smanjio R^2 , no prilagođeni R^2 se uvećao-indikacija da smo eliminirali neke nepotrebne regresore.

Dodatno pojednostavljenje modela smanjuje prilagođeni R^2

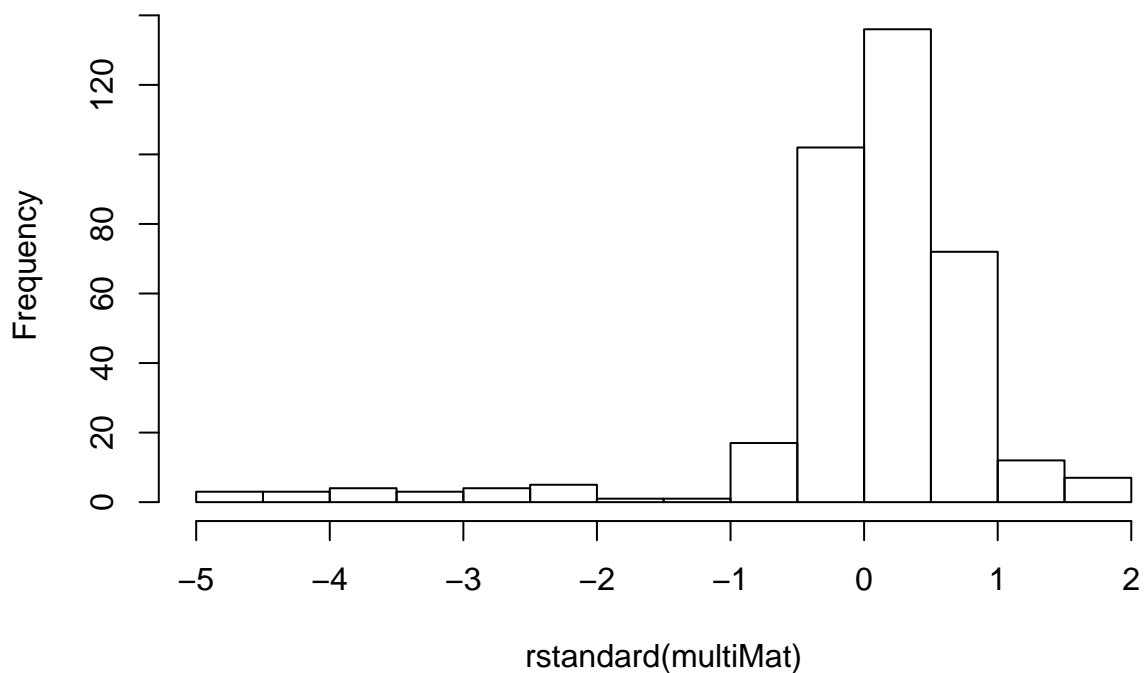
```
multiMat3 = lm(data = students_dummies, G3_mat ~ G2_mat + G1_mat + age)
summary(multiMat3)
##
## Call:
## lm(formula = G3_mat ~ G2_mat + G1_mat + age, data = students_dummies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.3833 -0.3680  0.2611  0.9880  3.4312
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.75584    1.50771   0.501  0.6164
## G2_mat       0.97551    0.05178  18.841 <2e-16 ***
## G1_mat       0.15411    0.05833   2.642  0.0086 **
## age         -0.14827    0.08643  -1.715  0.0871 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.928 on 366 degrees of freedom
## Multiple R-squared:  0.8264, Adjusted R-squared:  0.825
## F-statistic: 580.9 on 3 and 366 DF, p-value: < 2.2e-16
```

Provjerimo normalnost reziduala

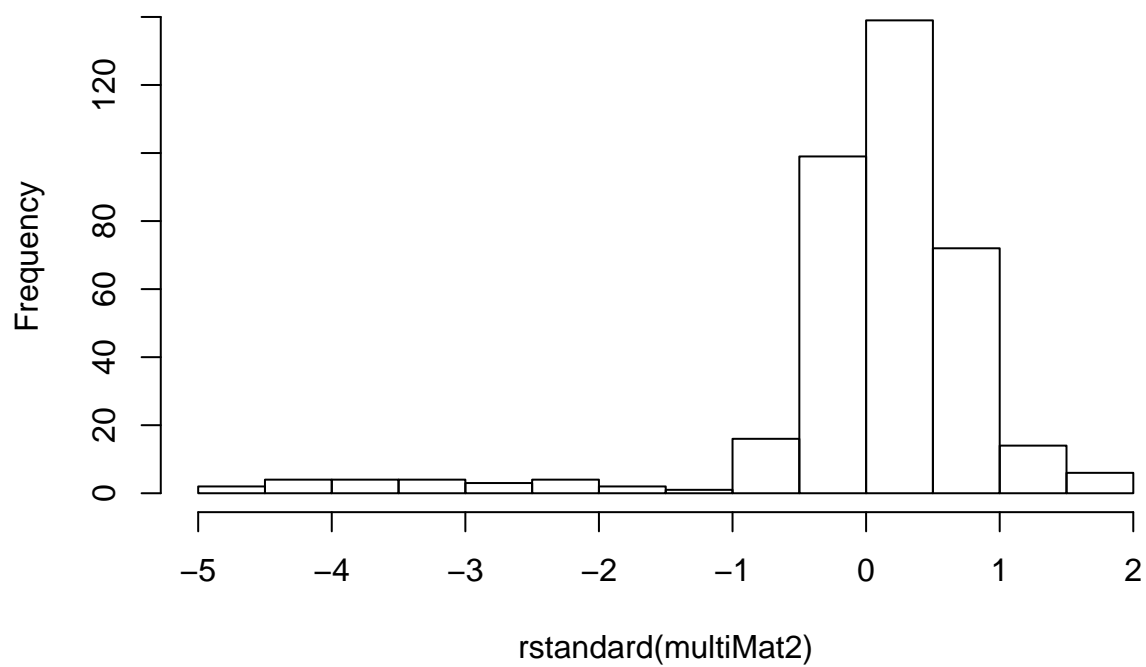
```
hist(rstandard(multiMat))
```

Histogram of rstandard(multiMat)



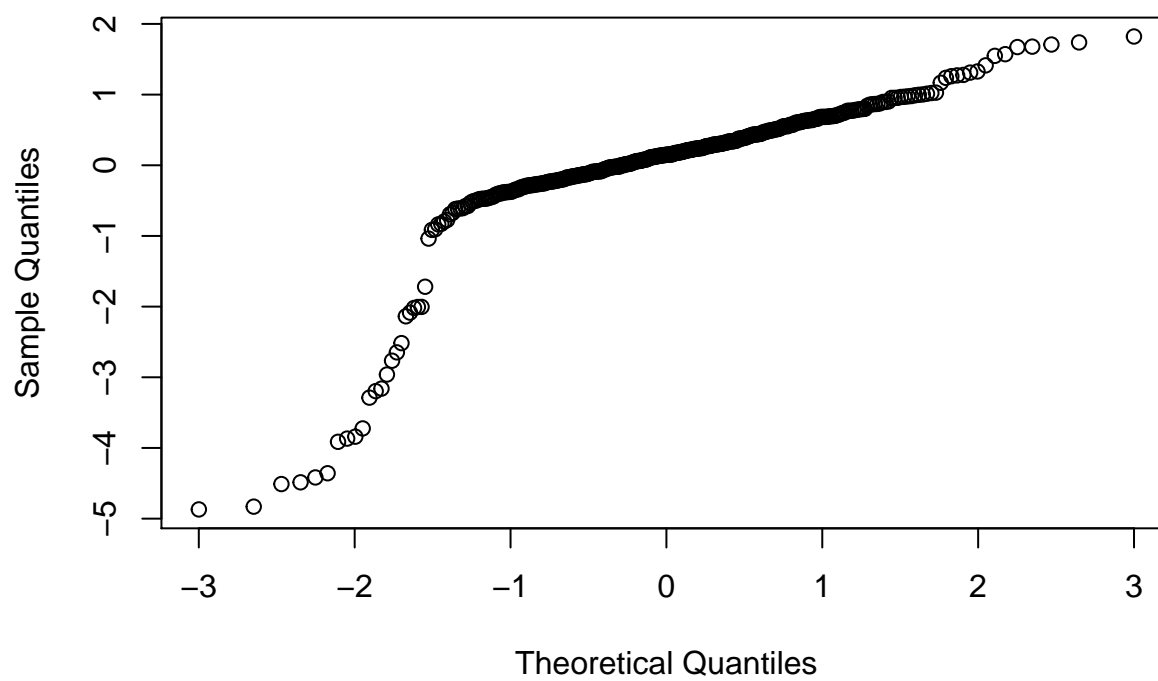
```
hist(rstandard(multiMat2))
```

Histogram of rstandard(multiMat2)



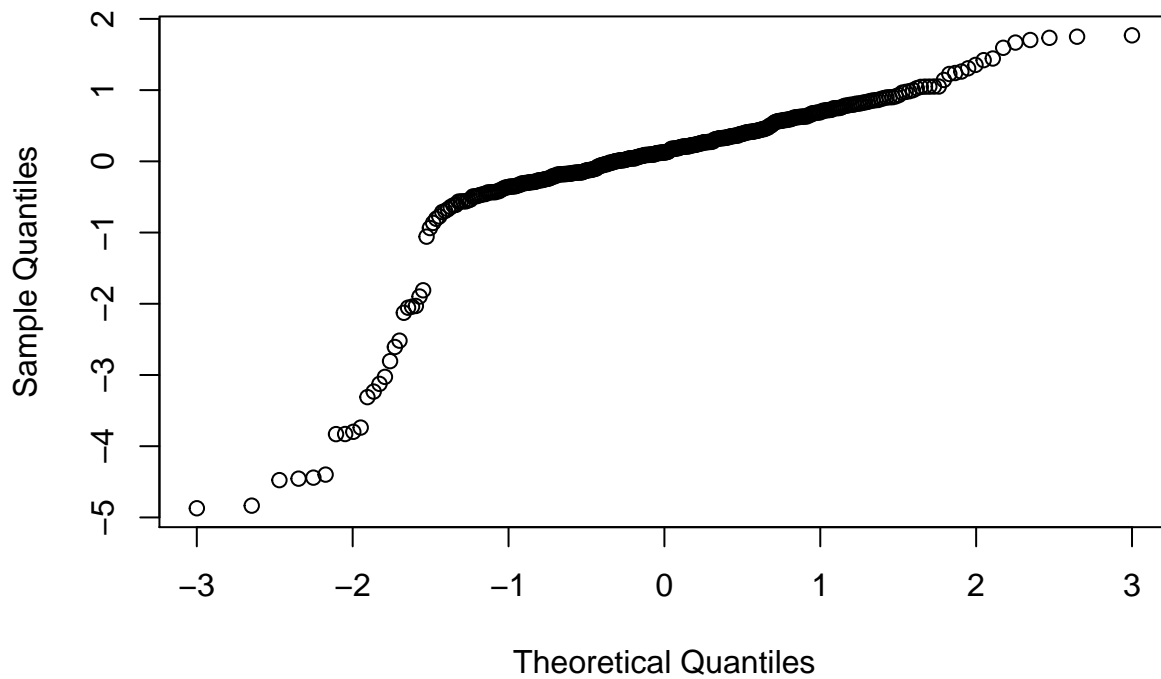
```
qqnorm(rstandard(multiMat2))
```

Normal Q-Q Plot



```
qqnorm(rstandard(multiMat2))
```

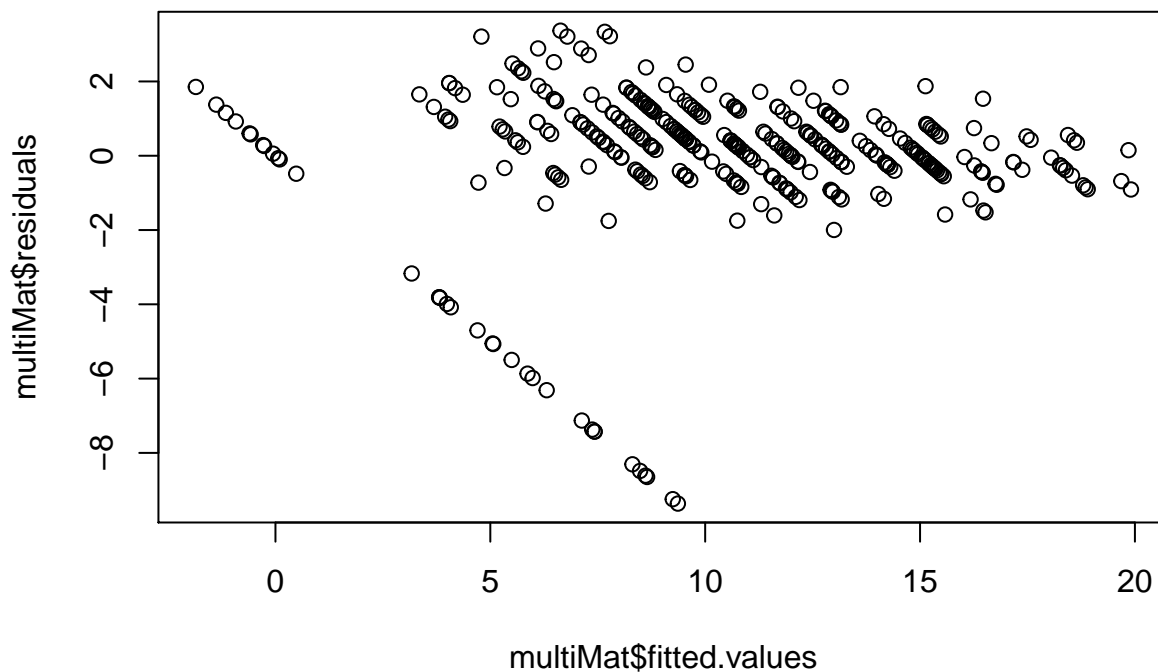
Normal Q-Q Plot



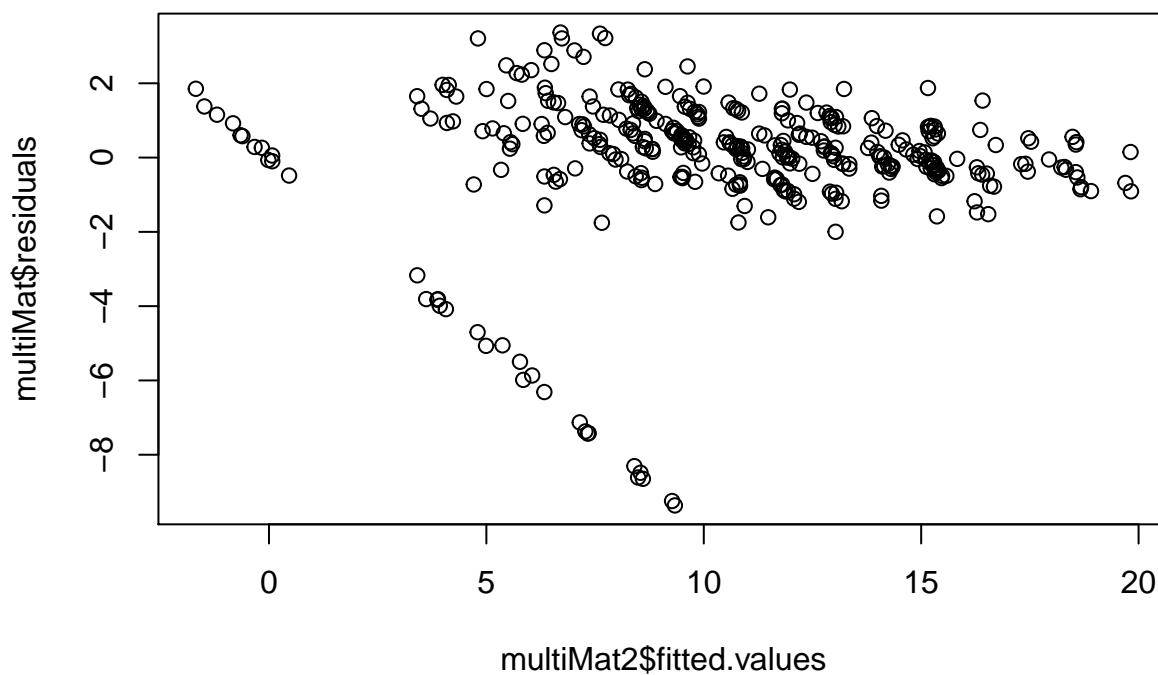
```
ks.test(rstandard(multiMat), "pnorm")
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(multiMat)
## D = 0.20494, p-value = 6.362e-14
## alternative hypothesis: two-sided
ks.test(rstandard(multiMat2), "pnorm")
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(multiMat2)
## D = 0.204, p-value = 8.438e-14
## alternative hypothesis: two-sided
```

Reziduali ne nalikuju normalnoj distribuciji. Promotrimo ih u ovisnosti o predviđenoj vrijednosti.

```
plot(multiMat$fitted.values, multiMat$residuals)
```



```
plot(multiMat2$fitted.values, multiMat$residuals)
```



Zanimljivo je još i pogledati koliko dobro možemo predvidjeti konačnu ocjenu iz matematike bez ikakvog znanja o drugim ocjenama, oslanjajući se na ostalih 6 od 10 najboljih prediktora

```
bezOcjenaMat = lm(data = students_dummies, G3_mat ~ failures_mat + higher_yes + Medu +
  age + absences_por + Fedu)
summary(bezOcjenaMat)
##
## Call:
## lm(formula = G3_mat ~ failures_mat + higher_yes + Medu + age +
##   absences_por + Fedu, data = students_dummies)
```



```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2296  -2.1146   0.2624   2.8734  11.1486
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.99438    3.67744   4.077 5.60e-05 ***
## failures_mat -1.87569    0.34079  -5.504 7.03e-08 ***
## higher_yes    1.46021    1.18270   1.235  0.2178
## Medu          0.58345    0.26284   2.220  0.0270 *
## age          -0.39114    0.19212  -2.036  0.0425 *
## absences_por -0.10462    0.04618  -2.265  0.0241 *
## Fedu         -0.06830    0.26292  -0.260  0.7952
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.192 on 363 degrees of freedom
## Multiple R-squared:  0.1861, Adjusted R-squared:  0.1727
## F-statistic: 13.84 on 6 and 363 DF,  p-value: 3.545e-14
```

Ovakav model objašnjava svega 18% varijance u promatranoj varijabli. Kao i očekivano same ocjene najbolji su prediktor konačne ocjene, ali i neke druge varijable nisu potpuno irelevantne.

Predviđanje konačne ocjene iz portugala

Pogledajmo koje su se varijable ispostavile najboljim prediktorima za G3_por (poredano po R^2 vrijednostima)

```
modelsPor[order(-modelsPor$rSquaredP), ]
##           varName      rSquaredP      pValueofFP
## 19           G2_por 0.7915470296 2.326282e-127
## 18           G1_por 0.6689900683 2.270811e-90
## 20          G3_total 0.6286216777 3.673507e-81
## 23           G_total 0.6244206858 2.920018e-80
## 22          G1_total 0.5674100983 6.010847e-69
## 21          G2_total 0.5496273344 1.011539e-65
## 16           G1_mat 0.3130606997 7.266864e-32
## 17           G2_mat 0.2654028560 1.805665e-26
## 6      failures_mat 0.1364018388 2.116052e-13
## 7      failures_por 0.1159509571 1.697080e-11
## 48          higher_yes 0.0940955134 1.676684e-09
## 11           Dalc 0.0775925031 5.088939e-08
## 5      studytime 0.0743472469 9.907920e-08
## 12           Walc 0.0568741710 3.502791e-06
## 25           sex_M 0.0410010252 8.765489e-05
## 2           Medu 0.0405516662 9.601895e-05
## 26          address_U 0.0399328014 1.088617e-04
## 24          school_MS 0.0307910847 6.983446e-04
## 13           health 0.0307821478 6.996189e-04
## 3           Fedu 0.0272816595 1.431045e-03
## 4      traveltime 0.0250405704 9.528638e-03
## 30          Mjob_other 0.0218326465 4.396327e-03
## 42      schoolsup_yes 0.0200072403 6.424402e-03
## 10           goout 0.0182977031 9.183912e-03
```

```
## 45      paid_por_yes 0.0182200648 9.334687e-03
## 38      reason_other 0.0160759079 1.466799e-02
## 39 reason_reputation 0.0154737362 1.666733e-02
## 32      Mjob_teacher 0.0152059363 1.764436e-02
## 15      absences_por 0.0143263204 2.128856e-02
## 29      Mjob_health 0.0132736154 2.669036e-02
## 46      activities_yes 0.0130844904 2.780209e-02
## 9       freetime 0.0113386701 4.064452e-02
## 14      absences_mat 0.0111220559 4.262370e-02
## 43      famsup_yes 0.0090146763 6.811285e-02
## 44      paid_mat_yes 0.0080125876 8.553334e-02
## 49      internet_yes 0.0068851590 1.110615e-01
## 36      Fjob_teacher 0.0067022942 1.159355e-01
## 28      Pstatus_T 0.0057079539 1.469445e-01
## 47      nursery_yes 0.0054854718 1.550896e-01
## 1       age 0.0038926213 2.312215e-01
## 34      Fjob_other 0.0027946130 3.105208e-01
## 41      guardian_other 0.0016055476 4.422206e-01
## 50      romantic_yes 0.0015392250 4.518122e-01
## 35      Fjob_services 0.0015298339 4.531965e-01
## 8       famrel 0.0014968789 4.581070e-01
## 27      famsize_LE3 0.0012632977 4.955024e-01
## 37      reason_home 0.0009382419 5.569766e-01
## 40      guardian_mother 0.0008942898 5.663672e-01
## 33      Fjob_health 0.0006006413 6.384293e-01
## 31      Mjob_services 0.0001345507 8.240217e-01
```

Razmotrit ćemo prvih 13 najboljih prediktora. Najprije provjerimo jesu li neke od tih varijabli visoko korelirane:

```
cor(cbind(students_dummies$G2_por, students_dummies$G1_por, students_dummies$G1_mat,
students_dummies$G2_mat, students_dummies$failures_mat, students_dummies$failures_por,
students_dummies$higher_yes, students_dummies$Dalc, students_dummies$studytime,
students_dummies$Walc, students_dummies$ssex_M, students_dummies$Medu, students_dummies$address_U))

##      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,] 1.0000000 0.8874806 0.59826670 0.57804548 -0.35088560 -0.2967996
## [2,] 0.8874806 1.0000000 0.58101916 0.53820225 -0.29128438 -0.2897208
## [3,] 0.5982667 0.5810192 1.00000000 0.85677052 -0.38630528 -0.1219066
## [4,] 0.5780455 0.5382023 0.85677052 1.00000000 -0.36862012 -0.1074594
## [5,] -0.3508856 -0.2912844 -0.38630528 -0.36862012 1.00000000 0.4855861
## [6,] -0.2967996 -0.2897208 -0.12190662 -0.10745935 0.48558611 1.0000000
## [7,] 0.2985681 0.2771680 0.22380010 0.21360606 -0.36598055 -0.2956059
## [8,] -0.2715077 -0.2466711 -0.08283751 -0.05289544 0.14114646 0.1993841
## [9,] 0.2666218 0.2577811 0.14209034 0.11349573 -0.18644632 -0.1965468
## [10,] -0.2328534 -0.2004985 -0.10744556 -0.07119278 0.16443018 0.1910275
## [11,] -0.1919346 -0.1819224 0.12463492 0.11910003 0.04030618 0.1419387
## [12,] 0.2150699 0.2009143 0.22205178 0.21293884 -0.22315525 -0.1899768
## [13,] 0.1961527 0.1845087 0.06894890 0.13062003 -0.05941857 -0.0571007
##      [,7]      [,8]      [,9]      [,10]      [,11]      [,12]
## [1,] 0.29856805 -0.27150767 0.26662180 -0.23285336 -0.19193462 0.21506989
## [2,] 0.27716795 -0.24667108 0.25778111 -0.20049854 -0.18192240 0.20091426
## [3,] 0.22380010 -0.08283751 0.14209034 -0.10744556 0.12463492 0.22205178
## [4,] 0.21360606 -0.05289544 0.11349573 -0.07119278 0.11910003 0.21293884
## [5,] -0.36598055 0.14114646 -0.18644632 0.16443018 0.04030618 -0.22315525
```

```
## [6,] -0.29560592 0.19938413 -0.19654679 0.19102755 0.14193871 -0.18997679
## [7,] 1.00000000 -0.09292552 0.16777164 -0.11633210 -0.14457663 0.15718279
## [8,] -0.09292552 1.00000000 -0.19163979 0.65415550 0.25606612 0.04130251
## [9,] 0.16777164 -0.19163979 1.00000000 -0.25430441 -0.28491033 0.05368487
## [10,] -0.11633210 0.65415550 -0.25430441 1.00000000 0.26642194 -0.02557592
## [11,] -0.14457663 0.25606612 -0.28491033 0.26642194 1.00000000 0.09005189
## [12,] 0.15718279 0.04130251 0.05368487 -0.02557592 0.09005189 1.00000000
## [13,] 0.04811237 -0.10087244 -0.01927545 -0.09195948 -0.02211323 0.13772096
##      [,13]
## [1,] 0.19615266
## [2,] 0.18450872
## [3,] 0.06894890
## [4,] 0.13062003
## [5,] -0.05941857
## [6,] -0.05710070
## [7,] 0.04811237
## [8,] -0.10087244
## [9,] -0.01927545
## [10,] -0.09195948
## [11,] -0.02211323
## [12,] 0.13772096
## [13,] 1.00000000
```

Otprije znamo za visoku koreiranost ocjena, a učimo još i visoku koreliranost razina konzumacija alkohola vikendom i radnim danima.

```
cor(students_dummies$Dalc, students_dummies$Walc)
## [1] 0.6541555
```

Zasad ne odbacujući nijedan regresor izradimo linearni model za prethodno izdvojenih 13 varijabli.

```
multiPor = lm(data = students_dummies, G3_por ~ G2_por + G1_por + G1_mat + G2_mat +
  failures_mat + failures_por + higher_yes + Dalc + studytime + Walc + sex_M +
  Medu + address_U)
summary(multiPor)
##
## Call:
## lm(formula = G3_por ~ G2_por + G1_por + G1_mat + G2_mat + failures_mat +
##     failures_por + higher_yes + Dalc + studytime + Walc + sex_M +
##     Medu + address_U, data = students_dummies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.5864 -0.4875 -0.0354  0.6254  5.6542
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.15002    0.58098  -0.258   0.7964
## G2_por         0.86135    0.06451  13.352 <2e-16 ***
## G1_por         0.12309    0.05982   2.058  0.0403 *
## G1_mat         0.09050    0.04275   2.117  0.0350 *
## G2_mat        -0.05491    0.03605  -1.523  0.1286
## failures_mat  -0.11933    0.12386  -0.963  0.3360
## failures_por  -0.36065    0.16981  -2.124  0.0344 *
## higher_yes     0.24102    0.37438   0.644  0.5201
```

```
## Dalc      -0.07723    0.10418   -0.741    0.4590
## studytime  0.05693    0.08871    0.642    0.5215
## Walc      0.01026    0.07215    0.142    0.8870
## sex_M     -0.15408    0.15683   -0.982    0.3265
## Medu     -0.01481    0.06755   -0.219    0.8266
## address_U  0.23738    0.17234    1.377    0.1693
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.316 on 356 degrees of freedom
## Multiple R-squared:  0.8072, Adjusted R-squared:  0.8001
## F-statistic: 114.6 on 13 and 356 DF,  p-value: < 2.2e-16
```

Pojednostavimo sad uzevši 8 varijabli s najnižim p-vrijednostima:

```
multiPor2 = lm(data = students_dummies, G3_por ~ G2_por + failures_por + G1_mat +
  G1_por + G2_mat + address_U + sex_M + failures_mat)
summary(multiPor2)
##
## Call:
## lm(formula = G3_por ~ G2_por + failures_por + G1_mat + G1_por +
##     G2_mat + address_U + sex_M + failures_mat, data = students_dummies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.6591 -0.4742 -0.0609  0.6474  5.3729
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.02710    0.41635   -0.065  0.9481
## G2_por        0.86980    0.06369   13.656 <2e-16 ***
## failures_por -0.38679    0.16620   -2.327  0.0205 *
## G1_mat        0.09229    0.04238    2.178  0.0301 *
## G1_por        0.12592    0.05941    2.120  0.0347 *
## G2_mat       -0.05733    0.03576   -1.603  0.1098
## address_U     0.23275    0.16954    1.373  0.1707
## sex_M       -0.21096    0.14718   -1.433  0.1526
## failures_mat -0.14099    0.11997   -1.175  0.2407
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.31 on 361 degrees of freedom
## Multiple R-squared:  0.8063, Adjusted R-squared:  0.802
## F-statistic: 187.8 on 8 and 361 DF,  p-value: < 2.2e-16
```

Ovaj skup varijabli ispostavlja se daje najveći prilagođeni R^2 :

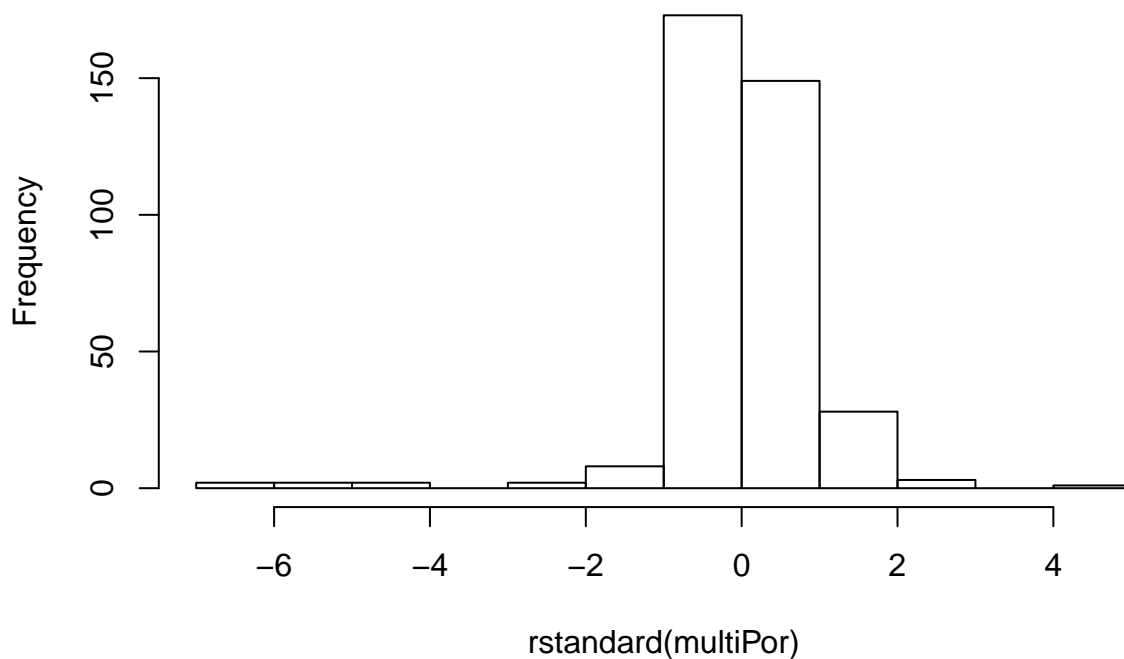
```
# npr ograničavajući na 6 regresora smanjuje se prilagođeni R^2
multiPor3 = lm(data = students_dummies, G3_por ~ G2_por + failures_por + G1_mat +
  G1_por + G2_mat + sex_M)
summary(multiPor3)
##
## Call:
## lm(formula = G3_por ~ G2_por + failures_por + G1_mat + G1_por +
##     G2_mat + sex_M, data = students_dummies)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.6223 -0.4462 -0.0743  0.6620  5.4957
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.07026    0.39699  -0.177  0.85963
## G2_por        0.88397    0.06328  13.968 < 2e-16 ***
## failures_por -0.47763    0.14722  -3.244  0.00129 **
## G1_mat        0.09189    0.04163   2.207  0.02792 *
## G1_por        0.12049    0.05899   2.043  0.04182 *
## G2_mat       -0.04862    0.03549  -1.370  0.17156
## sex_M        -0.20984    0.14731  -1.425  0.15515
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.312 on 363 degrees of freedom
## Multiple R-squared:  0.8045, Adjusted R-squared:  0.8013
## F-statistic: 249 on 6 and 363 DF, p-value: < 2.2e-16
```

Provjerimo još normalnost reziduala:

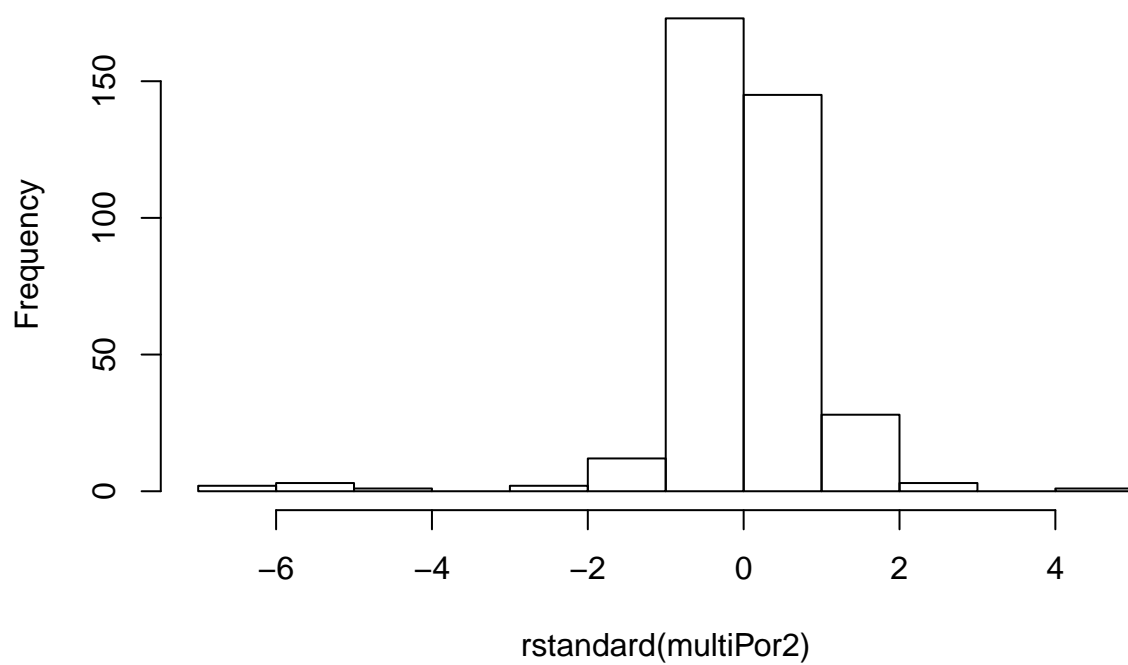
```
hist(rstandard(multiPor))
```

Histogram of rstandard(multiPor)



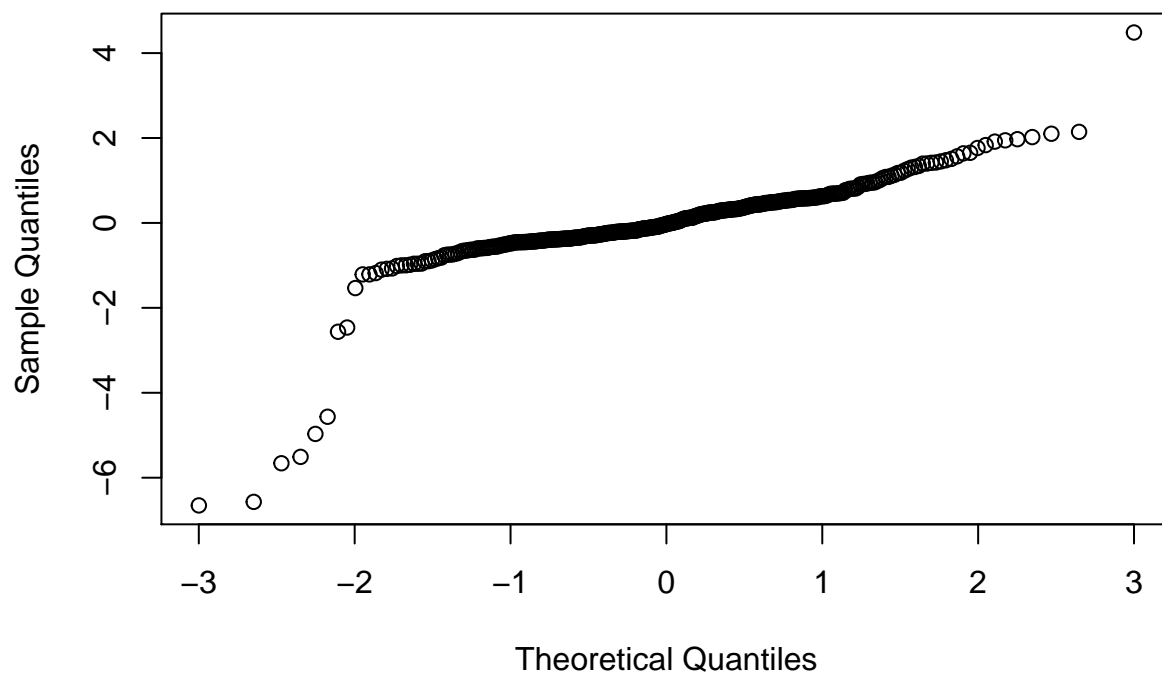
```
hist(rstandard(multiPor2))
```

Histogram of rstandard(multiPor2)



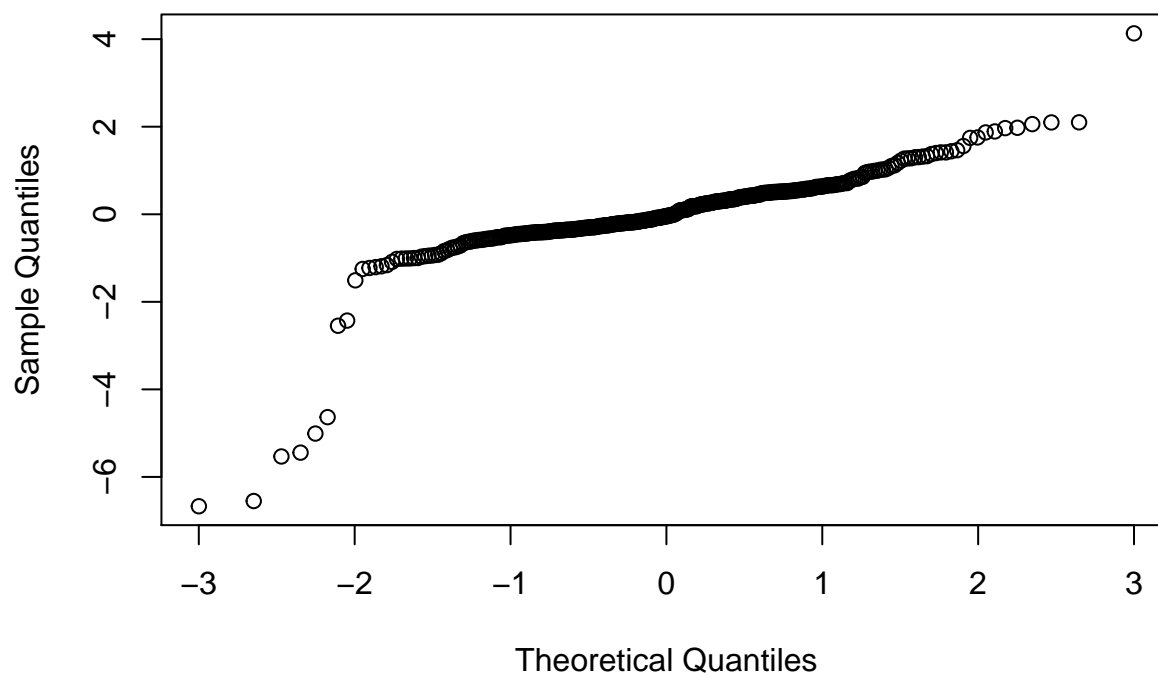
```
qqnorm(rstandard(multiPor2))
```

Normal Q-Q Plot



```
qqnorm(rstandard(multiPor2))
```

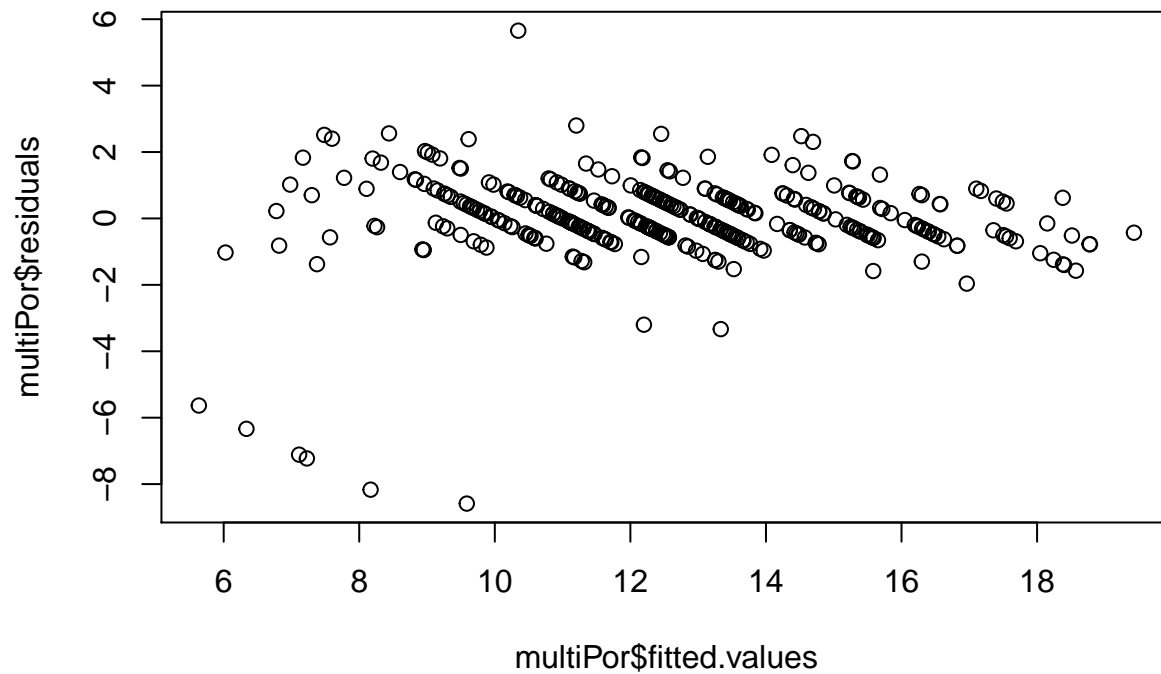
Normal Q-Q Plot



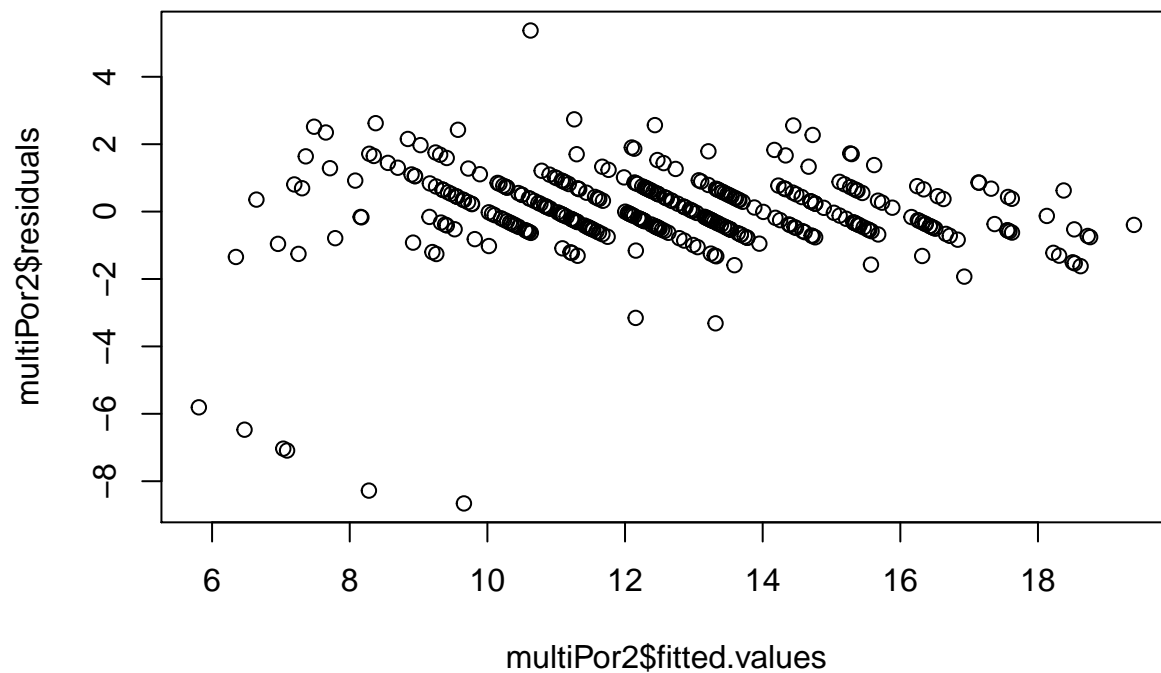
```
ks.test(rstandard(multiPor), "pnorm")
##
## One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(multiPor)
## D = 0.16271, p-value = 6.199e-09
## alternative hypothesis: two-sided
ks.test(rstandard(multiPor2), "pnorm")
##
## One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(multiPor2)
## D = 0.16601, p-value = 2.778e-09
## alternative hypothesis: two-sided
```

Reziduali nalikuju normalnoj distribuciji nešto više nego kod modela za konačnu ocjenu iz matematike, ali i dalje ne osobito. Promotrimo ih u ovisnosti o predviđenoj vrijednosti.

```
plot(multiPor$fitted.values, multiPor$residuals)
```



```
plot(multiPor2$fitted.values, multiPor2$residuals)
```



Promotrimo još koliko dobro možemo predvidjeti konačnu ocjenu iz portugalskog bez znanja o drugim ocjenama, oslanjajući se na ostalih 9/13 najboljih prediktora:

```
bezOcjenaPor = lm(data = students_dummies, G3_por ~ failures_mat + failures_por +
  higher_yes + Dalc + studytime + Walc + sex_M + Medu + address_U)
summary(bezOcjenaPor)
##
## Call:
## lm(formula = G3_por ~ failures_mat + failures_por + higher_yes +
##     Dalc + studytime + Walc + sex_M + Medu + address_U, data = students_dummies)
```



```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4097  -1.3974  -0.0087   1.5525   6.9747
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.4176     0.9259  10.171 < 2e-16 ***
## failures_mat  -0.7558     0.2211  -3.419 0.000701 ***
## failures_por  -0.6845     0.3143  -2.178 0.030072 *
## higher_yes     1.9206     0.7032   2.731 0.006618 **
## Dalc          -0.5468     0.1952  -2.802 0.005359 **
## studytime      0.4769     0.1664   2.866 0.004399 **
## Walc           0.0275     0.1367   0.201 0.840629
## sex_M         -0.5150     0.2840  -1.813 0.070605 .
## Medu          0.2870     0.1268   2.264 0.024192 *
## address_U      1.0420     0.3199   3.257 0.001234 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.5 on 360 degrees of freedom
## Multiple R-squared:  0.2962, Adjusted R-squared:  0.2786
## F-statistic: 16.83 on 9 and 360 DF,  p-value: < 2.2e-16
```

Model bez ocjena za portugalski objašnjava skoro 30% varijance u promatranoj varijabli. Značajno poboljšanje u odnosu na model bez ocjena za matematiku.