Vilniaus Universitetas

Kovariacinė analizė

Laboratorinis darbas

Darbą atliko:

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus

Duomenų Mokslas

3 kursas 2 gr.

Vilnius, 2021

Naudoti metodai

Darbas atliktas naudojant R ir SAS.

Naudoti R paketai:

tidyverse

janitor

car

rstatix

Duomenys ir jų šaltiniai

JAV moksleivių egzaminų balai pagal su šeima, mokymusi susijusius rodiklius.

Duomenų šaltinis - Kaggle. Prieiga per internetą: https://www.kaggle.com/rsasma/high-school-grad-performance

Duomenis sudaro šie stulpeliai:

„Gender“ – moksleivio lytis.

„Race“ – moksleivio rasė.

„Parental_Education“ – tėvų išsilavinimas.

„Special_Coaching“ – ar studentas laikė specialų pasiruošimą egzaminams.

„Attendance“ – lankomumas (proc. pamokų).

„DailyStudy_Hours“ – laikas, praleistas mokantis per dieną.

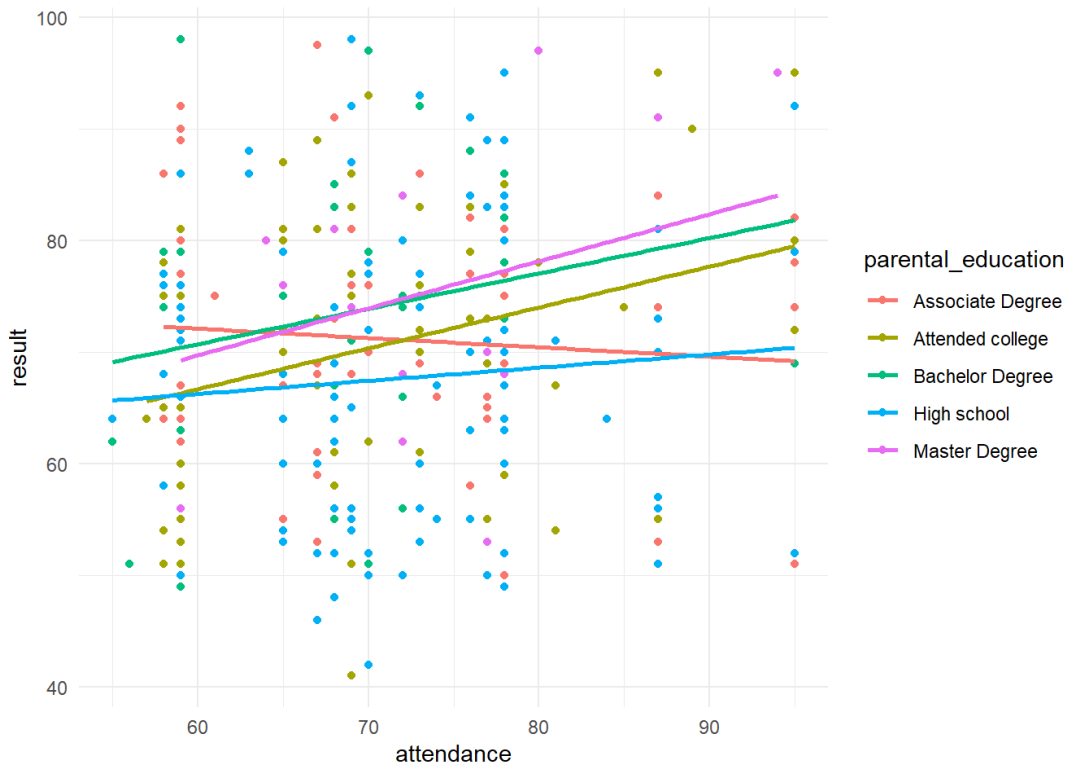
„Result“ – egzaminų rezultatų balas.

Atliktos analizės aprašymas

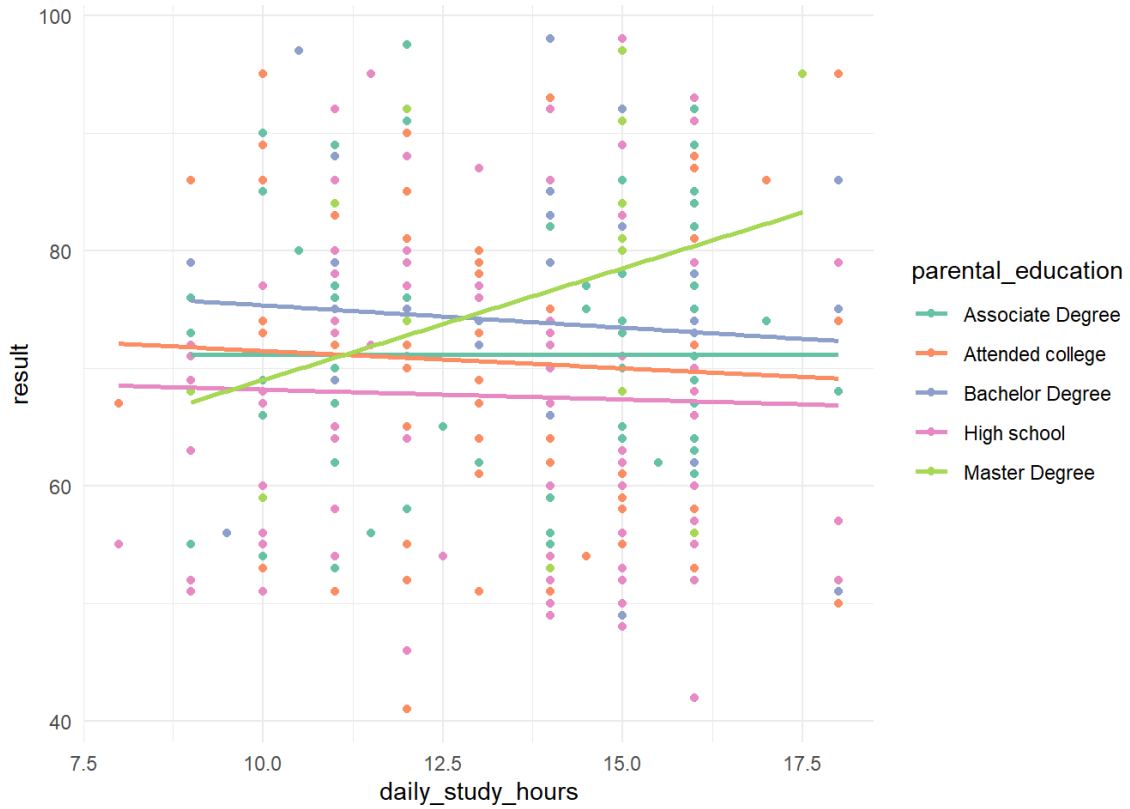
1. Naudojant R

Tikslas: rasti kokią įtaką egzaminų rezultatų vidurkiams turi tėvų išsilavinimas atsižvelgiant į papildomų kintamųjų įtaką.

```
library(tidyverse)
library(car)
library(readxl)
library(janitor)

x<-readxl::read_xlsx("HighSchool.xlsx",sheet = 1) %>% clean_names()
x <- x %>%
  drop_na()

# Duomenys išsaugomi į failą
write_csv(x,"high_school_modified.csv")


# Koeficiento lygybės skirtingiems faktoriaus lygmenims patikrinimas
ggplot(x,aes(attendance,result,color=parental_education)) + geom_point() + geom_smooth(method="lm",se=F
ALSE) +
  theme_minimal()
```

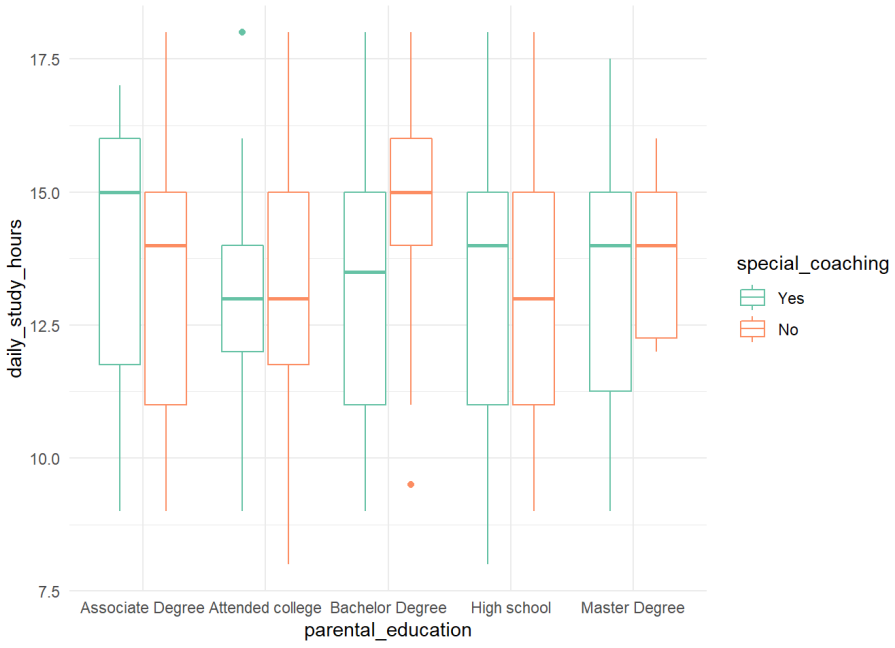


```
ggplot(x,aes(daily_study_hours,result,color=parental_education)) + geom_point() +
  geom_smooth(method="lm",se=FALSE) + theme_minimal() + scale_color_brewer(palette="Set2")
```

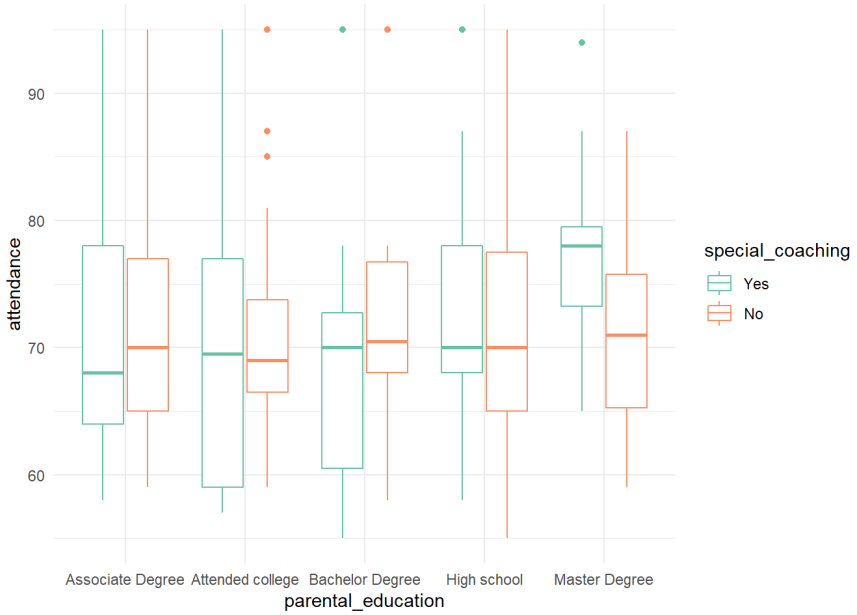


Pasirinktos papildomos kovariantės: valandų praleistų per dieną mokantis kiekis (stulp. „daily_study_hours“) ir lankomumas procentais (stulp. „attendance“). Tarp abiejų kovariančių rastas faktoriaus lygmuo, kurio krypties koeficientas skiriasi nuo likusių. Reikalingas patikrinimas ar šie skirtumai statistiškai reikšmingi.

```
ggplot(x,aes(x=parental_education,y=daily_study_hours,color=special_coaching)) + geom_boxplot() + theme
_minimal() +
scale_color_brewer(palette="Set2")
```



```
ggplot(x,aes(x=parental_education,y=attendance,color=special_coaching)) +
  geom_boxplot() + theme_minimal() + scale_color_brewer(palette="Set2")
```



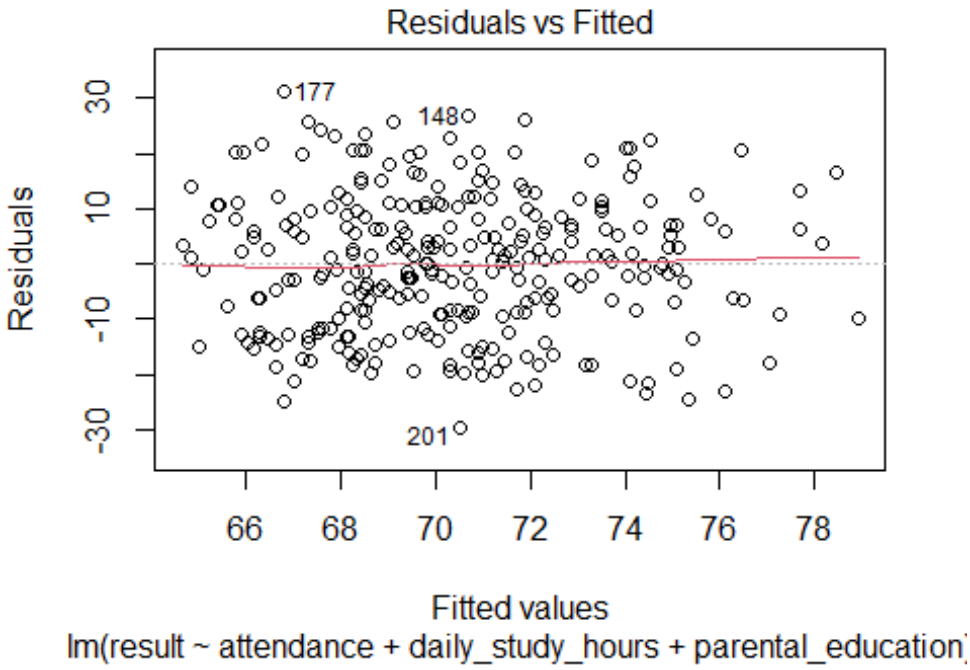
Modelio prieldaidos

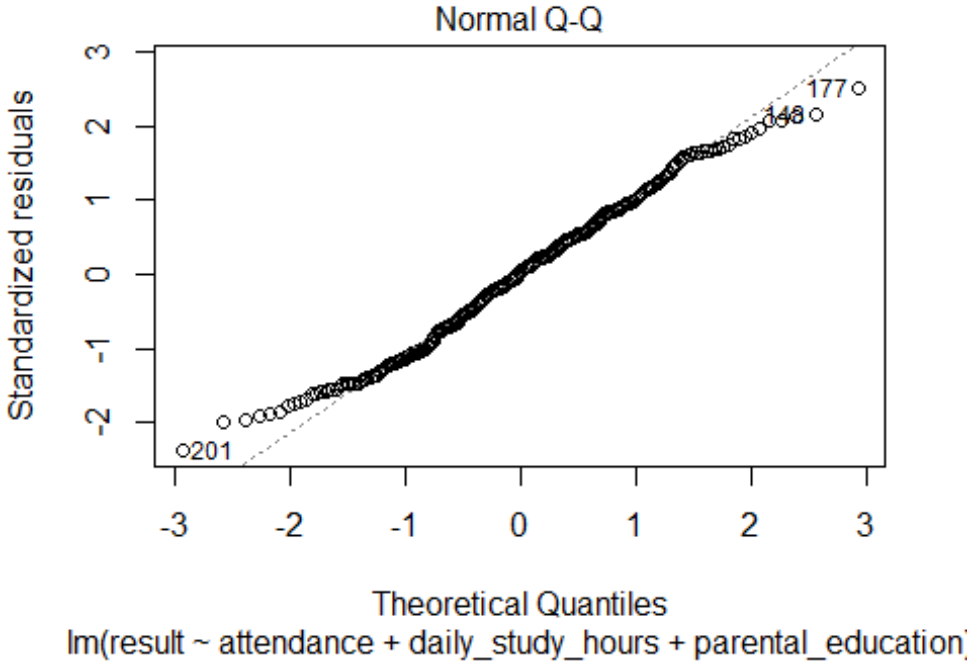
```
library(rstatix)

anova_test(result~attendance*parental_education + daily_study_hours*parental_education,data=x,type=3, d
etailed=TRUE) # Hipotezės apie koeficientų lygybę visiems faktoriaus lygmenims neatmetos
## ANOVA Table (type III tests)
##
##                                 Effect      SSn      SSd DFn DFd      F
## 1                           (Intercept) 7177.745 43739.79   1 279 45.784
## 2                            attendance  811.357 43739.79   1 279  5.175
## 3                    parental_education  547.638 43739.79   4 279  0.873
## 4                     daily_study_hours    9.655 43739.79   1 279  0.062
## 5         attendance:parental_education  781.693 43739.79   4 279  1.247
## 6 parental_education:daily_study_hours  316.544 43739.79   4 279  0.505
##          p p<.05      ges
## 1 7.75e-11     * 0.141000
## 2 2.40e-02     * 0.018000
## 3 4.80e-01       0.012000
## 4 8.04e-01       0.000221
## 5 2.91e-01       0.018000
## 6 7.32e-01       0.007000
# Kovariacinės analizės modelio sukūrimas
model <- anova_test(result~attendance + daily_study_hours + parental_education,data=x,type=3, detailed=
TRUE)
model
## ANOVA Table (type III tests)
##
##               Effect       SSn      SSd DFn DFd      F        p p<.05   ges
## 1        (Intercept) 14064.638 44881.75   1 287 89.937 9.78e-19     * 0.239
## 2         attendance   800.888 44881.75   1 287  5.121 2.40e-02     * 0.018
## 3  daily_study_hours    59.400 44881.75   1 287  0.380 5.38e-01       0.001
## 4 parental_education  1821.330 44881.75   4 287  2.912 2.20e-02     * 0.039


model_lm <- lm(result~attendance + daily_study_hours + parental_education,data=x)

# Dispersijų lygybė grupėms
leveneTest(result~parental_education,data=x)
```

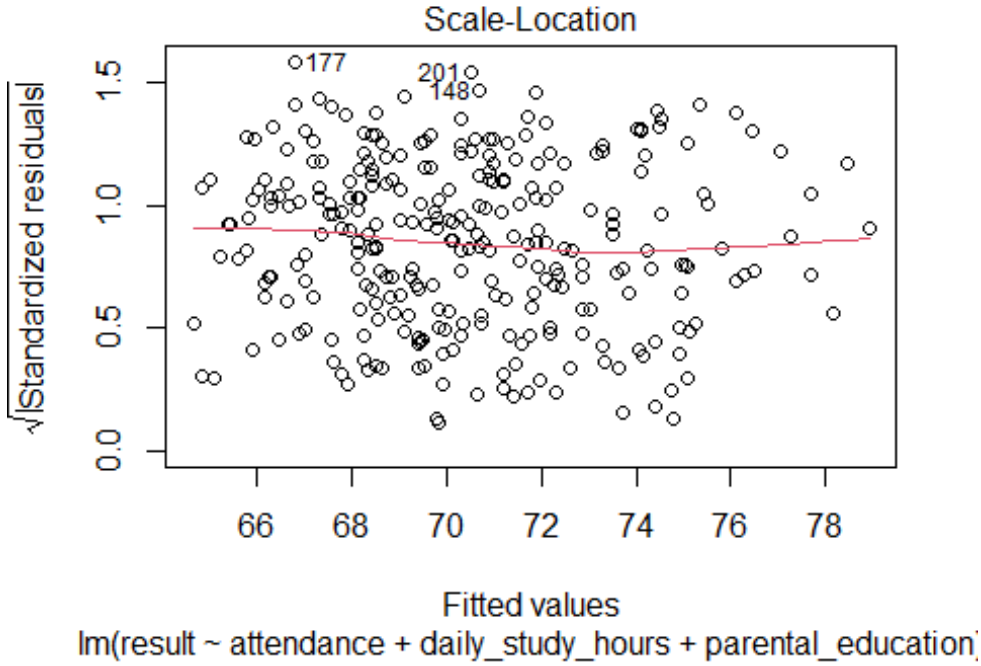
```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  0.5562 0.6946
##       289
# Liekanų normalumas
shapiro.test(resid(model_lm))
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_lm)
## W = 0.98791, p-value = 0.01471
plot(model_lm)
```

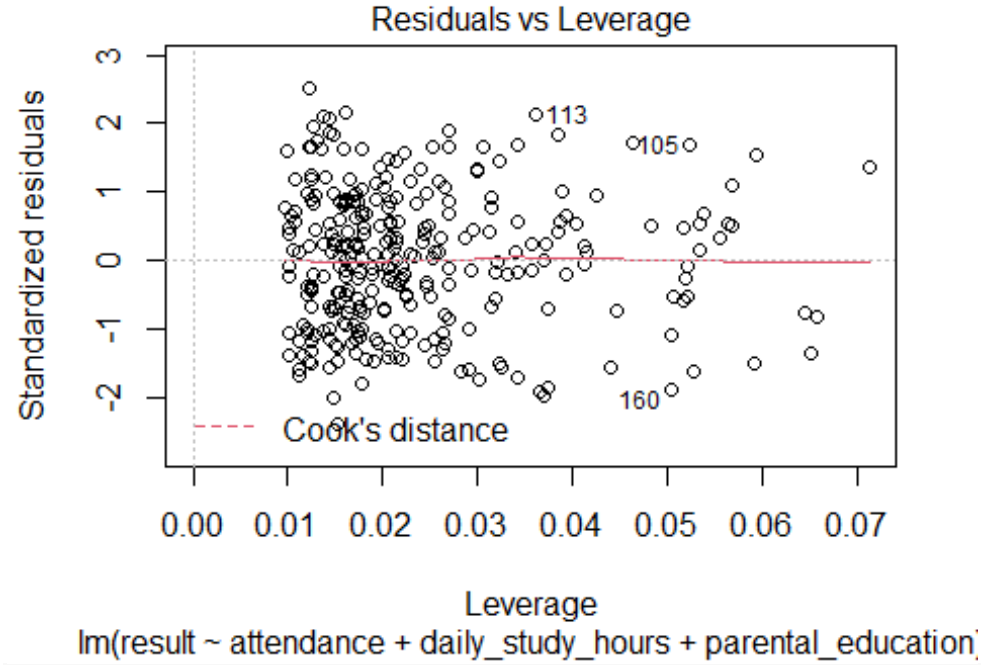
Hipotezės apie krypties koeficientų lygybę neatmetamos nei “attendance” (p=0.29), nei “daily_study_hours” (p=0.67) kovariantėms, todėl toliau taikomas kovariacinės analizės modelis.

Hipotezės apie dispersijų lygybę grupėms ir liekanų normalumą neatmetamos.

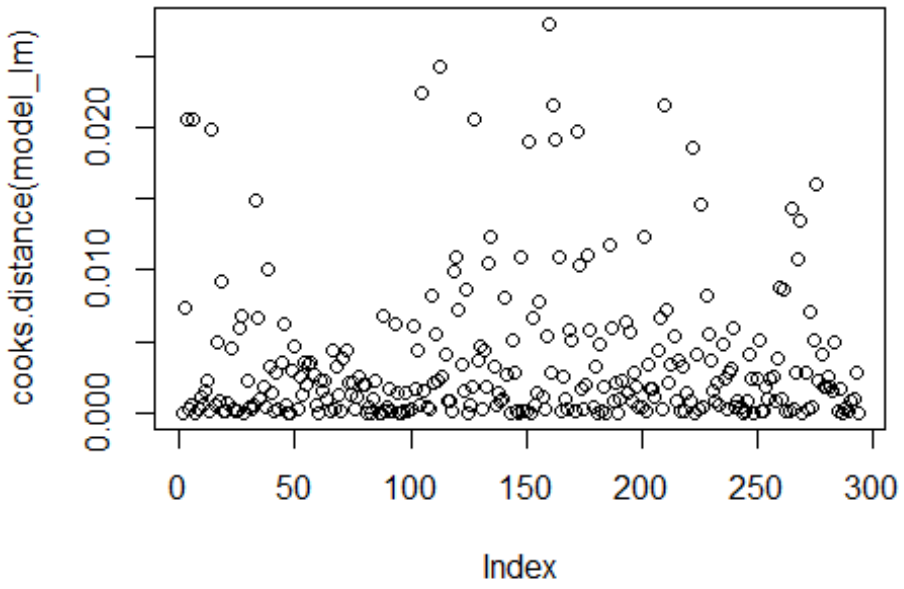

Residuals vs Fitted
Residuals
Fitted values
lm(result ~ attendance + daily_study_hours + parental_education)
177
148
201
30
10
-10
-30
66
68
70
72
74
76
78


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(result ~ attendance + daily_study_hours + parental_education)
177
148
201
3
2
1
0
-1
-2
-3

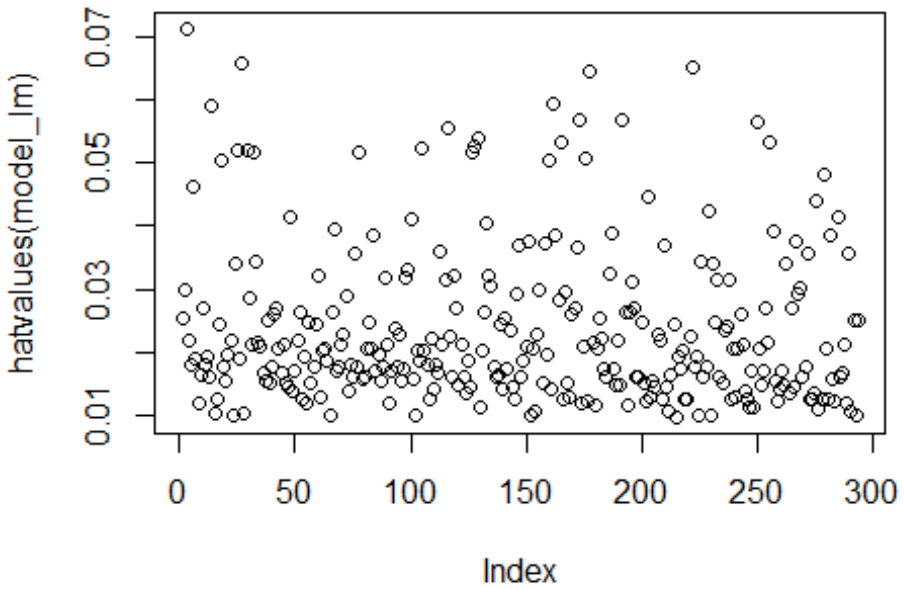





```r
# Išskirtys
plot(cooks.distance(model_lm))
```




```
plot(hatvalues(model_lm))
```



Post-hoc vidurkių palyginimai

Statistiškai reikšminga kovariantė “attendance” (p=0.24). Faktoriaus “parental_education” įtaka statistiškai reikšminga (p=0.02). Šiam faktoriui atlikti poriniai palyginimai naudojant Bonferroni pataisą siekiant atrasti statistiškai reikšmingas egzaminų rezultatų vidurkių pagal tėvų išsilavinimą skirtumų poras atsižvelgiant į kovariantes.

```
library(emmeans)

means <- emmeans_test(result~parental_education,covariate = c(daily_study_hours,attendance),p.adjust.me
thod="bonferroni",data=x)
means
```

```
## # A tibble: 10 x 9
##    term        .y.   group1  group2     df statistic       p  p.adj p.adj.signif
##  * <chr>       <chr> <chr>   <chr>   <dbl>     <dbl>   <dbl>  <dbl> <chr>
##  1 daily_stud~ resu~ Associ~ Attend~   287     0.261 0.794   1.00   ns
##  2 daily_stud~ resu~ Associ~ Bachel~   287    -1.12  0.264   0.953  ns
##  3 daily_stud~ resu~ Associ~ High s~   287     1.88  0.0612  0.468  ns
##  4 daily_stud~ resu~ Associ~ Master~   287    -1.26  0.210   0.906  ns
##  5 daily_stud~ resu~ Attend~ Bachel~   287    -1.33  0.184   0.869  ns
##  6 daily_stud~ resu~ Attend~ High s~   287     1.61  0.109   0.685  ns
##  7 daily_stud~ resu~ Attend~ Master~   287    -1.43  0.153   0.810  ns
##  8 daily_stud~ resu~ Bachel~ High s~   287     2.63  0.00908 0.0872 ns
##  9 daily_stud~ resu~ Bachel~ Master~   287    -0.277 0.782   1.00   ns
## 10 daily_stud~ resu~ High s~ Master~   287    -2.51  0.0127  0.120  ns
```

```
get_emmeans(means)
```

```
## # A tibble: 5 x 9
##   daily_study_hours attendance parental_education emmean    se    df conf.low
##               <dbl>      <dbl> <fct>                <dbl> <dbl> <dbl>    <dbl>
## 1              13.3       71.3 Associate Degree      71.2  1.52   287     68.2
## 2              13.3       71.3 Attended college      70.7  1.50   287     67.7
## 3              13.3       71.3 Bachelor Degree       74.2  2.22   287     69.9
## 4              13.3       71.3 High school           67.5  1.23   287     65.1
## 5              13.3       71.3 Master Degree         75.2  2.81   287     69.7
## # ... with 2 more variables: conf.high <dbl>, method <chr>
```

Visos egzaminų rezultatų vidurkių poros pagal tėvų išsilaivinimą atsižvelgiant į papildomas kovariantes statistiškai reikšmingas nesiskyrė.

Rezultatai

Koviariacine analize (ANCOVA) siekta rasti kokią įtaką moksleivių egzaminų rezultatams turi tėvų išsilavinimas atsižvelgiant į valandų, praleistų mokantis kiekį (stulp. „daily_study_hours“) ir lankomumą procentais (stulp. „attendance“).

Atsižvelgiant į anksčiau minėtas kovariantes rasta statistiškai reikšminga tėvų išsilavinimo įtaka (*F= 2.91 p=0.02*).

Post-hoc poriniai vidurkių palyginimai atlikti naudojant Bonferroni pataisą, tačiau statistiškai reikšmingų skirtumų nerasta.

Rasta statistiškai reikšminga lankomumo įtaka (stulp. „attendance“ *p=0.02*). Praleistų mokantis valandų įtaka nebuvo statistiškai reikšminga.

2. Naudojant SAS

```
PROC IMPORT DATAFILE='/home/u45871880/high_school_modified.csv'
       DBMS=CSV
       OUT=data;
       GETNAMES=YES;
RUN;


/* Hipotezė apie krypties koeficientų lygybę*/
PROC GLM DATA=data; CLASS parental_education special_coaching;
MODEL result = parental_education daily_study_hours attendance
daily_study_hours*parental_education attendance*parental_education / SS3;
RUN;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
parental_education	4	1821.330107	455.332527	2.91	0.0219
daily_study_hours	1	59.399971	59.399971	0.38	0.5382
attendance	1	800.888158	800.888158	5.12	0.0244

```
/* Modelio prielaidos */
/* Vidurkių palyginimai */
PROC GLM DATA=data plots=ALL;
 CLASS parental_education special_coaching;
MODEL result = parental_education daily_study_hours attendance / SS3;
LSMEANS parental_education / stderr pdiff cov adjust=bon;
RUN;
```

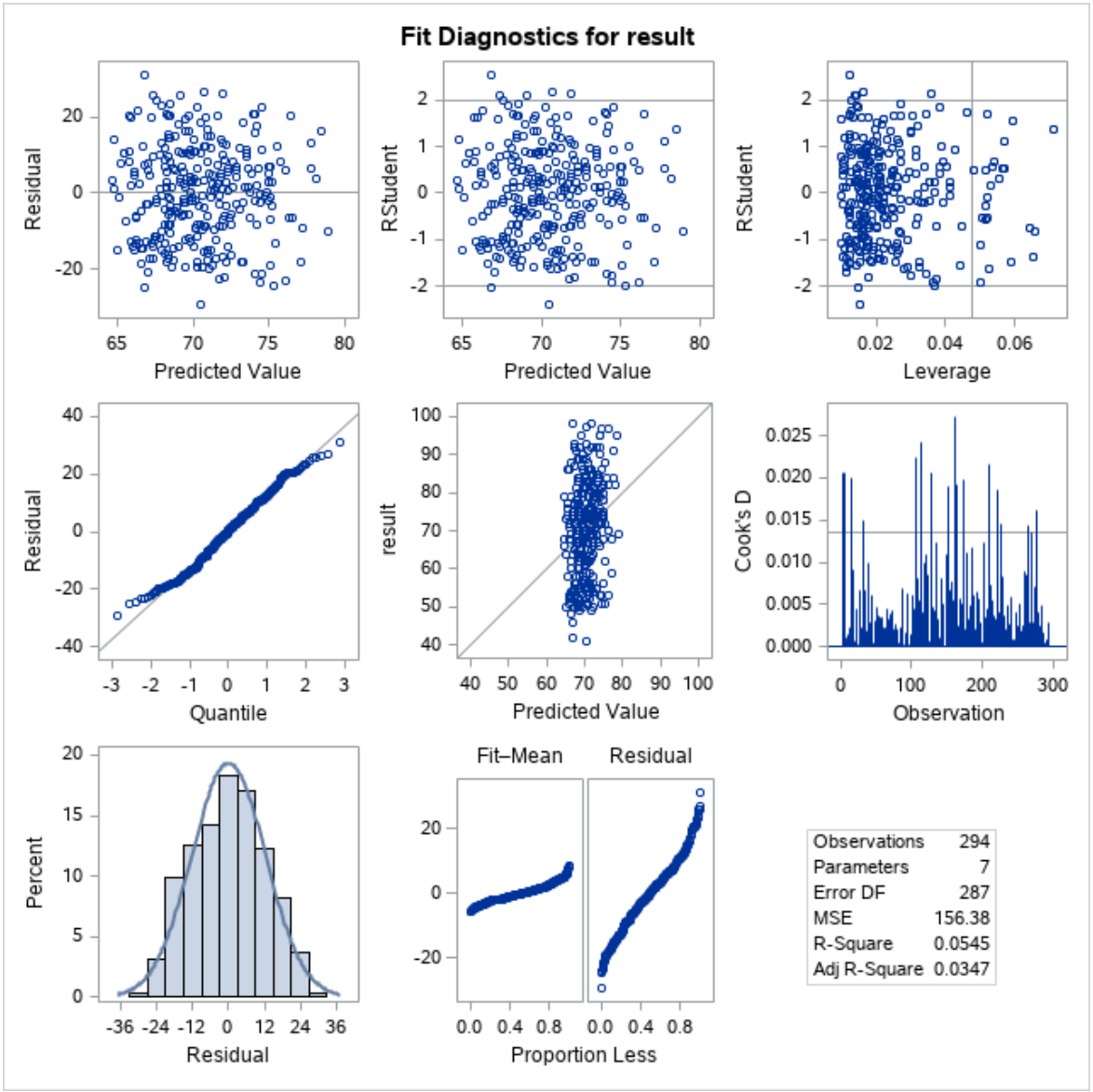
The GLM Procedure

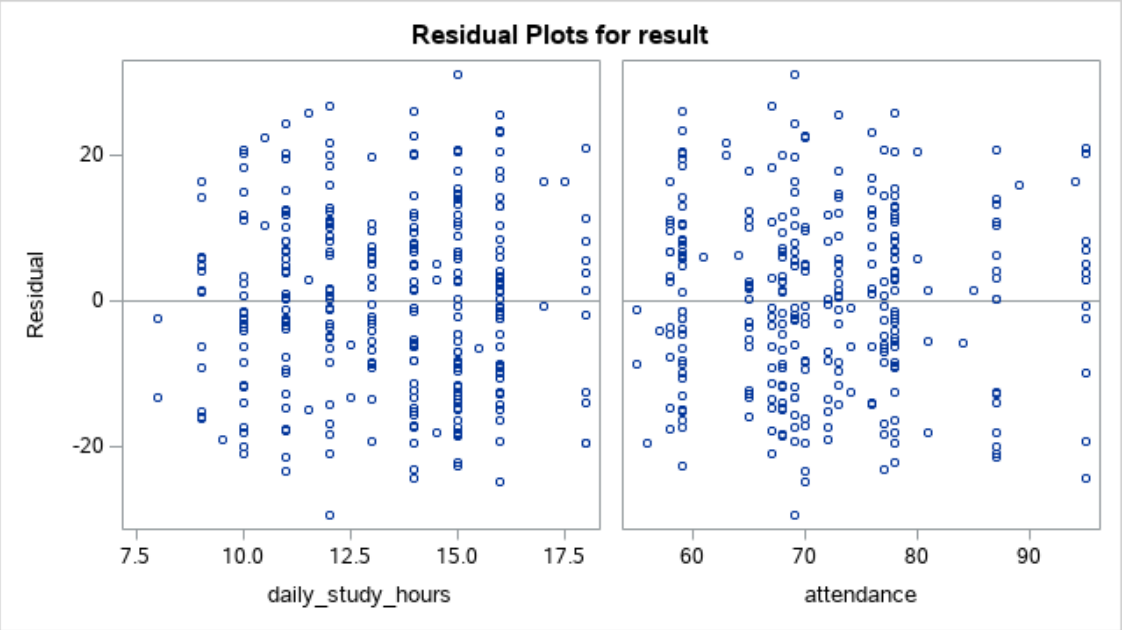
Dependent Variable: result

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	2584.68098	430.78016	2.75	0.0128
Error	287	44881.75184	156.38241		
Corrected Total	293	47466.43282			

R-Square	Coeff Var	Root MSE	result Mean
0.054453	17.76672	12.50530	70.38605

Source	DF	Type III SS	Mean Square	F Value	Pr > F
parental_education	4	1821.330107	455.332527	2.91	0.0219
daily_study_hours	1	59.399971	59.399971	0.38	0.5382
attendance	1	800.888158	800.888158	5.12	0.0244





The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni

parental_education	result LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
Associate Degree	71.2154950	1.5165929	<.0001	1
Attended college	70.6589177	1.4963209	<.0001	2
Bachelor Degree	74.2280414	2.2222210	<.0001	3
High school	67.5479992	1.2273452	<.0001	4
Master Degree	75.2216430	2.8053940	<.0001	5

Least Squares Means for effect parental_education
Pr > |t| for H0: LSMean(i)=LSMean(j)

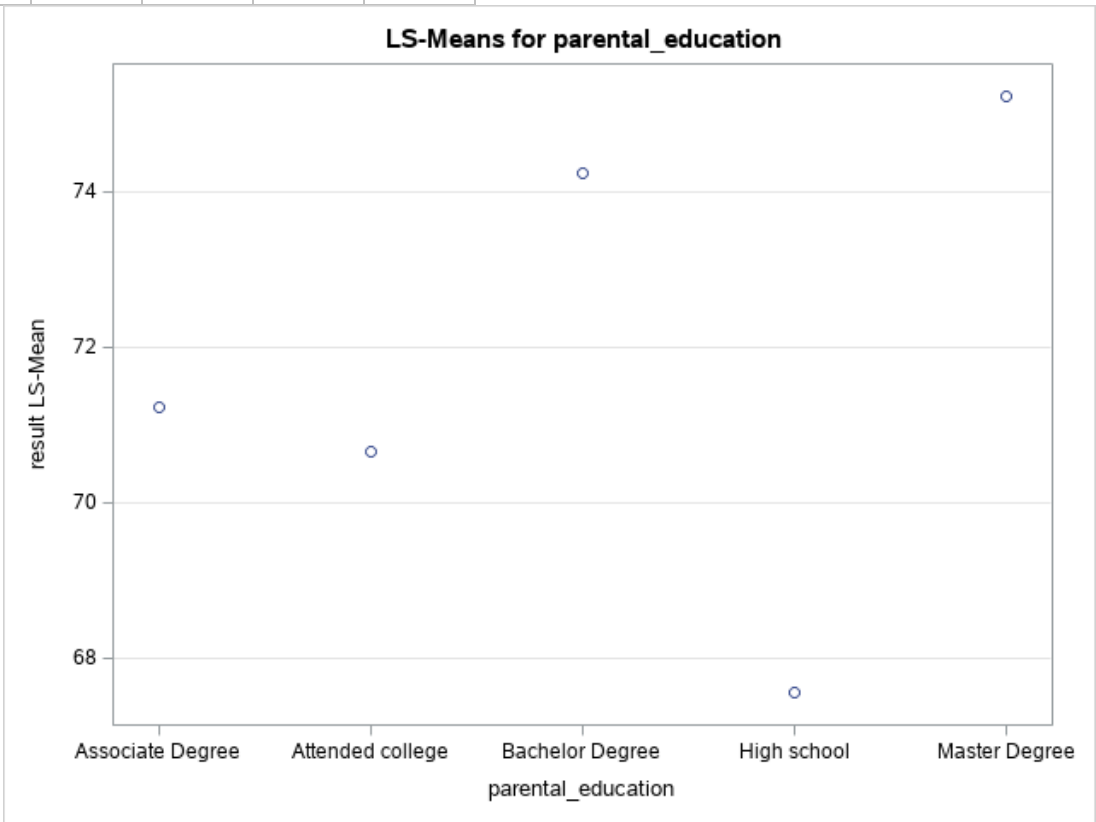
Dependent Variable: result

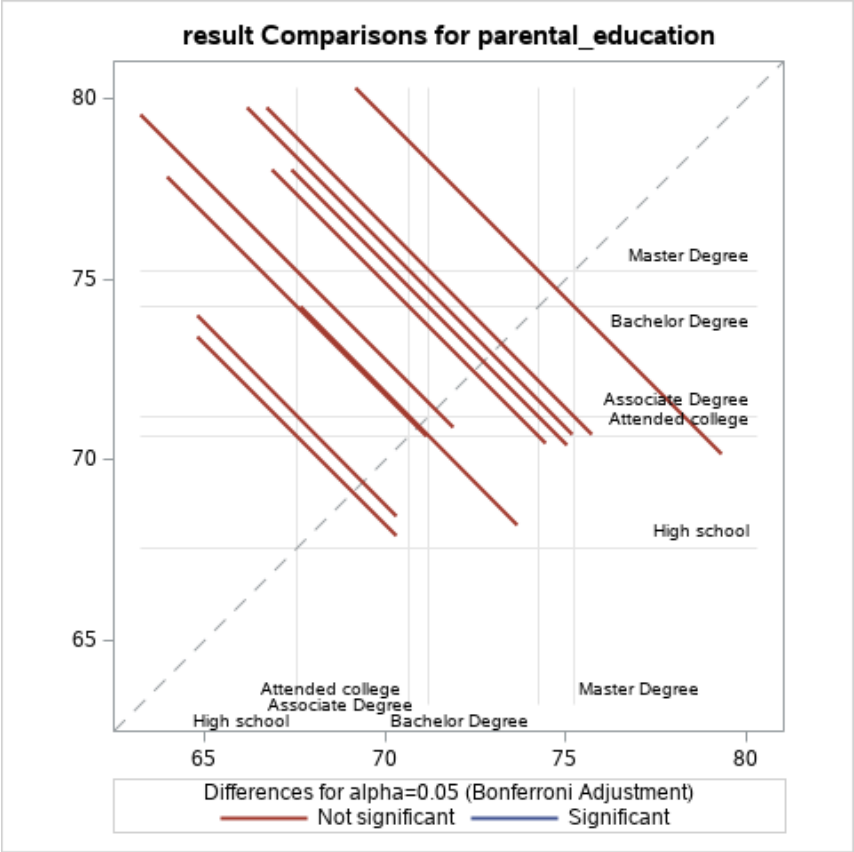
i/j	1	2	3	4	5
1		1.0000	1.0000	0.6118	1.0000
2	1.0000		1.0000	1.0000	1.0000
3	1.0000	1.0000		0.0908	1.0000
4	0.6118	1.0000	0.0908		0.1269

Least Squares Means for effect parental_education
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: result

i/j	1	2	3	4	5
5	1.0000	1.0000	1.0000	0.1269	





Kaip ir atlikus užduotį su R, poriniai vidurkių palyginimai atlikti naudojant Bonferroni pataisą, tačiau statistiškai reikšmingų skirtumų nerasta.

3. Naudojant Python