



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.

„*Test\_Prep\_Course*“ – ar studentas laikė pasiruošimą egzaminams.

„*Special\_Coaching*“ – ar studentas lanko papildomus mokymus.

„*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.

Naudojami du faktoriai: Tėvų išsilavinimas (stulp. „parental\_education“) ir pasiruošimo egzaminams laikymas (stulp. „test\_prep\_course“). Pasirinktos papildomos kovariantės: valandų praleistų per dieną mokantis kiekis (stulp. „daily\_study\_hours“) ir lankomumas procentais (stulp. „attendance“). Tarp abiejų kovariančių rasti faktoriaus lygmenys, kurių krypties koeficientai skiriasi nuo likusių. Reikalingi patikrinimai ar šie skirtumai statistiškai reikšmingi.

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

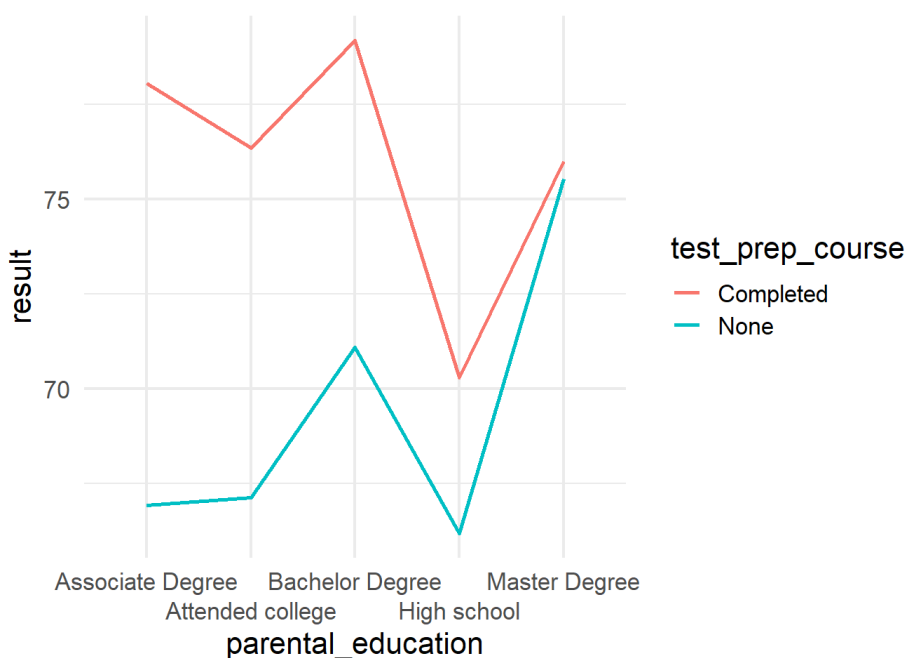
x <- readxl::read_xlsx("HighSchool.xlsx", sheet = 1) %>% clean_names()

x <- x %>%
  # sudaromas jungtinis faktorius
  mutate(combined = factor(paste(parental_education, test_prep_course))) %>%
  drop_na()

write_csv(x, "high_school_modified.csv")

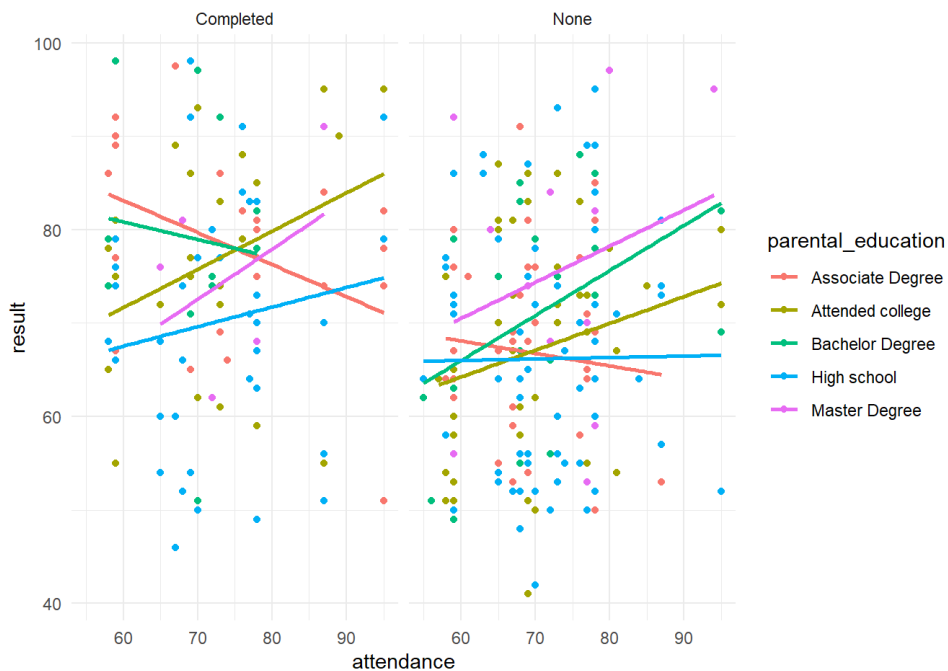
# Tiriamieji grafikai

# Faktorių efektai neatsižvelgiant į kovariantes
ggplot(x, aes(parental_education, result, color = test_prep_course, group = test_prep_course)) +
  stat_summary(fun = "mean", geom = "line", size = 1) +
  theme_minimal(base_size = 16) +
  guides(x = guide_axis(n.dodge = 2))
```

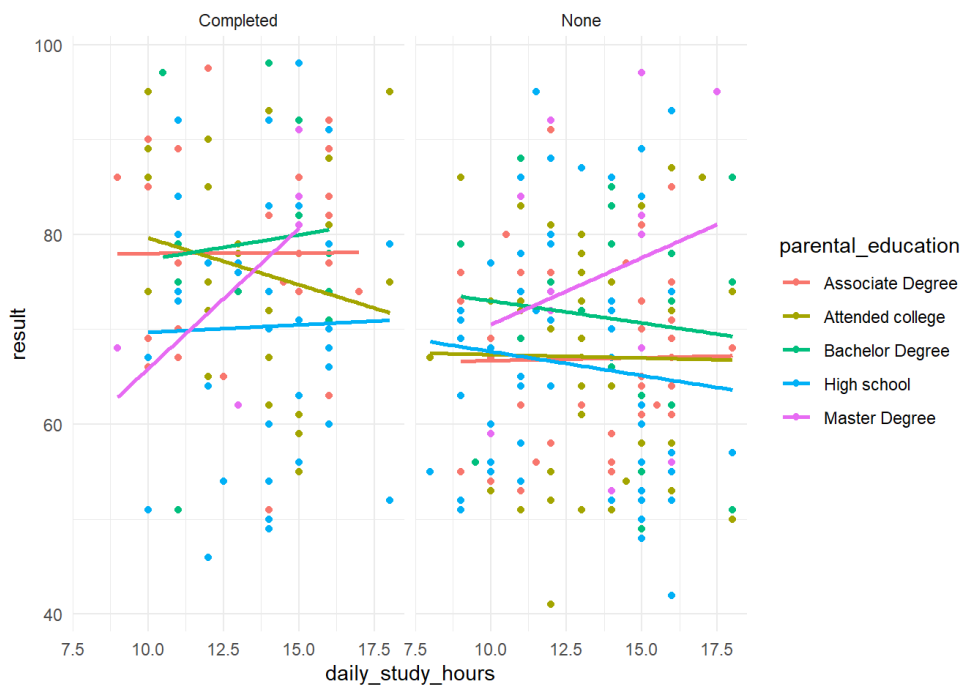


# Hipotezės apie koeficientų lygybę visiems faktorių lygmenims

```
ggplot(x, aes(attendance, result, color = parental_education)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(vars(test_prep_course)) +
  theme_minimal()
```

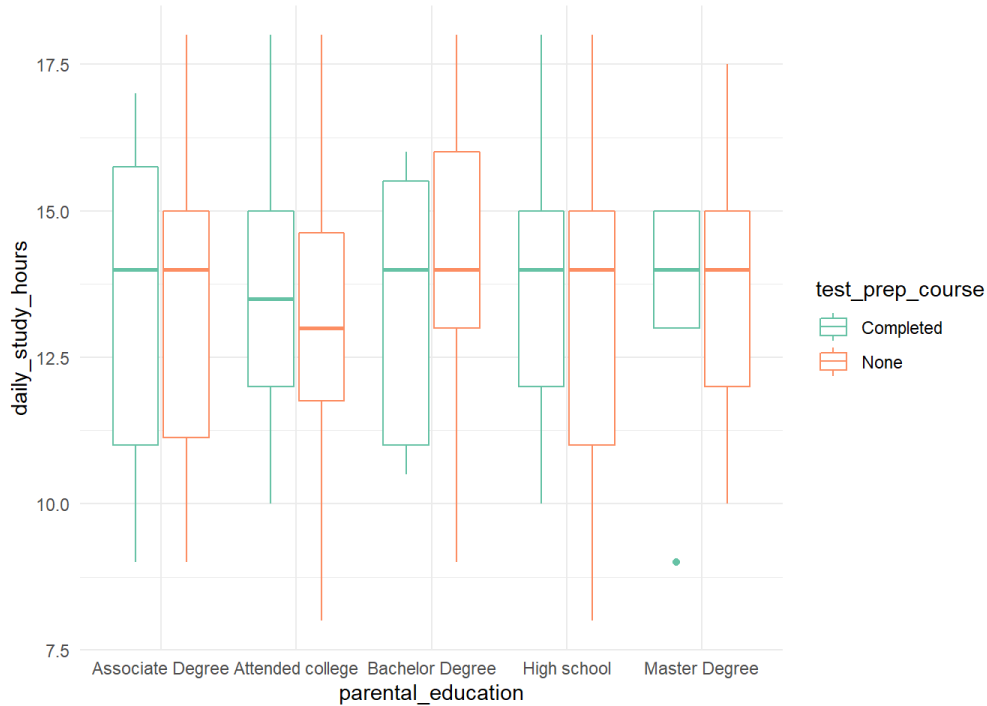


```
ggplot(x, aes(daily_study_hours, result, color = parental_education)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(vars(test_prep_course)) +
  theme_minimal()
```



*# Kovariančių pasiskirstymas pagal faktorių lygmenis*

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



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



```

library(rstatix)

# Hipotezė apie koeficientų lygybę neatmetama
anova_test(result ~ attendance * combined + daily_study_hours * combined, data = x, type = 3, detailed
= TRUE)

## ANOVA Table (type III tests)
##
##          Effect      SSn      SSd DFn DFd      F      p p<.05
## 1      (Intercept) 6366.834 38499.07   1 264 43.659 2.15e-10 *
## 2      attendance  294.059 38499.07   1 264  2.016 1.57e-01
## 3      combined 1286.732 38499.07   9 264  0.980 4.57e-01
## 4  daily_study_hours   74.310 38499.07   1 264  0.510 4.76e-01
## 5  attendance:combined 1848.959 38499.07   9 264  1.409 1.84e-01
## 6 combined:daily_study_hours 691.507 38499.07   9 264  0.527 8.55e-01
##      ges
## 1 0.142
## 2 0.008
## 3 0.032
## 4 0.002
## 5 0.046
## 6 0.018

# Hipotezė apie faktorių sąveikos nebuvimą neatmetama
anova_test(result ~ attendance + daily_study_hours + parental_education * test_prep_course, data = x, t
ype = 3, detailed = TRUE)

## ANOVA Table (type III tests)
##
##          Effect      SSn      SSd DFn DFd      F
## 1      (Intercept) 15651.089 41014.03   1 282 107.612
## 2      attendance  406.149 41014.03   1 282   2.793
## 3  daily_study_hours   63.175 41014.03   1 282   0.434
## 4  parental_education 1837.177 41014.03   4 282   3.158
## 5  test_prep_course 1888.974 41014.03   1 282  12.988
## 6 parental_education:test_prep_course 696.990 41014.03   4 282   1.198
##      p p<.05      ges
## 1 1.44e-21 * 0.276
## 2 9.60e-02 0.010
## 3 5.10e-01 0.002
## 4 1.50e-02 * 0.043
## 5 3.71e-04 * 0.044
## 6 3.12e-01 0.017

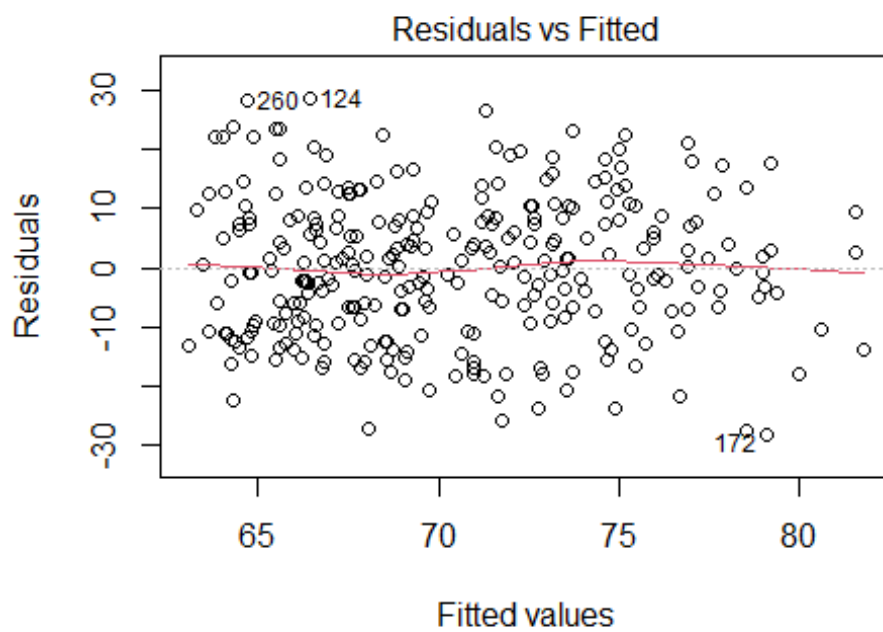
model <- anova_test(result ~ attendance + daily_study_hours + parental_education + test_prep_course, da
ta = x, type = 3, detailed = TRUE)
model

## ANOVA Table (type III tests)
##
##          Effect      SSn      SSd DFn DFd      F      p p<.05      ges
## 1      (Intercept) 15880.601 41711.02   1 286 108.889 8.20e-22 * 0.276
## 2      attendance  449.317 41711.02   1 286   3.081 8.00e-02 0.011
## 3  daily_study_hours   89.378 41711.02   1 286   0.613 4.34e-01 0.002
## 4  parental_education 1832.896 41711.02   4 286   3.142 1.50e-02 * 0.042
## 5  test_prep_course 3170.732 41711.02   1 286  21.741 4.79e-06 * 0.071

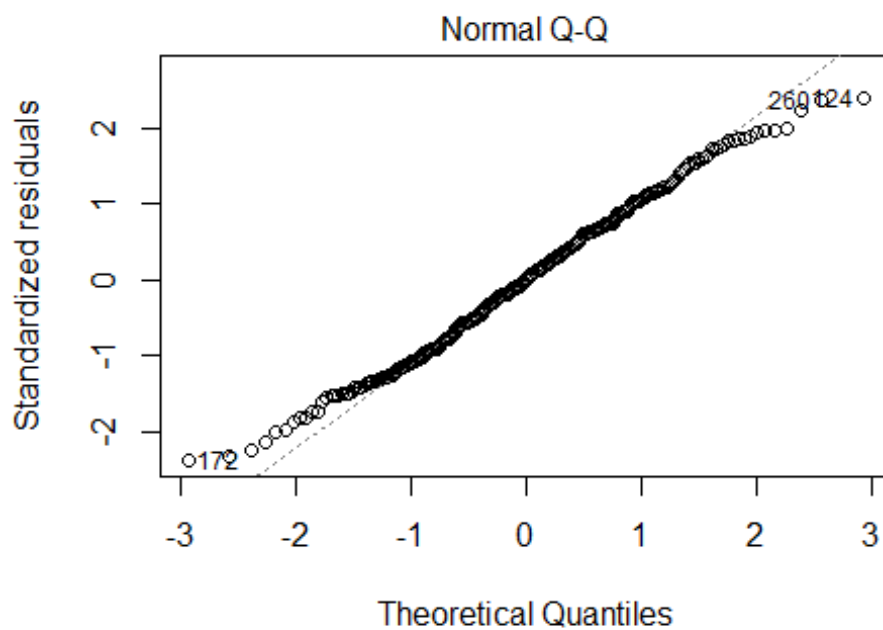
# Modelio prielaidų patikrinimas
model_aov <- aov(result ~ attendance + daily_study_hours + parental_education + test_prep_course, data
= x)

plot(model_aov)

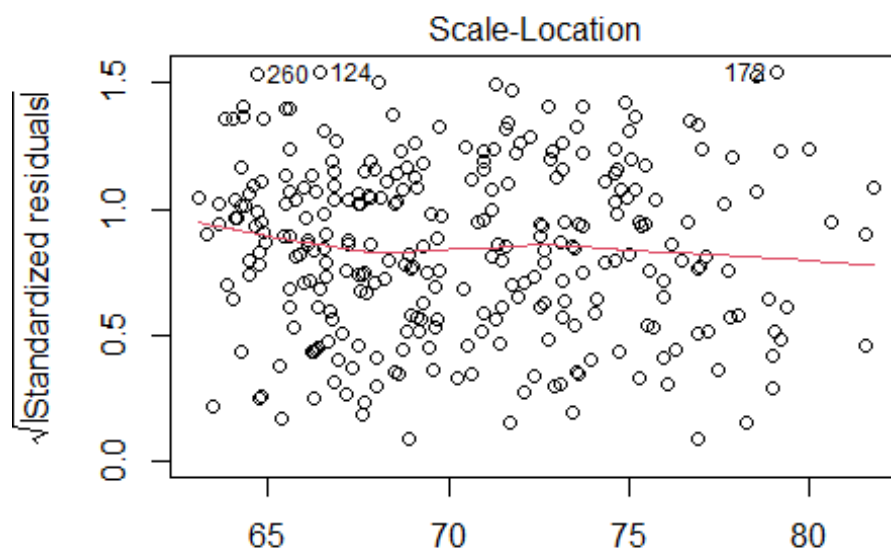
```



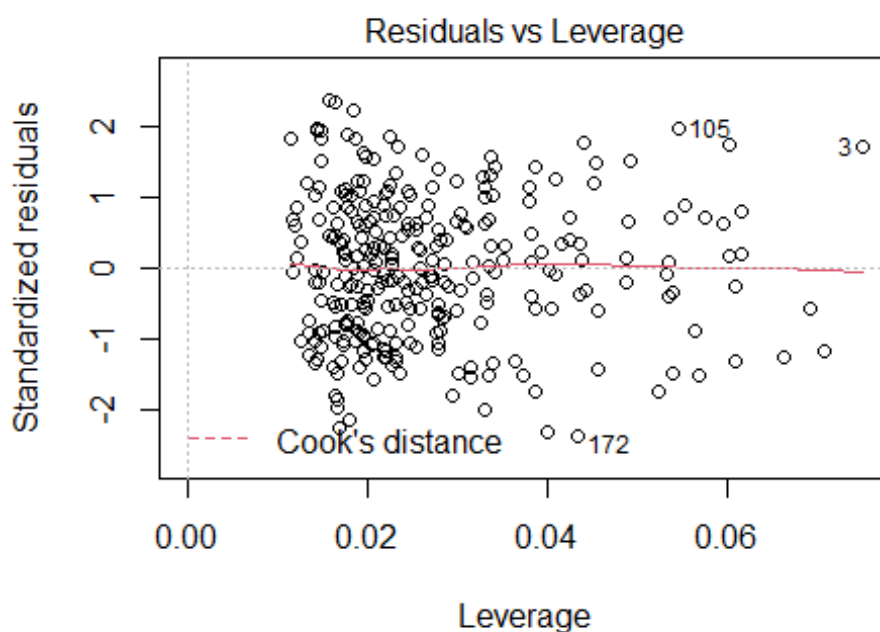
```
ov(result ~ attendance + daily_study_hours + parental_education + test
```



```
ov(result ~ attendance + daily_study_hours + parental_education + test
```



Fitted values  
`ov(result ~ attendance + daily_study_hours + parental_education + test`



Leverage  
`ov(result ~ attendance + daily_study_hours + parental_education + test`

```
leveneTest(result ~ combined, data = x, center = "mean")

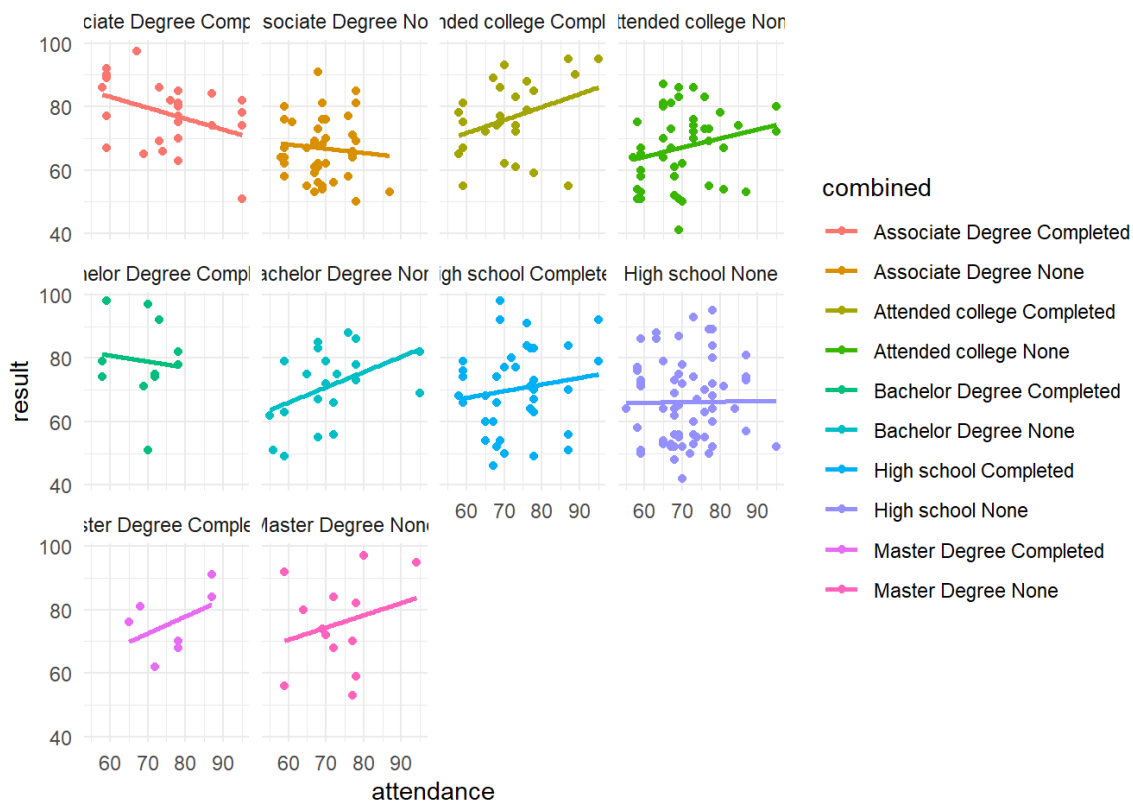
## Levene's Test for Homogeneity of Variance (center = "mean")
##      Df F value Pr(>F)
## group  9  0.9914 0.4472
##      284

shapiro.test(resid(model_aov))
```

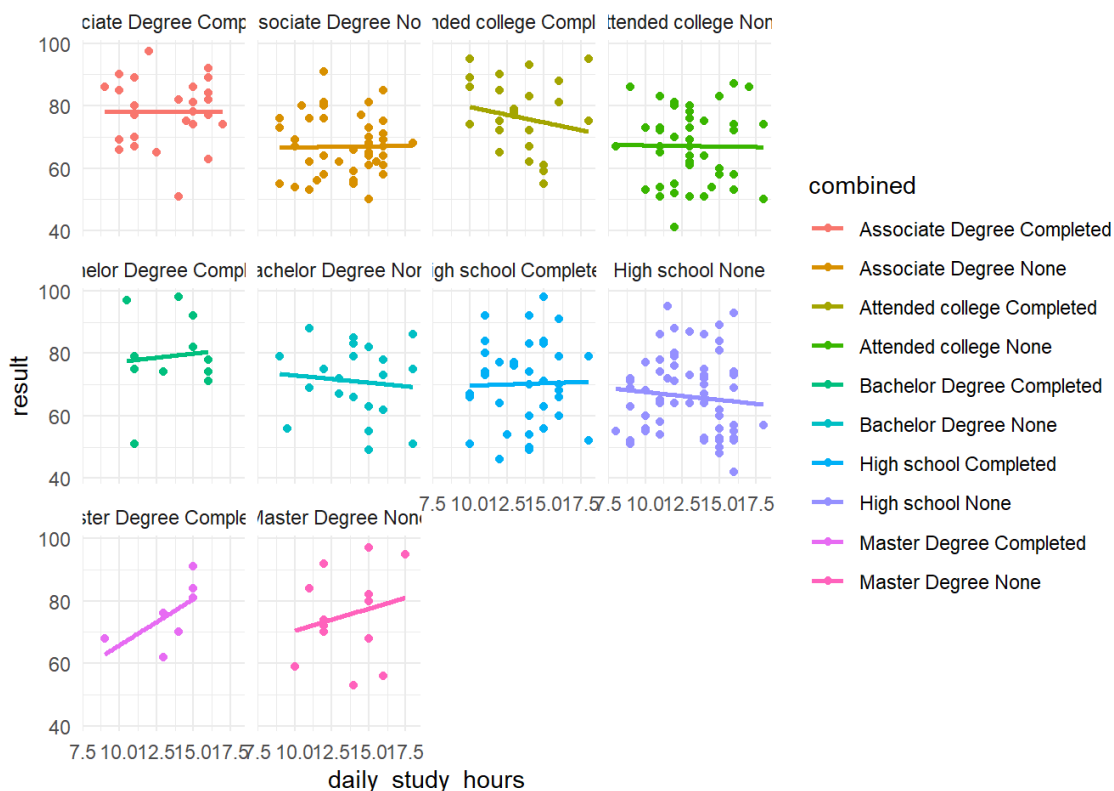


```
##
## Shapiro-Wilk normality test
##
## data: resid(model_aov)
## W = 0.99195, p-value = 0.1105

# Tiesinis ryšys tarp kovariančių ir priklausomo kintamojo
ggplot(x, aes(attendance, result, color = combined)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal() +
  facet_wrap(vars(combined))
```



```
ggplot(x, aes(daily_study_hours, result, color = combined)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal() +
  facet_wrap(vars(combined))
```



Hipotezės apie krypties koeficientų lygybę neatmetamos nei “attendance” ( $p=0.16$ ), nei “daily\_study\_hours” ( $p=0.48$ ) kovariantėms, todėl galimas taikyti kovariacinės analizės modelis.

Hipotezės apie dispersijų lygybę grupėms ir liekanų normalumą neatmetamos.

Abiejų kovariančių įtaka nėra statistiškai reikšminga, todėl toliau nagrinėjamas modelis be kovariančių.

Faktoriaus “test\_prep\_course” įtaka statistiškai reikšminga ( $p<0.01$ ). Kadangi šį faktorių sudaro du lygmenys, post-hoc palyginimai neatliekami.

Faktoriaus “parental\_education” įtaka irgi 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.

*# Kadangi abi kovariantės nebuvo statistiškai reikšmingos pasirenkamas*

*# modelis be kovariančių*

library(emmeans)

anova\_test(result ~ parental\_education + test\_prep\_course, data = x, type = 3, detailed = TRUE)

## ANOVA Table (type III tests)

##

##	Effect	SSn	SSd	DFn	DFd	F	p	p<.05
## 1	(Intercept)	1043116.905	42195.62	1	288	7119.640	3.97e-205	*
## 2	parental_education	1783.242	42195.62	4	288	3.043	1.80e-02	*
## 3	test_prep_course	3494.839	42195.62	1	288	23.854	1.73e-06	*

## ges

## 1 0.961

## 2 0.041

## 3 0.076

```

model_aov <- aov(result ~ parental_education + test_prep_course, data = x)

res <- x %>% emmeans_test(result ~ parental_education, model = model_aov)
res

## # 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 parental_education result Assoc~ Atten~  288    0.266 0.790    1      ns
## 2 parental_education result Assoc~ Bache~  288   -1.14 0.254    1      ns
## 3 parental_education result Assoc~ High ~  288    1.77 0.0780 0.780    ns
## 4 parental_education result Assoc~ Maste~  288   -1.54 0.124    1      ns
## 5 parental_education result Atten~ Bache~  288   -1.36 0.175    1      ns
## 6 parental_education result Atten~ High ~  288    1.49 0.137    1      ns
## 7 parental_education result Atten~ Maste~  288   -1.73 0.0855 0.855    ns
## 8 parental_education result Bache~ High ~  288    2.58 0.0104 0.104    ns
## 9 parental_education result Bache~ Maste~  288   -0.516 0.606    1      ns
## 10 parental_education result High ~ Maste~  288   -2.74 0.00660 0.0660 ns

get_emmeans(res)

## # A tibble: 5 x 7
##   parental_education emmean    se    df conf.low conf.high method
##   <fct>              <dbl> <dbl> <dbl>    <dbl>    <dbl> <chr>
## 1 Associate Degree    72.0  1.48  288    69.1     74.9 Emmeans test
## 2 Attended college    71.5  1.46  288    68.6     74.4 Emmeans test
## 3 Bachelor Degree     75.0  2.15  288    70.8     79.2 Emmeans test
## 4 High school         68.7  1.21  288    66.3     71.1 Emmeans test
## 5 Master Degree       76.8  2.72  288    71.4     82.1 Emmeans test

```

Visos egzaminų rezultatų vidurkių pagal tėvų išsilavinimą poros statistiškai reikšmingai tarpusavyje nesiskiria.

## Rezultatai

Naudojant koviariacinę analizę (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“).

Tyrimo metu rasta, kad anksčiau minėtos kovariantės nėra statistiškai reikšmingos, todėl toliau taikyta dvifaktorisė dispersinė analizė. Naudojant šį metodą rastos statistiškai reikšmingos pasiruošimo egzaminui kurso ( $F=23.6$ ,  $p<0.001$ ) ir tėvų išsilavinimo įtakos ( $F=3.11$   $p=0.02$ ).

Post-hoc poriniai vidurkių palyginimai pagal tėvų išsilavinimą atlikti naudojant Bonferroni pataisą, tačiau statistiškai reikšmingų skirtumų tarp porų nerasta.

## 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 combined;  
MODEL result = combined daily_study_hours*combined attendance*combined / SS3;  
RUN;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
parental_education	4	1837.177125	459.294281	3.16	0.0146
test_prep_course	1	1888.974487	1888.974487	12.99	0.0004
parental*test_prep_	4	696.989672	174.247418	1.20	0.3119
daily_study_hours	1	63.174982	63.174982	0.43	0.5104
attendance	1	406.149112	406.149112	2.79	0.0958

```
/* Hipotezė apie faktorių sąveikos nebuvimą*/  
PROC GLM DATA=data; CLASS parental_education test_prep_course;  
MODEL result = parental_education test_prep_course parental_education*test_prep_course  
daily_study_hours attendance / SS3;  
RUN;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
parental_education	4	1837.177125	459.294281	3.16	0.0146
test_prep_course	1	1888.974487	1888.974487	12.99	0.0004
parental*test_prep_	4	696.989672	174.247418	1.20	0.3119
daily_study_hours	1	63.174982	63.174982	0.43	0.5104
attendance	1	406.149112	406.149112	2.79	0.0958

```
/* Modelio prielaidos */  
/* Vidurkių palyginimai */  
PROC GLM DATA=data plots=ALL;  
CLASS parental_education test_prep_course;  
MODEL result = parental_education test_prep_course daily_study_hours attendance / SS3;  
LSMEANS parental_education / stderr pdiff adjust=bon;  
OUTPUT out=res residual=liekanos;  
RUN;
```

## The GLM Procedure

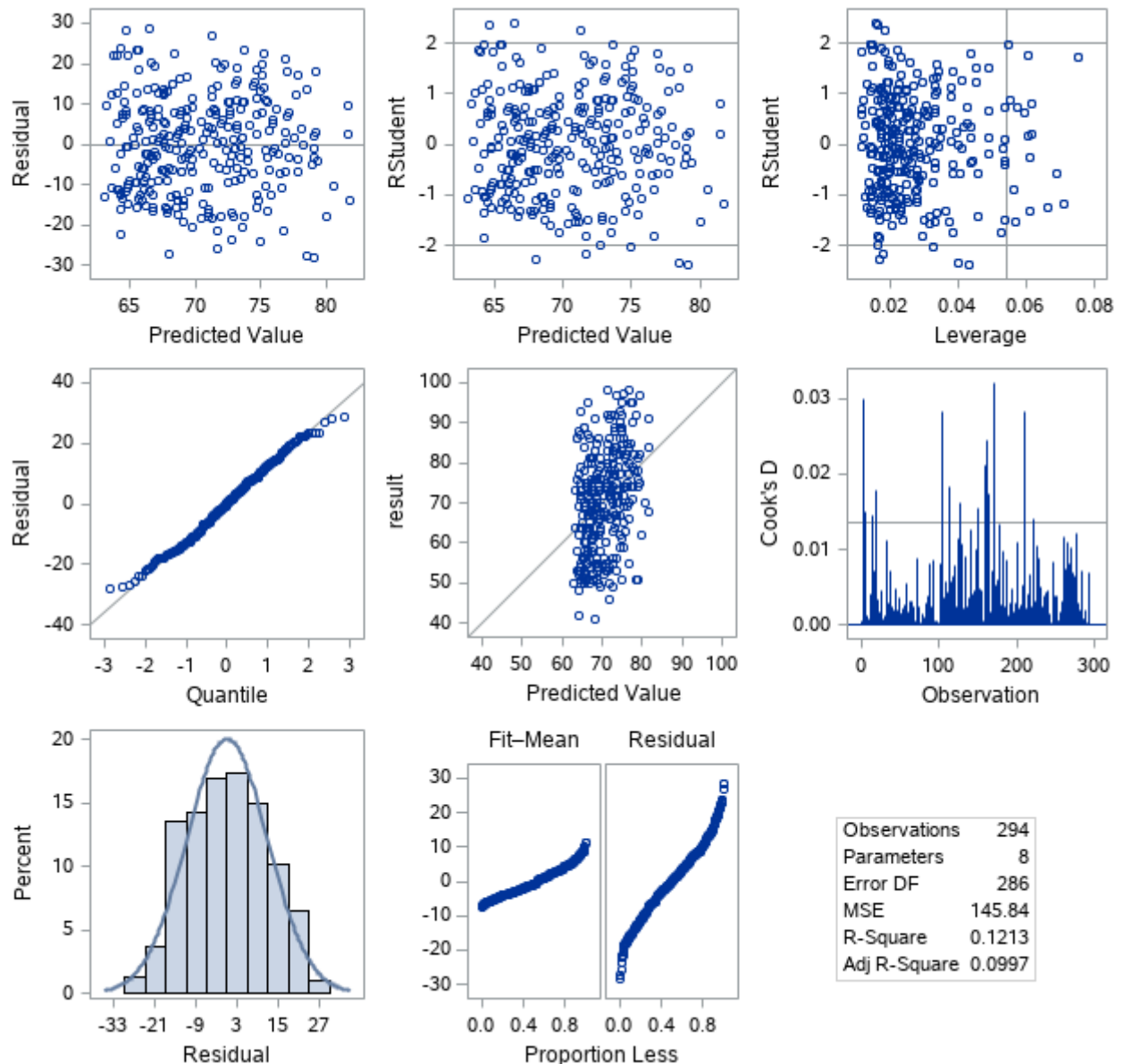
Dependent Variable: result

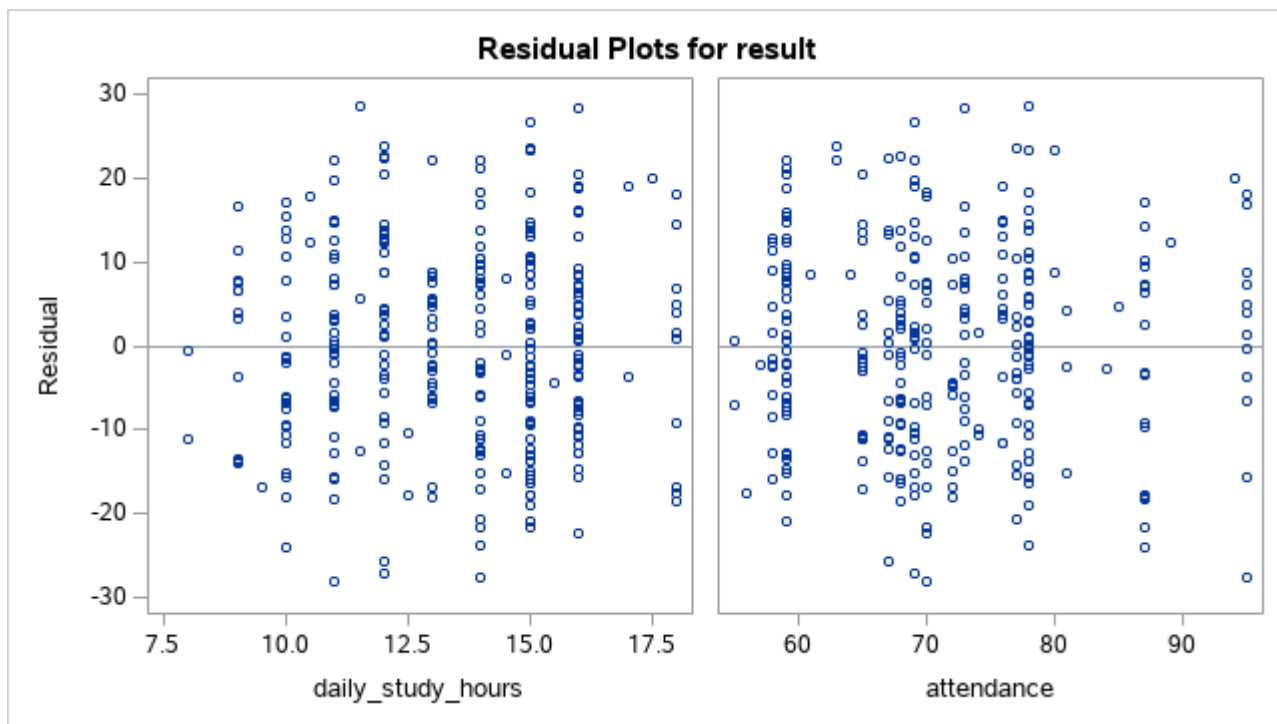
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	5755.41331	822.20190	5.64	<.0001
Error	286	41711.01951	145.84273		
Corrected Total	293	47466.43282			

R-Square	Coeff Var	Root MSE	result Mean
0.121252	17.15757	12.07654	70.38605

Source	DF	Type III SS	Mean Square	F Value	Pr > F
parental_education	4	1832.896011	458.224003	3.14	0.0150
test_prep_course	1	3170.732331	3170.732331	21.74	<.0001
daily_study_hours	1	89.377837	89.377837	0.61	0.4344
attendance	1	449.316585	449.316585	3.08	0.0803

### Fit Diagnostics for result





```
/* Normalumo testas */
PROC UNIVARIATE data=res normal;
VAR liekanos;
RUN;
```

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.991947	Pr < W	0.1105
Kolmogorov-Smirnov	D	0.040262	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.069686	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.502076	Pr > A-Sq	0.2133

```
/* Dispersijų lygybės testas */
PROC GLM DATA=data plots=none;
CLASS combined;
MODEL result = combined;
MEANS combined / HOVTEST=levене(type=abs);
RUN;
```

#### The GLM Procedure

Levene's Test for Homogeneity of result Variance ANOVA of Absolute Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
combined	9	394.5	43.8317	0.99	0.4472
Error	284	12555.7	44.2103		

```

/* Abi kovariantės statistiškai nereikšmingos */
/* Naudojamas modelis be kovariančių */
PROC GLM DATA=data plots=ALL;
CLASS parental_education test_prep_course;
MODEL result = parental_education test_prep_course/ SS3;
LSMEANS parental_education / stderr pdiff adjust=bon;
OUTPUT out=res residual=liekanos;
RUN;

```

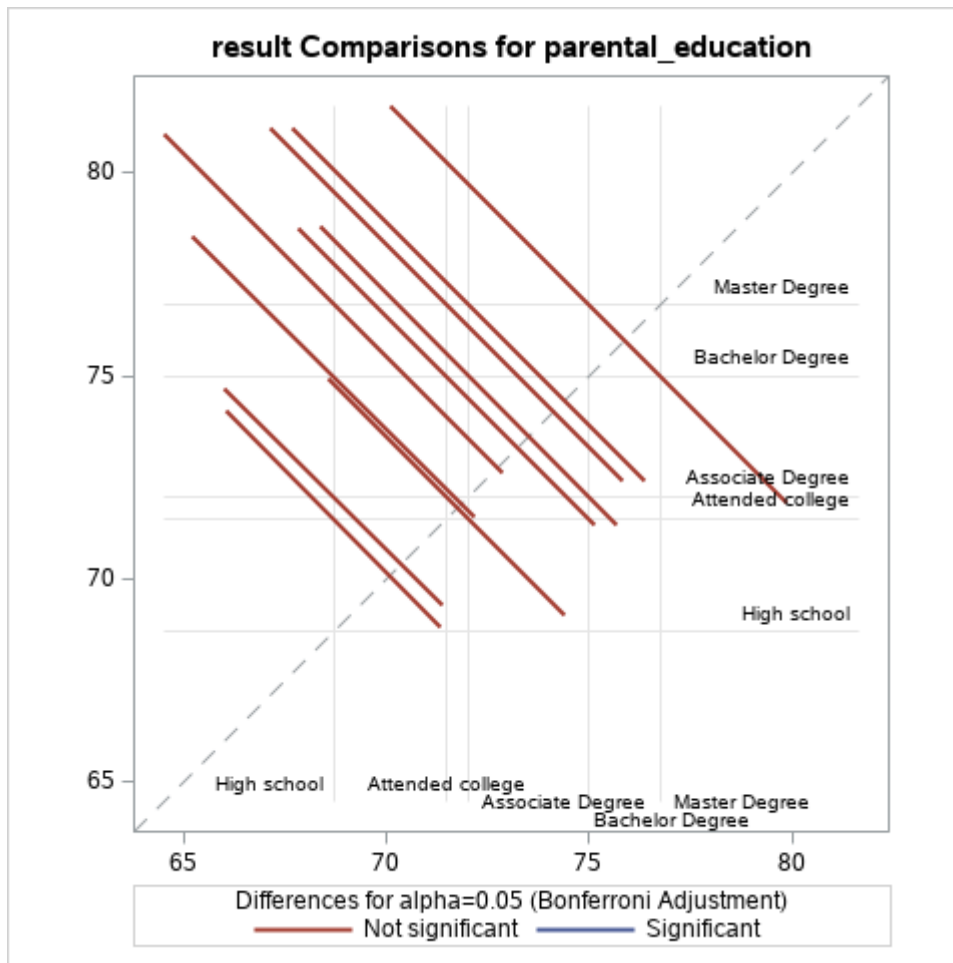
**The GLM Procedure**  
**Least Squares Means**  
**Adjustment for Multiple Comparisons: Bonferroni**

parental_education	result LSMEAN	Standard Error	Pr >  t	LSMEAN Number
Associate Degree	72.0271931	1.4779767	<.0001	1
Attended college	71.4788253	1.4589894	<.0001	2
Bachelor Degree	74.9951002	2.1520026	<.0001	3
High school	68.6877848	1.2056497	<.0001	4
Master Degree	76.7752962	2.7155299	<.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	0.7801	1.0000
2	1.0000		1.0000	1.0000	0.8547
3	1.0000	1.0000		0.1045	1.0000
4	0.7801	1.0000	0.1045		0.0660
5	1.0000	0.8547	1.0000	0.0660	



Kaip ir atlikus užduotį su R, atlikta dvifaktorinė dispersinė analizė, poriniai vidurkių palyginimai atlikti naudojant Bonferroni pataisą, tačiau statistškai reikšmingų skirtumų nerasta.