Statistical Learning - Fall 2019.
(Instructor: Kok-Seng Wong)

MATH 540

Assignment 2
Report

Kussainova, Assem

Due: November 15, 2019, Submitted: November 15, 2019

Abstract The analysis of the data is widely used in all fields of activity. Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusion and supporting decision-making. In this assignment was discussed some essential analyzing methods for Data Science: Normality test, Multiple regression, One-way ANOVA, Two-way ANOVA, Chi-squared test and etc. Moreover, the choice of particular tests was briefly discussed for chosen dataset analysis.

Keywords High School and Beyond dataset · Normality test · Multiple regression · One-way ANOVA · Two-way ANOVA · Chi-squared test

1 Introduction

Statistical analysis of data is widely used in all fields of activity: economics, management, sociology, political science, management, medicine, philology, technical sciences, etc. It allows you to find relationships and identify patterns in the studied processes and phenomena. This assignment will describe an overview of statistical analysis methods and as a part of it, data analysis tools will be used to describe data collected from High School and Beyond survey using software platforms such as RStudio and IBM SPSS.

2 Data

For data analysis purpose high school and beyond survey dataset was considered. The High School and Beyond data set (hsb) is from a large-scale longitudinal study conducted by the National Opinion Research Center (1980) under contract with the National Center for Education Statistics. Two hundred observations were randomly sampled from the High School and Beyond survey, and had formed a data frame with 200 observations and 11 variables. [1] Brief descriptions of the variables in the order in which they are in the le are given below.

Table 1. High School and Beyond Survey dataset entries

Variable Number	Variable Name	Variable Label	Coded Responses
1	id	Student identification number	(3-digit consecutive numbers)
2	female	Sex	0 = Male 1 = Female
3	race	Race or ethnicity	1 = Hispanic 2 = Asian 3 = Black 4 = White
4	ses	Socioeconomic status	1 = Low 2 = Medium

			3 = High
5	schtyp	School type	1 = Public 2 = Private
6	prog	High school program	1 = General 2 = Academic preparatory 3 = Vocational
7	read	Reading T-score	standardized to mean 50 and SD 10
8	write	Writing T-score	standardized to mean 50 and SD 10
9	math	Math T-score	standardized to mean 50 and SD 10
10	science	Science T-score	standardized to mean 50 and SD 10
11	socst	Social Studies T-score	standardized to mean 50 and SD 10

3 Methods and approach to study

For performing analysis RStudio and SPSS software were used.

Several key questions were chosen as a motivation for described dataset analysis:

- Does type of program influence on total test results of students?
- Is there a relationship between type of program students enrolled and their socioeconomic status? If yes, does this relationship affect the results of reading test?
- Can we predict writing scores for students according their gender and other test results?

4 Analysis

4.1 Effect of program on students' test results

Methods The analysis of variance model provides us with useful preliminary information about how much total variation in test scores occurs within different types of school program. It also can provide useful information about the reliability of each program's sample mean as an estimate of its true population mean.

Analysis When one-way ANOVA is chosen for data analysis, part of the process involves checking to make sure that the data can actually be analyzed using a one-way ANOVA. It is only appropriate to use a one-way ANOVA if data "passes" assumptions that are required for a one-way ANOVA to give a valid result. [2]

Main assumptions for one-way ANOVA:

1. Dependent variable should be approximately normally distributed for each category of the independent variable.
2. There needs to be homogeneity of variances.

Normality test:

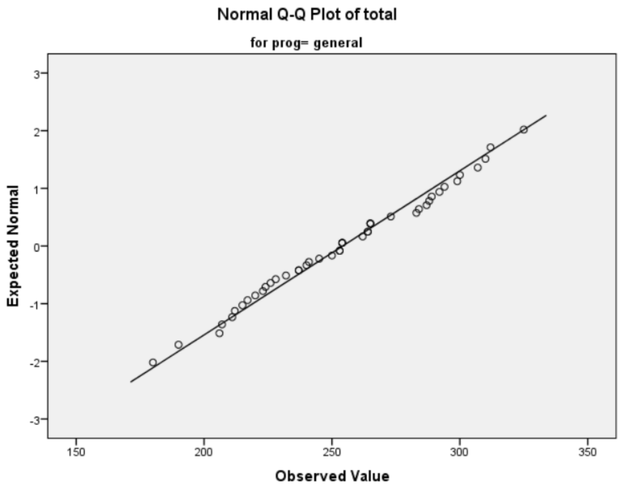

Fig. 1 Q-Q Plots for total scores from general school program

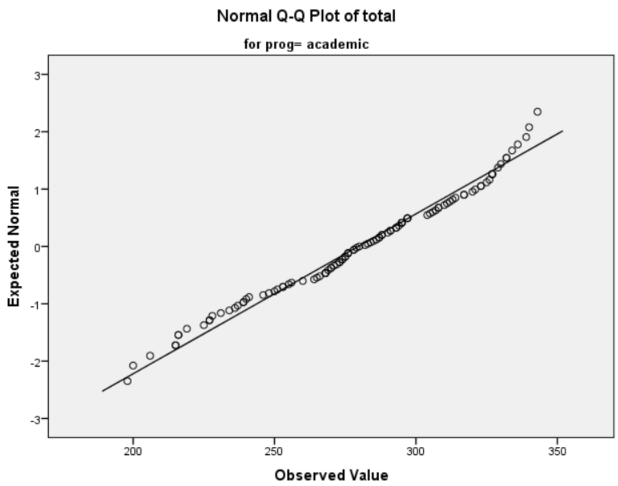

Fig. 2 Q-Q Plots for total scores from academic school program

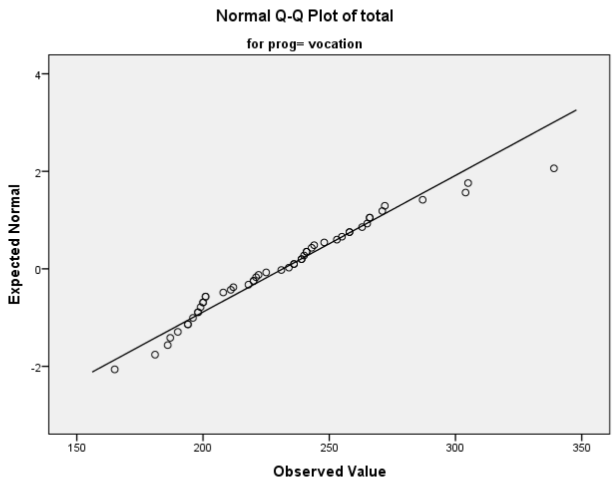

Fig. 3 Q-Q Plots for total scores from vocation school program

Q-Q Plots reflect that all data is normally distributed as all points on the graph are close to the fitting line. Next step is to check the assumption that data in all groups has equal homogeneity of variances.

Homogeneity of variances:

H_0: homogeneity of variance of students' total score within programs can be assumed

H_1: homogeneity of variance of students' total score within programs cannot be assumed

Test of Homogeneity of Variances

total

Levene Statistic	df1	df2	Sig.
.011	2	197	.989

Fig. 4 Levene's test results

sig. = 0.989 ($p > 0.05$)
Test is non-significant
Accept H0
Conclusion: homogeneity of variances of total test scores within different programs can be assumed

Assumptions for performing One-way ANOVA test are met. To examine re-search question, One-way ANOVA test is accomplished to see if there is a significant difference among total scores among different types of programs.

One-way ANOVA:

H_0: mean of total test scores doesn't differ between three program types
H_1: mean of total test scores differs between three program types

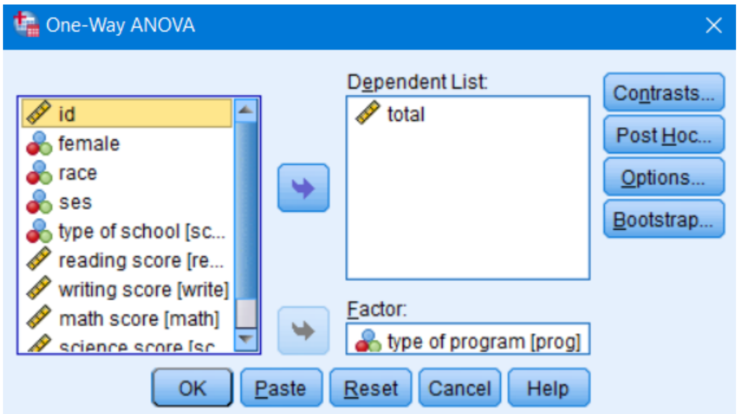

Fig. 5 Variables for one-way ANOVA in SPSS

In this case, independent variable is type of the program, and dependent variable is total test score of the student. In the SPSS interface the dependent variable, total, is transferred into the Dependent List: box and the independent variable, prog, into the Factor: box.

ANOVA

total

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	44429.664	3	14809.888	10.088	.000
Within Groups	287745.531	196	1468.089		
Total	332175.195	199			

Fig. 6 One-way ANOVA results

sig. = 0.000003 ($p < 0.05$)
Test is significant
Reject H0
Conclusion: mean of total test scores differs between three program types

The table that shows output of the ANOVA analysis and whether there is a statistically significant difference between programs' group means. The significance value is 0.000003, which is below 0.05 and, therefore, there is a statistically significant difference in the mean of total test scores differs between three program types. Specification of mean difference between three program types is not included by ANOVA test. For identification of precise mean difference of data Post Hock test is chosen.

Post Hock Test:

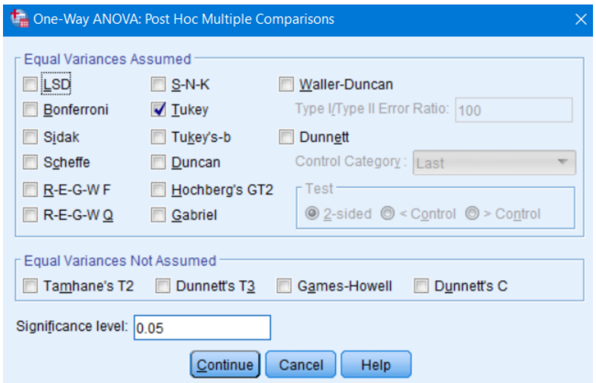

Fig. 7 Post Hock test types in SPSS

As seen from the Figure 7 above, Tukey Post Hock test was chosen for determining which means of different program types differ exactly.

Post Hoc Tests

Multiple Comparisons

Dependent Variable: total

Tukey HSD

(I) type of program	(J) type of program	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
general	academic	-25.49206*	6.35429	.000	-40.4981	-10.4860
	vocation	22.53556*	7.32813	.007	5.2297	39.8414
academic	general	25.49206*	6.35429	.000	10.4860	40.4981
	vocation	48.02762*	6.12786	.000	33.5563	62.4990
vocation	general	-22.53556*	7.32813	.007	-39.8414	-5.2297
	academic	-48.02762*	6.12786	.000	-62.4990	-33.5563

*. The mean difference is significant at the 0.05 level.

Fig. 8 Post Hock test results

Comparison 1:

$$H_0: \mu_{general} = \mu_{academic}$$
$$H_1: \mu_{general} \neq \mu_{academic}$$

sig. = 0.000251 ($p < 0.05$)
Test is significant
Reject H_0
Conclusion: means of general and academic programs are not equal

Comparison 2:

$$H_0: \mu_{general} = \mu_{vocation}$$
$$H_1: \mu_{general} \neq \mu_{vocation}$$

sig. = 0.007 ($p < 0.05$)
Test is significant
Reject H_0
Conclusion: means of general and vocation programs are not equal

Comparison 3:

$$H_0: \mu_{vocation} = \mu_{academic}$$

$$H_1: \mu_{vocation} \neq \mu_{academic}$$

sig. = 5.1007^{-9} (p < 0.05)
Test is significant
Reject H_0
Conclusion: means of vocation and academic programs are not equal

$$\mu_{general} \neq \mu_{academic} \neq \mu_{vocation}$$

These results show that differences in total test scores within all three types of programs in school are significant.

Conclusions From the performed tests it is known that type of program in school strongly influence on total test results of students. There was a statistically significant difference between groups as determined by one-way ANOVA. A Tukey Post Hock test revealed that the total scores were statistically significantly lower for those who studied at general program compared to the students from academic program (scores differ by 25.49). Moreover, there was statistically significant difference between the scores in general and vocation programs (scores differ by 22.53), as well as the big difference in scores of academic and vacation groups (scores differ by 48.0276).

4.2 Relationship between type of program and socioeconomic status

Methods The chi-square test for independence, also called Pearson's chi-square test or the chi-square test of association is used to discover if there is a relationship between two categorical variables. [3] For further analysis assuming that these two predictors affect the results of reading score test, two-way ANOVA test is proposed. The two-way ANOVA compares the mean differences between groups that have been split on two independent variables (called factors). The primary purpose of a Two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable. For example, you could use a Two-way ANOVA to understand if the significant difference among reading scores among different programs remains true within students with different socioeconomic status, where programs (general/academic/vocation) and socioeconomic status (low/middle/high) are your independent variables, and reading test score is your dependent variable. [4]

Analysis For analysis of data using a chi-square test for independence is needed to make sure that the data "passes" two assumptions. [3]

Main assumptions for Chi-squared test:

1. Two variables should be measured at an ordinal or nominal level (i.e., categorical data)
2. Two variables should consist of two or more categorical, independent groups.

Both variables meet established assumptions for Chi-squared test, so now this test can be used for analyzing relationship.

Chi-squared test:

$$H_0: There\ is\ no\ association\ between\ type\ of\ program\ and\ socioeconomic\ status$$
$$H_1: There\ is\ an\ association\ between\ type\ of\ program\ and\ socioeconomic\ status$$

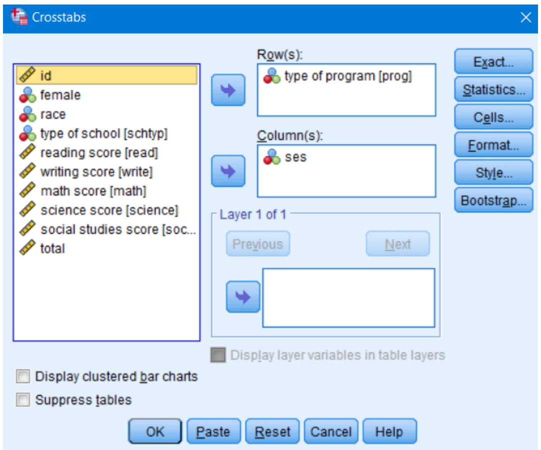

Fig. 9 Specifying independent variables in SPSS for analysis

From the Figure 9 above shown transferring one of the variables into the Row(s): box and the other variable into the Column(s): box. In this case, prog variable placed into the Row(s): box and ses variable into the Column(s): box.

type of program * ses Crosstabulation

			ses			Total
			low	middle	high	
type of program	general	Count	16	20	9	45
		Expected Count	10.6	21.4	13.1	45.0
	academic	Count	19	44	42	105
		Expected Count	24.7	49.9	30.5	105.0
	vocation	Count	12	31	7	50
		Expected Count	11.8	23.8	14.5	50.0
Total		Count	47	95	58	200
		Expected Count	47.0	95.0	58.0	200.0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	16.604[a]	4	.002
Likelihood Ratio	16.783	4	.002
Linear-by-Linear Association	.060	1	.807
N of Valid Cases	200		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.58.

Fig. 10 Chi-squared test output in SPSS

sig. = 0.002 ($p < 0.05$)
Test is significant
Reject H0
Conclusion: There is an association between type of program and socioeconomic status

Results of performed test state that there is a statistically significant association between socioeconomic status of students and type of program studied. Therefore, analysis can be continued to check the effect of this relationship on reading scores using two-way ANOVA.

For analysis of data using a two-way ANOVA test is data needed to pass two assumptions. [4]

Main assumptions for two-way ANOVA:

1. Dependent variable should be approximately normally distributed for each combination of the groups of the two independent variables.
2. There needs to be homogeneity of variances for each combination of the groups of the two independent

variables. [4]

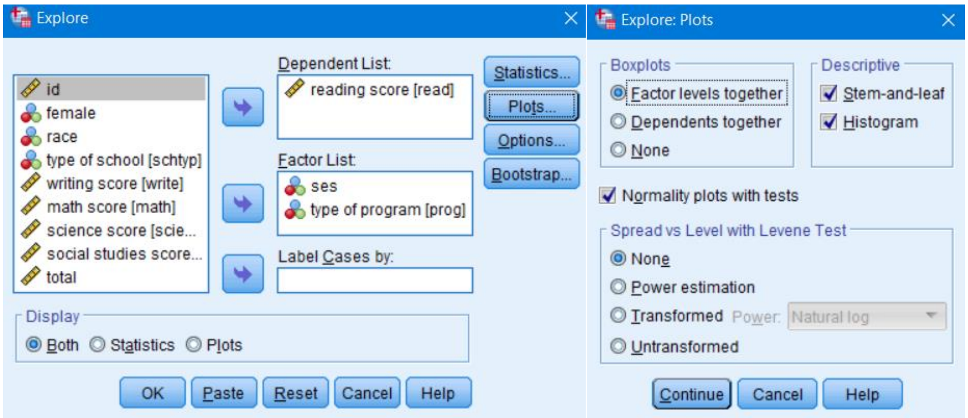

Fig. 11.1, 11.2 Variables for checking normality of data in SPSS

The settings above provide an output with several tests for normal distribution of data. Following are the results for checking the normality of reading score, factored by socioeconomic status and program type.

Reading scores grouped by socioeconomic status:

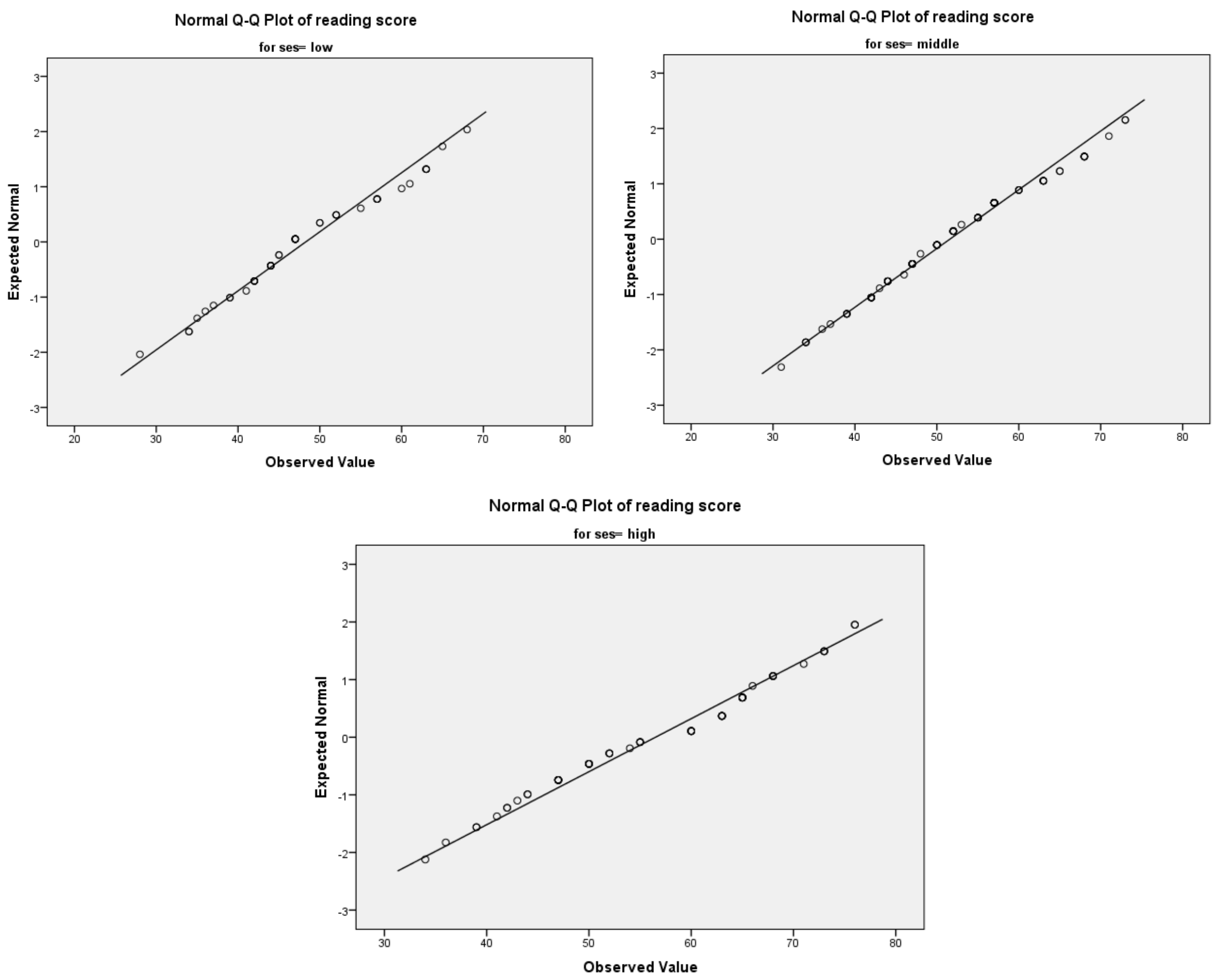

Fig. 12.1, 12.2, 12.3 Q-Q Plots of reading scores grouped by low, middle and high socioeconomic status

Reading scores grouped by socioeconomic status meet an assumption for normal distribution of data. Next reading scores are tested for normality within type of program.

Reading scores grouped by type of program:

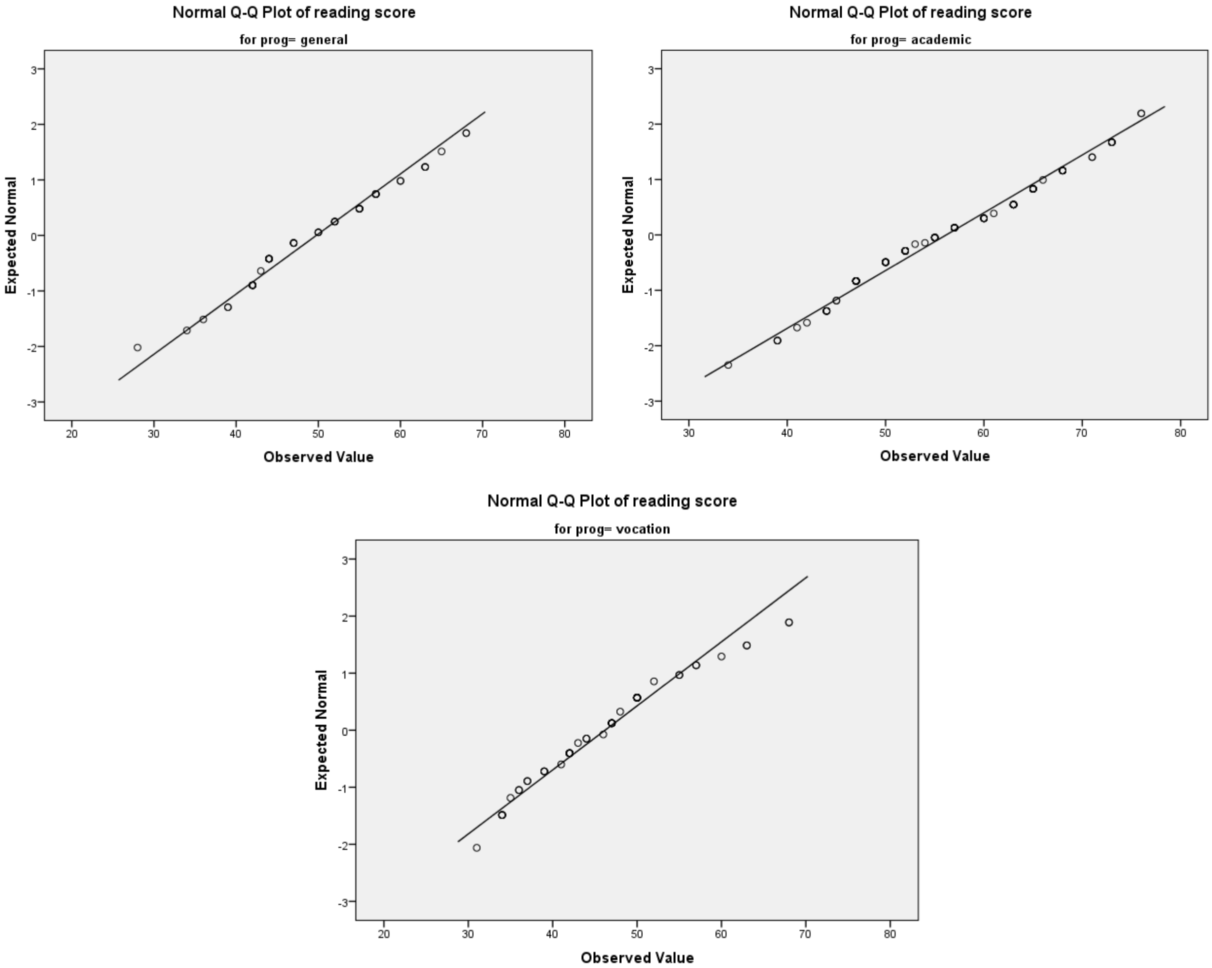

Fig. 13.1, 13.2, 13.3 Q-Q Plots of reading scores grouped by low, middle and high socioeconomic status

Normality assumption for reading score is also met within program types. Next step to check for homogeneity of variances of read variable within these 2 factors.

Homogeneity of variances test:

Homogeneity of variances of reading scores within types of program:

$$H_0: homogeneity\ of\ variances\ of\ reading\ scores\ within\ types\ of\ program\ can\ be\ assumed$$
$$H_1: homogeneity\ of\ variances\ of\ reading\ scores\ within\ types\ of\ program\ cannot\ be\ assumed$$

Test of Homogeneity of Variances

reading score

Levene Statistic	df1	df2	Sig.
.892	2	197	.411

Fig. 14 Levene’s test results for reading scores within types of program

sig. = 0.411 ($p > 0.05$)
Test is non-significant
Accept H0
Conclusion: homogeneity of variances of reading scores within types of program can be assumed

Homogeneity of variances of reading scores within socioeconomic status:

$$H_0: homogeneity\ of\ variances\ of\ reading\ scores\ within\ socioeconomic\ status\ can\ be\ assumed$$

$$H_1: homogeneity\ of\ variances\ of\ reading\ scores\ within\ socioeconomic\ status\ cannot\ be\ assumed$$

Test of Homogeneity of Variances

reading score

Levene Statistic	df1	df2	Sig.
2.443	2	197	.090

Fig. 15 Levene's test results for reading scores within socioeconomic status

sig. = 0.09 ($p > 0.05$)
Test is non-significant
Accept H0
Conclusion: homogeneity of variances of reading scores within socioeconomic status can be assumed

All assumptions for Two-way ANOVA test are met, therefore it is possible now to use a two-way ANOVA to understand if the significant difference among reading scores among different programs remains true within students with different socioeconomic status.
Two-way ANOVA test:

$$H_{01}: there\ is\ no\ difference\ in\ the\ means\ of\ reading\ scores\ by\ type\ of\ program$$
$$H_{02}: there\ is\ no\ difference\ in\ the\ means\ of\ reading\ scores\ by\ socioeconomic\ status$$
$$H_{03}: there\ is\ no\ interaction\ between\ type\ of\ program\ and\ socioeconomic\ status$$

$$H_{11}: there\ is\ a\ difference\ in\ the\ means\ of\ reading\ scores\ by\ type\ of\ program$$
$$H_{12}: there\ is\ a\ difference\ in\ the\ means\ of\ reading\ scores\ by\ socioeconomic\ status$$
$$H_{13}: there\ is\ an\ interaction\ between\ type\ of\ program\ and\ socioeconomic\ status$$

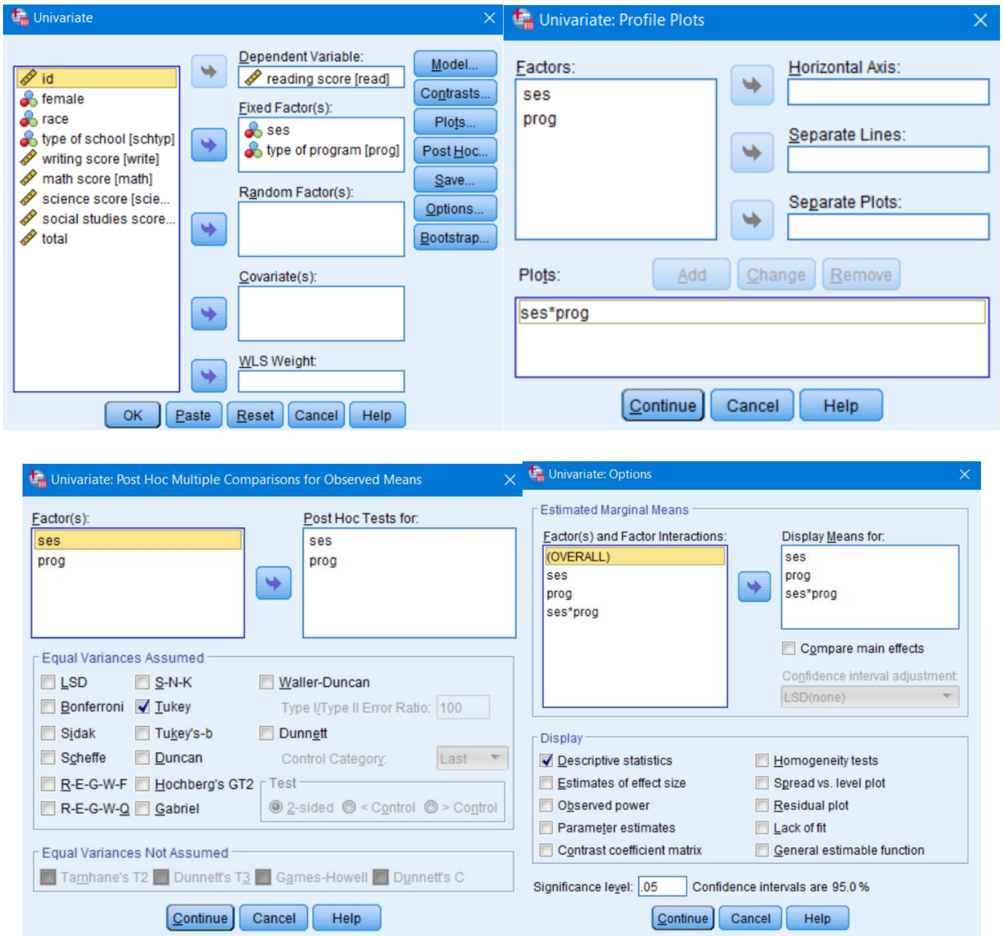

Fig. 16.1, 16.2, 16.3, 16.4 Settings for two-way ANOVA test in SPSS

From the figures above is seen that the dependent variable, read, is transferred into the Dependent Variable: box, and both independent variables, ses and prog, into the Fixed Factor(s): box. In the Plots window the independent variable, ses, transferred from the Factors: box into the Horizontal Axis: box, and the other independent variable, prog, into the Separate Lines: box for creating relationship of these variables on reading score. Then for both independent variables chosen Tukey Post Hock test in order to get precise difference of means. In the last window all variables are chosen for creating plot of dependence between reading score and affecting factors (socioeconomic status and type of program).

Results of two-way ANOVA test can be seen on next page.

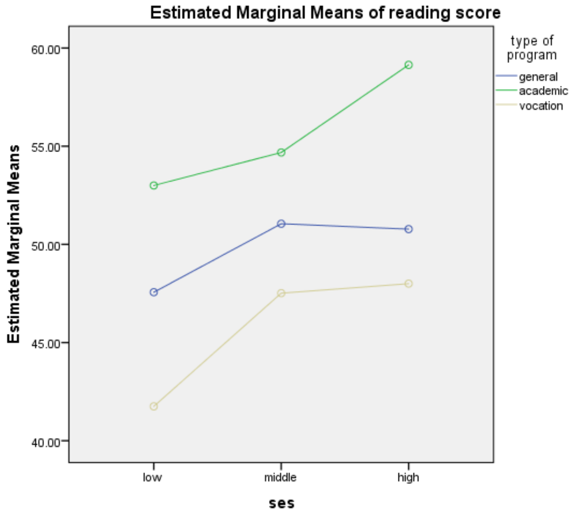

Fig. 17 Marginal means for all variables participated in two-way ANOVA

The plot of the mean reading score for each combination of groups of socioeconomic status and type of program are plotted in a line graph, as shown above.

Tests of Between-Subjects Effects

Dependent Variable: reading score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4810.297[a]	8	601.287	7.129	.000
Intercept	353811.382	1	353811.382	4195.013	.000
ses	600.479	2	300.240	3.560	.030
prog	2732.137	2	1366.068	16.197	.000
ses * prog	202.489	4	50.622	.600	.663
Error	16109.123	191	84.341		
Total	566514.000	200			
Corrected Total	20919.420	199			

a. R Squared = .230 (Adjusted R Squared = .198)

Fig. 18 Two-way ANOVA test results

The important results of two-way ANOVA are the ses, prog and ses*prog rows. These rows inform whether independent variables and their interaction have a statistically significant effect on the dependent variable. It is important to first look at the ses*prog interaction as this will determine interpretation of results. It is seen from the "Sig." column that we have not statistically significant interaction at the p = .0663 level. From the table above is shown that there was a statistically significant difference in mean interest in reading scores between socioeconomic statuses (p = .030), and there were statistically significant differences between types of program.

Post Hock Tests:

Multiple Comparisons

Dependent Variable: reading score

Tukey HSD

(I) ses	(J) ses	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
low	middle	-3.3024	1.63777	.111	-7.1710	.5663
	high	-8.2234*	1.80240	.000	-12.4809	-3.9659
middle	low	3.3024	1.63777	.111	-.5663	7.1710
	high	-4.9211*	1.53034	.004	-8.5359	-1.3062
high	low	8.2234*	1.80240	.000	3.9659	12.4809
	middle	4.9211*	1.53034	.004	1.3062	8.5359

Based on observed means.
The error term is Mean Square(Error) = 84.341.

*. The mean difference is significant at the .05 level.

Fig. 18 Post Hock test results for read within ses

Comparison 1:

$$H_0: \mu_{low} = \mu_{middle}$$
$$H_1: \mu_{low} \neq \mu_{middle}$$

sig. = 0.111 (p > 0.05)
Test is non-significant
Accept H_0
Conclusion: means of low and middle classes are equal

Comparison 2:

$$H_0: \mu_{low} = \mu_{high}$$
$$H_1: \mu_{low} \neq \mu_{high}$$

sig. = 0.000027 (p < 0.05)
Test is significant
Reject H_0
Conclusion: means of low and high classes are not equal

Comparison 3:

$$H_0: \mu_{middle} = \mu_{high}$$
$$H_1: \mu_{middle} \neq \mu_{high}$$

sig. = 0.004 (p < 0.05)
Test is significant
Reject H_0
Conclusion: means of middle and high classes are not equal

$$\mu_{high} \neq (\mu_{low} = \mu_{middle})$$

These results show that differences in reading test scores within high and low, high and middle are significant. Moreover, the means in reading test scores within low and middle statuses are equal.

Multiple Comparisons

Dependent Variable: reading score

Tukey HSD

(I) type of program	(J) type of program	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
general	academic	-6.4063*	1.63630	.000	-10.2715	-2.5412
	vocation	3.5556	1.88708	.146	-.9020	8.0131
academic	general	6.4063*	1.63630	.000	2.5412	10.2715
	vocation	9.9619*	1.57799	.000	6.2345	13.6893
vocation	general	-3.5556	1.88708	.146	-8.0131	.9020
	academic	-9.9619*	1.57799	.000	-13.6893	-6.2345

Based on observed means.
The error term is Mean Square(Error) = 84.341.

*. The mean difference is significant at the .05 level.

Fig. 18 Post Hock test results for read within type of program

Comparison 1:

$$H_0: \mu_{general} = \mu_{academic}$$
$$H_1: \mu_{general} \neq \mu_{academic}$$

sig. = 0.000368 ($p < 0.05$)
Test is significant
Reject H_0
Conclusion: means of general and academic programs are not equal

Comparison 2:

$$H_0: \mu_{general} = \mu_{vocation}$$
$$H_1: \mu_{general} \neq \mu_{vocation}$$

sig. = 0.146 ($p < 0.05$)
Test is non-significant
Accept H_0
Conclusion: means of general and vocation programs are equal

Comparison 3:

$$H_0: \mu_{vocation} = \mu_{academic}$$
$$H_1: \mu_{vocation} \neq \mu_{academic}$$

sig. = 1.0654^{-8} ($p < 0.05$)
Test is significant
Reject H_0
Conclusion: means of vocation and academic programs are not equal

$$\mu_{academic} \neq (\mu_{general} = \mu_{vocation})$$

These results show that differences in reading test scores within general and academic, vocation and academic are significant. Moreover, the means in reading test scores within general and vocation statuses are equal.

Conclusions From performed tests it is identified that there a relationship between type of program students enrolled and their socioeconomic status. Moreover, it is also known that this relationship doesn't affect the reading score, but 2 predictors make an impact independently from each other. A two-way ANOVA was conducted that examined the effect of socioeconomic status and type of program on reading score. There was not statistically significant interaction between the effects of socioeconomic status and type of program on reading score, p = .663. Simple main effects analysis showed that high socioeconomic statuses students gained significantly more reading scores than others, and there higher reading scores in academic program in comparison to the remaining ones.

4.3 Writing score prediction

Methods Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable). The variables used to predict the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regressor variables). [5] In this case multiple regression is performed to understand whether writing exam score can be predicted based on gender of student and his test scores.

Analysis For analysis of data using multiple regression is needed to make sure that the data "passes" assumptions. [3]

Main assumptions for multiple regression:

1. Data must be normal.
2. Data must not show **multicollinearity**, which occurs when you have two or more independent variables that are highly correlated with each other.

Check for normality:

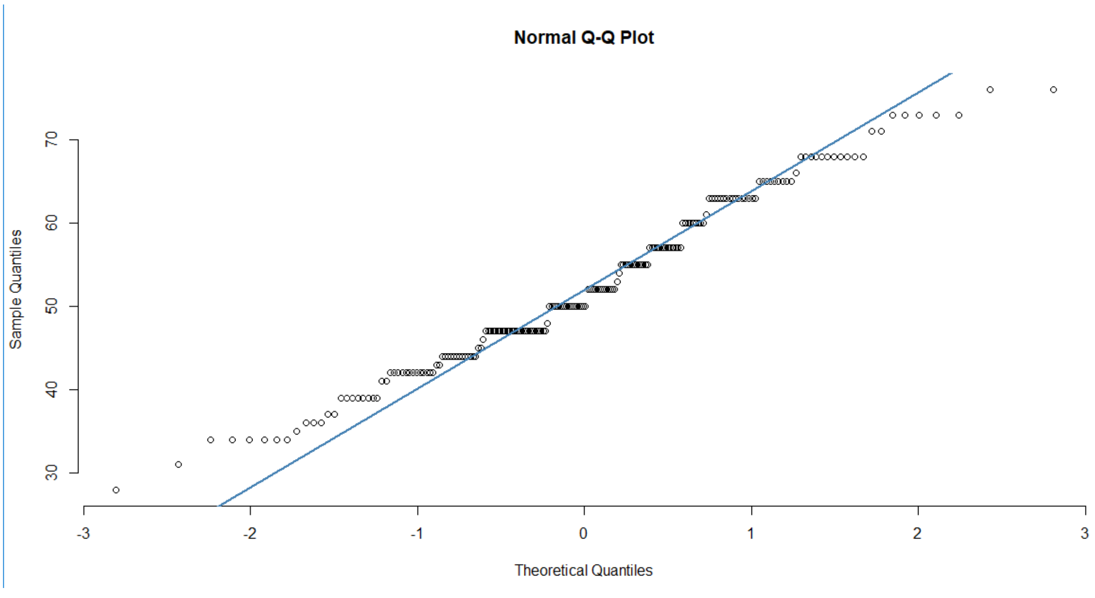

Fig. 19 Q-Q Plot for read score

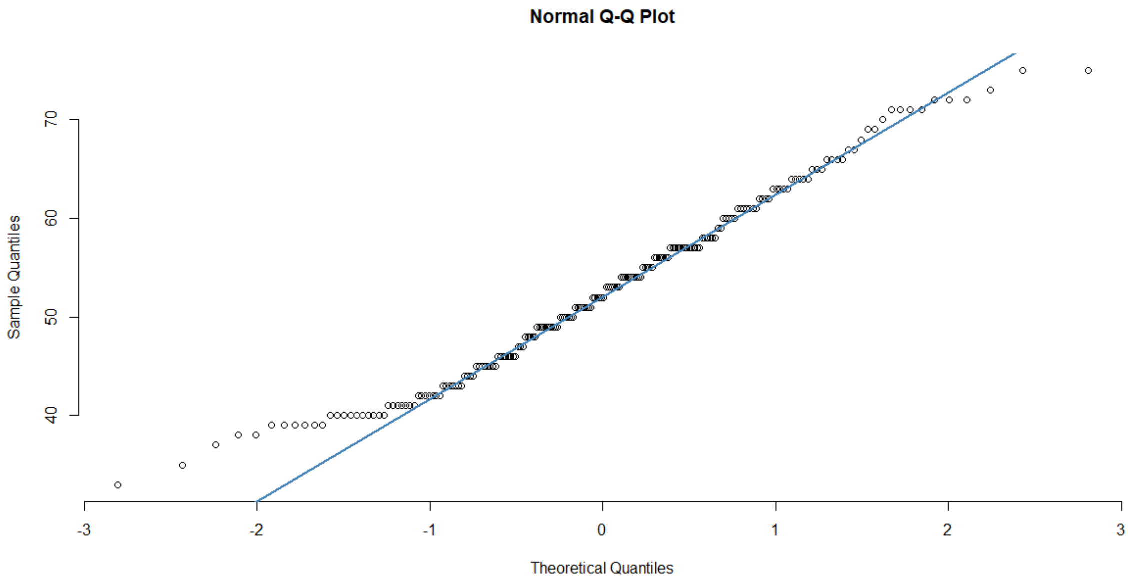

Fig. 20 Q-Q Plot for math score

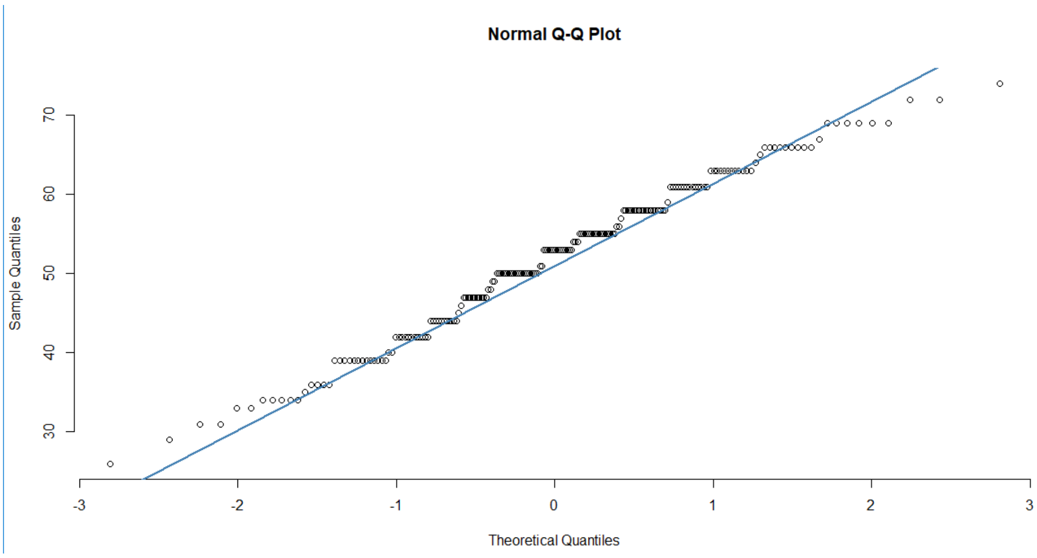

Fig. 21 Q-Q Plot for science score

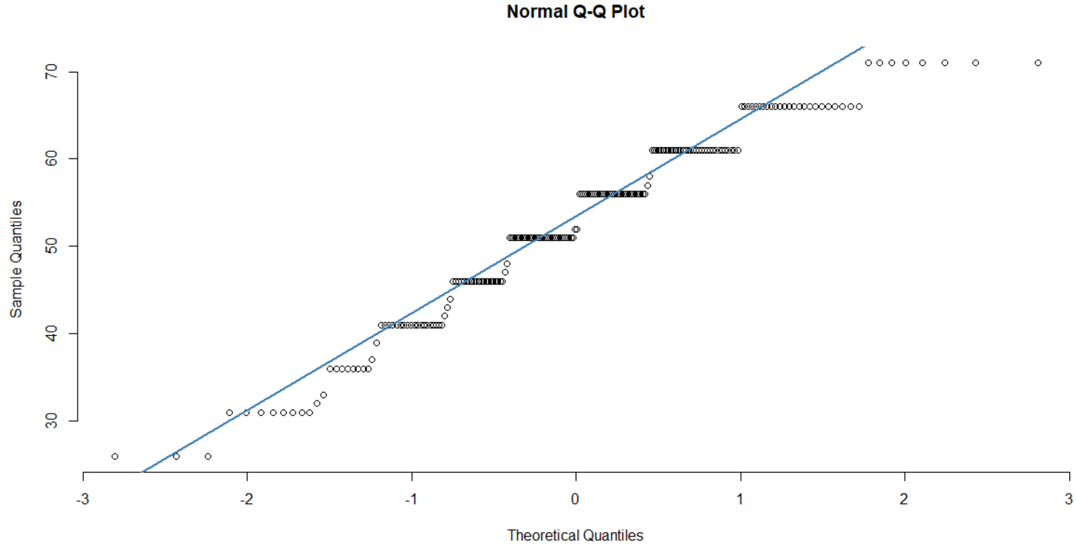

Fig. 22 Q-Q Plot for social study score

According to the Q-Q plots all data is assumed to be in a normal distribution. Next step is checking for multicollinearity.

Multicollinearity within model:

```
> vif(glm.fit1)
  female     read     math  science    socst
1.035324 2.404597 2.143782 1.957841 1.744156
```

Fig. 23 VIF values for predictors in built regression model

The numerical value for VIF tells (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells that the variance of a particular coefficient is 90% bigger than when expected if there was no multicollinearity — if there was no correlation with other predictors.

A rule of thumb for interpreting the variance inflation factor: 1 = not correlated, between 1 and 5 = moderately correlated, greater than 5 = highly correlated. From the Figure 23 above, it is seen that VIF values indicate low level of possible multicollinearity within predictors. Thereby, all assumptions for building multiple regression model are met.

Multiple regression:

```
> glm.fit1 <- lm(write ~ female + read + math + science + socst, data = f)
> summary(glm.fit1)

Call:
lm(formula = write ~ female + read + math + science + socst,
    data = f)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.040  -4.015  -0.264   3.938  14.989 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.13876    2.80842   2.186 0.030025 *  
female       5.49250    0.87542   6.274 2.24e-09 ***
read         0.12541    0.06496   1.931 0.054989 .  
math         0.23807    0.06713   3.547 0.000489 ***
science      0.24194    0.06070   3.986 9.51e-05 ***
socst        0.22926    0.05284   4.339 2.30e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.059 on 194 degrees of freedom
Multiple R-squared:  0.6017,	Adjusted R-squared:  0.5914 
F-statistic:  58.6 on 5 and 194 DF,  p-value: < 2.2e-16
```

Fig. 24 Multiple regression model results in R

As read predictor has low significance, it isn't considered in equation.

Y = 0.030025 + (2.24e-09*female) + (0.000489*math) + (9.51e-05*science) + (2.30e-05*socst)

$$p(write|female, math, science, socst) = \frac{e^{(0.030025\,+\,(2.24e-09*\mathrm{female})\,+\,(0.000489*\mathrm{math})\,+\,(9.51e-05*\mathrm{science})\,+\,(2.30e-05*\mathrm{socst}))}}{1 + e^{(0.030025\,+\,(2.24e-09*\mathrm{female})\,+\,(0.000489*\mathrm{math})\,+\,(9.51e-05*\mathrm{science})\,+\,(2.30e-05*\mathrm{socst}))}}$$

Female: A one-unit increase in female is associated with an increase in the Y by 2.24e-09 units
Math: A one-unit increase in math is associated with an increase in the Y by 0.000489 units.
Science: A one-unit increase in science is associated with an increase in Y by 9.51e-05 units.
Socst: A one-unit increase in socst is associated with an increase in the Y by 2.30e-05 units.

Now, the writing score can be predicted from the regression model.

```
> predict(glm.fit1, data.frame(female=0, read=49, math=45, science=57, socst=52), type="response")
       1 
48.70955 
```

Fig. 25 Prediction of writing score

Conclusions From performed tests it is possible now to predict writing score from gender of student and hos test scores. A multiple regression was run to predict writing score from gender and other test scores. These variables statistically significantly predicted writing score, despite reading score. All four variables added statistically significantly to the prediction, p < .05.

References

[1] UCLA Academic Technology Services. Reference: http://www.ats.ucla.edu/stat/data/hsb2.csv

[2] Laerd statistics, 2018. Reference: https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php

[3] Laerd statistics, 2018. Reference: https://statistics.laerd.com/spss-tutorials/chi-square-test-for-association-using-spss-statistics.php

[4] Laerd statistics, 2018. Reference: https://statistics.laerd.com/spss-tutorials/two-way-anova-using-spss-statistics.php

[5] Laerd statistics, 2018. Reference: https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php