Final assignment: Project report - a study of life expectancy and education

Martin Ottosson

2024-01-14

Introduction

This project aims at studying life expectancy and education in the form of years of schooling and literacy rate. The data comes from Gapminder(2024) and covers countries from the whole world for the years 2009 and 2011. The variables are defined as:

- life_exp = Life expectancy at birth. The number of years a newborn infant would live if the current mortality rates at different ages were to stay the same throughout its life.
- schooling_men = Mean years in school (men 25 years and older). The average number of years of school attended by all people in the age and gender group specified, including primary, secondary and tertiary.
- schooling_women = Mean years in school (women 25 years and older). The average number of years of school attended by all people in the age and - gender group specified, including primary, secondary and tertiary.
- literacy_rate = Literacy rate, adult total (% of people ages 15 and above). Adult literacy rate is the percentage of people ages 15 and above who can, with understanding, read and write a short, simple statement on their everyday life.

In the project the following research questions will be answered:

- Is there any correlation between life expectancy and years of schooling?
- Is there a difference between years of schooling for men and women?
- Is there a difference in life expectancy between continents?
- Is there an association between life expectancy and the literacy rate?

Kaplan, Spittel and Zeno (2014) raise a concern for the United States as life expectancy is falling behind countries of similar economic levels. One of their key points is that education is one of the best predictors of life expectancy, but it is hard to understand why. Eventhough this final project is not aiming at explaining the exact reasons, it will put some light on the issue and try to understand where some of the differences in life expectancy are when it comes to years of schooling and literacy rate.

The subject is interesting for social sciences since many factors may be behind why some countries have populations who are expected to live longer lives and it is an important topic for any decision maker who wants to invest in the well-being for its population. Kaplan et. al. point out that United States spend more than 17% of gross domestic product (GDP) on health care compared to around 10% for other similar countries, but still the U.S. is lagging behind. As an example from the gapminder data of life expectancy in 2009; the U.S. level was 78.6 years compared to example 81.7 in Sweden or 81.4% at its neighbor Canada.

Method

In part 1-3 the data comes from 2009 which is the last year with available data for mean number of school years. The dataset contains 174 observations. In part 4 literacy rate is also introduced and here data for 2011 is used since it is the last year with available data for literacy rate but primarily because the earlier years have a lot of missing data. The second dataset has 81 observations.

Countries were matched in two datasets in a spreadsheet program and if there was missing information for any of the variables, then the observation was removed. The datasets were then opened in a text editor, decimal signs were replaced with points and country names using special signs had these signs removed in order to avoid formatting issues. Some categorical variables were also created during the analysis both in the spreadsheet program and in R based on some conditions.

In the first dataset a variable called Continent was added with countries classified into the geographical continent it belongs to. Here the Central American countries were classified as North America (Britannica, 2024), Russia was classified as Asia and Turkey as Asia since their biggest landmasses are part of Asia.

In the second dataset variables for life expectancy and literacy rate intervals were created in R in order to treat them as categorical data. This was needed to perform analysis with frequency data and was the prefered way given the form of the available data.

The methods used in the project are:

Spearman's rank correlation was the preferred method for testing the relationship between life expectancy and years of schooling since the assumption of normally distributed data did not hold.

A test of difference between groups was performed with the Wilcoxon signed rank test for the sake of testing whether there is any difference between schooling years for men and women. Prior to the test a Shapiro-Wilk normality test came to the conclusion that a non-parametric test was really necessary due to the lack of normality.

Testing differences of more than two groups was made with the Kruskal-Wallis rank sum test after various transformations (arcsine, log10 and square root transformation) were tried for life expectancy without managing to fullfill necessary assumptions.

A chi-squared test and a Fisher's exact test was used for the analysis with frequency data to see if there is any association between life expectancy and the literacy rate.

Result

Part 1 - Correlations

The first part of the project looks into correlations between life expectancy and schooling.

Figure 1 shows the scatterplot between schooling for men and women. Even though not being one of the research questions it is still interesting to see it there is a relationship between schooling for men and women while we are at it. The plot shows a clear relationship and it is not very surprising that countries with many years in school for men also has it for women, but of course there is variation and in part 2 we will see if there is a difference of schooling between the genders.

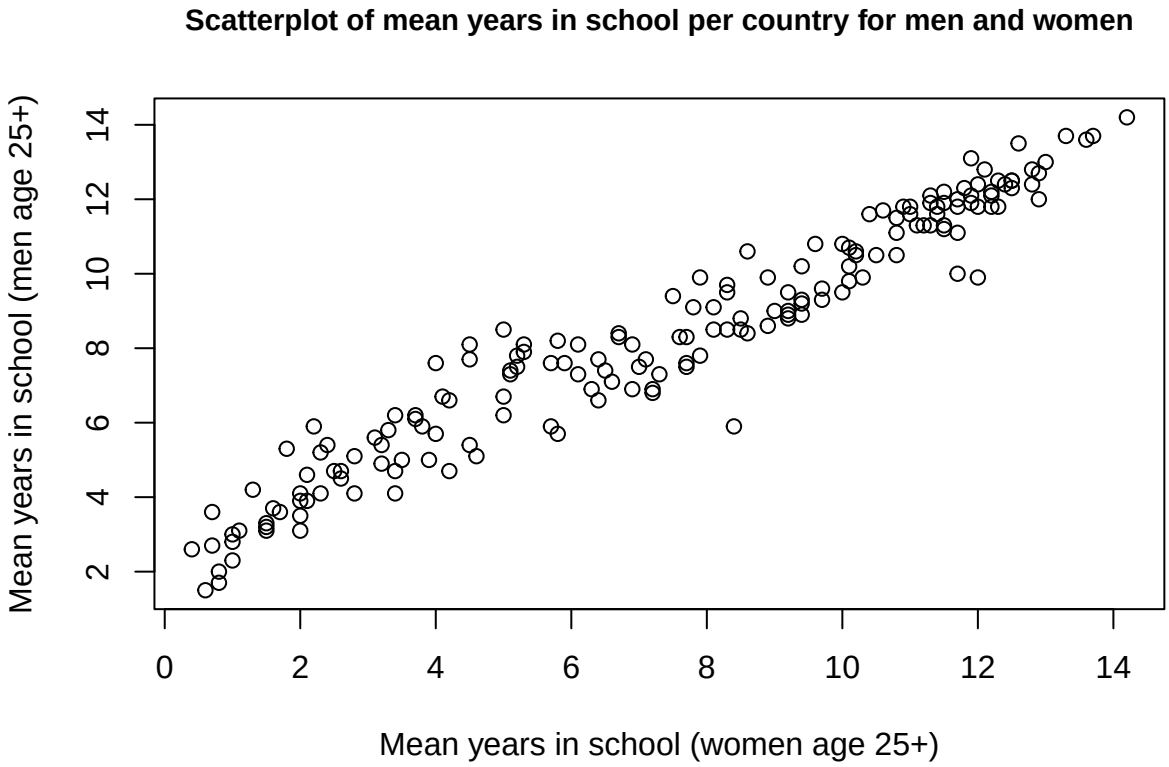

Figure 1

The relationship between schooling for men and life expectation in figure 2 is not very clear. But it looks like a positive relation.

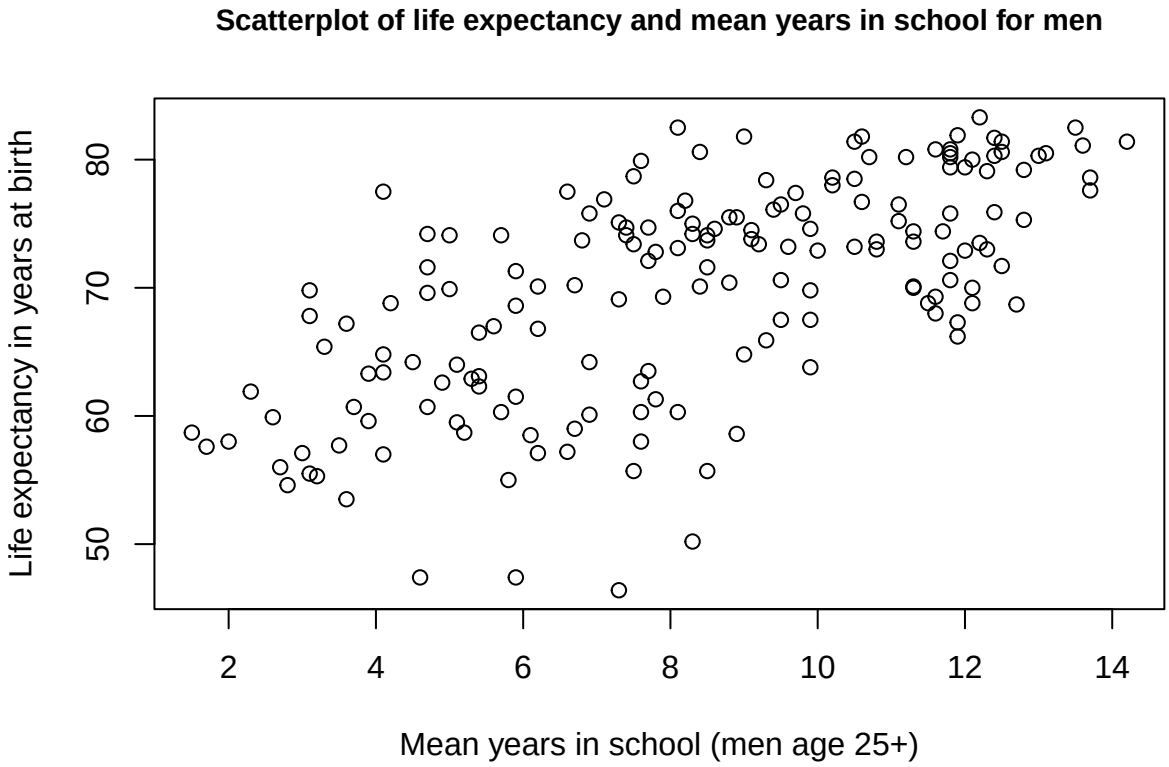

Figure 2

A similar pattern appears for women in figure 3, a positive relationship which may not be very strong.

Figure 3

Since we seem to have linear relationships it would be fair to test the Pearson's product-moment correlation to test if the correlations are significant, but first there are some other assumptions that also should be checked. Both variables need to be approximately normally distributed for the correlation test to be valid. Histograms show the distribution.

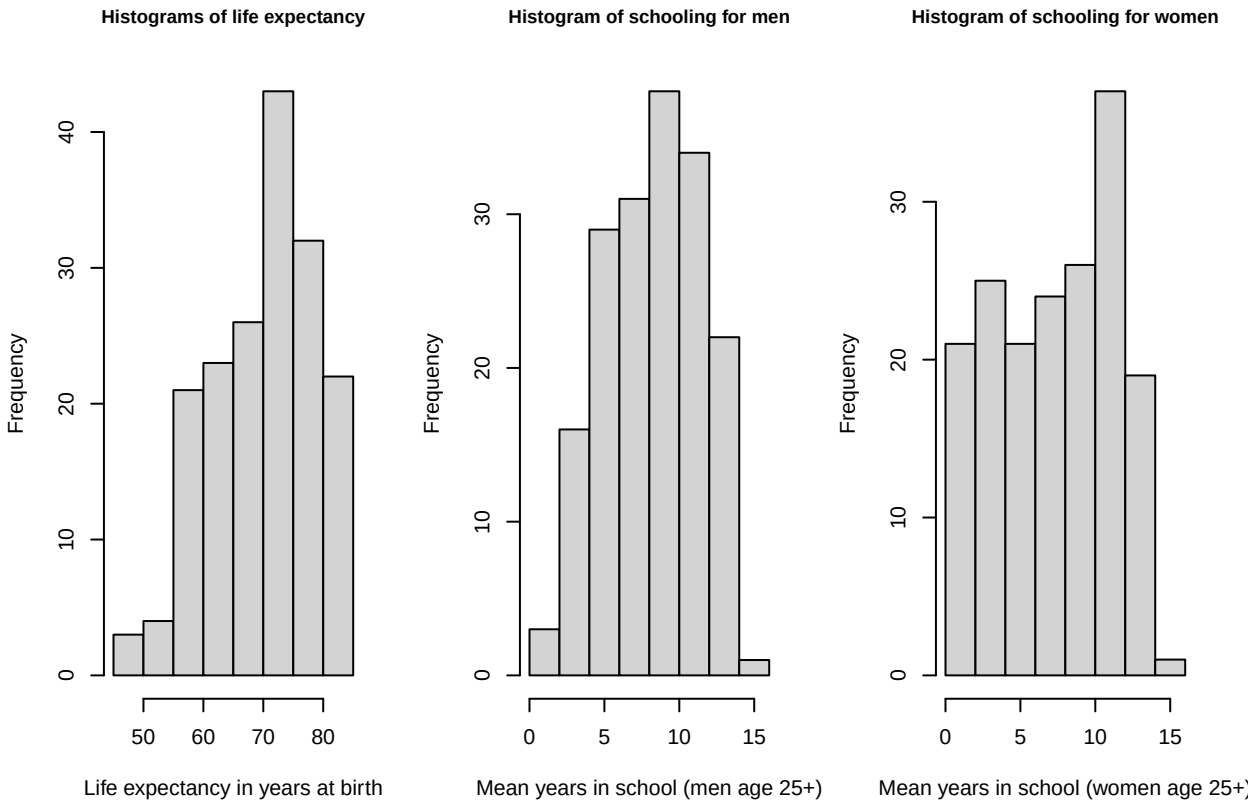

Figure 4, 5 and 6

The histograms in figure 4-6 do not show normality in a clear way for any of the three variables. The distribution for life expectancy looks a bit skewed and schooling for women looks almost evenly distributed. Schooling for men seems to be closest to an approximately normal distribution. Diagnostic plots for regression models might give us a clearer answer and can also be used to test the assumption of constant variation in the residuals (homoskedasticity).

Linear models are created where model 1 is schooling for men on the y-axis and schooling for women on the x-axis, model 2 and 3 have life expectancy on the y-axis and schooling for men and women on the x-axis respectively. Since the schooling variables are clearly correlated it is pointless to include both of them in a multiple regression model which would probably have problems with multicollinearity. Model 1 is not a realistic model because we cannot say anything about causality between schooling for men and women, but it is included as a way of controlling the assumptions for the correlation test.

The diagnostic plots in figure 7-9 show no signs of problems with heteroskedasticity in the residuals vs fitted plots (with the residuals as a function of fitted values). But when it comes to the Q-Q plots, it does not seem like all the points follow the straight line, especially not for model 2 and 3. So just as the histograms in figure 4-6 gave clues about not having approximately normal distributions, these plots amplify the conclusion about not fulfilling the assumption. The results from the regression models are not shown since we cannot be sure that we have valid results.

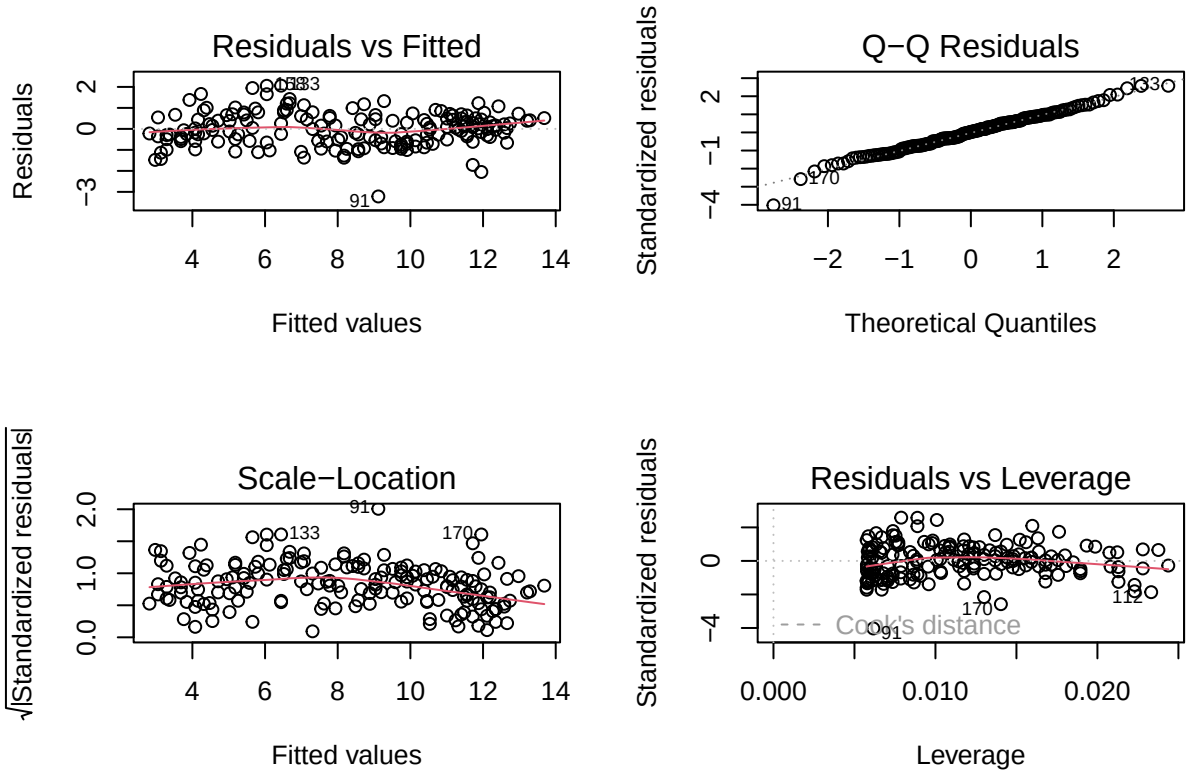

Figure 7 - Diagnostic plots for model 1 (y=schooling_men, x=schooling_women)

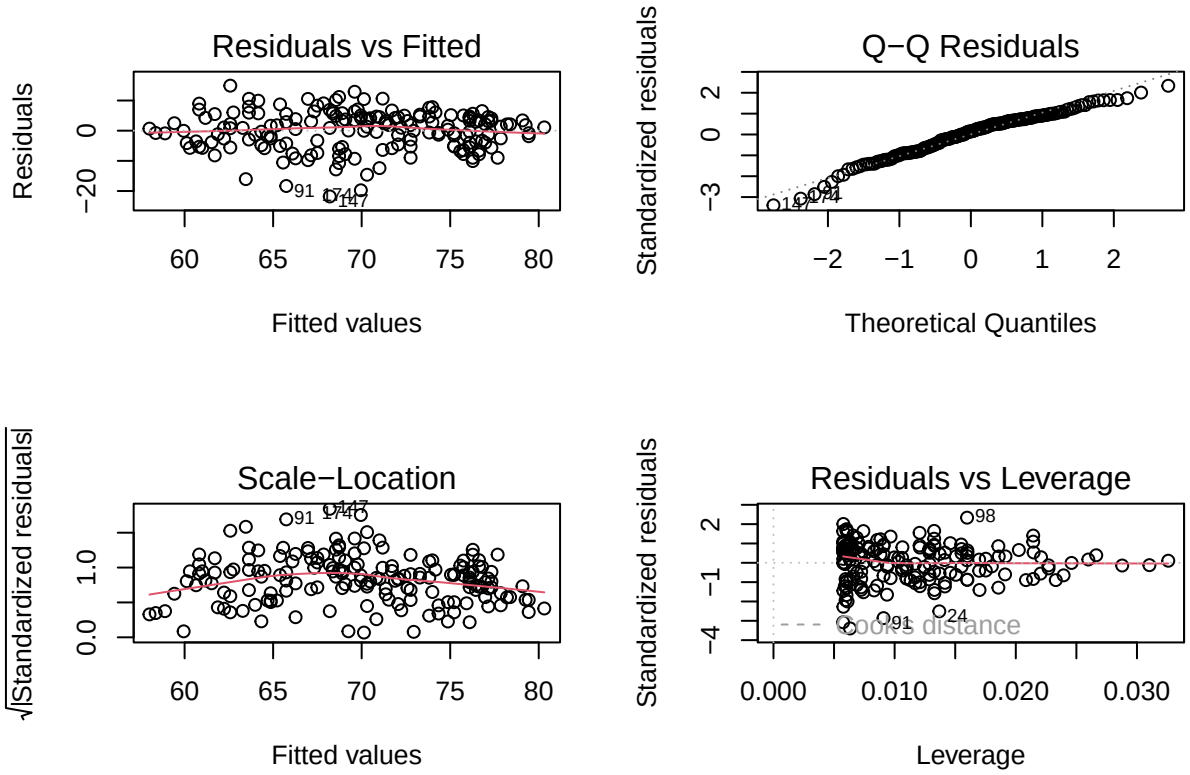

Figure 8 - Diagnostic plots for model 2 (y=life_exp, x=schooling_men)

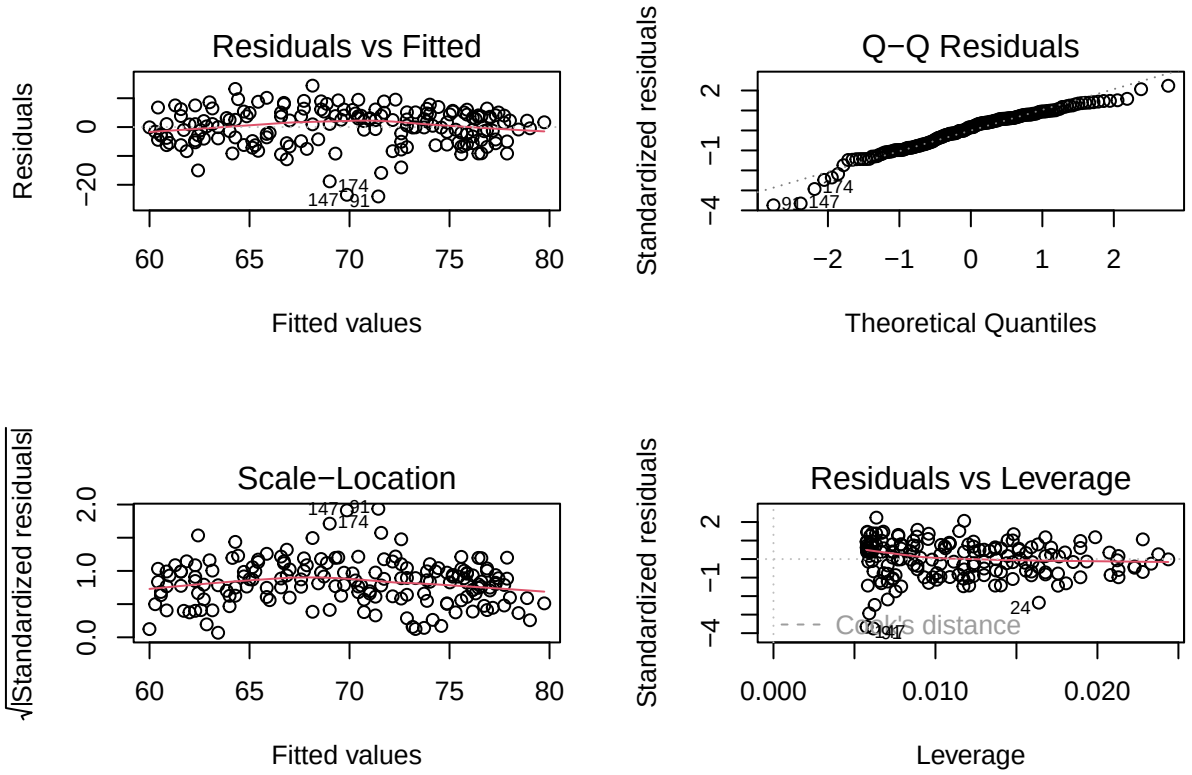

Figure 9 - Diagnostic plots for model 3 (y=life_exp, x=schooling_women)

Spearman's rank correlation tests are appropriate when any of the important assumptions do not hold. In table 1 the value of the test statistic is very high and clearly significant at the 5%-level. The null hypothesis of no correlation can be rejected. There is a positive correlation between mean years in school for men and women who are 25 years or older. The first impression from the scatterplot in figure 1 was correct. The test is two-sided, meaning it is a test for a correlation but not for the direction, but since the scatterplot was so clear and because the rho-value is 0.97, we can conclude there is a strong positive correlation. A value of rho equal to 1 means "a perfect positive relationship between two variables" (Statology, 2024).

Table 1 - Spearman's rank correlation test for mean years in school for men vs women

estimate	statistic	p.value	method	alternative
0.9700733	26274.87	0	Spearman's rank correlation rho	two.sided

The correlation between life expectancy and schooling for men is also significant, but with a lower value for the test statistic rho (table 2). We can conclude there is a correlation.

Table 2 - Spearman's rank correlation test for life expectancy and mean years in school for men

estimate	statistic	p.value	method	alternative
0.6560991	301936.4	0	Spearman's rank correlation rho	two.sided

The result of the correlation test for life expectancy and schooling for women in table 3 is very similar as the test above. There is a significant correlation. It can be concluded that the null hypothesis of no correlation can be rejected for all three tests and the alternative hypothesis of there being a correlation is accepted.

Table 3 - Spearman's rank correlation test for life expectancy and mean years in school for women

estimate	statistic	p.value	method	alternative
0.6589036	299474.1	0	Spearman's rank correlation rho	two.sided

Part 2 - Testing differences between groups

In this part we want to test if there is a difference between schooling years for men and women and as stated in part 1 there was a strong correlation between men and women. The variables are not independent since the data was gathered for the same countries and for this reason a paired test is necessary.

We saw previously that we could not assume approximately normally distributed data by merely looking at the histograms and Q-Q plots. To be sure we can check the normality assumption by doing a formal statistical test called Shapiro-Wilk normality test.

In all the three normality tests (table 4-6) the p-values are very low and the null hypothesis of normality can be rejected. The decision to not assume normality for the correlations tests in part 1 was appropriate. Consequently, a non-parametric test is needed for testing differences between groups. The Wilcoxon signed-rank test is appropriate since the data is paired.

Table 4 - Shapiro-Wilk normality test for life expectancy

statistic	p.value	method
0.9525699	1.37e-05	Shapiro-Wilk normality test

Table 5 - Shapiro-Wilk normality test for mean years in school (men age 25+)

statistic	p.value	method
0.9676321	0.000448	Shapiro-Wilk normality test

Table 6 - Shapiro-Wilk normality test for mean years in school (women age 25+)

statistic	p.value	method
0.9435009	2.2e-06	Shapiro-Wilk normality test

In the Wilcoxon test the alternative hypothesis is that schooling years for men is greater than schooling years for women. The boxplot in figure 10 gives the impression that men in general, world wide, have more years in school compared to women. The median for men is 8.35 and for women 7.7. With this in mind it is more interesting to also test for direction in the test rather than just stating there is a difference.

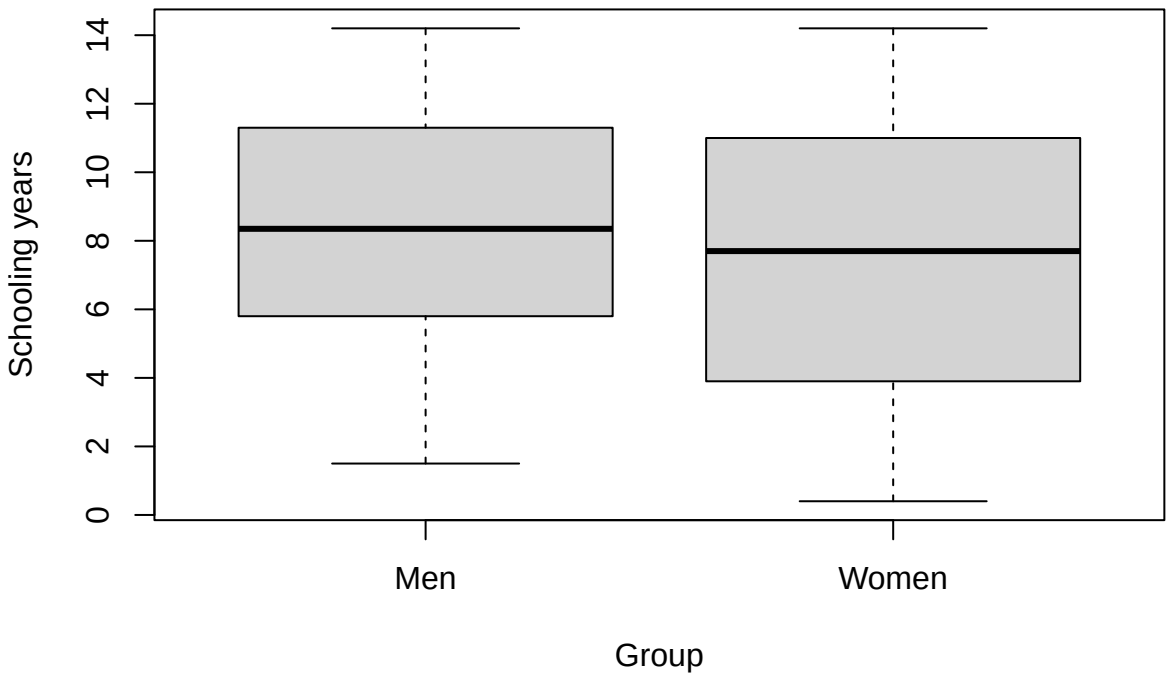

Figure 10 - Boxplots with schooling years for men and women (age 25+)

According to the results in table 7 with a very small p-value, the null hypothesis can be rejected and the alternative hypothesis that schooling years for men is greater than for women is accepted.

Table 7 - Wilcoxon signed rank test with continuity correction for testing the difference between schooling years for men and women

statistic	p.value	method	alternative
11321	0	Wilcoxon signed rank test with continuity correction	greater

Part 3 - Testing differences between more than two groups

Part 3 is focusing on whether there is a difference between continents when it comes to life expectancy. As mentioned in the method section the Central American countries are classified as North America and Russia and Turkey are classified as Asia. The reasons are geographical and it is important to have in mind that it can affect the results.

After running an ANOVA-test for testing differences between more than two groups and considering the diagnostic plots we can assume homoskedasticity. The spread of the largest group does not look like more than three times the spread compared to the smallest group in the Residuals vs Fitted plot (figure 11). But the Q-Q plot look suspicious and it should not be a surprise, because we could not assume normality for life expectancy earlier in the analysis. Subsequently, it is necessary to try some other approach.

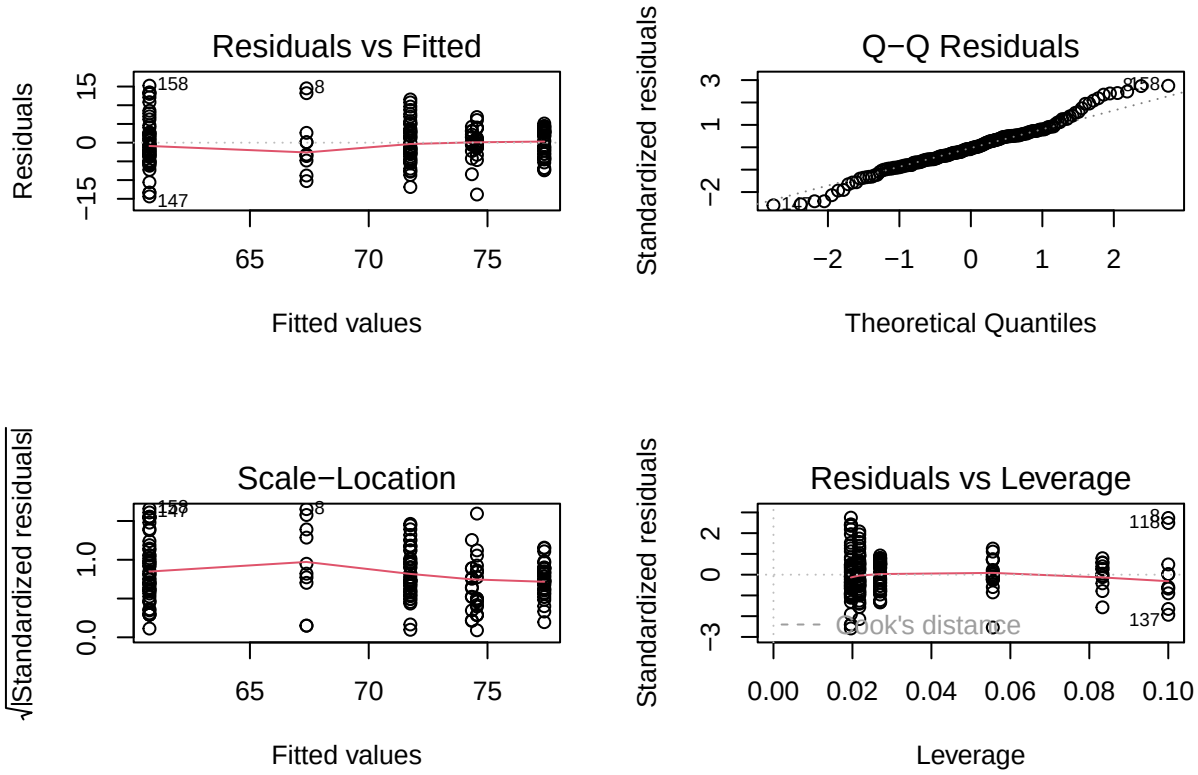

Figure 11 - Diagnostic plots from ANOVA test of life expectancy grouped in continents

The ANOVA-test could still be valid with some data transformation. Figure 12 presents the diagnostic plots with an arcsine transformation of life expectation. The Residuals vs Fitted plot looks ok, but it does still not look like a normal distribution as is shown in the Q-Q plot.

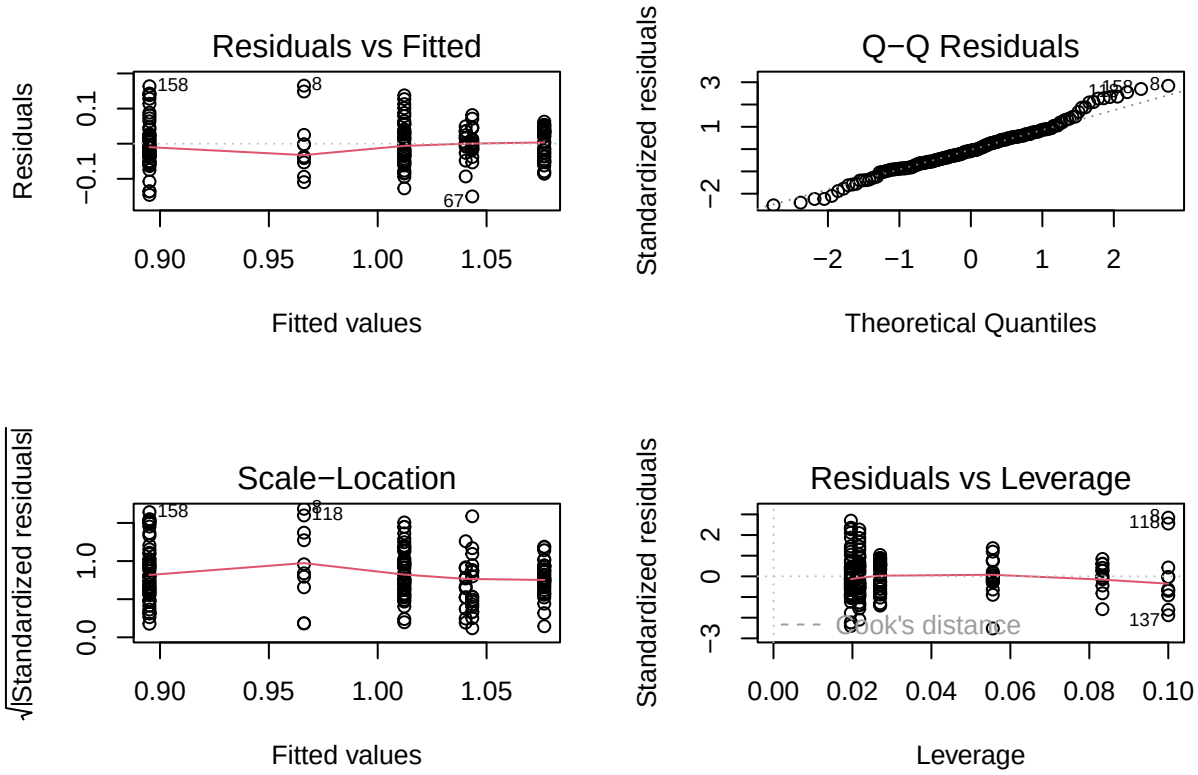

Figure 12 - Diagnostic plots from ANOVA test of arcsine transformed life expectancy grouped in continents

Figure 13 displays the diagnostic plots of the ANOVA-test with a log10 transformation of life expectancy and it is not decided to assume normality.

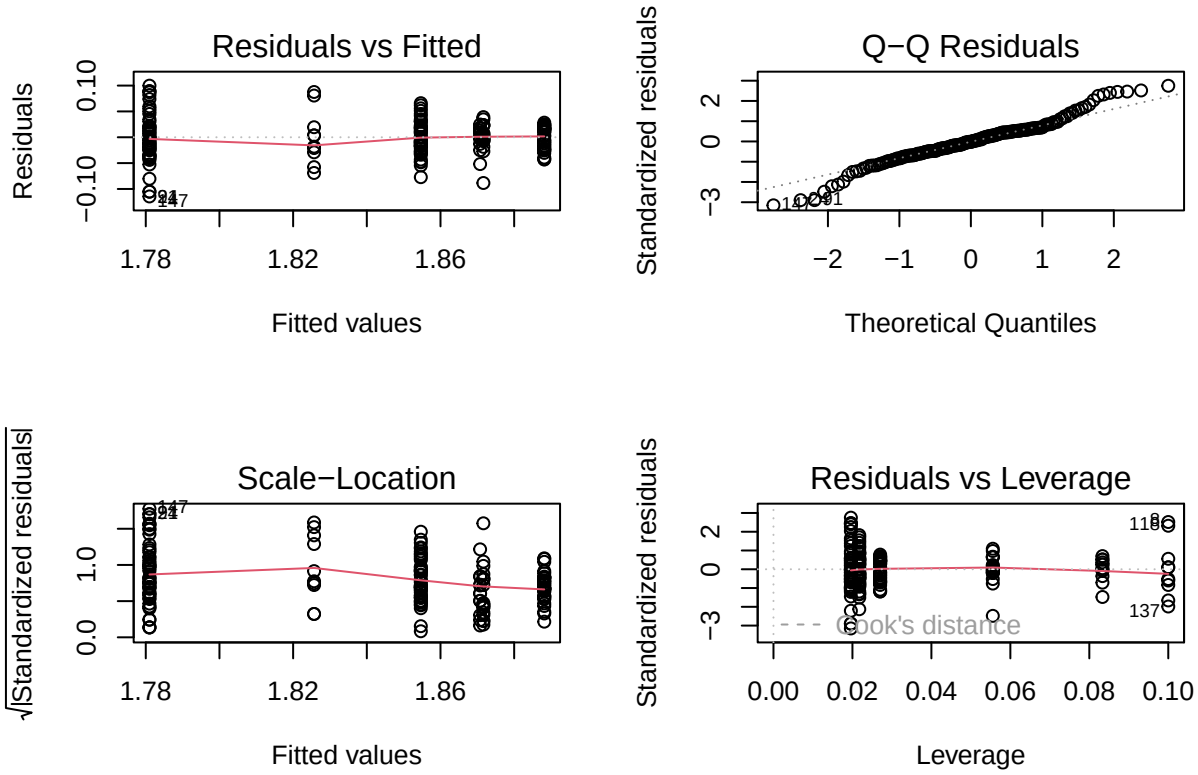

Figure 13 - Diagnostic plots from ANOVA test of log10 transformed life expectancy grouped in continents

Normality can also not be assumed with a square root transformation as the Q-Q plot in figure 14 reveals.

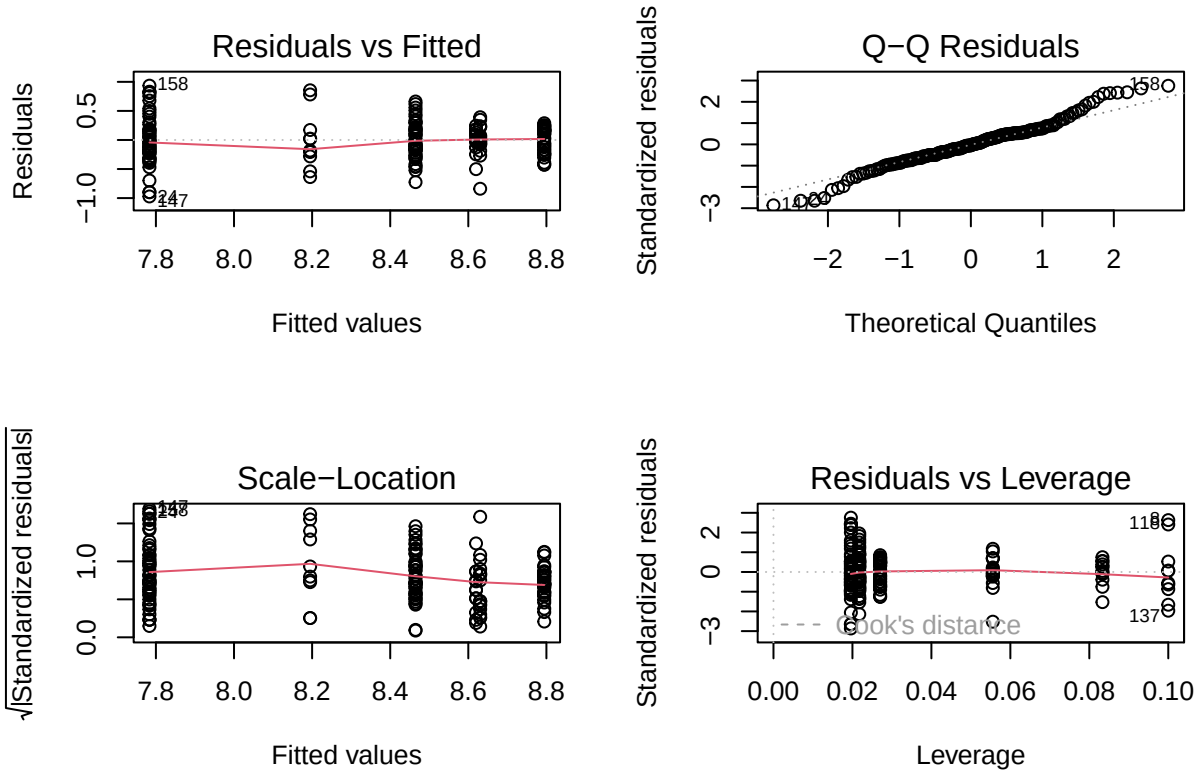

Figure 14 - Diagnostic plots from ANOVA test of square root transformed life expectancy grouped in continents

After trying with various transformations it is determined to continue with a non-parametric test called Kruskal-Wallis rank sum test.

The Kruskal-Wallis rank sum test in table 8 gives a high value for the test statistic, 5 degrees of freedom and a very low p-value. We conclude there is a difference between continents when it comes to life expectancy.

Table 8 - Kruskal-Wallis rank sum test for differences of life expectancy between continents

statistic	p.value	parameter	method
94.96082	0	5	Kruskal-Wallis rank sum test

With regard to the signficant result from the Kruskal-Wallis test we are interested in knowing more about where the differences are, i.e. between which continents is there a difference in the amount of life expectancy years at birth. A post hoc test has significant results for:

- Africa-Asia
- Africa-Europe
- Africa-North America
- Africa-South America
- Asia-Europe
- Europe-Oceania

There is a significant difference in life expectancy between these continents and the boxplots in figure 15 makes it easy to visually see these differences by comparing the medians between continents. Perhaps the

biggest difference is between Africa and Europe. North America and South America are at similar life expectancy levels according to the boxplots and the difference between them is not significant in the post hoc test. If this is surprising we have to keep in mind that the Central American countries are part of North America. Both United States and Canada are above the median, but as we know life expectancy in the U.S. is not as high as it could be with the size of its economy in mind. Costa Rica for example had 80.6 years in the data compared to 78.6 for the U.S. Regarding the difference between Europe and Oceania there are two outliers for Oceania; Australia (80.5 years) and New Zealand (80.6 years) and they are among the top levels of Europe.

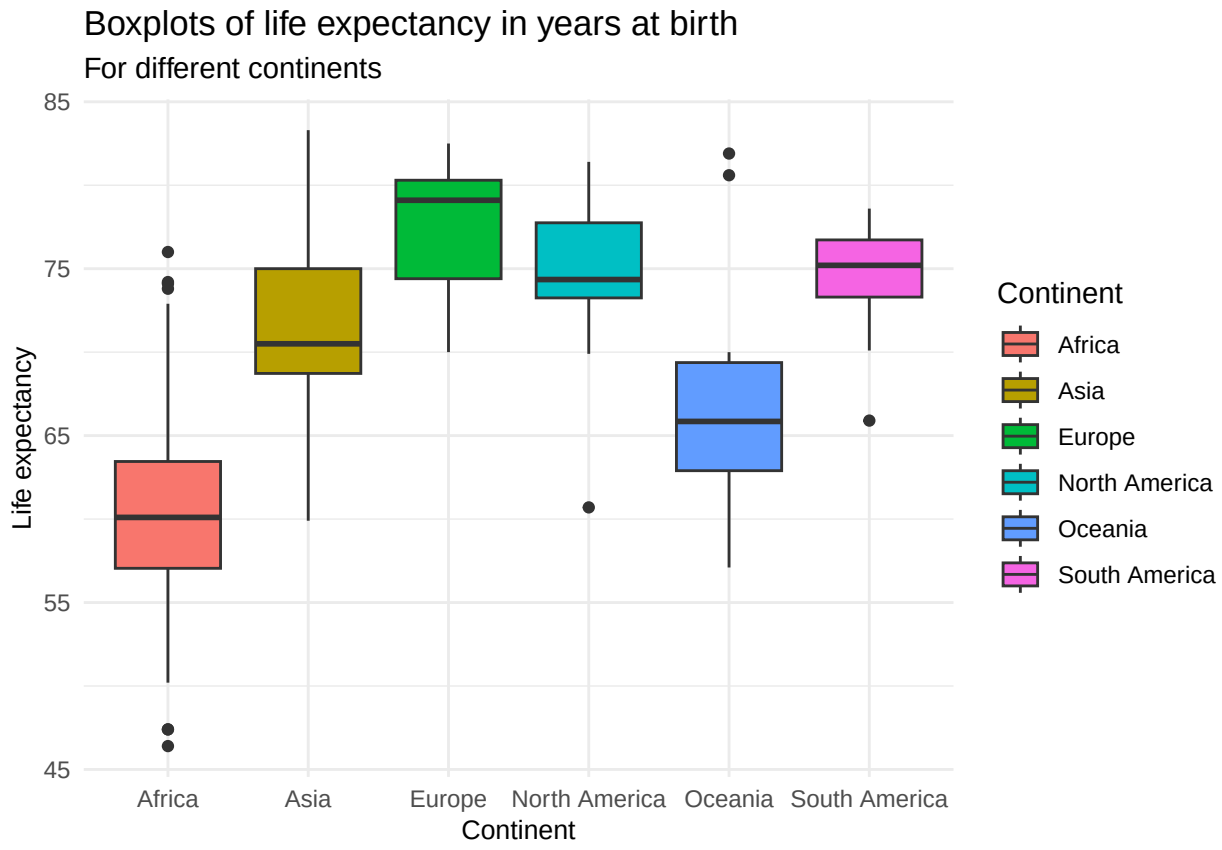

Figure 15

Part 4 - Tests with frequency data

In the last part of the project we are using another dataset from 2011. The variables are life expectancy and literacy rate as described in the introduction.

In this analysis we are doing non-parametric tests with frequency data. Since the dataset does not contain frequency data some data manipulation is necessary and it is done by creating categorical variables with intervals. Figure 16 shows the variables with a scatterplot and it is our starting point for deciding the intervals.

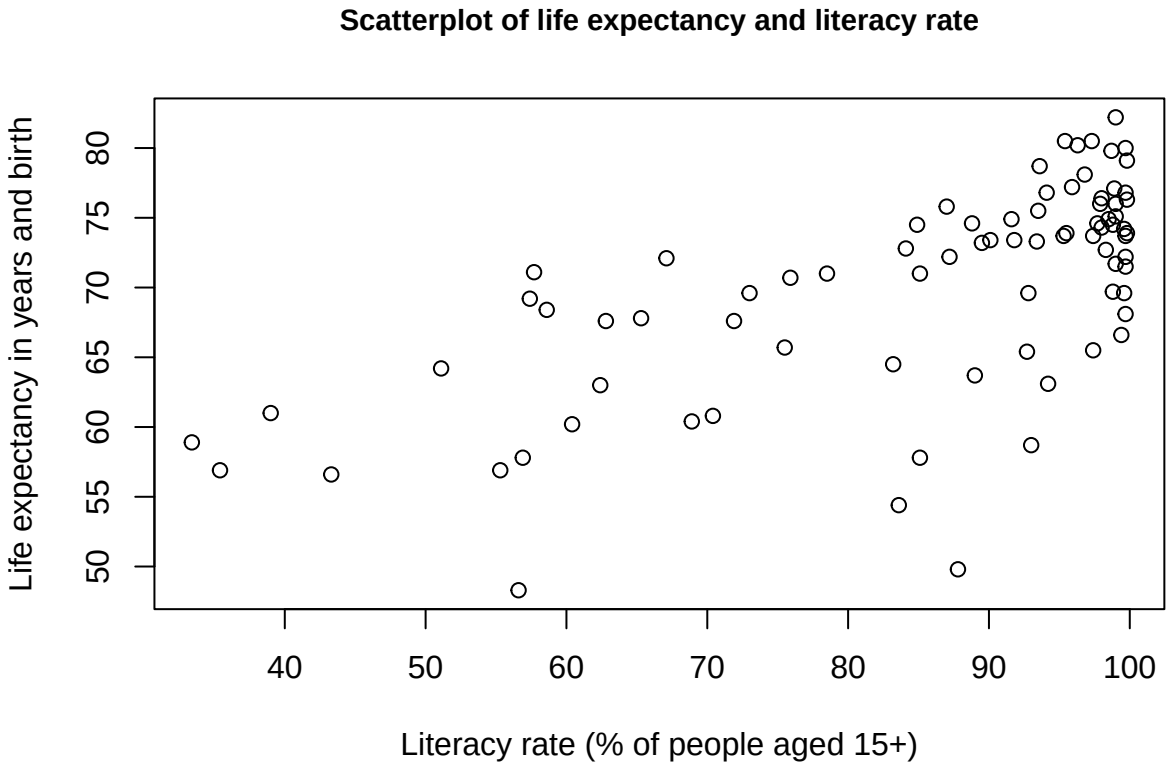

Figure 16

A first attempt was made with four intervals for life expectancy and five intervals for literacy rate, creating a 4 by 5 contingency table. However, there were several cells with no counts and an approach with a smaller table is taken.

We draw the line at the median, so categorical variables are created for life expectancy above or equal to 72.2 years, or below, and literacy rate above or equal to 93%, or below. This gives a 2 by 2 contingency table as seen in table 9 with two cells with more than 30 counts and two cells with 9 counts each.

Table 9 - Contingency table (2*2) for life expectancy and literacy rate

	Above median literacy rate	Below median literacy rate
Above median life expectancy	32	9
Below median life expectancy	9	31

Then a Chi-squared test for association is done, which is a non-parametric test suited for count data when we have no clear expectation beforehand about the result, and we can conclude there is an association between life expectancy and literacy rate because of the low p-value (table 10). A high literacy rate seem to be associated with a high life expectancy and vice versa, just by examining the contingency table and knowing that the chi-squared test is significant.

Table 10 - Chi-squared test of association between life expectancy and literacy rate

statistic	p.value	parameter	method
24.9939	6e-07	1	Pearson's Chi-squared test

An expected count is performed in table 11 and calculated as column sum * row sum / total sum. There is a clear difference between the actual frequency and expected frequency in all of the four cells.

Table 11 - Expected counts for the groups of life expectancy and literacy rate

	Above median literacy rate	Below median literacy rate
Above median life expectancy	20.75309	20.24691
Below median life expectancy	20.24691	19.75309

The chi-squared components are also calculated. The sum of each component is equal to the Chi-Squared test statistic approximately (but they are rounded in table 12).

Table 12 - Chi-squared components

	Above median literacy rate	Below median literacy rate
Above median life expectancy	6.10	6.25
Below median life expectancy	6.25	6.40

To show the distribution of the contributions in percentage values each value can be divided by the total value of the test statistic (table 13). We can see that they are quiet evenly distributed.

Table 13 - Chi-squared contributions

	Above median literacy rate	Below median literacy rate
Above median life expectancy	24.39	25.00
Below median life expectancy	25.00	25.62

Figure 17 shows the contributions visually. A larger contribution is shown as a larger and darker circle and the largest contribution comes from below median life. exp. and below median literacy rate. Above median life. exp. and above median literacy rate has no circle. It is not because it doesn't contribute. They all contribute with around 25%, but this one has the lowest share. The graph kind of overstates the difference in contributions, but the conclusion is that there is a significant association between life expectancy and literacy rate. They are dependent.

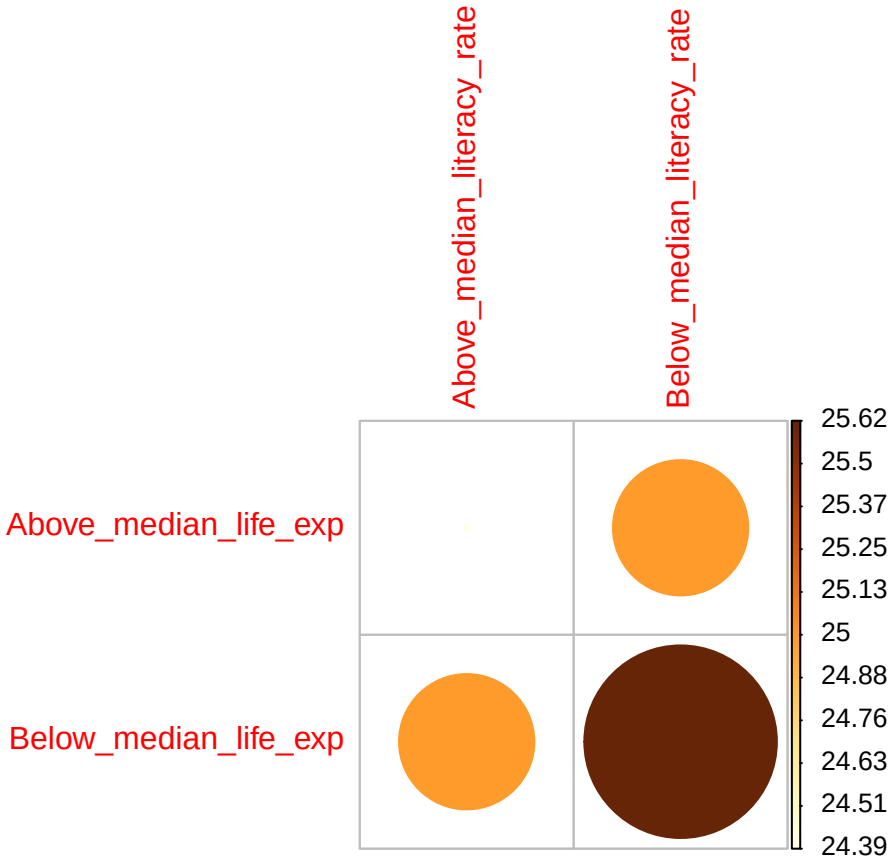

Figure 17 - Correlation plot of Chi-Squared contributions from associations

Another way to look at the result from the Chi-squared test is with the Pearson residuals which show the direction between the observed and expected counts. Table 14 shows positive residuals for high life expectancy and high literacy rate as well as for low life expectancy and low literacy rate and the other two cells are negative.

Table 14 - Pearson residuals for the Chi-squared test

	Above median literacy rate	Below median literacy rate
Above median life expectancy	2.468835	-2.499505
Below median life expectancy	-2.499505	2.530556

Figure 18 presents the residuals visually in another correlation plot with blue circles where the expected counts are higher than actual counts and red circles for the opposite.

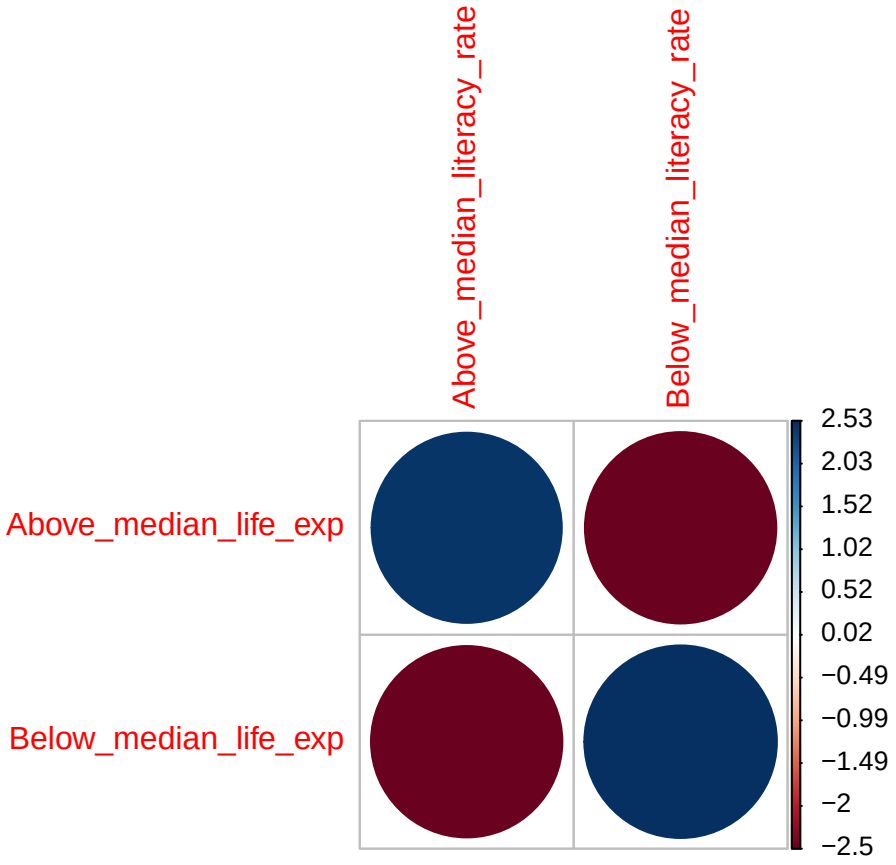

Figure 18 - Correlation plot of Pearson residuals for the Chi-squared test

We still want to make the test a little bit more interesting with more cells in the contingency table so new categorical variables are made with three levels each separated at the 33th and 66th percentile. The meaning of the intervals in table 15 are:

- Low life expectancy: Below the 33rd percentile (less than 68.22 years)
- Middle life expectancy: Between the 33rd and up to the 66th percentile (between 68.22 and 74.13 years)
- High life expectancy: Above or equal to the 66th percentile (between 74.14 and 82.2 years)
- Low literacy rate: Below the 33rd percentile (less than 84.98%)
- Middle literacy rate: Between the 33rd and up to the 66th percentile (between 84.98 and 97.37%)
- High literacy rate: Above or equal to the 66th percentile (between 97.38 and 99.8%)

Table 15 - Contingency table (3*3) for life expectancy and literacy rate

	High literacy rate	Middle literacy rate	Low literacy rate
High life expectancy	18	6	3
Middle life expectancy	8	9	9
Low life expectancy	1	11	16

There are a couple of cells with less than 5 counts as shown in table 16. When there are cells with few counts it is preferred to do a Fisher's exact test. The test is significant (table 16) at the 5%-level and we can conclude there is an association between life expectancy and literacy rate even when looking at smaller intervals.

Table 16 - Fisher's Exact Test for Count Data

p.value	method	alternative
9.6e-06	Fisher's Exact Test for Count Data	two.sided

Discussion

This study has focused on life expectancy and its connection to education in terms of years of schooling for men and women as well as literacy. Kaplan, Spittel and Zeno stated that education is one of the key predictors for life expectancy and this study has made an effort in investingating whether there is a relationship with education in the form of schooling years, where the differences are and if there is any association between life expectancy and literacy rate.

One of the main conclusions is that there indeed is a relationship between schooling and life expectancy with the support of significant non-parametric Spearman's rank correlation tests. Years in school is significantly correlated to life expectancy for both men and women and we could also show that there is a difference between men and women in this matter. Men have more years in school than women as the Wilcoxon signed rank test was significant. A common field of study within economics is wage discrimination between men and women; looking into unexplained wage differences. Surely, education would be an important control factor in such a study, but what if women do not have the same availability to education as men. The difference of schooling found in this study indicates that women might not have the same chance of going to school as much as men. We have not looked into salaries in this study, but we know that schooling is significantly related to life expectancy and if women could reach the same level of schooling as men, then it would most likely have a positive impact on life expectancy levels in these countries where there is a gap.

A Kruskal-Wallis rank sum test then gave a significant test statistic for differences of life expectancy between groups. Here we mentioned how the countries were classified in continents and that North America is not as high as one could have thought. Clearly much could be done in the world, just look at the life expectancy difference between Africa and Europe (figure 15). By comparing the medians there is a difference of (not exactly but almost) 20 years and with the dispersion shown by the boxplot for Africa we can tell there are some countries with very low levels. Leshoto (47.4 years) and Eswatini (46.4 years) are two outliers for Africa with extremely low levels. Some things that are worth to mention when the data is at country level is that countries differ in population and size of the economy, but in these statistical tests they are all treated equally, i.e. they are not weighted in any way.

Finally, a chi-squared test was made with count data and we could see a big difference in expected counts and actual counts in the contingency table with life expectancy above and below the median as well as the literacy rate above and below the median. The test showed an association between these two variables with this setup. A Fisher's exact test was also significant in a somewhat bigger contingency table with 9 cells, where the limits were set at the 33rd and 66th percentile for each variable. Both of the tests then came to the conclusion that life expectancy and the literacy rate are dependent.

To conclude, we have showed that education really is something closely related to life expectancy and more exactly in the form of total years in school and measured as the literacy rate. There is a difference between schooling for men and women and life expectancy differs between continents. We can therefore say that Kaplan, Spittel and Zeno were right when claiming that education is one of the best predictors for life expectancy. Education can however be measured in different ways and other variables such as if having a university degree or not, results in standardized tests etc. could also be of interest to test. Most interestingly would be to develop a more complex model, such as a multiple regression model, where other important predictors were included such as health care spend, gross domestic product, some measure for income equality

and so one. Then, all else equal, we could come to more exact conclusions and be able to give estimates for how much any specific measure for education predicts life expectancy.

References

Britannica, 2024. Central America. https://www.britannica.com/place/Central-America (Retrieved 2024-01-14).

Gapminder, 2024. Download the data - Choose individual indicators. https://www.gapminder.org/data/. (Retrieved 2024-01-05).

Kaplan, Robert M. Spittel, Michael L. Zeno, Tia L. (2014). Educational Attainment and Life Expectancy. https://journals.sagepub.com/doi/pdf/10.1177/2372732214549754 (Retrieved 2024-01-13).

Statology, 2024. How to Calculate Spearman Rank Correlation in R. https://www.statology.org/spearman-correlation-in-r/ (Retrieved 2024-01-10).

Appendix

In part 3 the analysis focused on differences between continents with regards to life expectancy. Another interesting question is if there are differences of life expectancy between different levels of schooling for women. We saw earlier that women indeed has a lower amount of schooling years than men, so we will focus on women in this extra analysis (not being one of the main research questions). The variable for years of schooling is not normally distributed so a non-parametric test is once again suitable.

First a new categorical variable with two levels is created; women with schooling above or equal to the median, which is 7.7 years and women with schooling below the median years. With this way of separating the variable in two parts we can test if there is a difference in life expectancy for women with a high level of schooling compared two a low level of schooling where high and low is considered as being more or less than what is normal in the world, i.e. the median.

A Kruskal-Wallis rank sum test (table 17) presents a high value of the test statistic, 1 degree of freedom and with a very low p-value. We accept that there is a difference between the two groups for life expectancy.

Table 17 - Kruskal Wallis rank sum test for differences of life expectancy between women with high or low amount of years in school

statistic	p.value	parameter	method
54.65737	0	1	Kruskal-Wallis rank sum test

There is no need for a post-hoc test since we only have two groups, but boxplots gives a clear picture of the difference. Figure 19 shows a difference of the medians of around 10 years in life expectancy for women with high amount of schooling years compared to the group with less schooling. There is of course variation but we can be confident enough to say that education is important for a woman to have an expected longer life. The analysis is not done with data on the individual level and it would be interesting to do the test with individual data within a country. Here we have looked at aggregate levels of life expectancy and mean years of school in each country.

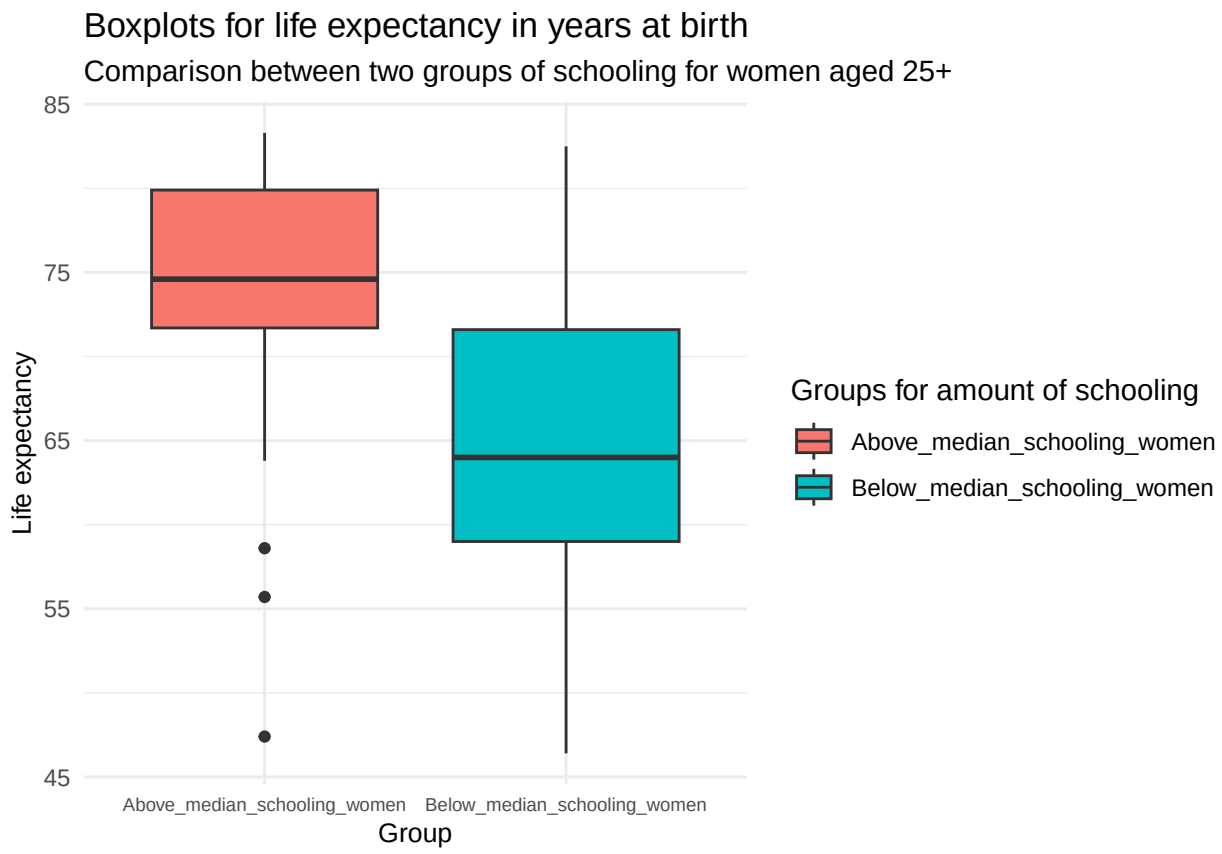

Figure 19

R-code

In this part of the appendix the R code from the study is presented.

```
### Libraries
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2) # Not sure if it is being used
library(pgirmess) # For Kruskal-Wallis post-hoc test
library(esquisse) # For doing boxplots
library(corrplot) # Used in Chi-Squared test
library(tables) # Used for flextables
library(flextable) # Used for flextables
library(knitr)
library(dplyr)
library(broom) # Used for kable tables

### Reading data for part 1 and 2
final_project_1 <- read.table("C:/Users/mart0/OneDrive/Dokument/Martin/Södertörns högskola/
Statistisk analys och visualisering i R - I/Part 7/Final_project/final_project_1.txt",
header = TRUE, sep = ",", dec = ".")

### Checking data
View(final_project_1)
str(final_project_1)
```

```
summary(final_project_1)

### Figure 1
attach(final_project_1)
plot(schooling_men ~ schooling_women,
     xlab = "Mean years in school (women age 25+)",
     ylab = "Mean years in school (men age 25+)",
     main = "Scatterplot of mean years in school per country for men and women",
     cex.main = 0.9)

### Figure 2
plot(life_exp ~ schooling_men,
     xlab = "Mean years in school (men age 25+)",
     ylab = "Life expectancy in years at birth",
     main = "Scatterplot of life expectancy and mean years in school for men",
     cex.main = 0.9)

### Figure 3
plot(life_exp ~ schooling_women,
     xlab = "Mean years in school (women age 25+)",
     ylab = "Life expectancy in years at birth",
     main = "Scatterplot of life expectancy and mean years in school for women",
     cex.main = 0.9)

### Figure 4, 5 and 6
par(mfrow=c(1,3))
hist(life_exp,
     xlab = "Life expectancy in years at birth",
     main = "Histograms of life expectancy",
     cex.main = 0.9)

hist(schooling_men,
     xlab = "Mean years in school (men age 25+)",
     main = "Histogram of schooling for men",
     cex.main = 0.9)

hist(schooling_women,
     xlab = "Mean years in school (women age 25+)",
     main = "Histogram of schooling for women",
     cex.main = 0.9)

### Figure 7
model_1 <- lm(schooling_men ~ schooling_women)
model_2 <- lm(life_exp ~ schooling_men)
model_3 <- lm(life_exp ~ schooling_women)
par(mfrow=c(2,2))
plot(model_1)

### Figure 8
par(mfrow=c(2,2))
plot(model_2)

### Figure 9
par(mfrow=c(2,2))
```

```
plot(model_3)

### Table 1
cor_test_1 <- cor.test(schooling_men, schooling_women, method = "spearman")
cor_test_1_table <- tidy(cor_test_1)
kable(cor_test_1_table)

### Table 2
cor_test_2 <- cor.test(life_exp, schooling_men, method = "spearman")
cor_test_2_table <- tidy(cor_test_2)
kable(cor_test_2_table)

### Table 3
cor_test_3 <- cor.test(life_exp, schooling_women, method = "spearman")
cor_test_3_table <- tidy(cor_test_3)
kable(cor_test_3_table)

### Table 4
shapiro_test_1 <- shapiro.test(life_exp)
shapiro_test_1_table <- tidy(shapiro_test_1)
kable(shapiro_test_1_table)

### Table 5
shapiro_test_2 <- shapiro.test(schooling_men)
shapiro_test_2_table <- tidy(shapiro_test_2)
kable(shapiro_test_2_table)

### Table 6
shapiro_test_3 <- shapiro.test(schooling_women)
shapiro_test_3_table <- tidy(shapiro_test_3)
kable(shapiro_test_3_table)

### Figure 10
par(mfrow=c(1,1))
final_project_1a <- c(final_project_1$schooling_men,final_project_1$schooling_women)
label <- factor(c(rep("Men",174), rep("Women",174)))
boxplot(final_project_1a~label, xlab = "Group", ylab = "Schooling years", outline = TRUE)

### Table 7
wilcox_test_1 <- wilcox.test(schooling_men, schooling_women, paired=TRUE, alternative = "greater")
wilcox_test_1
detach(final_project_1)
wilcox_test_1_table <- tidy(wilcox_test_1)
kable(wilcox_test_1_table)

### Reading data for part 3
final_project_2 <- read.table("C:/Users/mart0/OneDrive/Dokument/Martin/Södertörns högskola/
Statistisk analys och visualisering i R - I/Part 7/Final_project/final_project_2.txt",
header = TRUE, sep = "\t", dec = ".")

# Checking data
View(final_project_2)
str(final_project_2)
summary(final_project_2)
```

```
### Figure 11
# One-way ANOVA
anova_1 <- aov(life_exp ~ Continent, data = final_project_2)
anova_1
summary(anova_1)
par(mfrow=c(2,2))
plot(anova_1)

### Figure 12
# Creates a new variable in the final_project_2 dataset for arcsine transformed life_exp
final_project_2$life_exp_arc <- asin(sqrt(final_project_2$life_exp/100))
View(final_project_2)
anova_2 <- aov(life_exp_arc ~ Continent, data = final_project_2)
anova_2
summary(anova_2)
par(mfrow=c(2,2))
plot(anova_2)

### Figure 13
# log transformation
final_project_2$life_exp_log <- log10(final_project_2$life_exp)
View(final_project_2)
anova_3 <- aov(life_exp_log ~ Continent, data = final_project_2)
anova_3
summary(anova_3)
par(mfrow=c(2,2))
plot(anova_3)

### Figure 14
# Square root transformation
final_project_2$life_exp_sqrt <- sqrt(final_project_2$life_exp)
View(final_project_2)
anova_4 <- aov(life_exp_sqrt ~ Continent, data = final_project_2)
anova_4
summary(anova_4)
par(mfrow=c(2,2))
plot(anova_4)

### Table 8
# Non-parametric test called Kruskal-Wallis rank sum test
kruskal_test_1 <- kruskal.test(life_exp ~ Continent, data = final_project_2)
kruskal_test_1
summary(kruskal_test_1)
kruskal_test_1_table <- tidy(kruskal_test_1)
kable(kruskal_test_1_table)
attach(final_project_2)
kruskal_test_1_adhoc <- kruskalmc(life_exp, Continent, probs = 0.05, cont=NULL)
detach(final_project_2)

### Figure 15
#esquisser()
ggplot(final_project_2) +
 aes(x = Continent, y = life_exp, fill = Continent) +
```

```
 geom_boxplot() +
 scale_fill_hue(direction = 1) +
 labs(x = "Continent", y = "Life expectancy", title = "Boxplots of life expectancy in years at birth",
 subtitle = "For different continents") +
 theme_minimal() +
 theme(axis.title.y = element_text(size = 10L),
 axis.title.x = element_text(size = 10L))

### Reading data for part 4
final_project_3 <- read.table("C:/Users/mart0/OneDrive/Dokument/Martin/Södertörns högskola/
Statistisk analys och visualisering i R - I/Part 7/Final_project/final_project_3.txt",
header = TRUE, sep = "\t", dec = ".")

# Checking data
View(final_project_3)
str(final_project_3)
summary(final_project_3)
sd(final_project_3$life_exp)
sd(final_project_3$literacy_rate)

### Figure 16
plot(final_project_3$life_exp ~ final_project_3$literacy_rate,
     xlab = "Literacy rate (% of people aged 15+)",
     ylab = "Life expectancy in years and birth",
     main = "Scatterplot of life expectancy and literacy rate",
     cex.main = 0.9)

### Table 9
final_project_3$life_exp_intervals <- ifelse(final_project_3$life_exp >= 72.2,
"Above_median_life_exp","Below_median_life_exp")

final_project_3$literacy_rate_intervals <- ifelse(final_project_3$literacy_rate >= 93,
"Above_median_literacy_rate","Below_median_literacy_rate")
View(final_project_3)
Table_1 <- table(final_project_3$life_exp_intervals, final_project_3$literacy_rate_intervals)
# Creating data frame as a matrix
Table_1_matrix <- data.frame(
  "Above_median_literacy_rate" = c(Table_1[1, 1], Table_1[2, 1]),
  "Below_median_literacy_rate" = c(Table_1[1, 2], Table_1[2, 2])
)
# Add a new column with row names
row_names <- c("Above median life expectancy", "Below median life expectancy")
Table_1_matrix <- cbind(Row_Names = row_names, Table_1_matrix)
# Create a flextable
ft1 <- flextable(Table_1_matrix)
# Set column names
col_names <- c("", "    Above median literacy rate  ", "Below median literacy rate")
ft1 <- set_header_labels(ft1, values = col_names)
# Set column widths
ft1 <- flextable::set_table_properties(ft1, width = .5, layout = "autofit")
# Align values to the center
ft1 <- flextable::align(ft1, align ="center", part = "all")
# Print the table
ft1
```

```
### Table 10
Test1_chisq <- chisq.test(Table_1, correct=FALSE)
Test1_chisq_table <- tidy(Test1_chisq)
kable(Test1_chisq_table)

### Table 11
exp_freq1 <- Test1_chisq$expected
# Creating data frame as a matrix
exp_freq1_matrix <- data.frame(
  "Above_median_literacy_rate" = c(exp_freq1[1, 1], exp_freq1[2, 1]),
  "Below_median_literacy_rate" = c(exp_freq1[1, 2], exp_freq1[2, 2])
)
# Add a new column with row names
row_names <- c("Above median life expectancy", "Below median life expectancy")
exp_freq1_matrix <- cbind(Row_Names = row_names, exp_freq1_matrix)
# Create a flextable
ft2 <- flextable(exp_freq1_matrix)
# Set column names
col_names <- c("", "    Above median literacy rate  ", "Below median literacy rate")
ft2 <- set_header_labels(ft2, values = col_names)
# Set column widths
ft2 <- flextable::set_table_properties(ft2, width = .5, layout = "autofit")
# Align values to the center
ft2 <- flextable::align(ft2, align ="center", part = "all")
# Print the table
ft2

### Table 12
components1 <- round(Test1_chisq$residuals^2, 2)
# Creating data frame as a matrix
components1_matrix <- data.frame(
  "Above_median_literacy_rate" = c(components1[1, 1], components1[2, 1]),
  "Below_median_literacy_rate" = c(components1[1, 2], components1[2, 2])
)

# Add a new column with row names
row_names <- c("Above median life expectancy", "Below median life expectancy")
components1_matrix <- cbind(Row_Names = row_names, components1_matrix)
# Create a flextable
ft3 <- flextable(components1_matrix)
# Set column names
col_names <- c("", "    Above median literacy rate  ", "Below median literacy rate")
ft3 <- set_header_labels(ft3, values = col_names)
# Set column widths
ft3 <- flextable::set_table_properties(ft3, width = .5, layout = "autofit")
# Align values to the center
ft3 <- flextable::align(ft3, align ="center", part = "all")
# Print the table
ft3

### Table 13
contrib1 <- 100*Test1_chisq$residuals^2/Test1_chisq$statistic
contrib1_round <- round(contrib1, 2)
```

```
# Creating data frame as a matrix
contrib1_round_matrix <- data.frame(
  "Above_median_literacy_rate" = c(contrib1_round[1, 1], contrib1_round[2, 1]),
  "Below_median_literacy_rate" = c(contrib1_round[1, 2], contrib1_round[2, 2])
)

# Add a new column with row names
row_names <- c("Above median life expectancy", "Below median life expectancy")
contrib1_round_matrix <- cbind(Row_Names = row_names, contrib1_round_matrix)
# Create a flextable
ft4 <- flextable(contrib1_round_matrix)
# Set column names
col_names <- c("", "    Above median literacy rate  ", "Below median literacy rate")
ft4 <- set_header_labels(ft4, values = col_names)
# Set column widths
ft4 <- flextable::set_table_properties(ft4, width = .5, layout = "autofit")
# Align values to the left
ft4 <- flextable::align(ft4, align ="center", part = "all")
# Print the table
ft4

### Figure 17
corrplot(contrib1, is.cor = FALSE, cl.align.text = "l")

### Table 14
Test1_chisq_residuals <- Test1_chisq$residuals
# Creating data frame as a matrix
Test1_chisq_residuals_matrix <- data.frame(
  "Above_median_literacy_rate" = c(Test1_chisq_residuals[1, 1], Test1_chisq_residuals[2, 1]),
  "Below_median_literacy_rate" = c(Test1_chisq_residuals[1, 2], Test1_chisq_residuals[2, 2])
)
# Add a new column with row names
row_names <- c("Above median life expectancy", "Below median life expectancy")
Test1_chisq_residuals_matrix <- cbind(Row_Names = row_names, Test1_chisq_residuals_matrix)
# Create a flextable
ft5 <- flextable(Test1_chisq_residuals_matrix)
# Set column names
col_names <- c("", "    Above median literacy rate  ", "Below median literacy rate")
ft5 <- set_header_labels(ft5, values = col_names)
# Set column widths
ft5 <- flextable::set_table_properties(ft5, width = .5, layout = "autofit")
# Align values to the left
ft5 <- flextable::align(ft5, align ="center", part = "all")
# Print the table
ft5

### Figure 18
corrplot(Test1_chisq$residuals, is.cor = FALSE, cl.align.text = "l")

### Table 15
# Calculate the percentiles
percentile_33_life_exp <- quantile(final_project_3$life_exp, probs = 0.33)
percentile_66_life_exp <- quantile(final_project_3$life_exp, probs = 0.66)
percentile_33_life_exp
```

```
percentile_66_life_exp
boxplot(final_project_3$life_exp)
percentile_33_literacy_rate <- quantile(final_project_3$literacy_rate, probs = 0.33)
percentile_66_literacy_rate <- quantile(final_project_3$literacy_rate, probs = 0.66)
percentile_33_literacy_rate
percentile_66_literacy_rate
boxplot(final_project_3$literacy_rate)
plot(final_project_3$life_exp ~ final_project_3$literacy_rate)
# Creating new variables in order to have count data
# Now the limits are based on the 33 and 66 percentiles
final_project_3$life_exp_intervals3 <- ifelse(final_project_3$life_exp < 68.22,
"Below_68.22_years",
ifelse(final_project_3$life_exp >= 68.22 & final_project_3$life_exp < 74.14,
"Between_68.22_and_74.13_years","Between_74.14_and_82.2_years"))
final_project_3$literacy_rate_intervals3 <- ifelse(final_project_3$literacy_rate < 84.98,
"Below_84.98%",
ifelse(final_project_3$literacy_rate >= 84.98 & final_project_3$literacy_rate < 97.38,
"Between_84.98_and_97.37%","Between_97.38_and_99.8%"))
View(final_project_3)
summary(final_project_3)
Table_2 <- table(final_project_3$life_exp_intervals3, final_project_3$literacy_rate_intervals3)
Table_2_matrix <- data.frame(
  "Low literacy rate" = c(Table_2[1, 1], Table_2[2, 1], Table_2[3, 1]),
  "Middle literacy rate" = c(Table_2[1, 2], Table_2[2, 2], Table_2[3, 2]),
  "High literacy rate" = c(Table_2[1, 3], Table_2[2, 3], Table_2[3, 3])
)
# Add a new column with row names
row_names <- c("High life expectancy", "Middle life expectancy", "Low life expectancy")
Table_2_matrix <- cbind(Row_Names = row_names, Table_2_matrix)
# Create a flextable
ft6 <- flextable(Table_2_matrix)
# Set column names
col_names <- c("", "    High literacy rate  ", "Middle literacy rate  ", "Low literacy rate")
ft6 <- set_header_labels(ft6, values = col_names)
# Set column widths
ft6 <- flextable::set_table_properties(ft6, width = .5, layout = "autofit")
# Align values to the left
ft6 <- flextable::align(ft6, align ="center", part = "all")
# Print the table
ft6

### Table 16
Test2_fisher <- fisher.test(Table_2)
Test2_fisher_table <- tidy(Test2_fisher)
kable(Test2_fisher_table)

### Table 17
summary(final_project_2)
final_project_2$schooling_women_above_median <- ifelse(final_project_2$schooling_women >= 7.7,
                                                       "Above_median_schooling_women",
                                                       "Below_median_schooling_women")
kruskal_test_2 <- kruskal.test(life_exp ~ schooling_women_above_median, data = final_project_2)
kruskal_test_2
summary(kruskal_test_2)
```

```
kruskal_test_2_table <- tidy(kruskal_test_2)
kable(kruskal_test_2_table)

### Figure 19
#esquisser()
ggplot(final_project_2) +
 aes(x = schooling_women_above_median, y = life_exp, fill = schooling_women_above_median) +
 geom_boxplot() +
 scale_fill_hue(direction = 1) +
 labs(x = "Group", y = "Life expectancy", title = "Boxplots for life expectancy in years at birth",
 subtitle = "Comparison between two groups of schooling for women aged 25+",
 fill = "Groups for amount of schooling") +
 theme_minimal() +
 theme(axis.title.y = element_text(size = 10L),
 axis.title.x = element_text(size = 10L),
 axis.text.x = element_text(size = 7L))
```