Assignment 1 by Martin Ottosson

Part 1. Correlation between GDP and access to personal computers in the year of 2005

Author: Martin Ottosson

Introduction

Method

Result

Is there a relationship between GDP/capita and the number of personal computers per 100 individuals measured in 2005?

First the data called GDP_PC is checked.

```
View(GDP_PC)
str(GDP_PC)
typeof(GDP_PC)
```

It can be seen that both GDPperCAPITA and PC_per_100 (number of computers per 100 habitants) are of type character.
Therefore they are converted to numeric.

```
attach(GDP_PC)
gdp_per_capita <- as.numeric(GDPperCAPITA)
gdp_per_capita
typeof(gdp_per_capita)

pc_per_100 <- as.numeric(PC_per_100)
pc_per_100
typeof(pc_per_100)
```

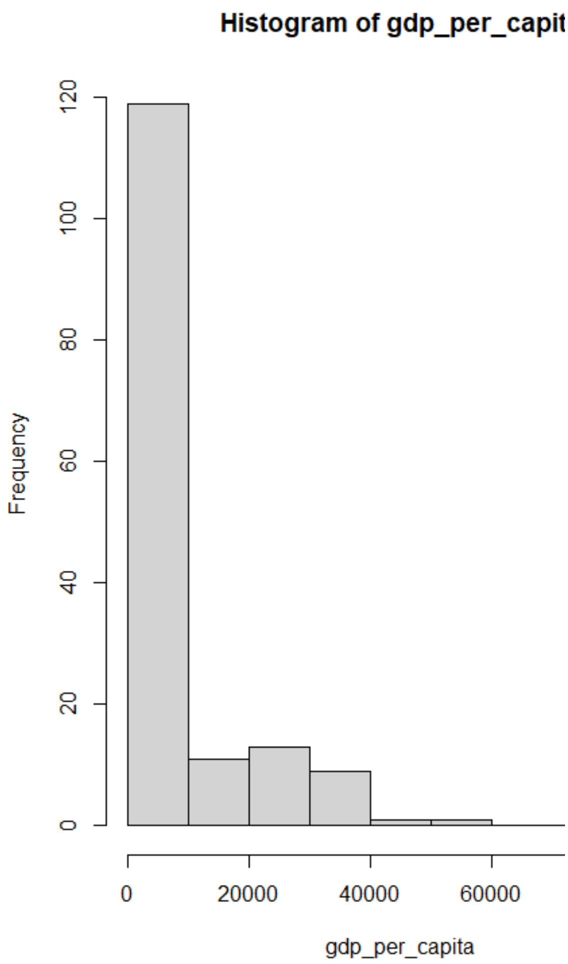
Then summary statistics are calculated.

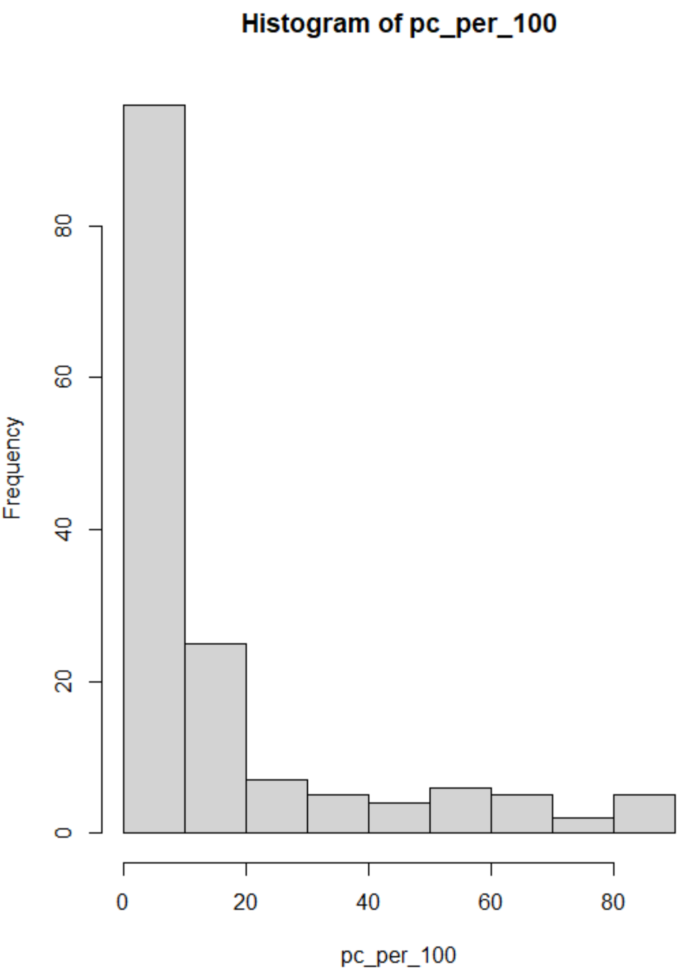
```
summary(gdp_per_capita)
summary(pc_per_100)
sd(gdp_per_capita)
sd(pc_per_100)
```

	GDP per capita	PC per 100
Mean	7963.5	15.781
Median	2196.2	5.9
Standard deviation	12328.81	22.12

In order to choose which test is appropriate it is important to check if the variables are approximately normally distributed. A histogram is good way to see the distribution of the data.

```
hist(gdp_per_capita)
```





```
hist(pc_per_100)
```

The variables do not seem to be approximately normally distributed according to the histograms. The mean and the median are far apart as can be seen in the table and the observations are not evenly distributed according to the scatterplot.

Therefore the Spearman's rank correlation is better to use.

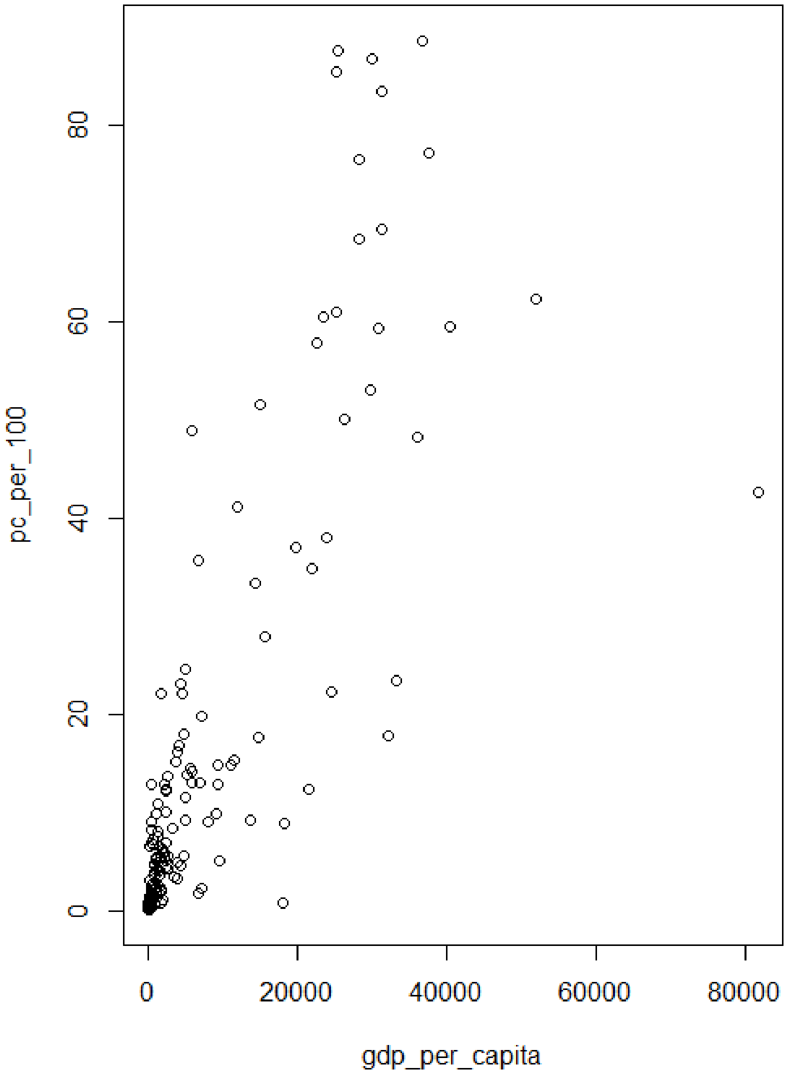
```
cor.test(gdp_per_capita, pc_per_100, method = "spearman")
```

Spearman's rank correlation rho

S	90137
Correlation coefficient rho	0.85
df	Not reported, but n – 2 = 155 – 2 = 153
P-value	< 0.01 (2.2e-16)

There is a significant positive correlation between GDP per capita and PC_per_100.

```
plot(pc_per_100 ~ gdp_per_capita)
detach(GDP_PC)
```



Discussion

As stated in the results section the Spearman's rank correlation test was chosen since an approximal normal distribution could not be seen for GDP per capita or PC per 100. A formal way to do hypothesis testing is to follow these steps:

1. Formulate a null hypothesis H0.
2. Calculate the test statistic
3. Calculate the significance probability
4. Reject or accept the null hypothesis

In this case the hypotheses are the following:

H0: There is no correlation between GDP per capita and PC per 100

H1: There is correlation between GDP per capita and PC per 100

As already stated the null hypothesis was rejected and the research hypothesis was accepted. There is a positive correlation. Normally alpha = 0.05 is chosen, which is the probability of making a Type 1 error of rejecting a null hypothesis which is actually true. The p-value in this test was very small and we can be very confident in the result.

References

Part 2. Regression analysis

Author: Martin Ottosson

Introduction

Method

Result

Has the electricity generation per capita in China increased from 1990 to 2005?

First the data is checked.

```
attach(El_China)
El_China
View(El_China)
str(El_China)
typeof(El_China)
```

Here we do not need to convert datatypes and we can make the regression.

```
model <- lm(El_China$El_China ~ El_China$Year)
model
summary(model)
```

The estimated model is

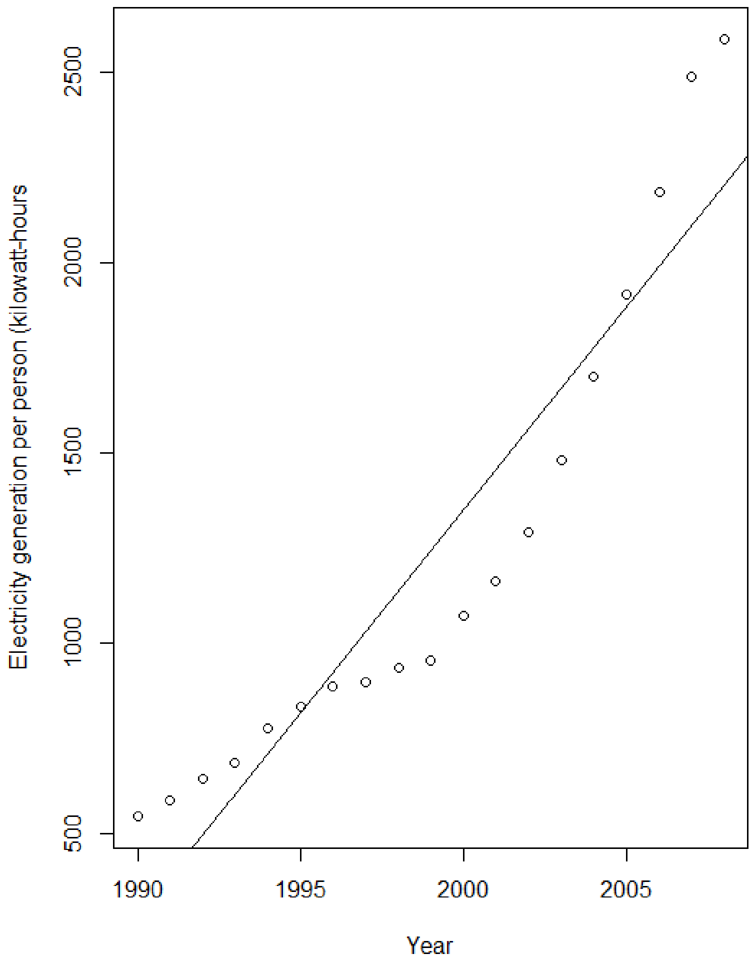
$$\mathit{Electricity} = -211961.6 + 106.7\mathit{Year}$$

For every year there is on average an increase of 106.7 in electricity generation per person (kilowatt-hours).

R2 is 0.88, the p-value for F is < 0.01 and the p-value for the Year coefficient is < 0.01.

Before interpreting these results we must check if the model is well adapted to the dataset using diagnostic plots and a scatterplot.

```
plot(El_China$El_China ~ El_China$Year,
    xlab = "Year", ylab = "Electricity generation per person (kilowatt-hours")
abline(model)
```

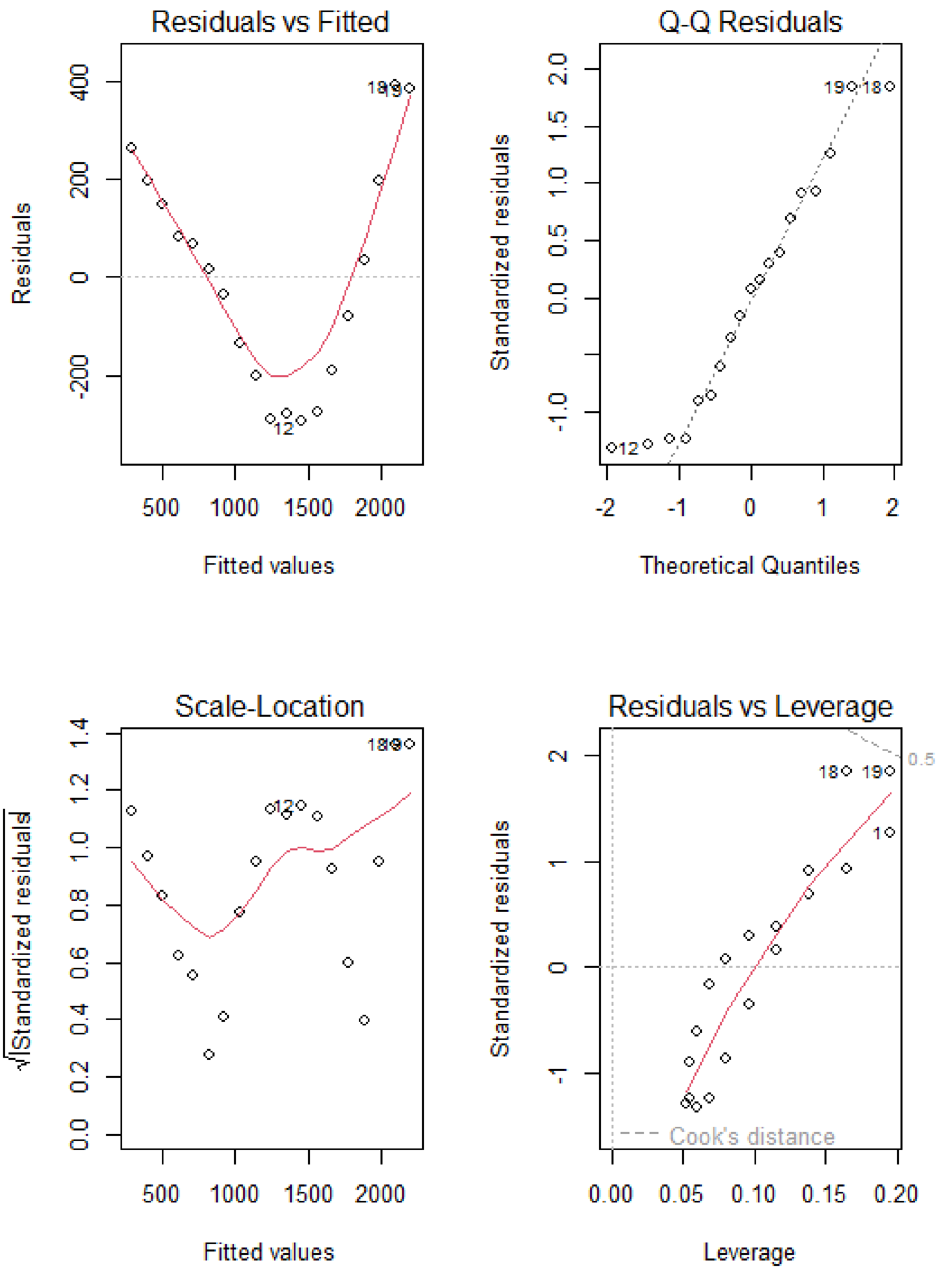


Looking at the scatterplot it can be seen that the observations are not evenly distributed around the regression line. There seem to be a pattern.

```
par(mfrow=c(2,2))
plot(model)
```

The plot with fitted values against residuals show a parabola and the model does not seem to capture a linear relationship. This could be a sign of heteroskedasticity, i.e. non-constant variance in the residuals.

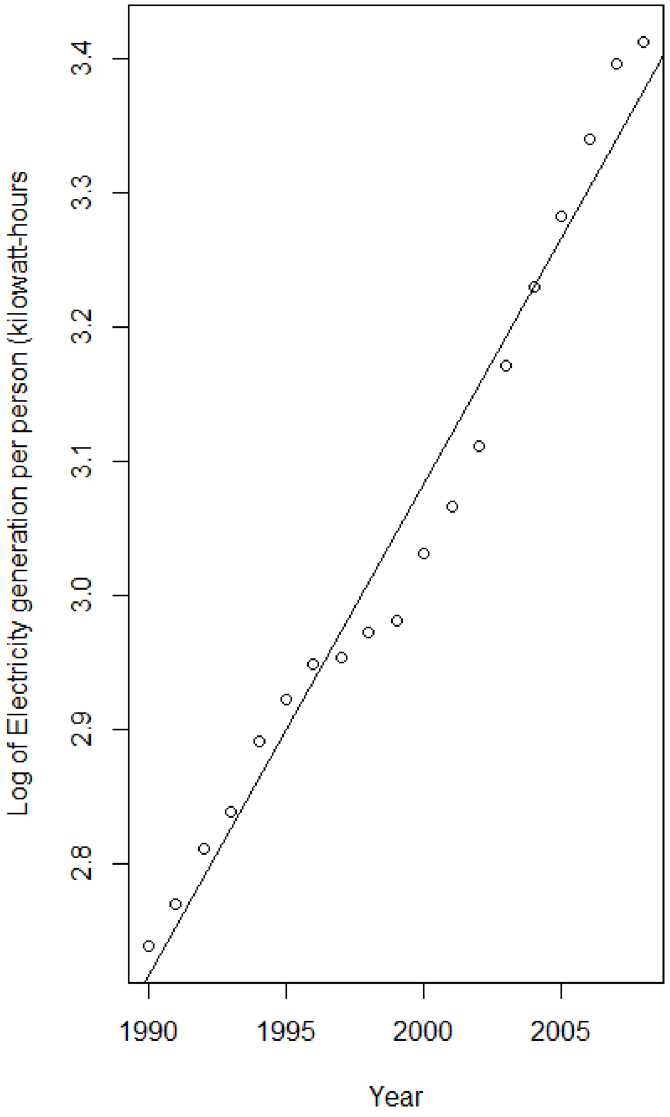
The Q-Q residuals plot shows appromximately normally distributed residuals since most of the points follow a straight line. There does not seem to be a problem with influential outliers according to the Residuas vs Leverage plot.



In the second model the response variable is transformed using log10 transformation. The R2 is now 0.97 and the t-values are significant at the 0.01 level for the F-statistic and the t-statistic.

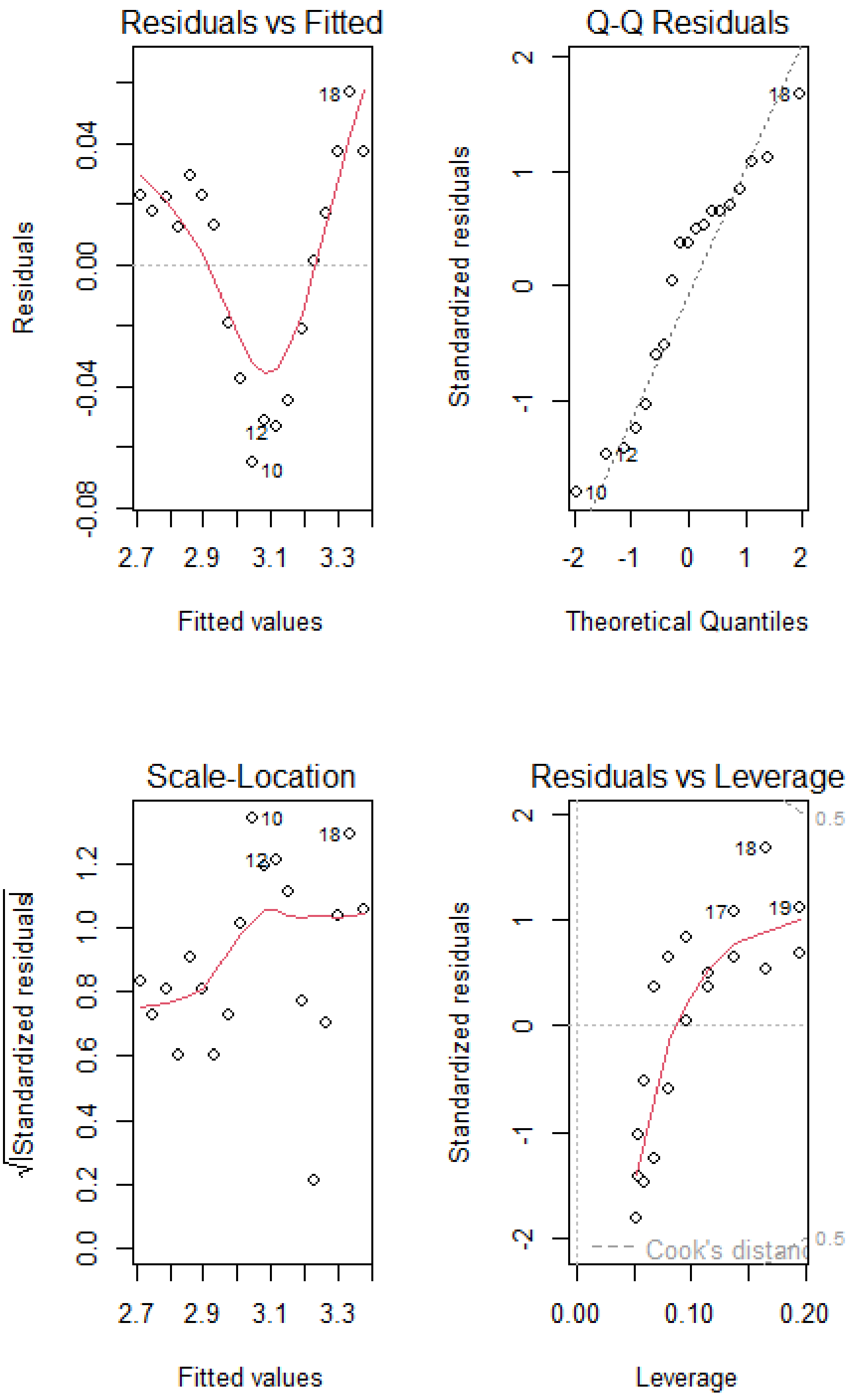
```
par(mfrow=c(1,1))
model_2 <- lm(log10(El_China) ~ Year, data = El_China)
model_2
summary(model_2)
```

The scatterplot looks a bit better, but there still seem to be a pattern.



```
par(mfrow=c(2,2))
plot(model_2)
```

There still seem to be a problem with heteroscedasticity and now the normality assumption can also be questioned.



In the third model the response variable is transformed with square root transformation.

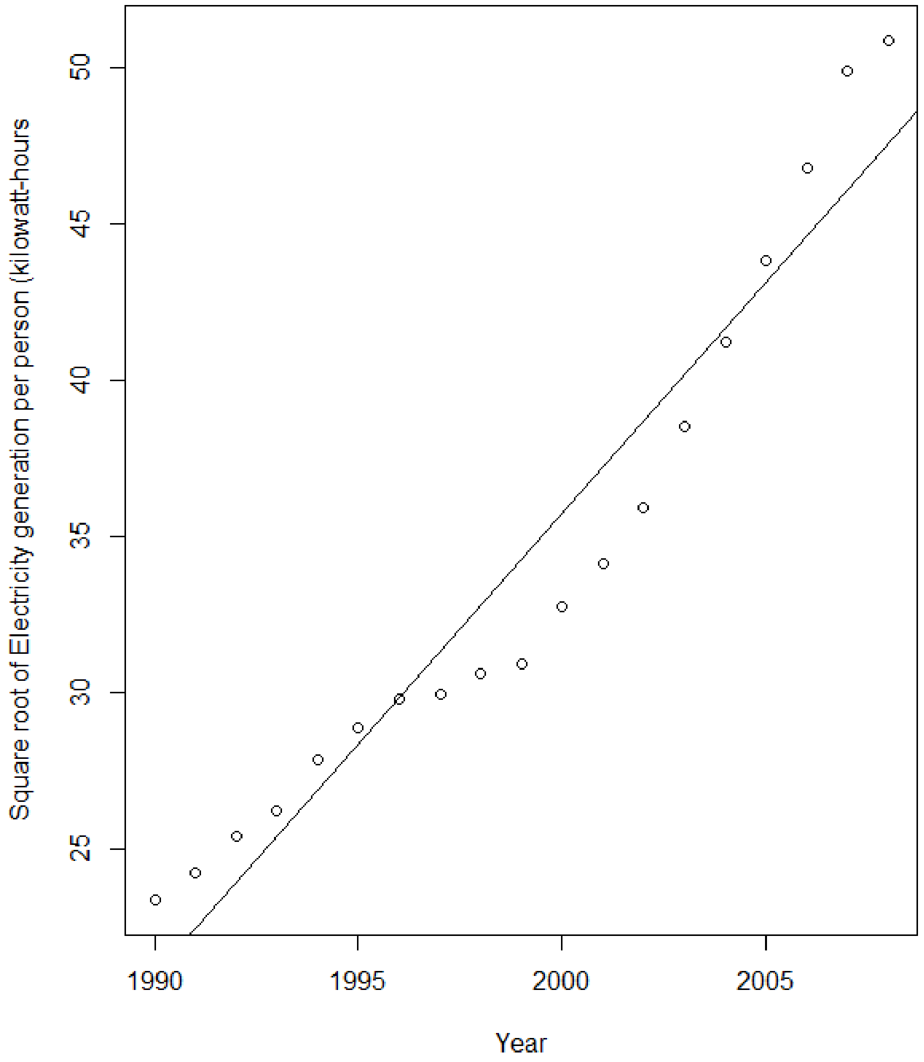
```
model_3 <- lm(sqrt(El_China) ~ Year, data = El_China)
model_3
summary(model_3)
```

The correlation of determination, R2, is 0.93 and the p-values for F and t are less than 0.01 which is clearly significant.

```
par(mfrow=c(1,1))
```

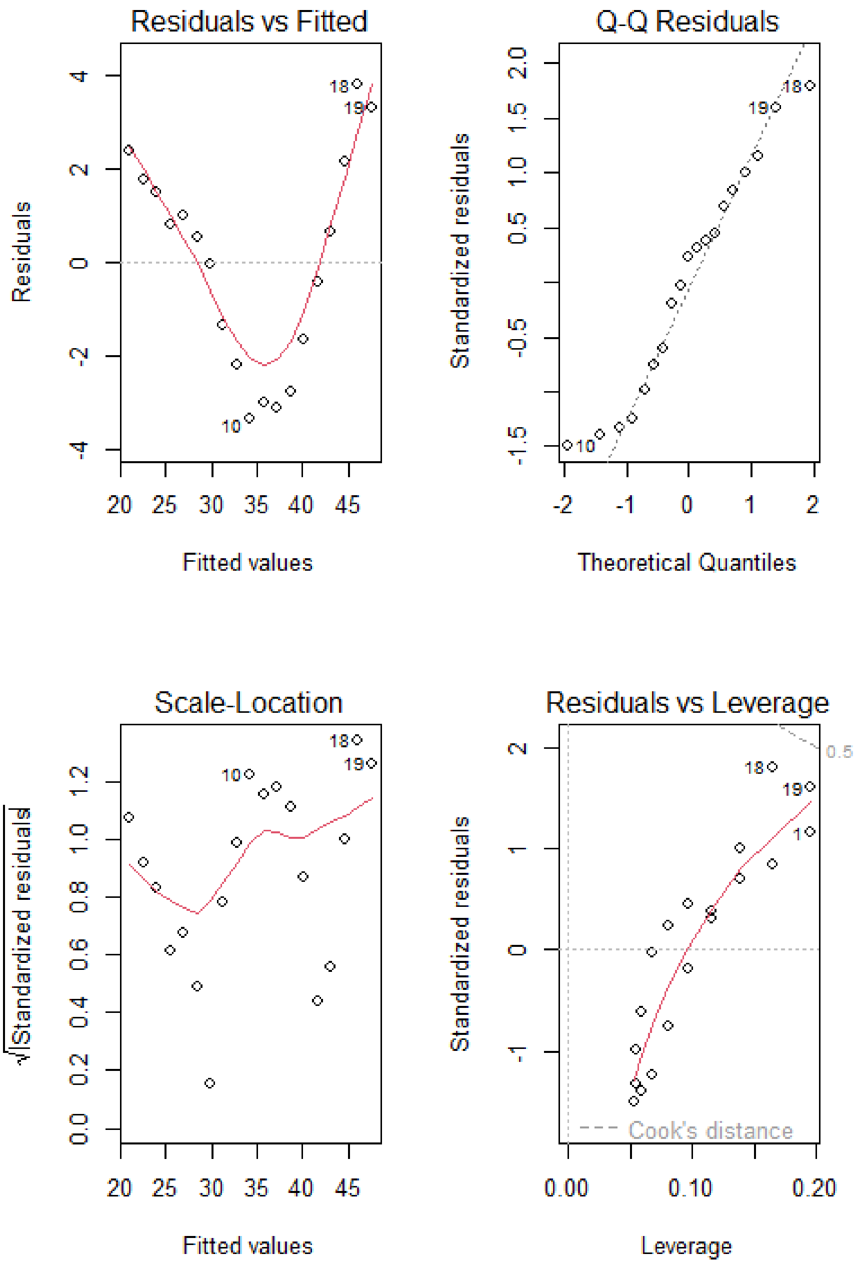
```
plot(sqrt(El_China$El_China) ~ El_China$Year,
    xlab = "Year", ylab = "Square root of Electricity generation per person (kilowatt-hours")
abline(model_3)
```

Comparing with the other scatterplots the square root transformation of Y seem to be the second best choice. The log10 transformation have the points closer to the regression line, but square root transformation seem to be better than the original model.



```
par(mfrow=c(2,2))
plot(model_3)
```

With square root transformation the normality looks better than for log10 transformation, but the residuals vs fitted show heteroscedasticity.



Discussion

The model with log10 transformation have observations which are scattered closest to the regression line.

All models show a parabola shape in the residuals vs fitted plot which is a sign of heteroscedasticity.

Regarding the normality assumption the best choices seem to be the original model or the square root transformation model.

All models have significant p-values for F and for t.

The log10 transformation has the highest R2 with 0.97, second comes the square root transformation with 0.93 and the original model has 0.88.

I would choose the model with square root transformation since the assumption of normally distributed residuals seem to hold for that model.

R2 of 0.93 is very high and clearly good enough. 93% of the variance in electricity capacity per capita can be "explained" by the variable year. However, there is probably something else behind the real explanation since time cannot produce anything in itself.

The best chosen model is:
Square root of Electricity = -2.914e+03 + 1.475e+00*Year*

b: 1.475e+00
SEb: 9.678e-02
t-value for Year-coefficient: 15.24
df: 17
p-value for Year-coefficient: 2.40e-11
We can see some very small values in these results (the ones with *e*).

References

Part 3. Testing difference between groups

Author: Martin Ottosson

Introduction

Method

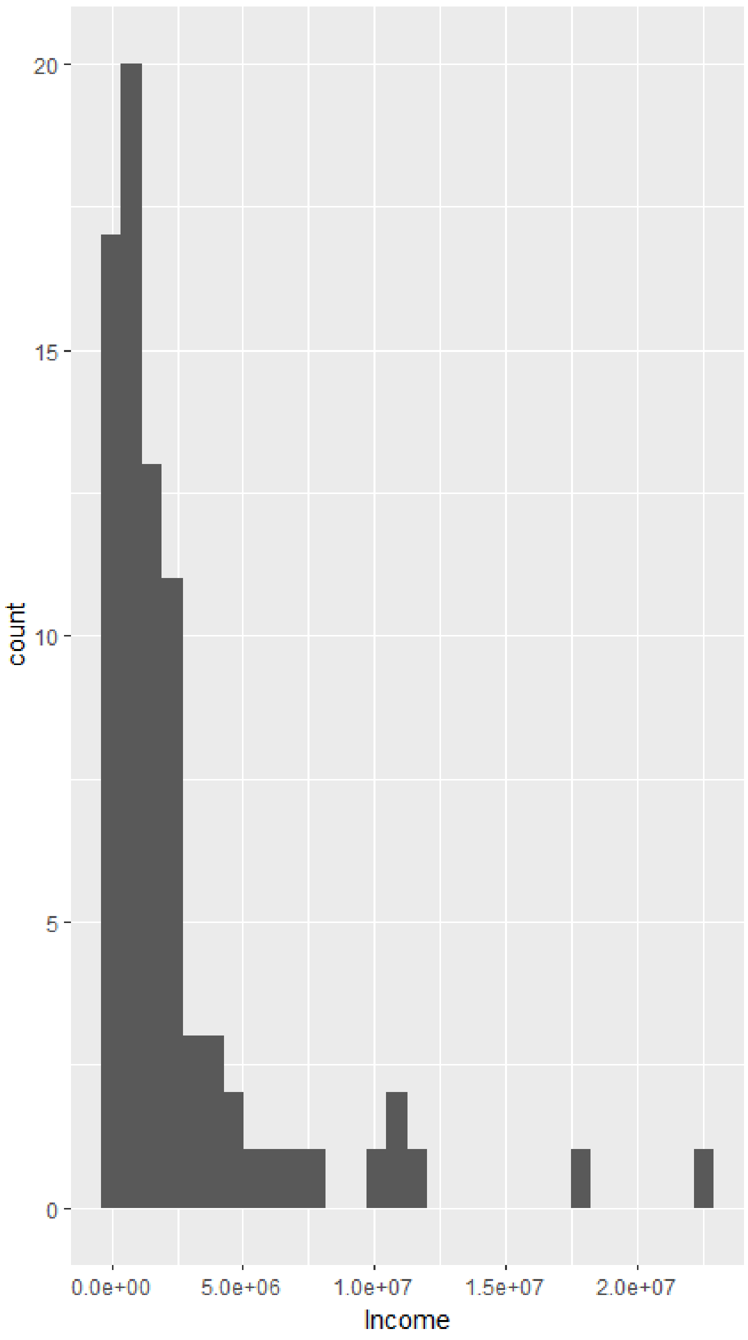
Result

Is there a difference in income between the New York districts, Manhattan and Brooklyn?

First the distribution of the income data is checked with a histogram.

```
library(ggplot2)
ggplot(Lander_HousingNew, aes(x=Income)) + geom_histogram()
```

It does not look approximately normally distributed, therefore the Wilcoxon's rank-sum test can be used. There is a tail to the right.



```
attach(Lander_HousingNew)
wilcox.test(Income ~ Boro, alternative = "two.sided")
detach(Lander_HousingNew)
```

There is a significant different in mean income between Brooklyn and Manhattan. The p-value is very low.

The data set is separated in two subsets in order to calculate the median, mean and standard deviation. One for Brooklyn and one for Manhattan.

```
brooklyn_data <- subset(Lander_HousingNew, Boro == "Brooklyn")
manhattan_data <- subset(Lander_HousingNew, Boro == "Manhattan")
```

```
summary(brooklyn_data$Income)
```

The mean income in Brooklyn is 639 548 and the median is 443 851.

```
summary(manhattan_data$Income)
```

The mean income in Manhattan is 3 344 961 and the median is 1 742 515.

```
sd(brooklyn_data$Income)
```

The standard deviation in Brooklyn is 598 728.4

```
sd(manhattan_data$Income)
```

The standard deviation in Manhattan is 4 345 457.

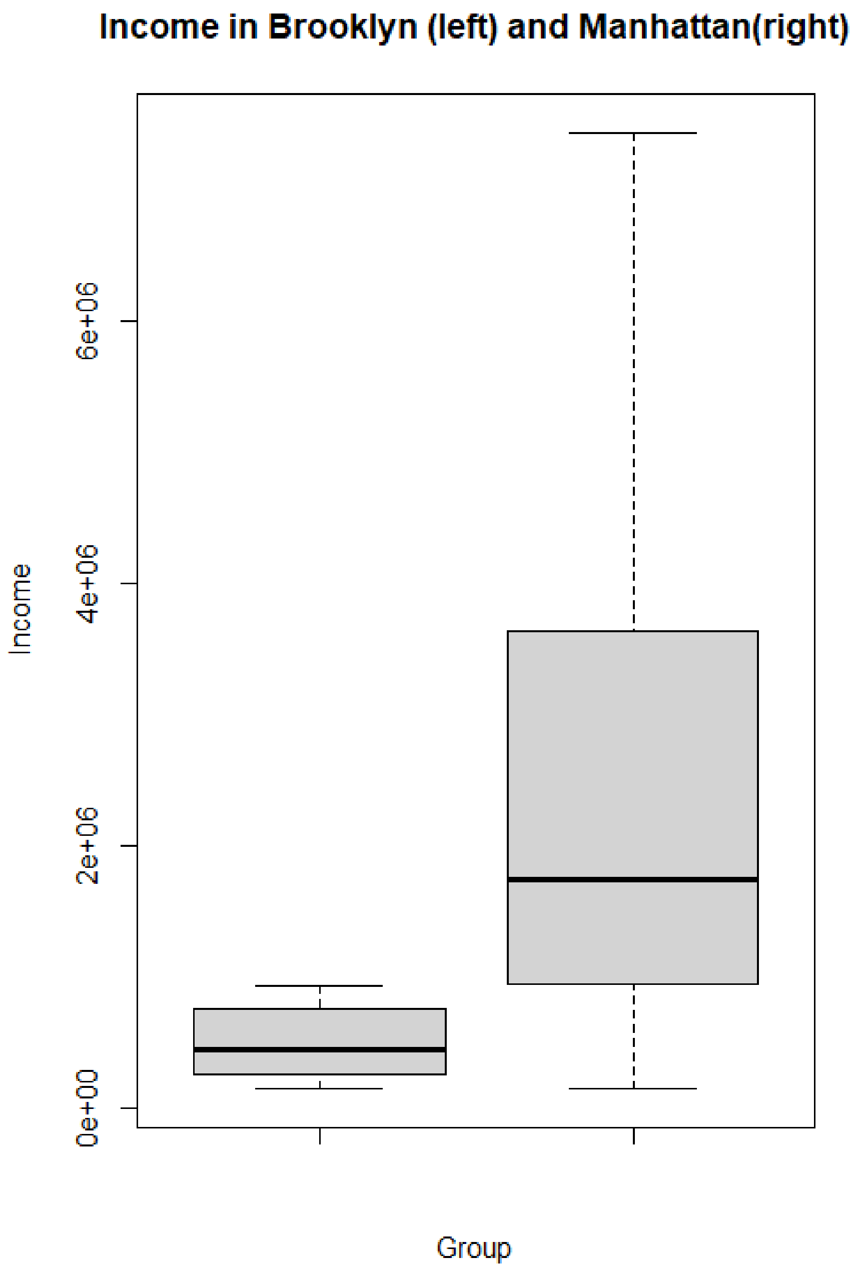
A boxplot shows the central tendency and dispersion of income for Brooklyn and Manhattan.

```
boxplot(brooklyn_data$Income, manhattan_data$Income, xlab="Group", ylab = "Income", main =
"Income in Brooklyn (left) and Manhattan(right)", outline=FALSE)
```

Discussion

The results of the Wilcoxon rank sum test:
This test was chosen because we could not assume normally distributed data for income
W: 223
p-value: 1.019e-05 (which is very small)
n: 79

There is a significant different in mean income between Brooklyn and Manhattan.



Result

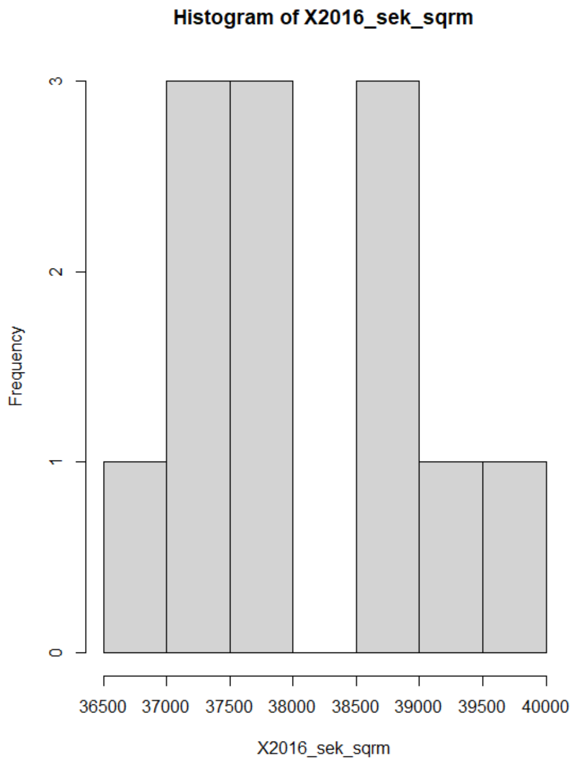
Are there differences in house pricing (SEK/m2) in Sweden between 2016 and 2017?

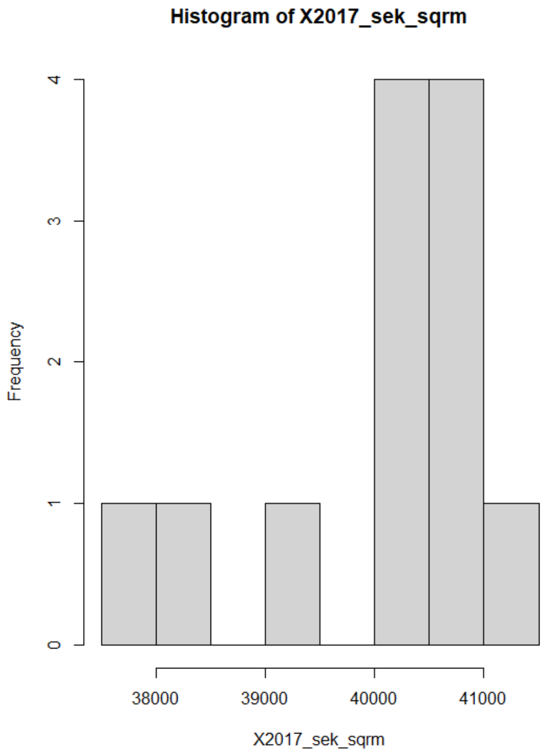
For this question we will perform a paired samples test of median or mean difference among treatments.

Histograms show the distribution of the data for the two years.

```
attach(Housepricing_sweden)
hist(X2016_sek_sqrm)
hist(X2017_sek_sqrm)
```

In this case we can do a paired t-test. There seems to be approximatley a normal distribution for sek per square meter for 2016 and 2017.





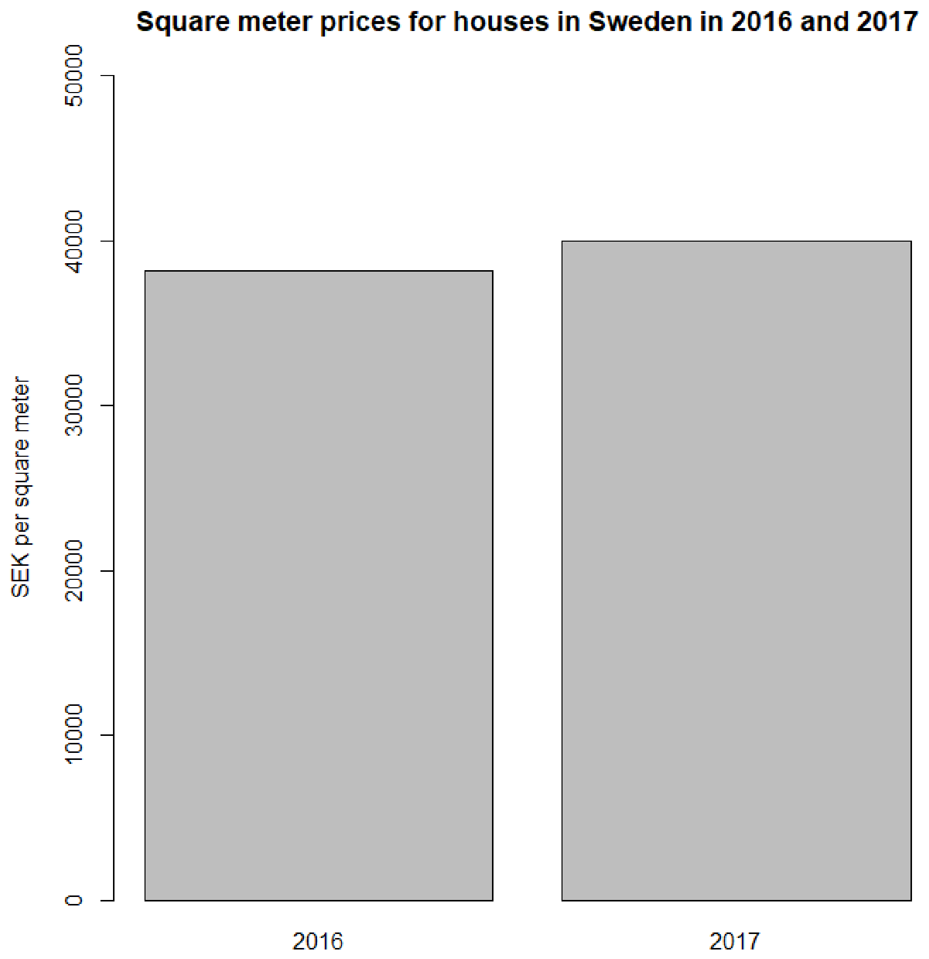
There is a significant mean difference between 2016 and 2017 and the p-value is 0.002885. We use a paired samples test because there are repeated measurements between two years.

```
t.test(X2017_sek_sqrm, X2016_sek_sqrm, paired = TRUE)
```

The mean difference is 1848. The mean price per square meter in 2016 is 38 128.5 and in 2017 39 976.5 sek.

```
mean_2016 <- mean(X2016_sek_sqrm)
mean_2017 <- mean(X2017_sek_sqrm)
year <- c("2016", "2017")
values <- c(mean_2016, mean_2017)
```

```
barplot(height = values, names.arg = year, ylim = c(0, 50000), ylab = "SEK per square meter", main = "Square meter prices for houses in Sweden in 2016 and 2017")
```



Discussion

The results of the paired t-test:
t-value: 3.8116
df: 11
p-value: 0.002885
There is a significant difference in square meter prices between 2016 and 2017 in Sweden.

The null hypothesis of no difference is rejected.

References