Lucas Faijdherbe(2594812) & Ruben van der Ham(2592271) – CS40

**Assignment 4**

**Exercise 4.1**

We are testing to see if an r = 0.876 based on n = 40 pairs are significantly different from 0, so, if the r value indicates that the Weight Watcher program is effective in reducing weight.

**Hypothesis:**

*Null hypothesis:*H0: ρ = 0

*Alternative hypothesis:*

Ha: ρ ≠ 0

**Significance level:**

α = 0.01

**Data:**

Data yields r = 0.876, n = 40.

**Test statistic:**

Tρ has a t-distribution with n – 2 = 40 – 2 = 38 degrees of freedom.

Observed value tρ = r = 11.1961483659

**Critical values:**

We have a two-tailed test with α = 0.01 and n = 40, so we get the critical values

-t38,0.01 and t38,0.01, which gives: -2.712 and 2.712.

Since tρ = 11.196 > 2.712, we reject H0.

**Conclusion:**

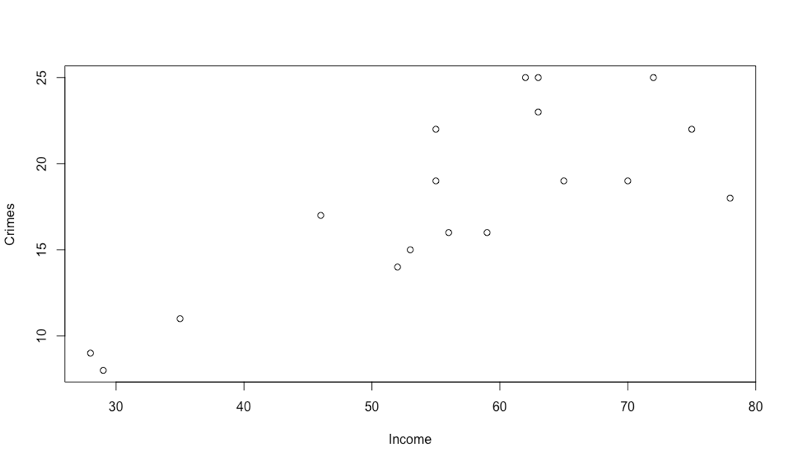
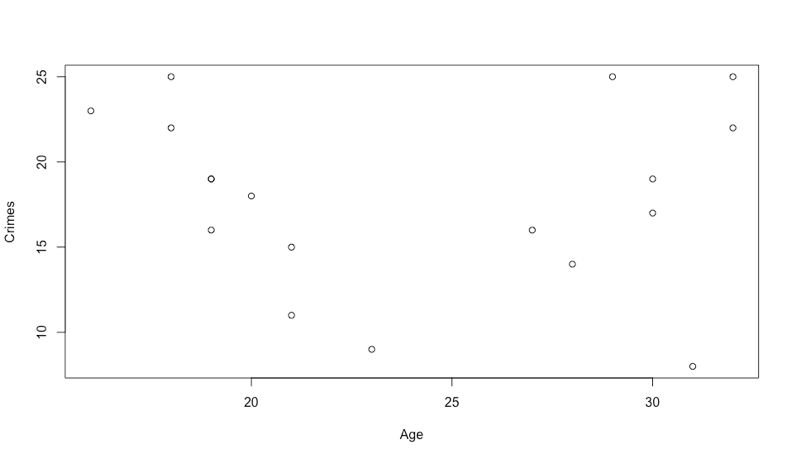
There is enough evidence to reject the claim that there is no linear correlation between the before weight and the after weight. So, one could argue that the value of r indicates that the Weight Watcher program is effective in reducing weight.

**Exercise 4.2**

In his book Outliers, author Malcolm Gladwell argues that more baseball players have birthdates in the months immediately following July 31, because that was the cutoff date for non-school baseball leagues. Here is a sample of frequency counts of months of birthdates of American-born major league baseball players starting with January: 387,329,366,344,336,313,313,503,421,434,398,371. Using a 0.05 significance level, is there sufficient evidence to warrant the rejection of the claim that American-born major league baseball players are born in different months with the same frequency?

**Exercise 4.3**

a)

Scatterplot of variables *age* and *crimes.*  
The sample linear coefficient is -0.071 (rounded).   
  
If we take a look at the scatterplot, it does not seem like there is any relationship between the two variables. The sample linear coefficient is -0.071 - very close to 0, which also suggest that there is no linear relationship between the two.  
  
b)   
Scatterplot of variables *income* and *crimes.* The sample linear coefficient is 0.792 (rounded).

If we take a look at the scatterplot, we do see signs of a positive linear relationship. Since the sample coefficient is 0.792, which is very close to 1. This suggests that there is indeed a positive linear relationship.

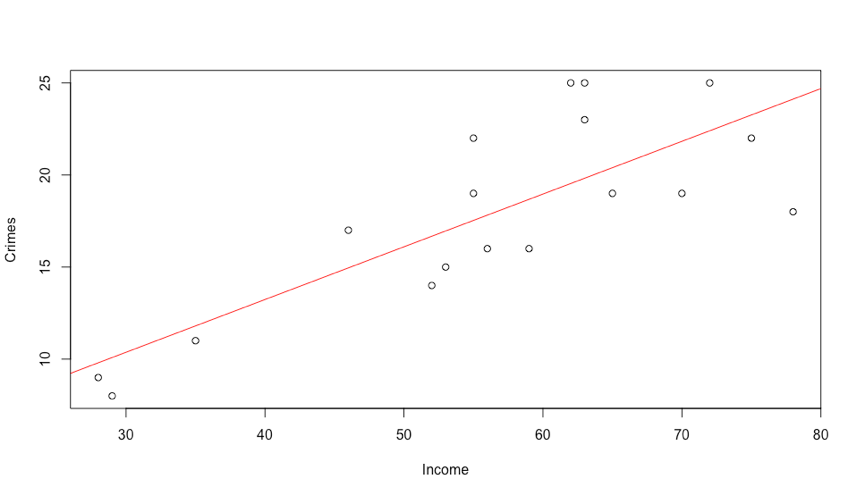
c) The fitted regression equation is given by

The R function *lm(crime$crimes ~ crime$income, data = crime)*yields

Intercept: b0 = 1.781

Slope: b1 = 0.286

So, the fitted regression equation becomes =



Scatterplot of variables *income* and *crimes*, with also plotted.

d) Claim: There is no linear relationship between the variables *income* and *crimes*.

**Hypothesis:**

*Null hypothesis:*

*Alternative hypothesis:*

**Significance level:**

α = 0.05

**Data:**

Data yields b1 = 0.286 and = 0.055

**Test statistic:**

The test statistic = b1 / has a t-distribution with n-2 = 18 – 2 = 16 degrees of freedom under H0. The observed value: = 0.286 / 0.055 = 5.2

**Critical values:**

We have a two-tailed test with α = 0.05 and n = 18, so we get the critical values –t16,0.05 and t16,0.05 which gives: -2.120 and 2.120.

Since = 5.2 > 2.120, we reject H0.

**Conclusion:**

The is sufficient evidence to warrant the rejection of the claim that there is no linear relationship between the variables *income* and *crimes*.

e) The requirements that have to be met are:

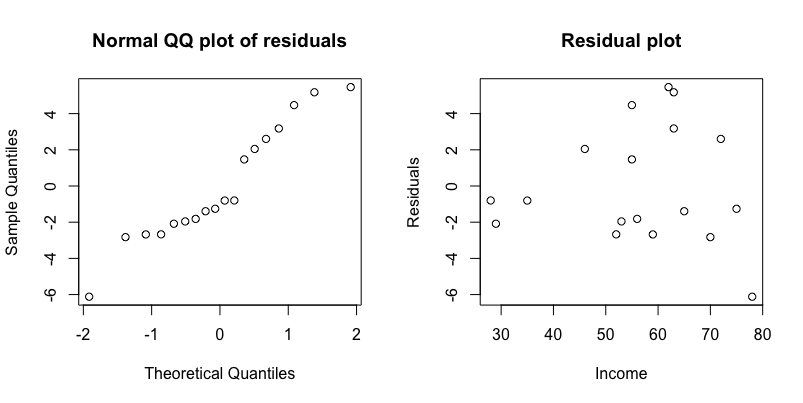
Errors are should be:

- Independent, something we can assume to be true

- Normally distributed

- Have a fixed standard deviation

With R, using qqnorm and plot for the residuals, we get the following plots:



From the QQ plot, we see that the plot follows approximately a straight line, so this probably comes from a normal distribution.

For the residual plot, there is no obvious pattern in the residuals. This has to be the case, otherwise there is something wrong: because our residuals look randomly placed, we can say that we have a fixed standard deviation. So we can positively say that the requirements for testing linearity are met.

**Exercise 4.4**

a)

**Appendix**

4.3 a)

> crime = read.table("/Users/lucasfaijdherbe/Library/Mobile Documents/com~apple~CloudDocs/Computer Science/Statistical Methods/Assignments/Assignment 4/Excersises/crimemale.txt", header = T)

> plot(crime$age, crime$crimes, xlab = 'Age', ylab = 'Crimes')

> cor(crime)

age income crimes

age 1.00000000 -0.4145025 -0.07095301

income -0.41450249 1.0000000 0.79155727

crimes -0.07095301 0.7915573 1.00000000

b)

> plot(crime$income, crime$crimes, xlab = 'Income', ylab = 'Crimes')

> cor(crime)

age income crimes

age 1.00000000 -0.4145025 -0.07095301

income -0.41450249 1.0000000 0.79155727

crimes -0.07095301 0.7915573 1.00000000

c and d)

> lmsim = lm(crime$crimes ~ crime$income, data = crime)

> summary(lmsim)

Call:

lm(formula = crime$crimes ~ crime$income, data = crime)

Residuals:

Min 1Q Median 3Q Max

-6.117 -2.054 -1.031 2.462 5.465

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.78111 3.21597 0.554 0.587

crime$income 0.28636 0.05527 5.181 9.1e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.315 on 16 degrees of freedom

Multiple R-squared: 0.6266, Adjusted R-squared: 0.6032

F-statistic: 26.85 on 1 and 16 DF, p-value: 9.097e-05

d) > lmsim = lm(crime$crimes ~ crime$income, data = crime)

> par(mfrow=c(1,2))

> qqnorm(lmsim$residuals, main="Normal QQ plot of residuals")

> plot(crime$income, lmsim$residuals, main = "Residual plot", ylab = "Residuals", xlab = "Income")