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?

**Hypothesis:**

*Claimed Distribution:*

For each category month, the expected value of baseball players is the same.

Therefore, the Pi of every category is also equal to all other Pi

*Null hypothesis:*H0: The above claimed distribution is agreed by the frequency counts.

*Alternative hypothesis:*

Ha: The above claimed distribution is not agreed by the frequency counts.

**Significance level:**

α = 0.05

**Data:**

Data yields n = 387+329+366+344+336+313+313+503+421+434+398+371 = 4515.

Pi= 4515/12 = 376.25

**Test statistic:**

X2 has a chi-squared distribution with n – 1 = 4515 – 1 = 4514 degrees of freedom.

X2 93.0717607973422

**Critical values:**

We have a right-tailed test with α = 0.05 and n = 4515, so we get the critical values

X24514,0.05, The closest amount of degree of freedom found in table 4 is 100, which gives: 140.169

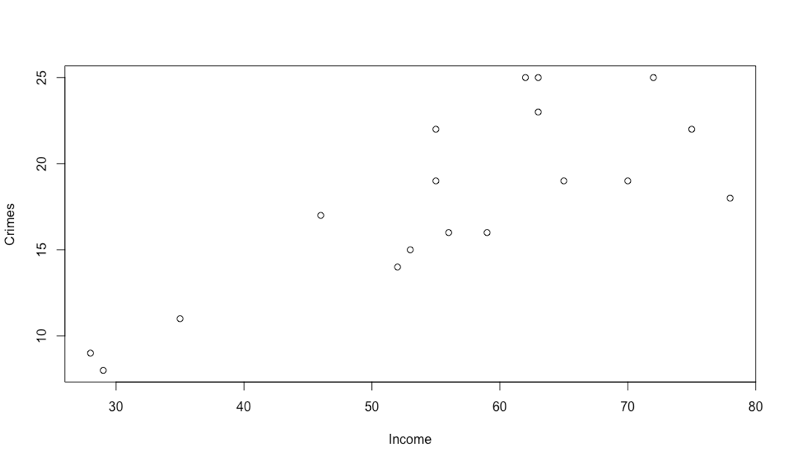
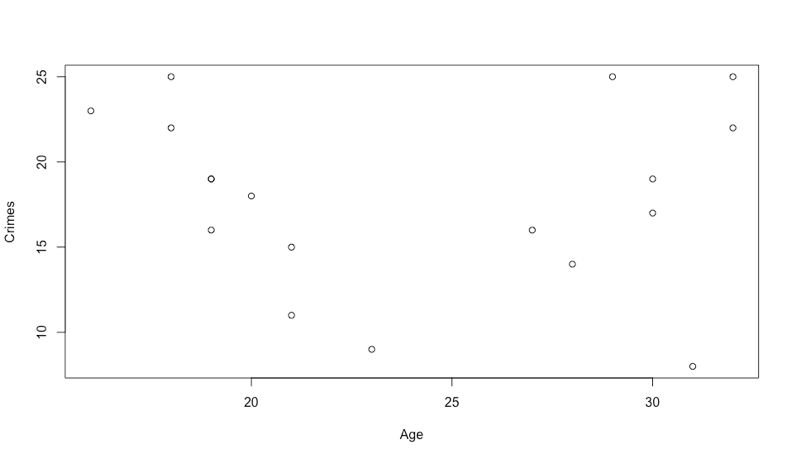
Since X2= 93.072 < 140.169, we fail to reject H0.

**Conclusion:**

There is not enough evidence to reject the claim that the frequency counts agree the claimed distribution (American-born major league baseball players are born in different months with the same frequency). Therefore we can reject the claim that “there is 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.

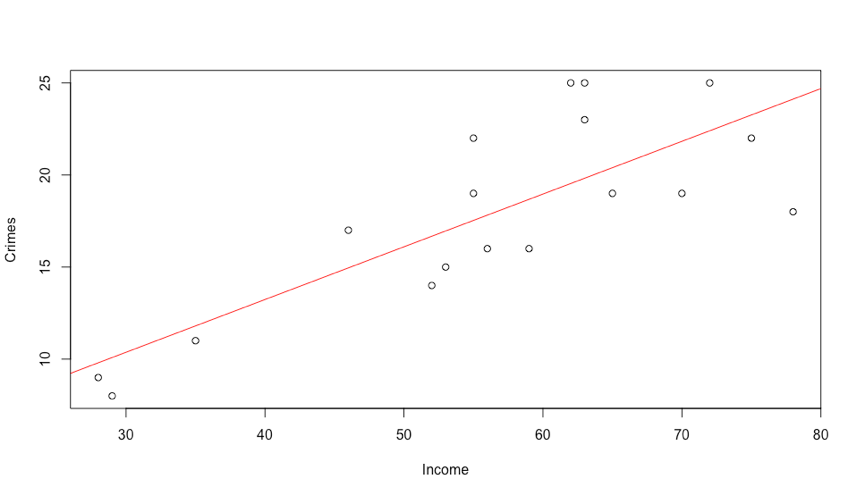
c) The fitted regression equation is given by

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

e)

**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)

> 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