**Quantitative Methods**

**List of Exercises N. 8**

**Selected Exercises from McClave (2014) – Chapter 12**

**12.3 Evaluating Overall Model Utility**

**Exercise 1. (2). Excel was used to fit the model E(y) = β0 + β1 x1+ β2 x2  to n = 20 data points, and the printout shown below was obtained:**

**Output: Regression equation is Y = 506.35 – 941.9 X1 – 429.1 X2**

**Predictor**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor** | **Coef** | **SE Coef** | **T** | **P** |
| **Constant** | **506.346** | **45.17** | **11.21** | **0.000** |
| **X1** | **-941.900** | **275.08** | **-3.42** | **0.003** |
| **X2** | **-429.060** | **379.83** | **-1.13** | **0.274** |
| **S = 94.251** |  |  |  |  |
| **R2 =45.9%** | **Adjusted R2 = 39.6%** |  |  |  |

**Analysis of Variance:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Squares** | **F** | **Pr > F** |
| **Regression** | **2** | **128329** | **64165** | **7.22** | **0.005** |
| **Residual Error** | **17** | **151016** | **8883** |  |  |
| **Total** | **19** | **279345** |  |  |  |

1. **What are the sample estimates of β0 , β1 and β2?**

If we look at the regression analysis, we can see that:

B0 is your constant or intercept, in the table: B0=506.346

B1 is the coefficient indicated by X1: B1=-941.900

B2 is the coefficient indicated by x2: B2=-429.06

1. **What is the least squares prediction equation?**

This is the least squares prediction equation:

E(y)=B0+B1x1+B2x2

Therefore using the numbers from exercise a, the least squares prediction equation for this problem is:

= 506.346-941.900x1-429.060x2

1. **Find SSE, MSE and s. Interpret the standard deviation in the context of the problem.**

SSE is the sum of the squared residuals.

MSE is mean squared error

s is standard deviation of your sample

SSE=151,016

MSE=8,883

s=94.251

We expect about 95% of the y-values to fall within 2s or 2(94.251) of 188.502 units of the fitted regression equation.

1. **Test H0: β1 = 0 against H0: β1 ≠ 0. Use α = 5%.**

Use the t-test statistic:

See page 705 for more information of this equation:

ts=b1/SSE\*b1

ts=-941.900/275.08

t=-3.424095

Use alpha = 5%. Now we look at the t-table, under two-tail probability. The degrees of freedom in a multiple regression = n-k-1, where k is the number of variables.

df=20-2-1

df=17

You are conducting a two-tail test. This is because your alternative hypothesis is B1 is not equal to 0. This means it can be both below and above zero. This is two directions and not one.

Look at your t-table under two-tailed alpha = 0.05 and df = 17. Then we get 2.110. Your rejection region is: t<-2.110 or t>2.110. The t-test statistic found for this sample was -3.42

Since the observed value for the test statistic falls in the rejection region t=-3.42 < -2.110), H0 is rejected. There is sufficient evidence to indicate B1 =/ 0 at alpha 0.05.

1. **Use a 95% confidence interval to estimate β2.**

A 95% confidence interval means that you must conduct a two-tailed test as there is 2.5% on one side and 2.5% on the other. Your df is the same as in exercise d. As your df and alpha are the same as in exercise d. Your t is also = 2.110.

See page 705 for more information of this equation:

-429.060+(2.110\*379.83)

-429.060-(2.110\*379.83)

Result: =(-1230.501, 372.3813)

1. **Find R2 and R2adjusted and interpret these values.**

You can find R^2 and R^2 adjusted in the given information.

R2 = R-squared =.459, which is 45.9% of the total sample variation of the y values is explained by the model containing x1 and x2.

R2 adjusted = R-Squared(adj)=.396. 39.6% of the total sample variation of the y values is explained by the model containing x1 and x2, adjusted for the sample size and the number of parameters in the model.

Adjusted R2 is more reliable than R^2 alone. This is because it takes more elements into consideration. Therefore also, it is usually smaller than R^2.

1. **Find the test statistic for testing β1 = β2 = 0.**

Testing B1 and/or B2=0 is essentially the same as testing if at least one of these variables is significant in predicting y (your line/model).

If B1 and/or B2=0 it means it does not have an effect on your line (y). If B1 and/or B2=0 they do not contribute to your model.

Remember your regression line equation:

=506.35-941.9x1-429.1x2

Here it says x1 and x2 instead of B1 and B2 but it is the same. If any of these values where 0 it mean that they do not contribute to the value of y at the other end of the equation.

For this test we will use the F-test. The F-ratio measures the improvement due to fitting the model i.e. the group means versus the grand mean of scores for all participant and compares this against the error remaining in the model, which is the difference between the actual scores and the respective means of the groups. Therefore, the F-test is the ratio of systematic variance: unsystematic variance, so higher scores are better.

In sum, when testing if your regression model is a good fit, use the F-test. Look at the F-table. To determine if at least one of the independent variables is significant in predicting y, we test:

The F from our table given at the beginning = 7.22. To look this up you must calculate v1 and v2 (degrees of freedom):

v1 = k

v2 = n-k-1 = 20-2-1=17

Remember that k is number of variables in your model. You can also see it in the given data. The rejection region is: F> 3.5915

Since the observed value of the test statistic falls in the rejection region (F=7.22>3.5915), H0 is rejected. There is sufficient evidence to indicate at least one of the variables, B1 or B2, is significant in predicting y at alpha = 0.05.

1. **Find the observed significance level of the test, part g. Interpret the result.**

The observed significance level of the test is p=0.005. Since the p-value is so small. We will reject H0 for most reasonable values of alpha. There is sufficient evidence to indicate at least one of the variables, x1 or x2, is significant in predicting y at alpha > 0.05.

Overall test of significance:

The Anova has a high F-value (7.22) which has a very low p-value of 0.005. This implies that this model is statistically significant.

Test of significance for the parameters:

B0 = 506.346, t-test of 11.21 with a very low p-value of less than 1%. This mean that this coefficient is significant at the 1% significance level.

B1=-941.9, t-test: t-stat of -3.42 with a very low p-value of 0.003, which is far below 1%, 5% or 10% of significance. This means that this coefficient is statistically significant.

B2=-429.060, t-test: t-stat of -1.13 with a p-value of 0.274, which is far above 1%, 5% or 10% of significance. This mean that this coefficient is not statistically significant.

**Exercise 2. (17, BDYIMG). *Reality TV and cosmetic surgery*. How much influence does the media, especially reality television programs, have on one’s decision to undergo cosmetic surgery? This was the question of interest to psychologists who published an article in Body Image: An International Journal of Research (March 2010). In the study, 170 college students answered questions about their impressions of reality TV shows featuring cosmetic surgery, level of self-esteem, satisfaction with one’s own body, and desire to have cosmetic surgery to alter one’s body. The variables analyzed in the study were measured as follows: DESIRE – scale ranging from 5 to 25, where the higher the value, the greater the interest in having cosmetic surgery; gender – 1 if male, 0 if female. SELFESTM – scale ranging from 4 to 40, where the higher the value, the greater the level of self-esteem; BODYSAT – scale ranging from 1 to 9, where the higher the value, the greater the satisfaction with one’s own body; and IMPREAL – scale ranging from 1 to 7, where the higher the value, the more one believes reality television shows featuring cosmetic surgery are realistic. The data for the study (simulated based on statistics reported in the journal article) are saved in the file. Selected observations are listed in the next table. The psychologists used multiple regression to model desire to have cosmetic surgery (y) as a function of gender (X1), self-esteem (X2), body satisfaction (X3) and impression of reality TV (X4).**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **1st and last 5 observations** | | | | | |
| **Student** | **DESIRE** | **GENDER** | **SELFESTM** | **BODYSAT** | **IMPREAL** |
| **1** | **11** | **0** | **24** | **3** | **4** |
| **2** | **13** | **0** | **20** | **3** | **4** |
| **3** | **11** | **0** | **25** | **4** | **5** |
| **4** | **11** | **1** | **22** | **9** | **4** |
| **5** | **18** | **0** | **8** | **1** | **6** |
| **…** | **…** | **…** | **…** | **…** | **…** |
| **166** | **18** | **0** | **25** | **3** | **5** |
| **167** | **13** | **0** | **26** | **4** | **5** |
| **168** | **9** | **1** | **13** | **5** | **6** |
| **169** | **14** | **0** | **20** | **3** | **2** |
| **170** | **6** | **1** | **27** | **8** | **3** |

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data: library(readxl)

L8E2 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/L8/L8E2.xlsx")

View(L8E2)

attach(L8E2)

Create variables: STUDENT <- STUDENT

SELFESTM <- SELFESTM

BODYSAT <- BODYSAT

IMPREAL <- IMPREAL

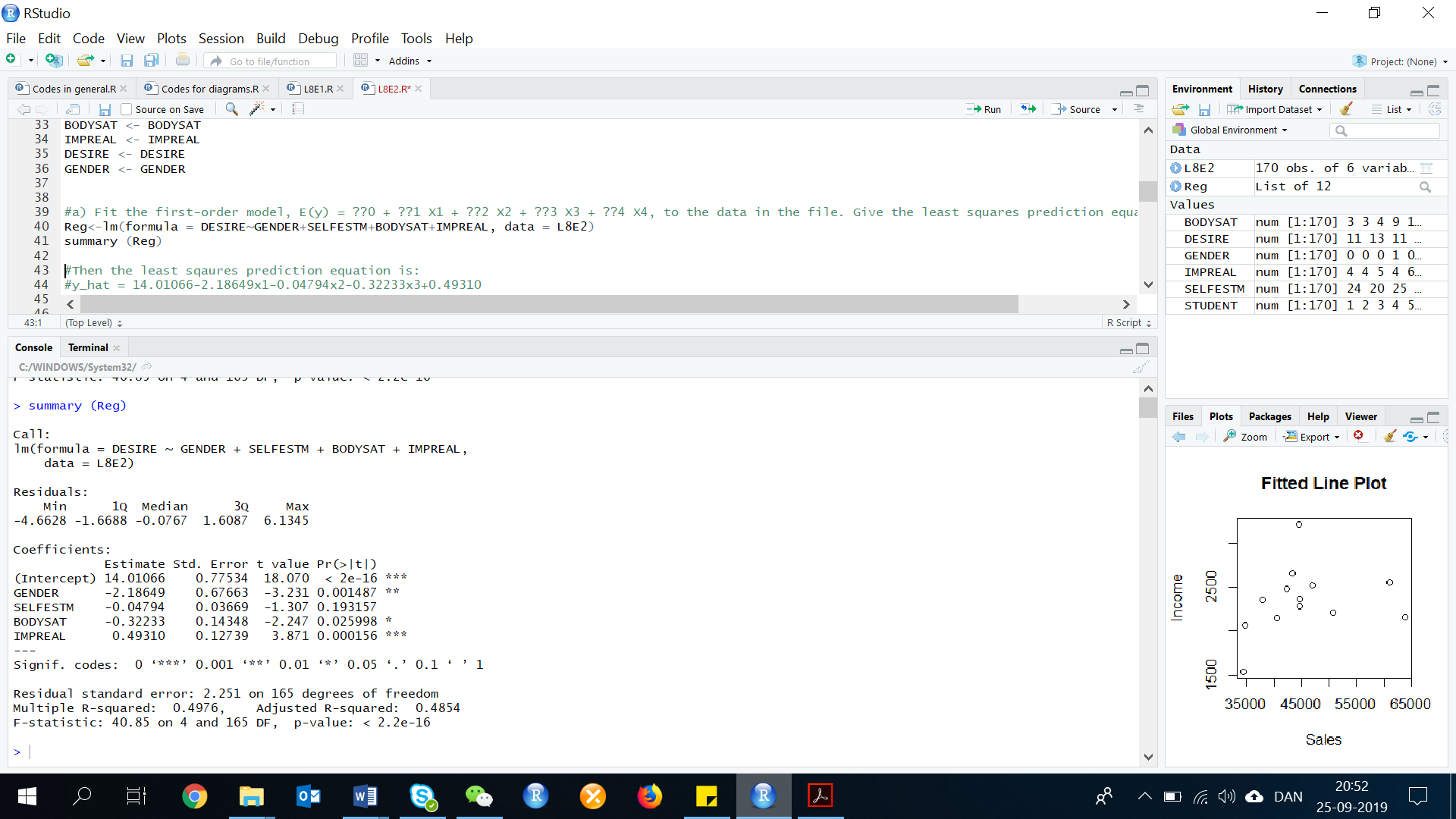
DESIRE <- DESIRE

GENDER <- GENDER

1. **Fit the first-order model, E(y) = β0 + β1X1 + β2X2 + β3X3 + β4 X4, to the data in the file. Give the least squares prediction equation.**

Reg<-lm(formula = DESIRE~GENDER+SELFESTM+BODYSAT+IMPREAL, data = L8E2)

summary (Reg)



Then the least squares prediction equation is:

= 14.01066-2.18649x1-0.04794x2-0.32233x3+0.49310x4

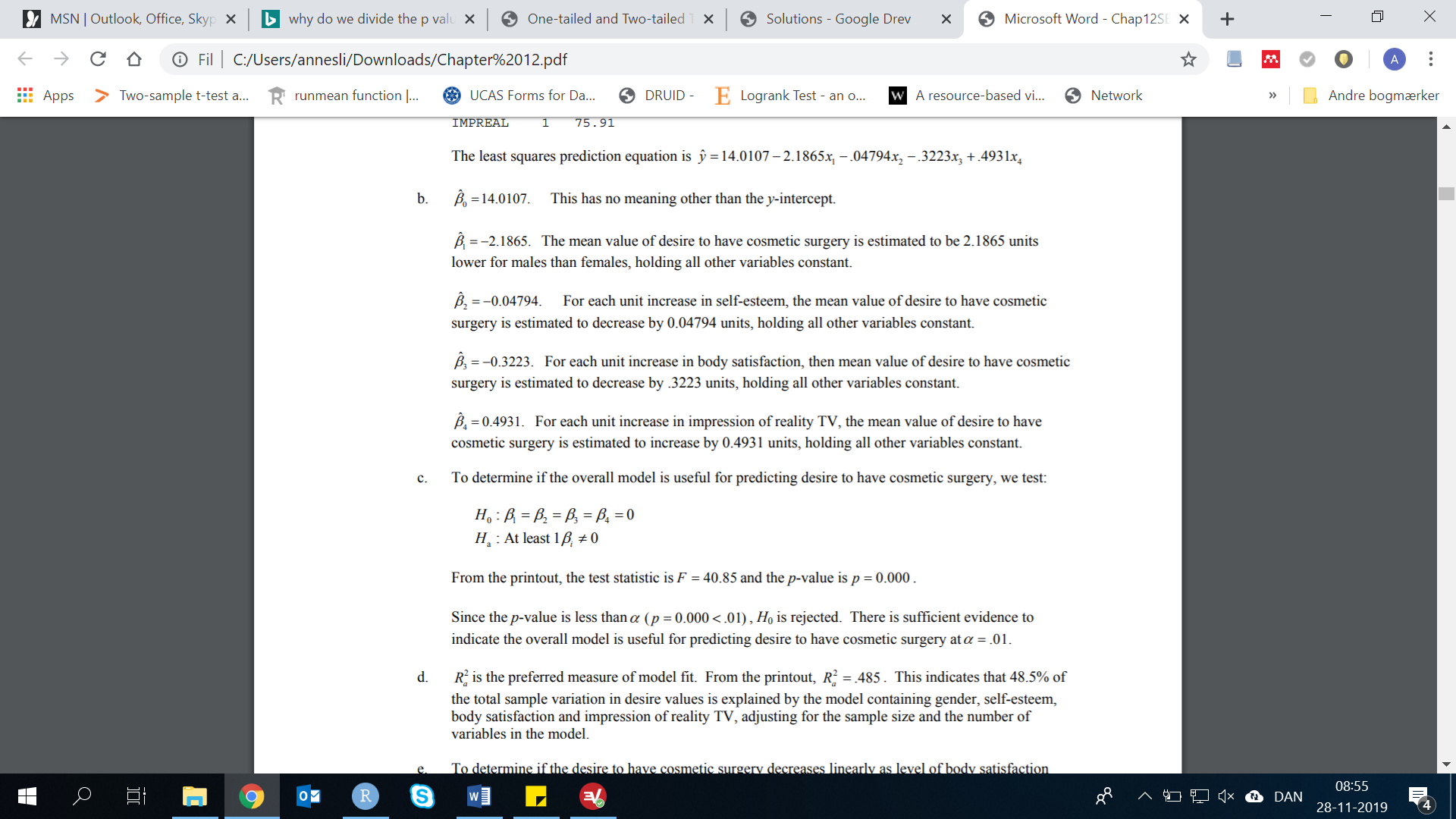
This equation is made of the values from the regression output, coefficients.

1. **Interpret the β estimates in the words of the problem.**

=14.01066. This is the constant, and has no meaning other than the y-intercept.

=-2.18649. The mean value of desire to have comestic surgery is estimated to be 2.18649 units lower for males than females, holding all other variables constant.

=-0.04794. For each unit increase in self-esteem, the mean value of desire to have cosmetic surgery is estimated to decrease by 0.04794 units, holding all other variables constant.



1. **Is the overall model statistically useful for predictiong desire to have cosmetic surgery? Test using α = 1%.**

To determine if the overall model is useful for predicting desire to have cosmetic surgery, we test:

From the printout, the test statistic is F=40.85, and the p-value of p=0.000.

Since the p-value is less than alpha (p=0.000 < 0.01). H0 is rejected. There is sufficient evidence to indicate the overall model is useful for predicting desire to have cosmetic surgery at alpha = 0.01.

1. **Which statistic R2 or R2 adjusted is the preferred measure of the model fit? Practically interpret the value of this statistic.**

R2-adjusted is the preferred measure of model fit. from the printout, R2a = 0.4854, which indicates that 48.5 % of the total sample variation is desire values is explained by the model containing gender, self-esteem, body satisfaction and impression of reality TV, adjusting for the sample size and the number of variables in the model.

1. **Conduct a test to determine whether desire to have cosmetic surgery decreases linearly as level of body satisfaction increases. Use α = 5%.**

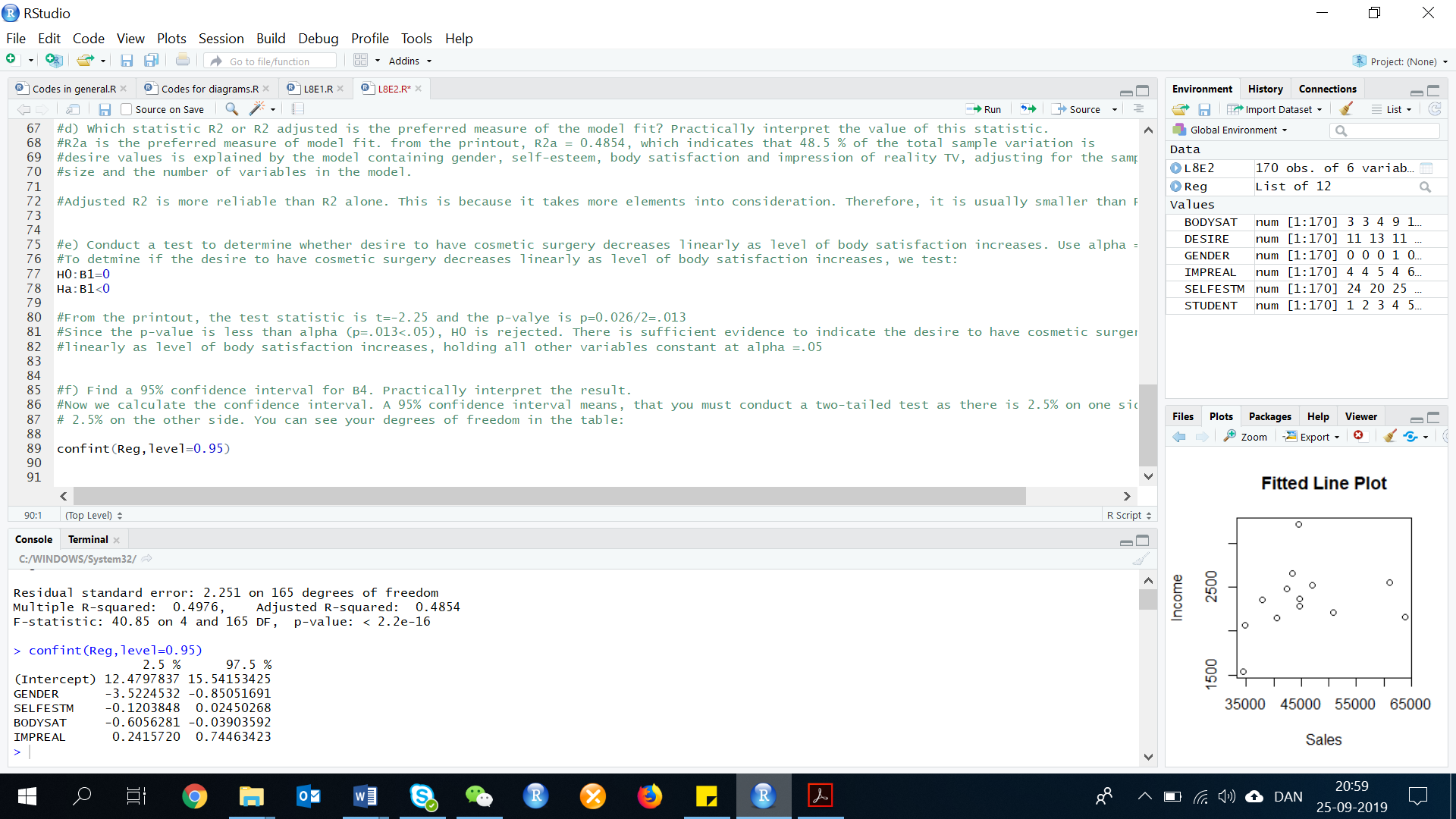
To determine if the desire to have cosmetic surgery decreases linearly as level of body satisfaction increases, we test:

From the printout, the test statistic is t=-2.247 and the p-value is p=0.026/2=.013

Since the p-value is less than alpha (p=.013<.05), H0 is rejected. There is sufficient evidence to indicate the desire to have cosmetic surgery decreases linearly as level of body satisfaction increases, holding all other variables constant at alpha =.05

1. **Find a 95% confidence interval for β4. Practically interpret the result.**

Now we calculate the confidence interval. A 95% confidence interval means, that you must conduct a two-tailed test as there is 2.5% on one side and 2.5% on the other side. You can see your degrees of freedom in the table:

confint(Reg,level=0.95)

**12.5 Interaction Models**

**Exercise 3. (43, BDYIMG). *Reality TV and cosmetic surgery*. Refer to the Body Image: AN International Journal of Research (March 2010) study of the impact of reality TV shows on a college student’s decision to undergo cosmetic surgery. The data for the study (simulated based on statistics reported in the journal article) are saved in the file. Consider the interaction model E(y) = β0 + β1x1 + β2x2 + β3x1x2, where y = desire to have cosmetic surgery (25-point scale), X1 = 1 if male and 0 if female, X2 = impression of reality TV (7-point scale). The model was fit to the data and the resulting Excel printout appears below.**

**Model Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **R** | **R squared** | **Adj. R squared** | **St. Error of the estimate** |
| **1** | **0.670** | **0.449** | **0.439** | **2.350** |

**Anova**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Sum of Squares** | **DF** | **Mean Square** | **F** | **Sig** |
| **Regression** | **747.001** | **3** | **249.000** | **45.86** | **0.000** |
| **Residual** | **916.787** | **166** | **5.523** |  |  |
| **Total** | **1663.788** | **169** |  |  |  |

**Dependent Variable: DESIRE**

**Predictors: Constant, GENDER, IMPREAL, GENDER\_IMPREAL**

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Unstandardized Coefficients** | | **Standardized Coefficients** | **t** | **Sig** |
|  | **B** | **Std. Error** | **Beta** |
| **Constant** | **11.779** | **0.674** |  | **17.486** | **0.000** |
| **Gender** | **-1.972** | **1.179** | **-0.303** | **-1.672** | **0.096** |
| **Impreal** | **0.585** | **0.162** | **0.258** | **3.617** | **0.000** |
| **Gender\_Impreal** | **-0.553** | **0.276** | **-0.378** | **-2.004** | **0.047** |

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

1. **Give the least squares prediction equation.**

According to the given information, the least sqaures prediction equation is:

1. **Find the predicted level of desire (y) for a male college student with an impression-of-reality-TV-sclae score of 5.**

The gender is male, therefore x1=1. the score on the reality TV-scale = 5, therefore x4=4. Now we must type it into our equation:

Result:

The predicted level of desire for a male college student with an impression of reality TV-scale score of 5 is eqaul to 9.967.

1. **Coduct a test of overall model adequacy. Use α = 10%.**

The model shows us, that:

The test statistic is F=45.086 and the p-value is p=.000.

Since the p-value is less than alpha (p=.000<.10), H0 is rejected. There is sufficient evidence to indicate the model is adequate in predicting desire to have cosmetic surgery at alpha = 0.10.

1. **Give a practical interpretation of the adjusted R2**

R2 adjusted = .439, 43.9% of the sample variation in the desire to have cosmetic surgery around its mean is ecplained by the model containing gender, impression of reality TV and the interaction of the two variables.

1. **Give a practical interpretation of s.**

The tables shows us, that s = 2.350. Most of the observed values of desire will fall within 2s=2(2.350)=4.70 units of their predicted values. 2s are about 95% of your data if your data is normal distributed.

1. **Conduct a test (at α = 10%) to determine if gender (X1) and impression of reality TV show X4 interact in the prediction of the level of desire for cosmetic surgery (Y).**

To determine if gender and impression of reality TV interact, we test:

The test statistic is t=-2.004 and the p-value is p=.047.

Since the p-value is less than alpha (p=.047>.10), H0 is rejected. There is sufficient evidence to indicate gender and impression of reality TV interact to affect desire to have cosmetic surgery at alpha = 0.10.

**12.7 Qualitative (Dummy) Variable Models**

**Exercise 4. (75). *Do blondes raise more funds?* During fundraising, does the physical appearance of the solicitor impact the level of capital raised? An economist at the University of Nevada-Reno designed an experiment to answer this question and published the results in Economic Letters (Vol. 100, 2008). Each in a sample of 955 households was contacted by a female solicitor and asked to contribute to the Center for Natural Hazards Mitigation Research. The level of contribution (in dollars) was recorded as well as their hair color of the solicitor (blond Caucasian, brunette Caucasian, or minority female).**

1. **Consider a model for the mean level of contribution, E(y), that allows for different means depending on the hair color of the solicitor. Create the appropriate number of dumy variables for the hair color. (use minority female as the base level).**

RECALL: *A Dummy variable or Indicator Variable is an artificial variable created to represent an attribute with two or more distinct categories/levels. Why is it used? Regression analysis treats all independent (X) variables in the analysis as numerical.*

*Numerical variables are interval or ratio scale variables whose values are directly comparable, e.g. '10 is twice as much as 5', or '3 minus 1 equals 2'. Often, however, you might want to include an attribute or nominal scale variable. In this case, hair color and race, which was no numerical value. Say you have three types of hair color + race '1', '2' and '3'. In this case, '3 minus 1' doesn't mean anything. you can't subtract blond from brunette. The numbers here are used to indicate or identify type of hair color and race and do not have intrinsic meaning of their own. Dummy variables are created in this situation to 'trick' the regression algorithm into correctly analyzing attribute variables. Often, this is used for gender where you code Female=1 and Male=0 (or the other way around)*

ANSWER:

Two dummy variables would be appropriate:

Let x1 = 1 if blonde Caucasian, 0 if otherwise.

Let x2 = 1 if brunette Caucasian, 0 if otherwise.

1. **Write the equation of the model, part a, incorporating the dummy variables.**

The model would be:

1. **In terms of the β’s in the model, what is the mean level of contribution for households contacted by a blond Caucasian solicitor?**

The mean for a blonde Caucasian solicitor would be:

E(y)=B0+B1\*(1)+B2\*(0) = B0+B1

Remember:

Let x1 = 1 if blonde Caucasian, 0 if otherwise.

Let x2 = 1 if brunette Caucasian, 0 if otherwise.

1. **In terms of the β’s in the model, what is the difference between the mean level of contribution for households contacted by a blond solicitor and those contacted by a minority female?**

The difference in the mean level of contribution between a blonde solicitor and a minority female solicitor is B1.

Then would x1 = 0 and x2 = 0, because they are the codes for minority female. The mean for a blonde Caucasian solicitor would be:

E(y)=B0+B1\*(0)+B2\*(0) = B0

This would be the model for minority female. Thus the difference between the mean level of contribution would be B1.

1. **One theory posits that blond solicitors will achieve the highest mean contribution level, but that there will be no difference between the mean contribution levels attained by brunette Caucasian and minority females. If this theory is true, give the expected signs of the β‘s in the model.**

If the theory is correct, then B0 (mean for minority female solicitors) will be positive, B2 will be positive (mean for female Caucasian solicitors is greater than the means of the groups), and B2 will be close to 0 (no different in the means for minority female solicitors and brunette Caucasian solicitors).

1. **The researcher found the β-estimate for the dummy variable blond Caucasian to be positive and significantly different from 0 (p-value < 0.01). The β-estimate for the dummy variable for brunette Caucasian was also positive, but not significantly different from 0 (p-value > 0.10). Do these results support the theory, part e?**

Yes. The B-estimate for the dummy variable for blonde Caucasian solicitors should be positive and significantly different from 0. The B-estimate for the dummy variable for brunette Caucasian solicitors should be close to 0. In this case, it is not statistically different from 0.

**Exercise 5. (79, REPELL). *Comparing mosquito repellents*. Which insect repellents protect best against mosquitoes? Consumer Reports (June 2000) tested 14 products that all claim to be an effective mosquito repellent. Each product was classified as either lotion / cream or aerosol / spray. The cost of the product (in dollars) was divided by the amount of the repellent needed to cover exposed areas of the skin (about 1/3 ounce) to obtain a cost-per-use value. Effectiveness was measured as the maximum number of hours of protection (in half-hour increments) provided when human testers exposed their arms to 200 mosquitoes. The data from the report are listed in the table below:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Insect Repellent** | **Type** | **Cost / Use** | **Maximum Protection** |
| **Amway Hour Guard 12** | **Lotion / Cream** | **2.08 USD** | **13.5 hours** |
| **Avon Skin-So-Soft** | **Aerosol / Spray** | **0.67** | **0.5** |
| **Avon BugGuard** | **Lotion / Cream** | **1** | **2** |
| **Ben’s Backyard Formula** | **Lotion / Cream** | **0.75** | **7** |
| **Bite Blocker** | **Lotion / Cream** | **0.46** | **3** |
| **BugOut** | **Aerosol / Spray** | **0.11** | **6** |
| **Cutter Skinsations** | **Aerosol / Spray** | **0.22** | **3** |
| **Cutter Unscented** | **Aerosol / Spray** | **0.19** | **5.5** |
| **Muskoll Ultra 6 Hours** | **Aerosol / Spray** | **0.24** | **6.5** |
| **Natrapel** | **Aerosol / Spray** | **0.27** | **1** |
| **Off! Deep Woods** | **Aerosol / Spray** | **1.77** | **14** |
| **Off! Skintastic** | **Lotion / Cream** | **0.67** | **3** |
| **Sawyer Deet Formula** | **Lotion / Cream** | **0.36** | **7** |
| **Repel Permanone** | **Aerosol / Spray** | **2.75** | **24** |

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data: library(readxl)

L8E5 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/L8/L8E5.xlsx")

View(L8E5)

attach(L8E5)

Create variable: Insect\_Rep <- Insect\_Rep

Type <- Type

Cost <- Cost

Hours <- Hours

1. **Suppose you want to use repellent type to model the cost per use (y). Create the appropriate number of dummy variables for repellent type and write the model.**

Only 1 is needed. Let x = 1 if Lotion/cream, 0 if otherwise.

1. **Fit he model, part a, to the data.**

First, create a dummy variable, where Lotion/Cream represent x=1.

Type1 <- c(1,0,1,1,1,0,0,0,0,0,0,1,1,0)

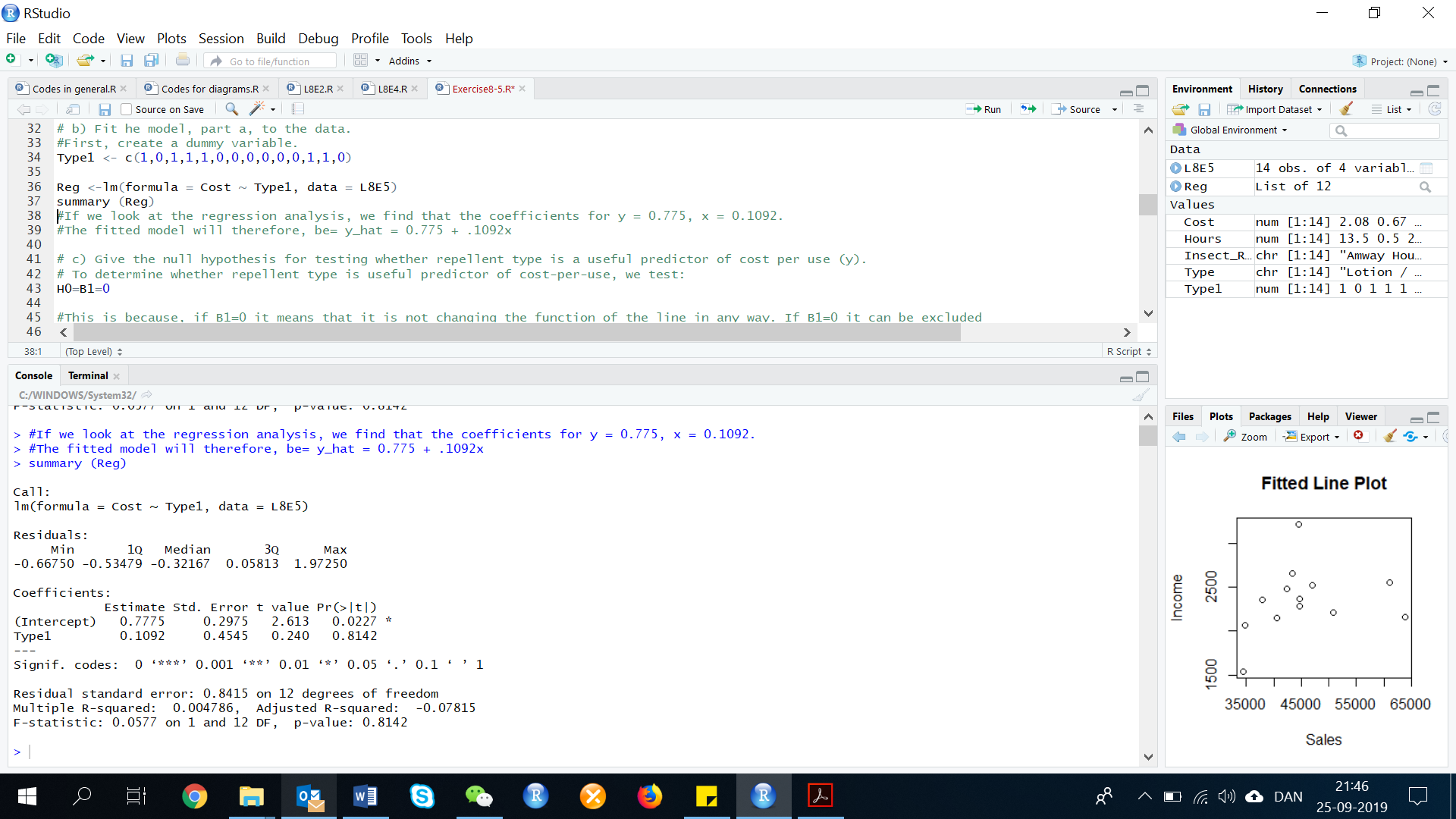
Note: If we made part a) differently, so Aerosol/Spray would be x = 1, and 0 if otherwise, the dummy would look differently:

Type2 <- c(0,1,0,0,0,1,1,1,1,1,1,0,0,1)

Then run the regression model:

RegB <-lm(formula = Cost ~ Type1, data = L8E5)

summary (RegB)



If we look at the regression analysis, we find that the coefficients for B0 = 0.775, B1 = 0.1092. The fitted model will therefore, be:

1. **Give the null hypothesis for testing whether repellent type is a useful predictor of cost per use (y).**

To determine whether repellent type is useful predictor of cost-per-use, we test:

This is because, if B1=0 it means that it is not changing the function of the line in any way. If B1=0 it can be excluded from the equation and there would be no difference.

1. **Conduct the test, part c, and give the appropriate conclusion. Use α = 10%.**

The alternative hypothesis is:

The test statistic is t=0.24, and the p-value is p=.814.

Since the p-value is not less than alpha (p=.814>.10), H0 is not rejected. There is insufficient evidence to indicate that repellent type is a useful predictor of cost-per-use at alpha = .10.

1. **Repeat parts a-d if the dependent variable is the maximum number of hours of protection (y).**

Do another regression with maximum number of hours as the y.

RegE <-lm(formula = Hours ~ Type1, data = L8E5)

summary (RegE)

If we look at the regression analysis, we find that the coefficients for y = 0.775, x = 0.1092. The fitted model will therefore, be:

To determine whether repellent type is a useful predictor of cost-per-use,

The test statistic is t=-0.46 and the p-value is p=.653. Since the p-value is not less than alpha (.653 < .10), H0 is not rejected. There is insufficient evidence to indicate that repellent type is a useful predictor of maximum number of hours of protection at alpha = .10.

**12.8 Models with Both Quantitative and Qualitative Variables**

**Exercise 6. (82). Consider a multiple regression model for a response y, with one quantitative independent variable X1 and one qualitative variable at 3 levels.**

1. **Write the first-order model that relates the mean response E(y) to the quantitative independent variable.**

This is a linear regression with E(y) being the dependent variable and a intercept + independent variable.

1. **Add the main effect terms for the qualitative independent variable to the model of part a. Specify the coding scheme you use.**

The model is an extension of the first model and then it is adding the 2 extra terms meaning that there will be: beta 1, beta 2, and beta 3.

Where:

x2 = 1 if level 2, 0 if otherwise.

x3 = 1 if level 3, 0 if otherwise.

1. **Add terms to the model of part b to allow for interaction between the quantitative and qualitative independent variables.**

Here you allow the terms to interact which gives the beta the following extensions, as in the exercises below:

To allow for interactions, the model would be:

1. **Under what circumstances will the response lines of the model in part c be parallel?**

The response lines will be parallel if:

1. **Under what circumstances will the model in part c have only one response line?**

There will be one response line if:

**Exercise 7. (92, WAGAP). *Agreeableness, gender, and wages*. Do agreeable individuals get paid less, on average, than those who are less agreeable on the job? And is this gap greater for males than for females? These questions were addressed in the Journal of Personality and Social Psychology (Feb. 2012). Several variables were measured for each in a sample of individuals enrolled in the National Survey of Midlife Development in the US. Three of these variables are: (1) level of agreeableness score (where higher scores indicate a greater level of agreeableness), (2) gender (male or female), and (3) annual income (dollars). The researchers modeled mean income, E(y), as a function of both agreeableness score (X1) and a dummy variable for gender (X2 = 1 if male, 0 if female). Data for a sample of 100 individuals (simulated, based on information provided in the study) are saved in the file. The first 10 observations are listed in the accompanying table.**

**Data for First 10 individuals in Study:**

|  |  |  |
| --- | --- | --- |
| **Income** | **Agree Score** | **Gender** |
| **44,770** | **3.0** | **1** |
| **51,480** | **2.9** | **1** |
| **39,600** | **3.3** | **1** |
| **24,370** | **3.3** | **0** |
| **15,460** | **3.6** | **0** |
| **43,730** | **3.8** | **1** |
| **48,330** | **3.2** | **1** |
| **25,970** | **2.5** | **0** |
| **17,120** | **3.5** | **0** |
| **20,140** | **3.2** | **0** |

**Consider the model, E(y) = β0 + β1 X1 + β2 X2. The researchers theorized that for either gender, income would decrease as agreeableness score increases. If this theory is true, what is the expected sign of β1 in the model?**

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data: library(readxl)

L8E7 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/L8/L8E7.xlsx")

View(L8E7)

attach(L8E7)

Create variables: Income <- Income

Agree <- Agree

Gender <- Gender

1. **The researchers also theorized that the rate of the decrease of income with agreeableness score would be steeper for males than for females (i.e. the income gap between males and females would be greater the less agreeable the individuals are). Can this theory be tested using the model, part a? Explain.**

The expected sign of would be negative.

1. **Consider the interaction model, E(y) = β0 + β1 X1 + β2 X2 + β3 X1 X2. If the theory, part b, is true, give the expected sign of β1 and of β3.**

No. In order to test this theory, the interaction between gender and agreeableness score would have to be added.

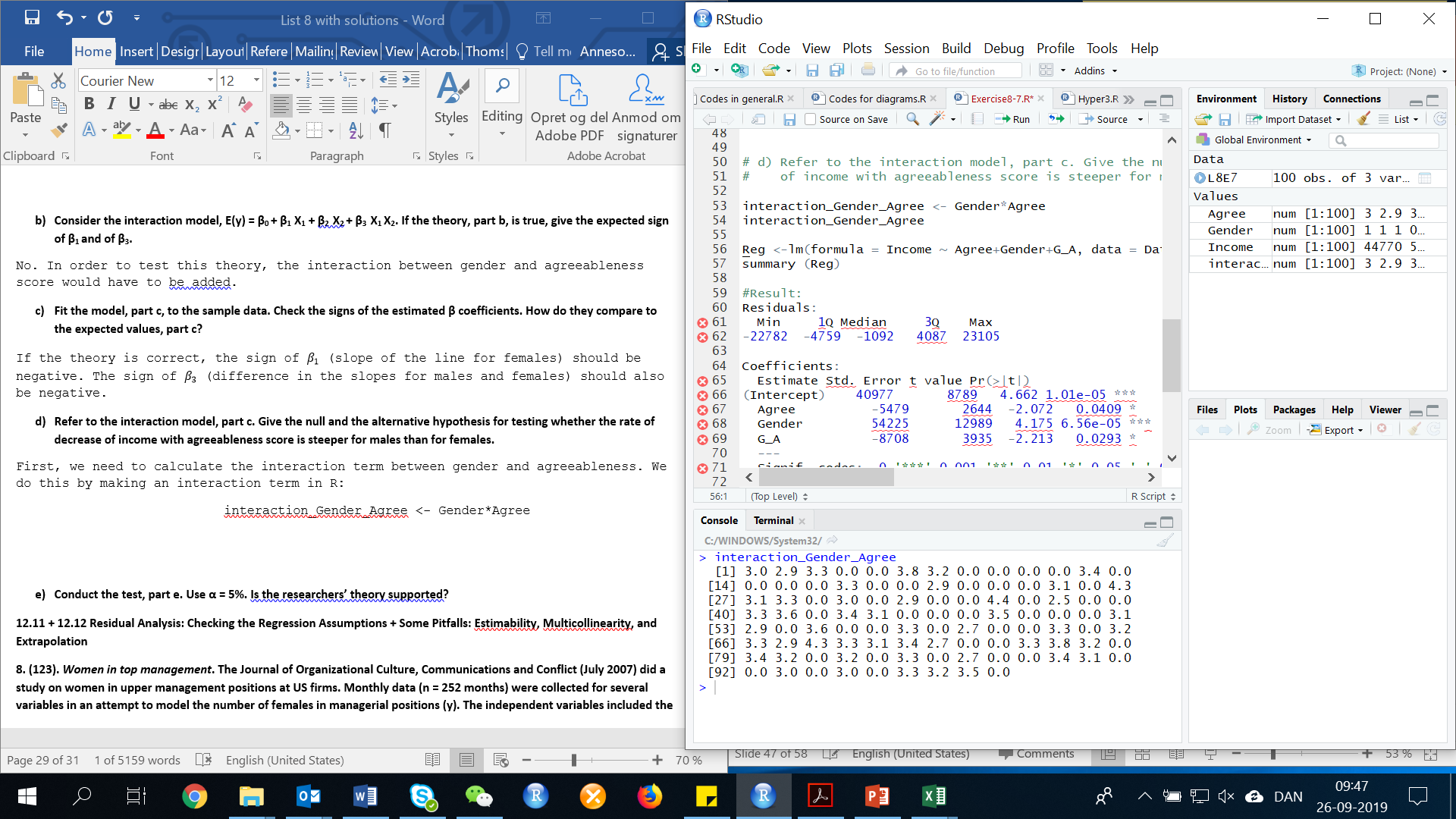
1. **Fit the model, part c, to the sample data. Check the signs of the estimated β coefficients. How do they compare to the expected values, part c?**

If the theory is correct, the sign of (slope of the line for females) should be negative. The sign of (difference in the slopes for males and females) should also be negative.

1. **Refer to the interaction model, part c. Give the null and the alternative hypothesis for testing whether the rate of decrease of income with agreeableness score is steeper for males than for females.**

First, we need to calculate the interaction term between gender and agreeableness. We do this by making an interaction term (multiplying the two variables) in R:

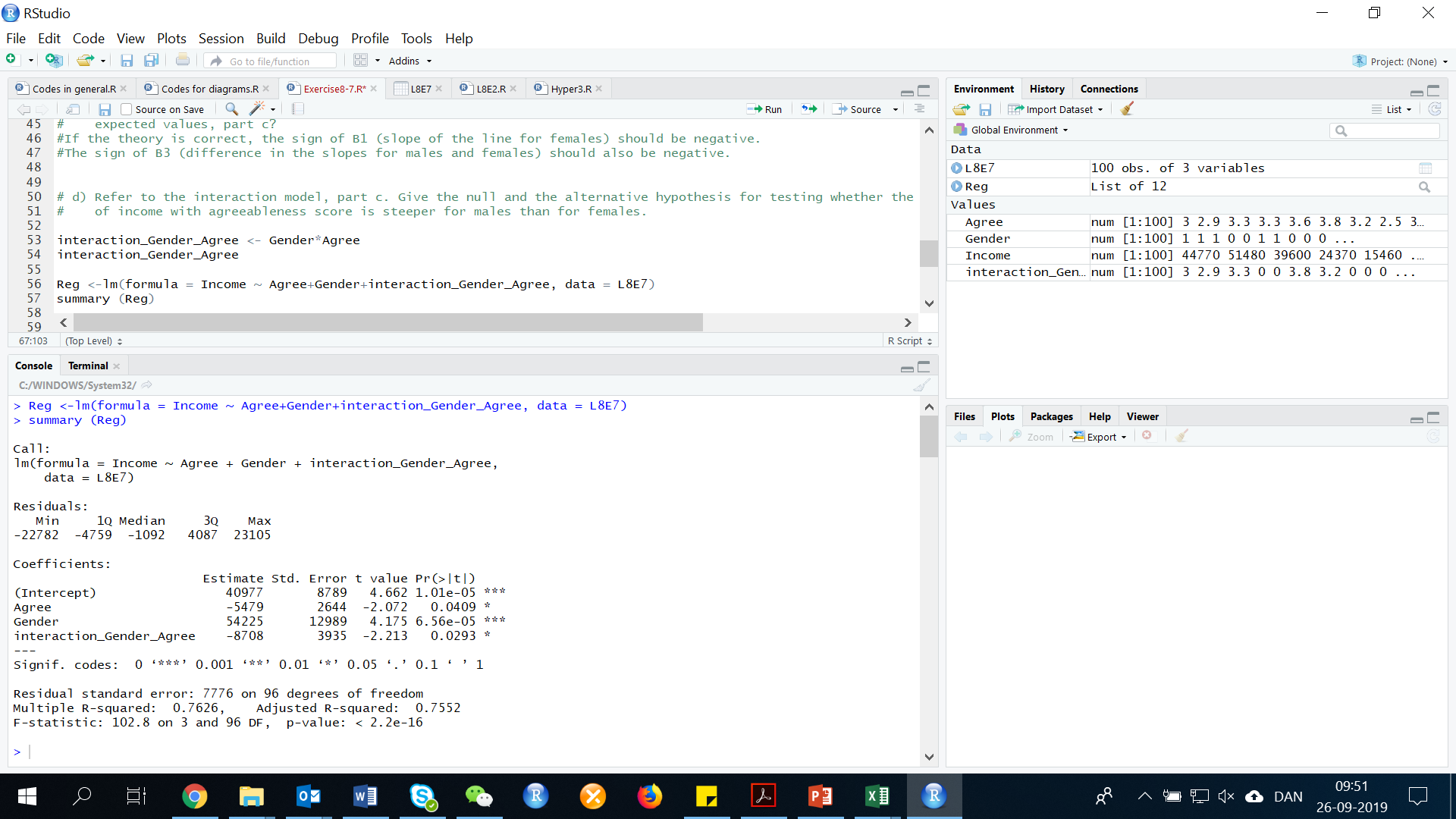
interaction\_Gender\_Agree <- Gender\*Agree



We can use this information to make the regression analysis:

Reg <-lm(formula = Income ~ Agree+Gender+interaction\_Gender\_Agree, data = L8E7)

summary (Reg)



As we expected and are both negative as expected.

1. **Conduct the test, part e. Use α = 5%. Is the researchers’ theory supported?**

To determine if the rate of decrease of income with agreeableness is steeper for males than for females, we test:

The test statistic is t = -2.21 and the p-value is p=.029/2=.0145. Since the p-value is less than alpha (p=.0145 <.05), H0 is rejected. There is sufficient evidence to indicate the rate of decrease of income with agreeableness is steeper for males than for females at alpha =.05.

**12.11-12 Residual Analysis: Checking the Regression Assumptions + Some Pitfalls: Estimability, Multicollinearity, and Extrapolation**

**Exercise 8. (123). *Women in top management*. The Journal of Organizational Culture, Communications and Conflict (July 2007) did a study on women in upper management positions at US firms. Monthly data (n = 252 months) were collected for several variables in an attempt to model the number of females in managerial positions (y). The independent variables included the number of females with a college degree (X1), the number of female high school graduates with no college degree (X2), the nuber of males in managerial positions (X3), the number of males with a college degree (X4), and the number of male high school graduates with no college degree (X5). The correlations are given in each part. Determine which of the correlations results in a potentia multicollinearity problem for the regression analysis.**

1. **The correlation relating number of females in managerial positions and number of females with a college degree: r = 0.983.**

The number of females in managerial positions is the dependent variable. The correlation between it and the independent variables does not imply multicollinearity.

1. **The correlation relating number of females in managerial positions and number of female high school graduates with no college degree: r = 0.074.**

Again, the number of females in managerial positions is the dependent variable. The correlation between it and the independent variables does not imply multicollinearity.

1. **The correlation relating number of males in managerial positions and number of males with a college degree: r = 0.722.**

Since the absolute value of the correlation coefficient is .722, this would imply there is a moderate potential for multicollinearity.

1. **The correlation relating number of males in managerial positions and number of male high school graduates with no college degree: r = 0.528.**

Since the absolute value of the correlation coefficient is .528, this would imply there is a moderate potential for multicollinearity.

**Exercise 9. (126, BDYIMG). *Reality TV and cosmetic surgery.* Refer to the Body Image: An International Journal of Research (March 2010) study of the influence of reality TV shows on one’s desire to undergo cosmetic surgery. Simulated data for the study is saved in the file. Fit the first-order model, E(y) = β0 + β1 X1 + β2 X2+ β3 X3+ β4  X4, where y = desire to have cosmetic surgery, X1 is a dummy variable for gender, X2 = level of self-esteem, X3 = level of body satisfaction, and X4 = impression of reality TV.**

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data: library(readxl)

L8E9 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/L8/L8E9.xlsx")

View(L8E9)

attach(L8E9)

Create variables: DESIRE <- DESIRE

SELFESTM <- SELFESTM

BODYSAT <- BODYSAT

IMPREAL <- IMPREAL

GENDER <- GENDER

1. **Check the data for multicollinearity. If you detect multicollinearity, what modifications to the model do you recommend?**

In this exercise, we need to detect for multicollinearity. In order to understand, what multicollinearity is, please use the link below as guidance:

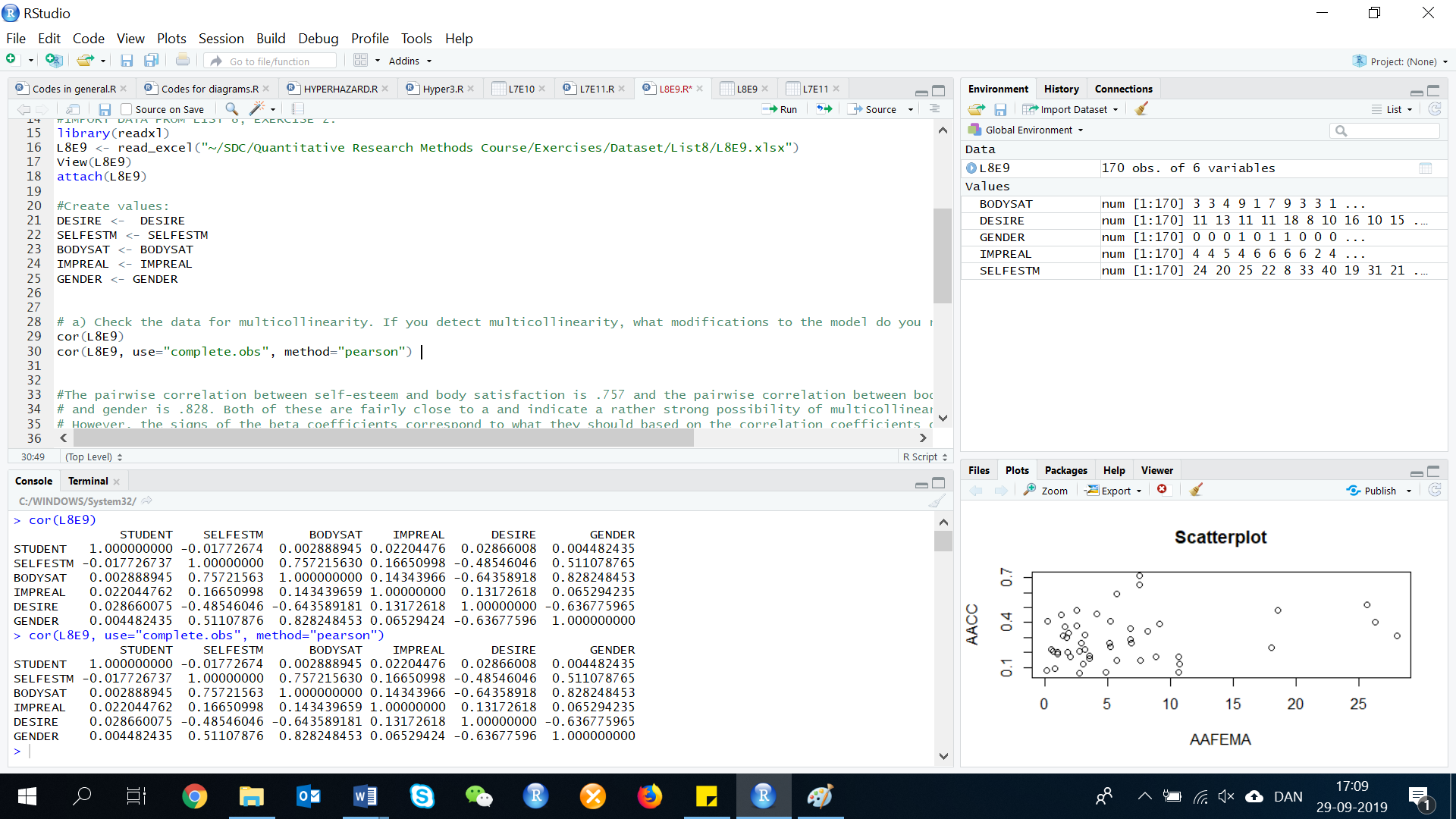
<http://www.real-statistics.com/multiple-regression/collinearity/>

Multicollineartiy is simply the independent variables overlapping in short. More detailed means that the and variable may attempt to explain the same information in the data set hence they move in the same pattern. In R there is many different ways to do this, this is one of them:

cor(DATA)

cor(L8E2)

Correlations can be calculated for a pair or matrix of variables. By default in this code, the pearson correlations are provided. There are other variants such as Spearman. You can specify the method, by doing as followed:

cor(L8E2, use="complete.obs", method="pearson")

The pairwise correlation between self-esteem and body satisfaction is .757 and the pairwise correlation between body satisfaction and gender is .828. Both of these are fairly close to a and indicate a rather strong possibility of multicollinearity. However, the signs of the beta coefficients correspond to what they should based on the correlation coefficients of the independent variables and the dependent variable. One might not want to include both variables in the pairs where the correlation is above .7.

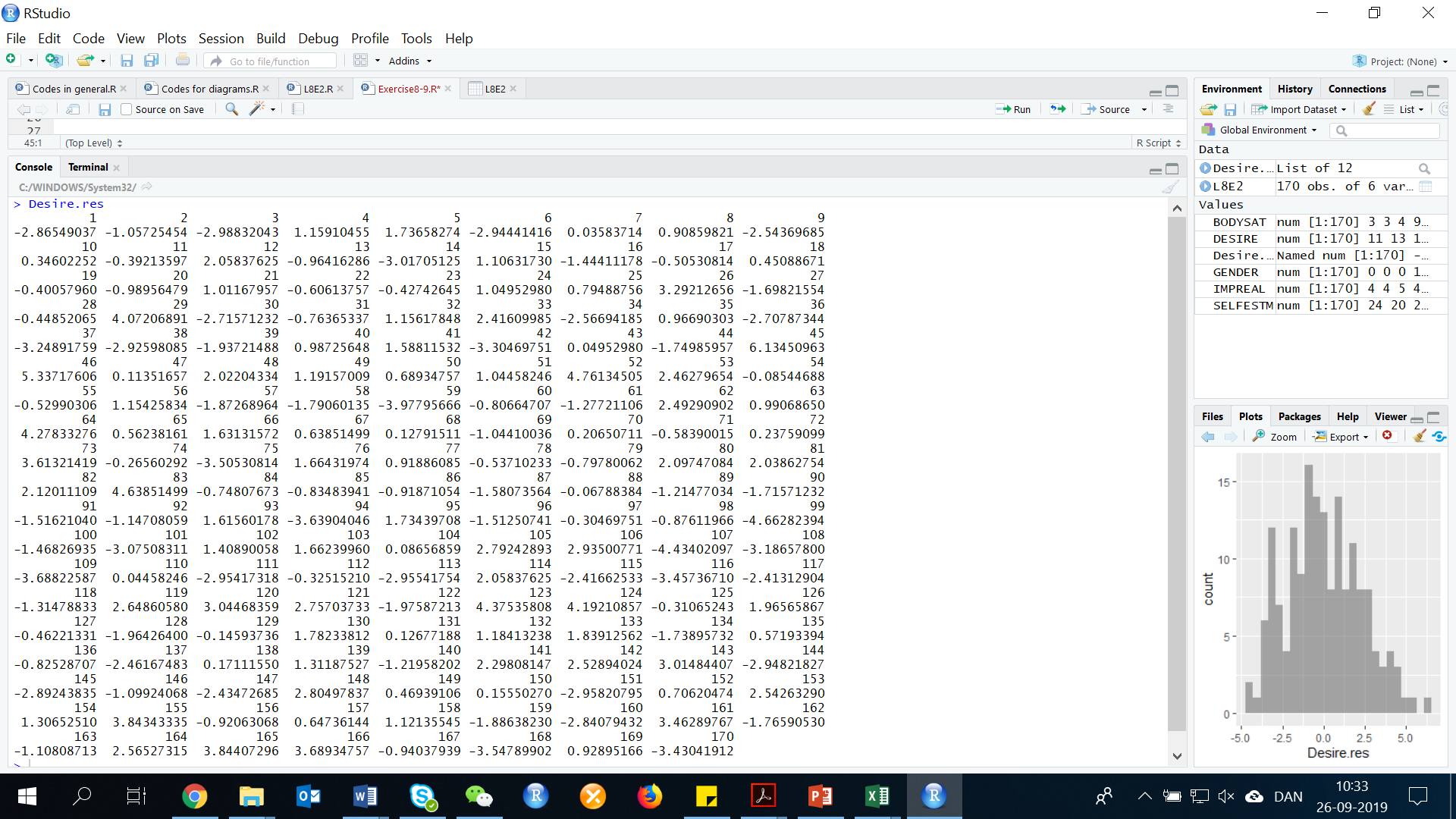
1. **Conduct a complete residual analysis for the model. DO you detect any violations of the assumptions? If so, what modifications to the model do you recommend?**

There is many different ways to make a residual analysis for the model. We use the following way:

Desire.lm <- lm(DESIRE ~ SELFESTM+BODYSAT+IMPREAL+GENDER, data=L8E9)

Desire.res <- resid(Desire.lm)

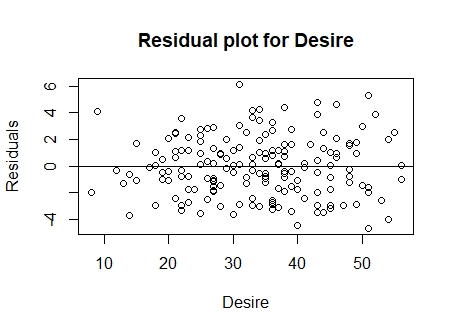
Desire.res

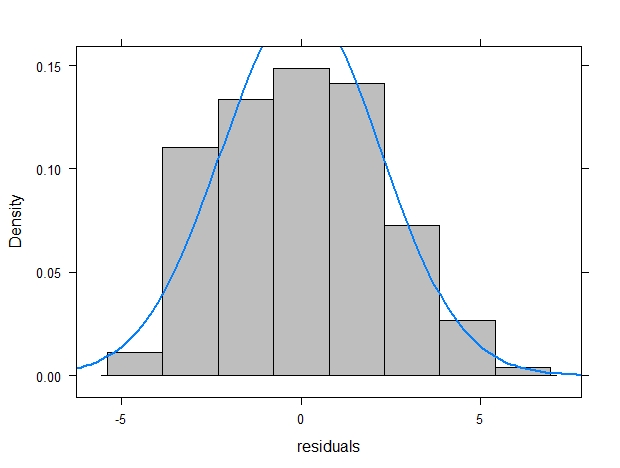


We now need to make some plots, that can give us a graphical explanation of the residuals.

Residual plot:

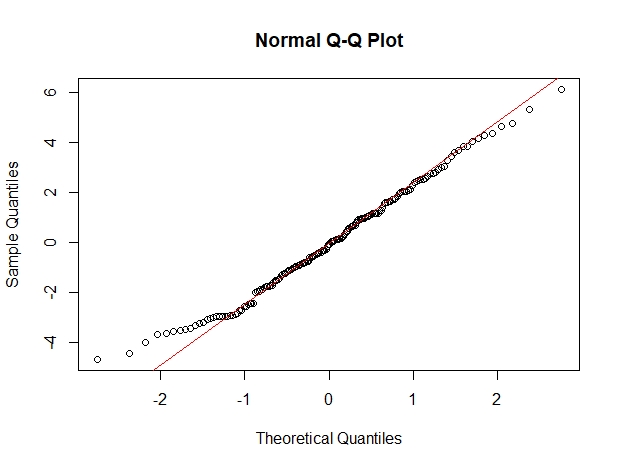
plot(L8E9$SELFESTM+BODYSAT+IMPREAL+GENDER, Desire.res, ylab="Residuals", xlab="SELFESTM+BODYSAT+IMPREAL+GENDER", main="Residual plot for Desire")

abline(0, 0)

Histogram: histogram(~Desire.res, xlab="residuals", fit="normal", data=L8E9, col="grey", border="black")

Normal probability plot:

qqnorm(Desire.res)

qqline(Desire.res, col="red")

From the normal probability plot and the histogram, the error terms appear to be fairly normally distributed. From the residuals plot there is no evidence of non-constant variance for the error terms. From the same plot, all of the standardized residuals have values less than 2.5 in absolute value. Thus, there are no outliers. Since the data were not collected sequentially, the plot of the residuals versus time is meaningless and we cannot check for independence of the error terms.

**Exercise 10. (127, WELLS). Arsenic in groundwater. Refer to the Environmental Science & Technology (Jan. 2005) study of the reliability of a commercial kit to test for arsenic in groundwater. Fit a first-order model for arsenic level (y) as a function of latitude (X1), longitude (X2) and depth (X3) to data saved in the file. Conduct a residual analysis of the data. Based on the results, comment on each of the following:**

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data: library(readxl)

L8E10 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List8/L8E10.xlsx")

View(L8E10)

attach(L8E10)

Create variables: ARSENIC <- ARSENIC

LATITUDE <- LATITUDE

LONGITUDE <- LONGITUDE

DEPTH <- DEPTH

To answer the questions in this exercise, we need to make a residual analysis:

arsenic.lm <- lm(ARSENIC ~ LATITUDE+LONGITUDE+DEPTH, data=L8E10)

arsenic.res <- resid(arsenic.lm)

arsenic.res

The code for each plot will be showed here, and the pictures of the plots will appear below:

Residuals plot:

plot(L8E10$LATITUDE+LONGITUDE+DEPTH, arsenic.res, ylab="Residuals", xlab="LATITUDE+LONGITUDE+DEPTH", main="Residual plot for Arsenic")

abline(0, 0)

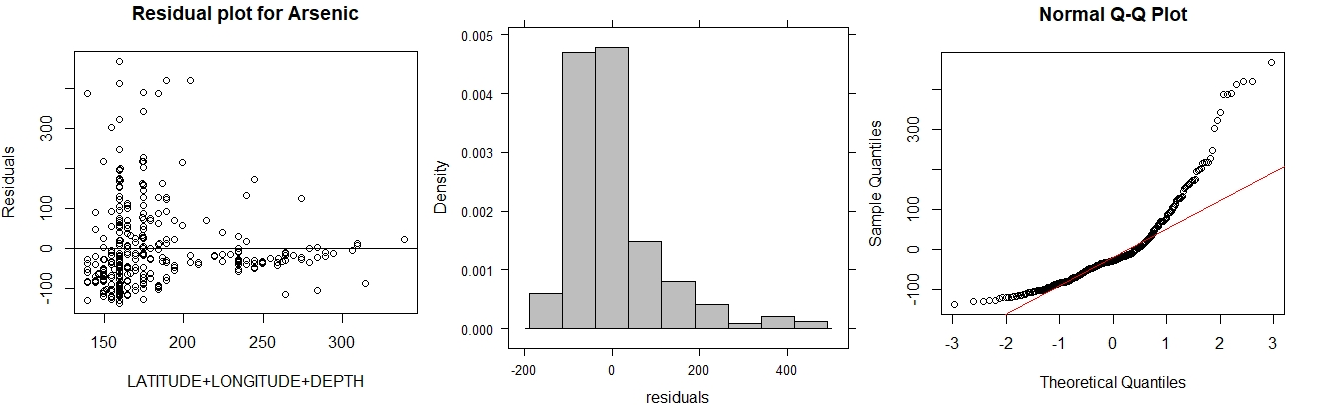
Histogram:

histogram(~arsenic.res, xlab="residuals", data=L8E10, col="grey", border="black")

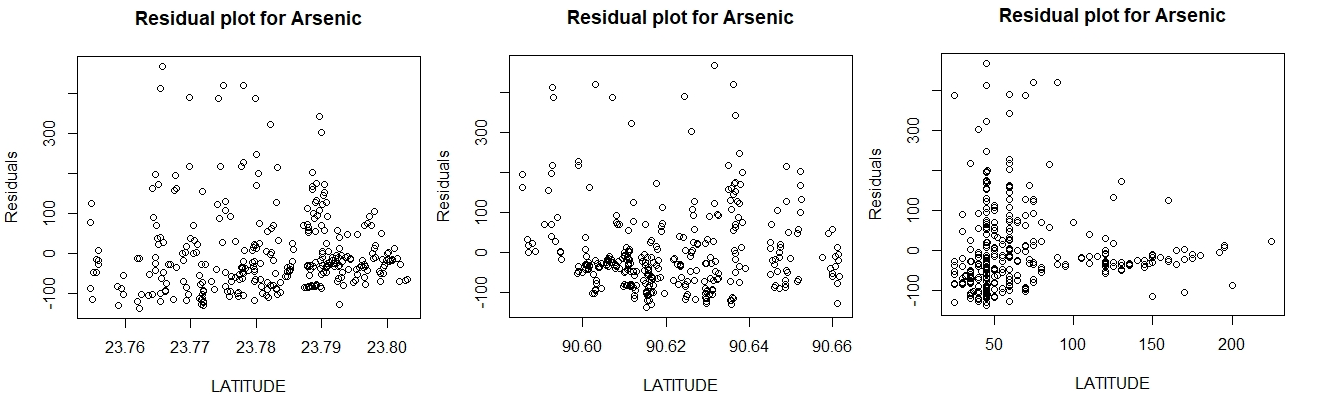
Normal probability plot:

qqnorm(arsenic.res)

qqline(arsenic.res, col="red")



Scatterplot of Lattitude, Longitude, and Depth vs. Arsenic:



1. **assumption of mean error = 0**

From the histogram of the standardized residuals, it appears that the mean of the residuals is close to 0. Thus, the assumption that the mean error is 0 appears to be met.

1. **assumption of constant error variance**

From the plot of the standardized residuals versus the fitted values, it appears that the spread of the residuals increases as the fitted values increase. Thus, it appears that the assumption of constant variance is violated.

1. **Outliers**

From the plots of the standardized residuals versus the fitted values, it appears that there are some outliers. there are several observations with standardized residuals of 4 or more.

1. **assumption of normally distributed errors**

From the normal probability plot, the data do not form a straight line. Thus, it appears that the assumption of normal error terms is violated.

1. **multicollinearity**

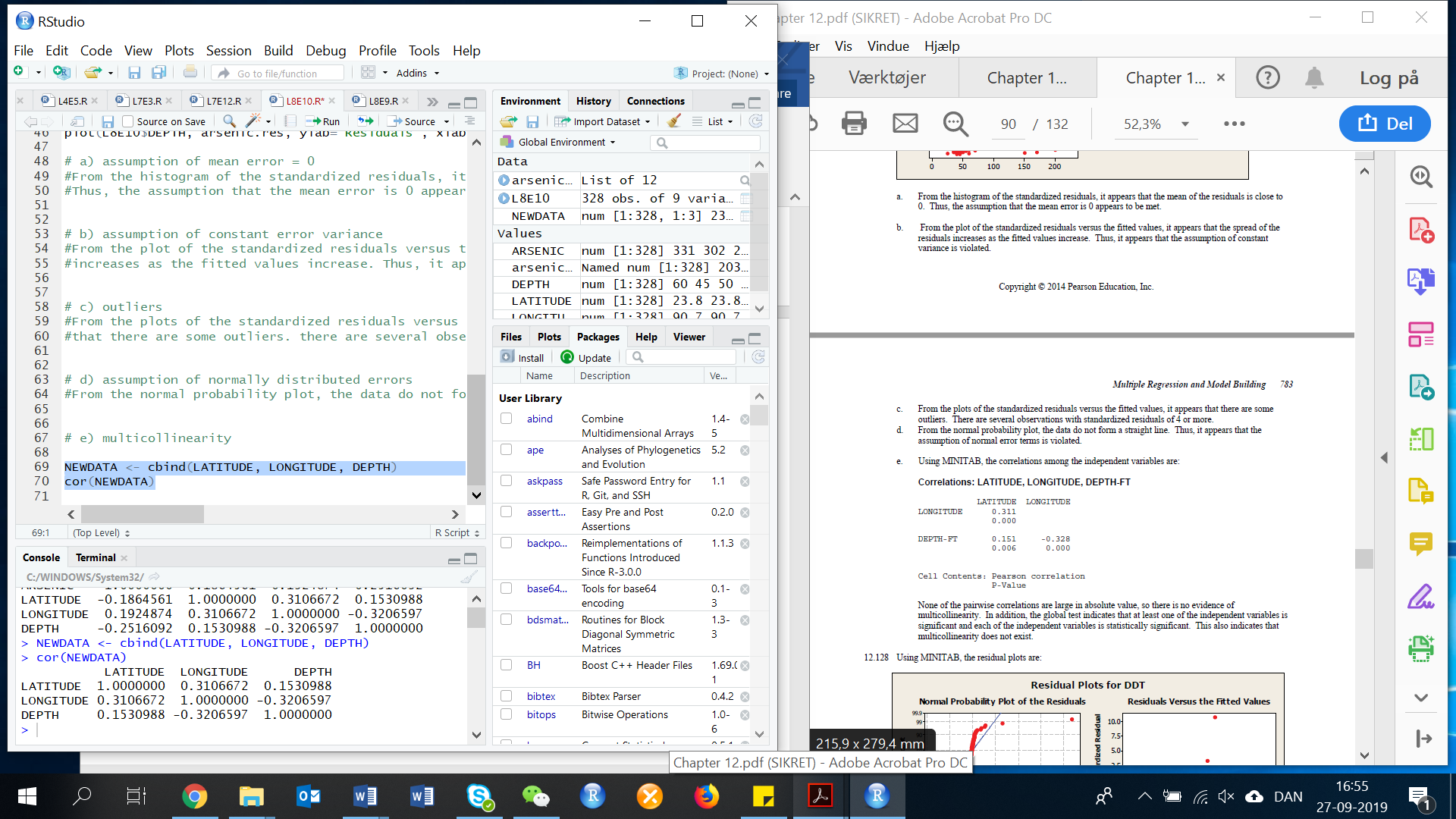
To find the correlations, we first need to make a new dataset, that only contains the three independent variables:

NEWDATA <- cbind(LATITUDE, LONGITUDE, DEPTH)

Now we can calculate the correlations:

cor()

cor(NEWDATA)



None of the pairwise correlations are large in absolute value, so there is no evidence of multicollinearity.