Hannah Schaefer

NRE 538

This exam is open book and open internet but you are NOT allowed to work with anyone else or ask anyone other than Meha or Oscar any questions about the exam. It is due at noon on Sunday, April 23.

Please answer the following questions by analyzing the associated datasets. For all tests, please:

* check whether the data meet the requirements/assumptions of the test you plan to run
* complete any transforms needed to make the data meet the required assumptions
* run the test
* interpret the results (do not include only the R output)
* check model fit in the case of linear regressions and/or glms
* if you have the option between running a linear model with a transformed y variable or a glm, choose the linear model with a transformed y variable. only run a glm when you have to.

Provide all answers in R or R markdown (similar to the take home quiz 4). Use the following scripts to load the datasets. The dataset to be used for each question is provided in bold at the end of the question.

Dataset Please use the following scripts to load in the data from GitHub

flying = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/flying.csv",header=TRUE, sep=",")

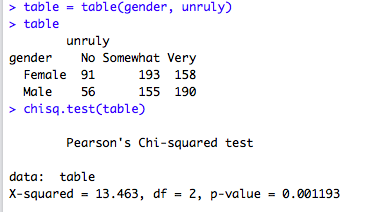
college = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/college.csv",header=TRUE, sep=",")

happy = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/happy.csv",header=TRUE, sep=",")

cancer = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/cancer.csv",header=TRUE, sep=",")

1. **Is there a significant association between gender (gender) and whether people think it’s rude to bring an unruly child on the plane (unruly\_child)? If yes, which gender tends to think that bringing an unruly child is more rude? Flying**

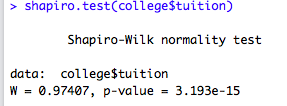
This data requires a chi-square test because both variables are categorical. After assuming the data observations were independent, I looked at the data and saw that some of the values were *NA*. Because one of the assumptions of the chi-square test is that all of the boxes in the table are filled, I removed the rows that had observations with NA. This left 582 observations from the original 1040. I made a contingency table with the gender and unruly child factors after I gave them new labels of “gender” and “unruly.” I ran a chi-square test for independence using that table and found that there was a significant association between gender and whether people think it is rude to bring an unruly child on the plane. The p-value was 0.001193 which also represents how much type 1 error we are willing to accept. The male gender is more likely to think bringing an unruly child is rude.

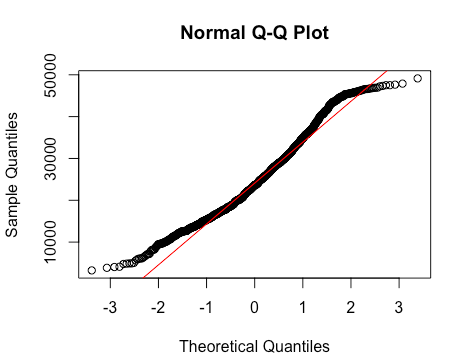
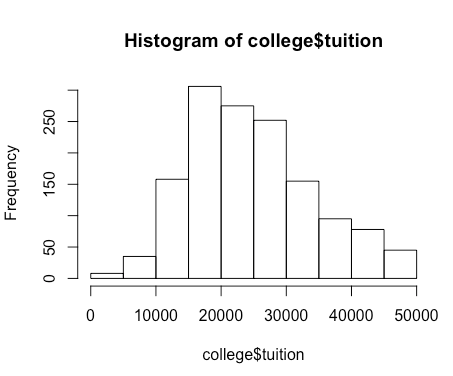


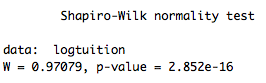
1. **Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition? college**

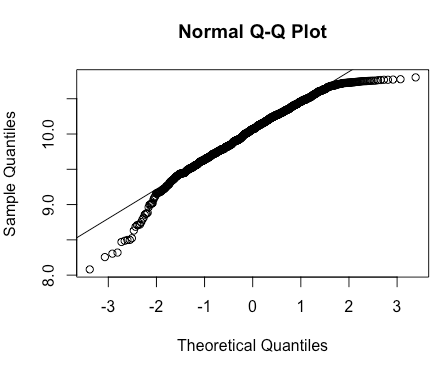
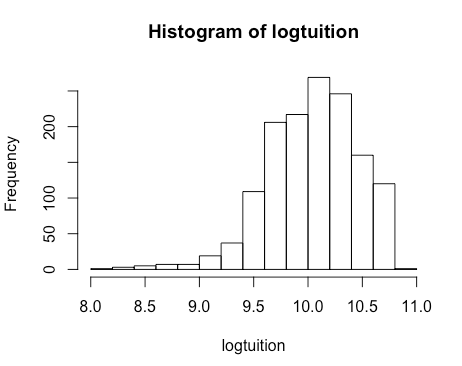
This requires a two-sample t-test because we are testing if the tuition amounts are significantly different between two types of institutions. The dependent variable is continuous, and the independent variable of type of institution is categorical. It is a two-sample because we are testing public and private nonprofit institutions both sampled in the “type” variable. We are assuming the data are randomly selected and the observations are independent. To test if the tuition data were normal, I ran a Shapiro-Wilk normality test. The results were significant with a p-value of 3.193e-15, which means that the we reject the null hypothesis that the data are normal, which means the data are not normally distributed. After looking at the histogram, the data has a slight right skew. The qqplot also showed the data not very well aligned towards the beginning and end values of the data. To try and make the data have a normal distribution, I ran a couple of transformations. First I did a log transformation, but this made the data more skewed and increased the p-value when I ran another Shapiro-Wilk test. When I used a square root transformation, the p-value from the Shapiro-Wilk test decreased and the qqplot looked more normal. The sample size of the data is also large, so this assumption could be slightly relaxed.

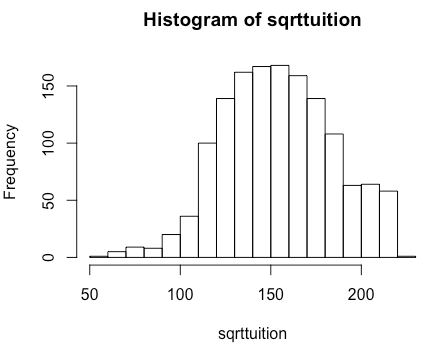
I then added a column for the square root of tuition data to the “college” data frame and then separated the private college data from the public college data. Then I ran an F test to test the variance and found that because the p-value was significant, it meant that there was unequal variance between the data. I decided to run a Welch’s t-test because it can run for data with unequal variance. The results from the Welch’s t-test had a very small p-value at 2.2e-16, meaning we can reject the null hypothesis. For this test, the null hypothesis was that the means between tuition costs for private institutes were different from those of public institutes. Looking at the means for each, private institutions have a significantly higher cost for tuition than public colleges. However, it is important to remember that the data have been transformed, and any interpretation must take the square root value to get the actual numerical value.

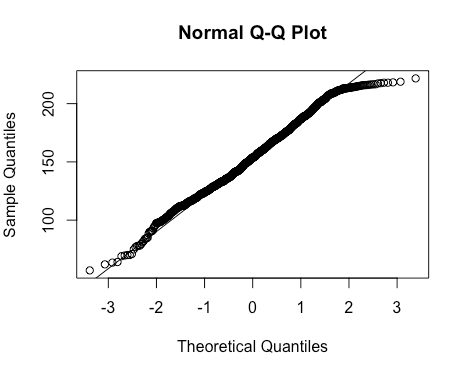


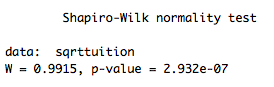


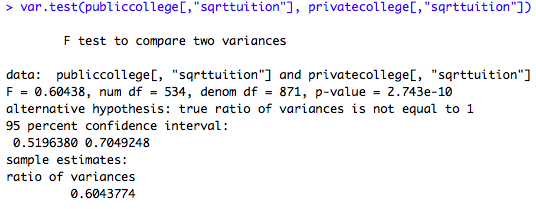


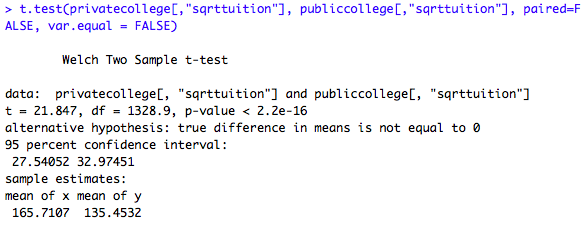








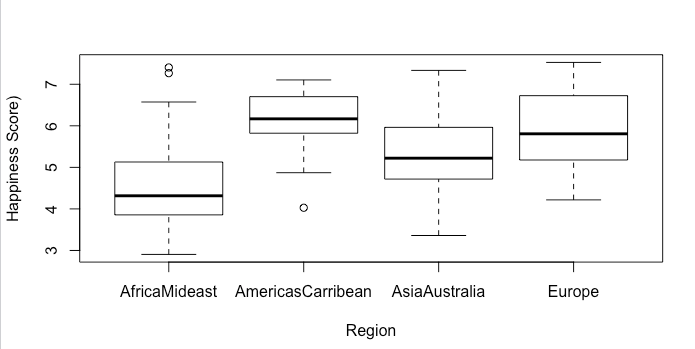




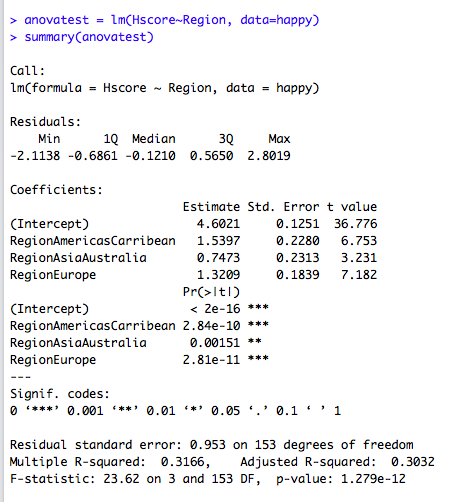
1. **Is there a significant difference in happiness (Hscore) by region (Region)? Happy**

First I made a boxplot to visualize the data. From the boxplot we see that the third and first quartiles of the AfricaMideast and AmericasCaribbean regions do not overlap, respectively, and the medians are significantly far from each other. The median of AfricaMideast does not overlap with any of the quartiles of the other regions. The median of the Europe region lies within the third quartile of AsiaAustralia and the first quartile of AmericasCaribbean. The median of AmericasCaribbean falls within the third quantile of Europe, and the median of AsiaAustralia is very close to the first quartile of Europe. The test required is an ANOVA. To make sure the data had a normal distribution, I plotted a histogram. The data looked normal for the dataset in its entirety. When I plotted histograms for all 4 regions and their Hscore, only the AmericasCaribbean data was skewed, and it was left-skewed. The sample size was small, and I didn’t think it would work well with a transformation.

Then I ran a Shapiro-Wilk test for the Hscore data for “happy”, but the results showed a significant p-value of 0.012 which suggests the data are not normally distributed. There is a fairly large sample size of 157, and the data were very close to passing the Shapiro-Wilk test (it is often very conservative) so I decided to move forward. I assumed the data were independent of one another. I ran the ANOVA with Hscore dependent on Region. AfricaMideast was the intercept value because it is first alphabetically. It was significantly different from 0 with a coefficient of 4.60. The AmericasCaribbean region was significantly differently from the Africa region with a coefficient of 1.54. The AsiaAustralia region had a coefficient of 0.0015, and the Europe region had a coefficient of 2.81e-11. These coefficient values would be added to the intercept value to see where they fall on the graph. With an adjusted R squared value of 0.30, the model explains about 30% of the variability in the data. With a p-value of 1.279e-12, there is a significant difference in happiness by region.



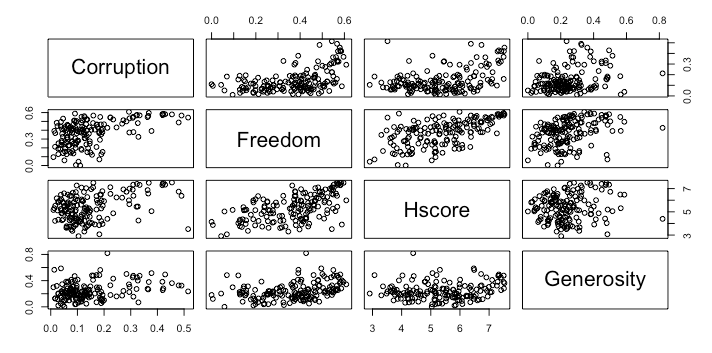


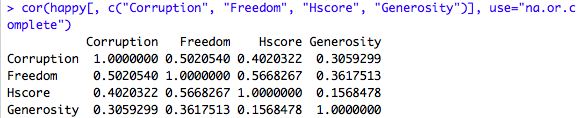


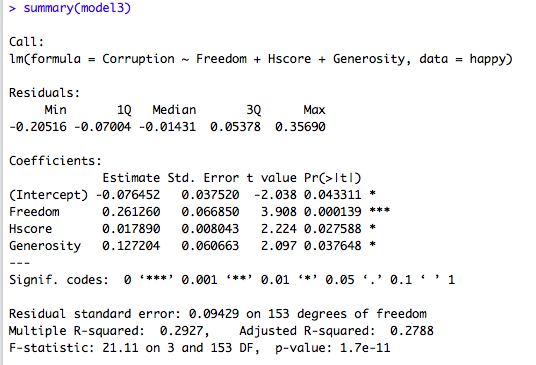
1. **What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model. happy**

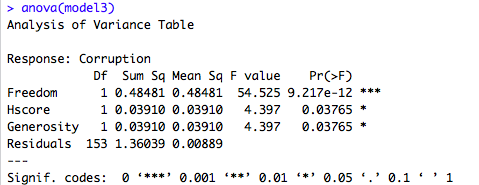
I chose the 3 continuous variables “Freedom,” “Hscore” and “Generosity” to include in my linear model. I assumed the data were independent from one another. One of the assumptions is that for the linear model there is a constant variance in the errors, but when I plotted my model later on I found there was some heteroscedasticity. This assumption can be flexible with linear models. The qqplot showed mostly normal data points. Before running the model I used “pairs” and “cor” to make sure the variables were not significantly correlated. The highest correlation I found was between “Freedom” and “Corruption” with a correlation value of about 0.50. I thought that this correlation was low enough that I would be able to use the variable without trouble.

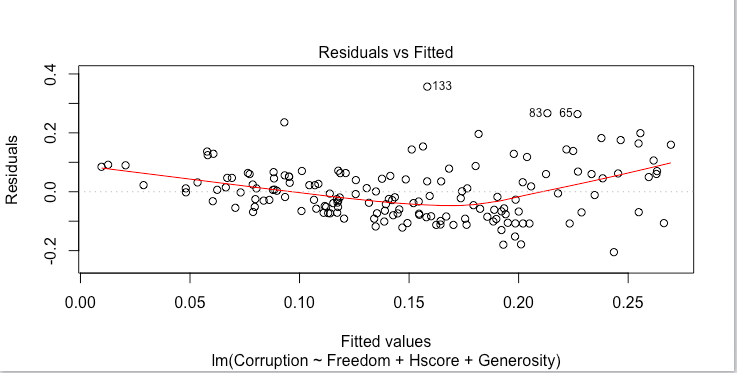
All three continuous independent variables were significantly different from the dependent variables “Corruption” according to my linear model. The MSE was calculated to be 0.0089 and was checked by using the “anova” code and the “Analysis of Variance Table.” The variables all had a linear relationship with the dependent variable. By using the “plot” function in R, the errors had some heteroscedasticity. The interpretation of the data before transformation would be that as freedom increases by 1, corruption increases by about 0.26. As happiness (Hscore) increases by 1, corruption increases by 0.018. As generosity increases by 1, corruption increases by 0.128. The adjusted R squared value of the model was 0.279, so it explains about 28% of the variance in the data. This is not a very good R squared value. I decided to try a log transformation and a square root transformation with the dependent variable. Neither of them produced data with a true normal distribution. Because of this, and since the sample size was fairly large, I stuck with my initial interpretation.

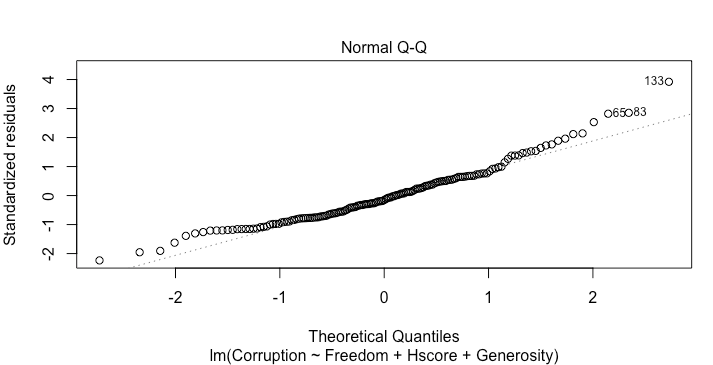


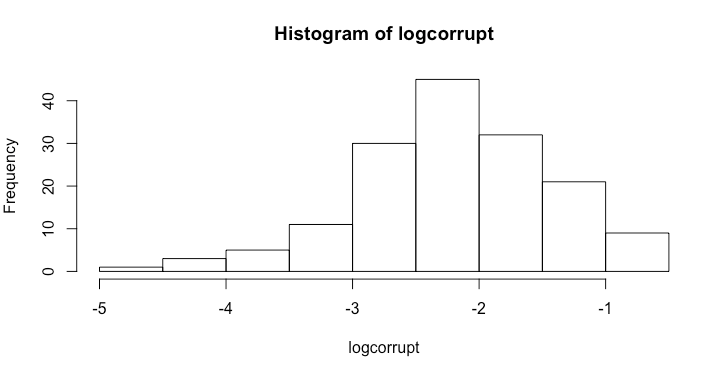


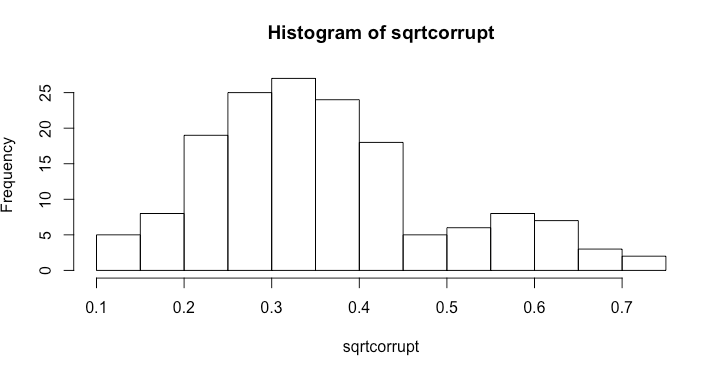






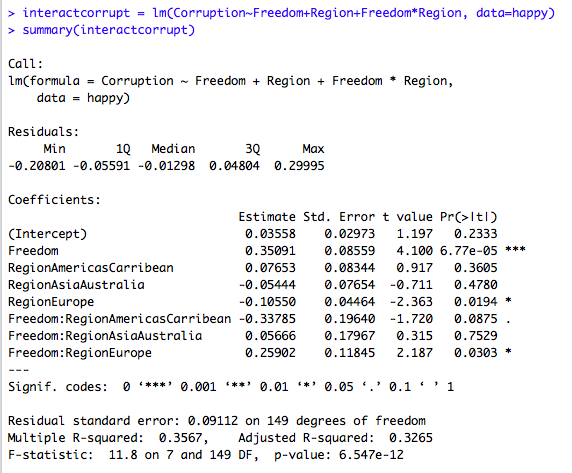






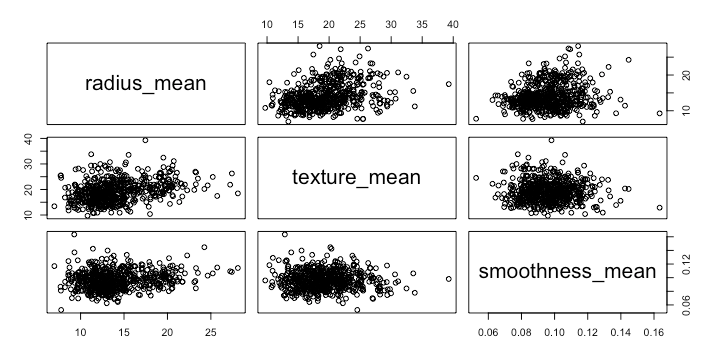
1. **Choose one of the continuous independent variables that was significant in the model for Question 4 and interact it with region (Region) to predict corruption (Corruption). This model should only include one continuous independent variable and its interaction with region. Does the influence of your continuous variable on corruption vary by region? If yes, how do you interpret the interaction? happy**

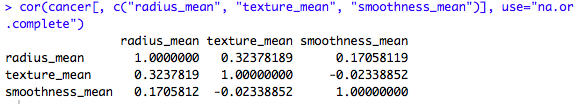
I choose to interact the “Freedom” variable with “Region” in my model. The intercept (“Corruption”) was not significantly different from 0. The effect of “Freedom” is significant, so every increase in Freedom leads to an increase of 0.35 of corruption. The variables for RegionAmericasCaribbean, RegionAsiaAustralia, Freedom:RegionAmericasCaribbean, and Freedom:RegionAsiaAustralia were not significant. For RegionAmericasCaribbean and RegionAsiaAustralia, this means that region has no effect on Corruption when Freedom is 0. For Freedom:RegionAmericasCaribbean,and Freedom:RegionAsiaAustralia, this means that the slope for the effect of Freedom on Corruption is not significantly different for these regions. The p-value for RegionEurope was 0.0194 which means it was significant. The interpretation would be that when Freedom is 0, Corruption for the region of Europe is -0.1055. There was also significance with the Freedom:RegionEurope coefficient. This means that the slope for the effect of Freedom on Corruption is significantly different for Europe compared to the other 3 regions. The slope for Europe would be 0.3558 plus 0.259.

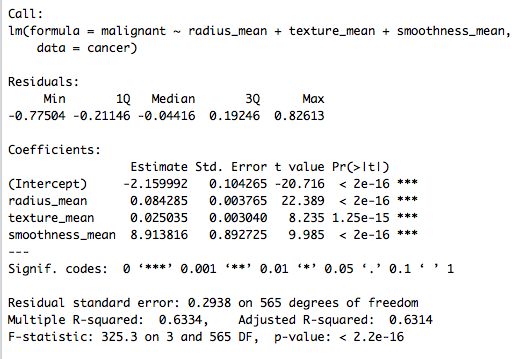


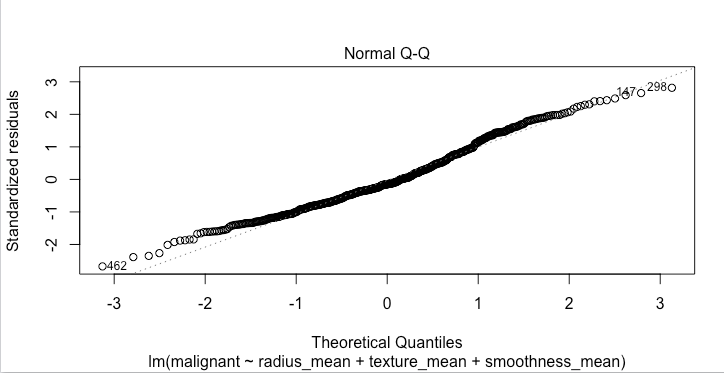
1. **Which factors are significantly associated with whether a breast cancer tumor is malignant or not? Choose three continuous independent variables to include in your model. cancer**

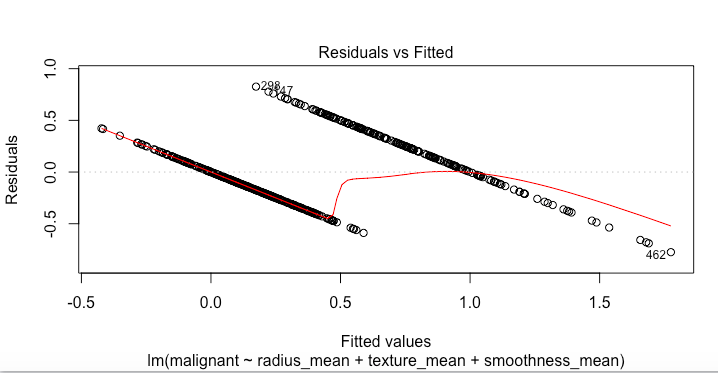
I decided to run a linear model with the data. I am assuming the data are independent and there is a linear relationship between the dependent and independent variables. The 3 independent variables were “radius\_mean,” “smoothness\_mean,” and “texture\_mean.” When I tested for correlation, these were the only three variables that did not have significant correlation with each other. According to the model, all three variables were statistically significant with p values below 0.05. The intercept, or malignant value, was significantly different from 0 at -2.16. As mean radius increases by 1, the change of the tumor being malignant increases by 0.08. As the value for mean texture increases by 1, the value for a malignant tumor increases by 0.025. As the mean smoothness value increases by 1, the value for a malignant tumor increases by 8.91. The adjusted R square value for the model is about 0.63 which means 63% of the variance in the data can be explained by the model, which is usually pretty good. I used the “plot” code in R to see that the qqplot showed a good fit to the data. The residuals plot was interesting, however. It is likely there was a relationship with the data that showed correlation with the errors. To solve this, robust sampling errors could be implemented.











1. BONUS/EXTRA CREDIT: Which independent variables are the most important in explaining whether a breast cancer tumor is malignant or not? Use the same 3 continuous independent variables you chose for question 6. **cancer.**

From looking at my three variables, smoothness\_mean has the largest effect on the malignant value in the model and is also very significant.