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=",")

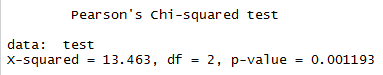
1. Is there a significant association between gender (gender) and whether people think it’s rude to bring an unruly child on the plane (unruly\_child)? If yes, which gender tends to think that bringing an unruly child is more rude? **Flying**

Test to use: Chi-Square Test for Independence

Assumptions:

* Random Sample: I am assuming that each case is randomly sampled from the larger population (this is a step within the sampling process, thus I cannot test for it using R)
* Independence: The observations are independent of one another (again, this cannot be tested using R, so I will assume the observations are independent of one another)
* Sample is small enough: Given that there are millions of members of each gender that could have possibly been asked their opinions on bringing an unruly child on a plane, and there are just 1,040 participants in this study, I can assume that the number of observations is less than 10% of the total number of possible observations in the overall population.
* Enough Expected and Observed Values: There are 843 observations (after removing NA data), which is an adequate number of values. Furthermore, my contingency table shows an adequate number of observations in each cell:
  + 

Running the test

* 
* The p-value is less than 0.05, thus I can reject the null hypothesis that the two categorical variables are independent. There does seem to be a significant association between gender and whether people think it is rude to bring an unruly child onto a plane.

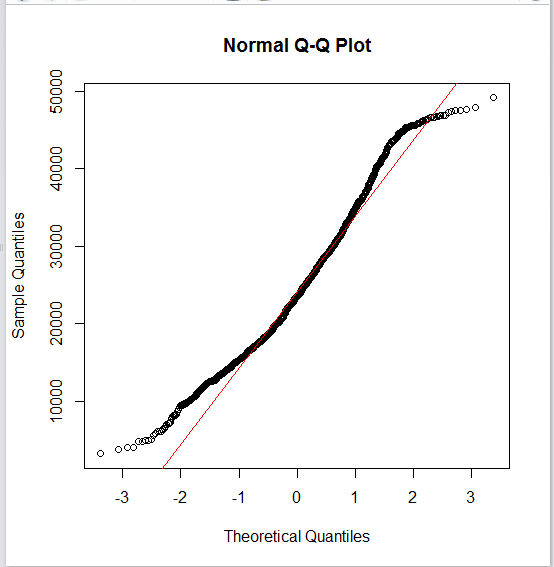
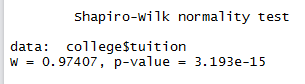
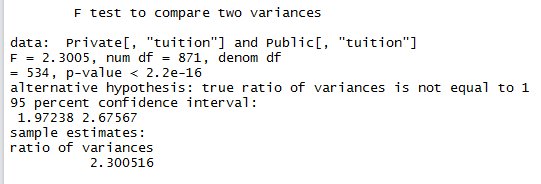
Which gender thinks it is ruder:

* Male

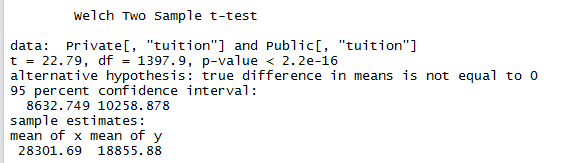
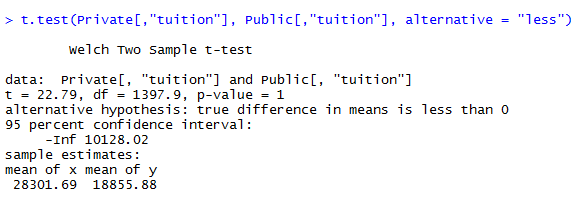
1. Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition? **College**

Test to use: Unpaired, Welch’s two-sample t-test

Assumptions:

* Dependent data are continuous: yes, tuition is a continuous variable
* Sample is randomly selected from population: I am assuming this is true (since there is no test to determine this)
* Values are nearly normal OR sample size is large enough: The sample size is over 30 so this automatically satisfies this test. After conducting tests for normality (qqplot and Shapiro tests), I found that the data is not normally distributed, however, since my sample size is over 30 I can continue on with the test.
  + Normality:
    - 
    - 
      * Significant p-value, thus reject the null hypothesis that the population is normally distributed.
* Equal variances between 2-populations: Fails this test! The p-value is significant, meaning we reject the null hypothesis that variances are equal. Thus, I need to use a Welch’s two-sample t-test
  + 

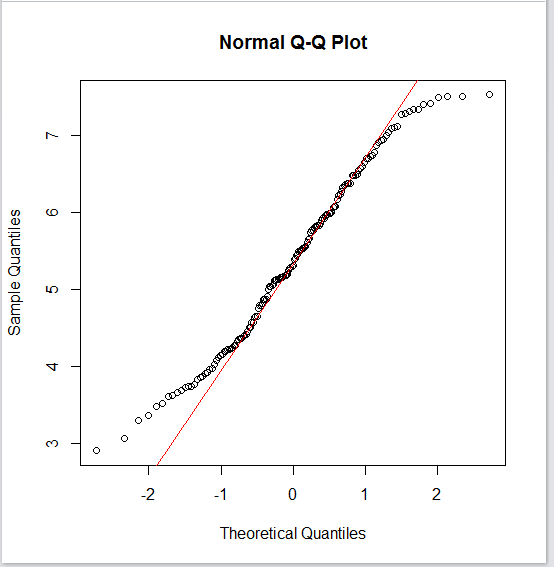
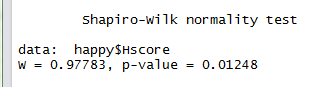
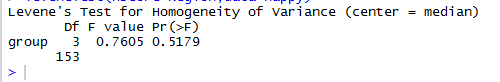
Running the test:

* + 
  + There is a significant difference between tuition at public and private institutions (since the p-value is less than .05, I can reject the null hypothesis that there is no difference).
* Which has higher tuition? Private.
  + I determined this by making my test a one-tailed t-test
  + 
    - The p-value is greater than 0.05, meaning we fail to reject the null hypothesis that private tuition is less than public tuition. Thus, I can infer that private tuition is higher than public tuition.

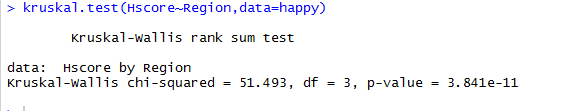
1. Is there a significant difference in happiness (Hscore) by region (Region)? **Happy**

Test to use: One- way ANOVA (Kruskal Wallis)

Assumptions:

* The populations are normally distributed
  + - Overall normality of dependent variable:
      * 
      * 
      * Significant p-value, thus reject the null hypothesis that the population is normally distributed. Thus, I need to use a Kruskal Wallis test.
* Samples must be independent of each other: Since there is no test for this I am assuming my data passes this assumption.
* Each population has the same variance: Given that the p-value is not significant, I fail to reject the null hypothesis that variances are equal. Thus, it passes.
  + 

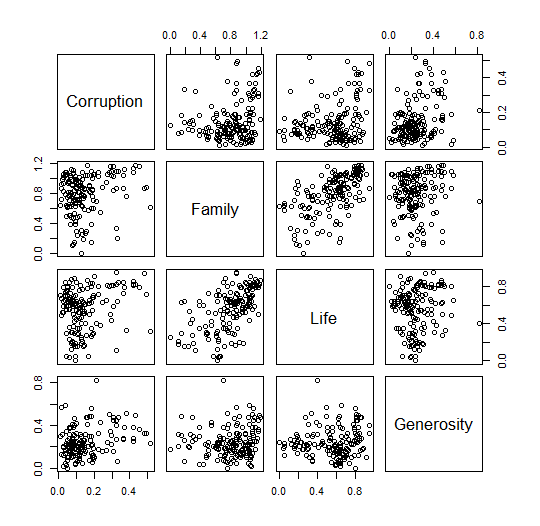
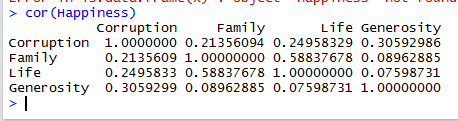
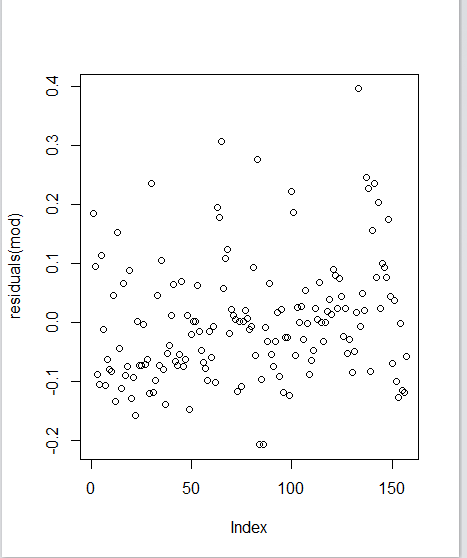
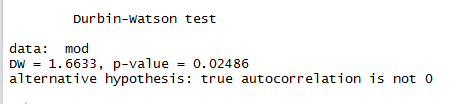
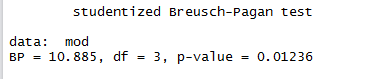
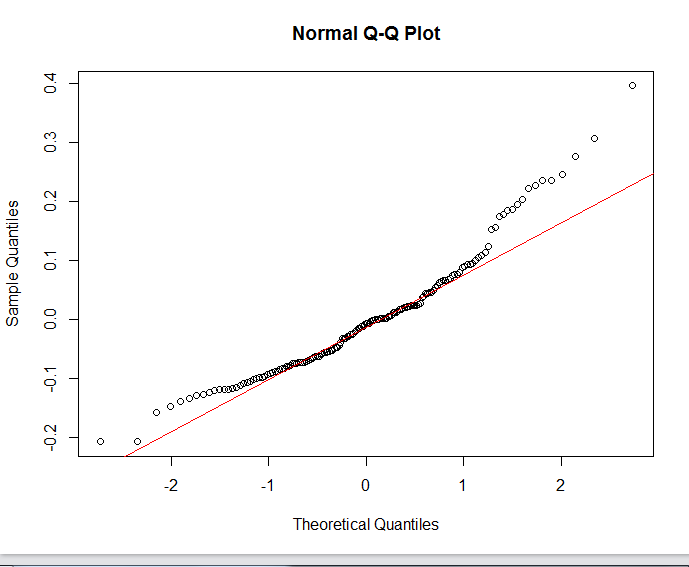
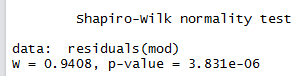
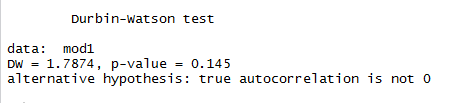
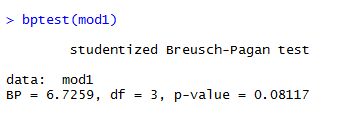
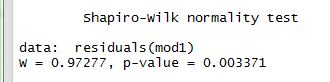
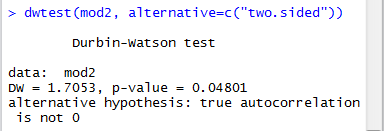
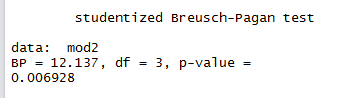
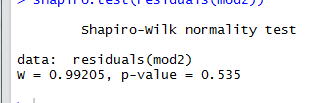
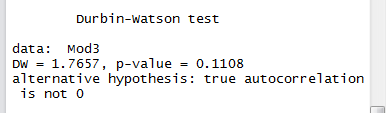
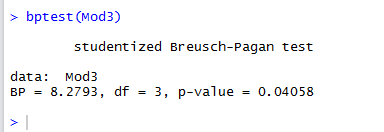
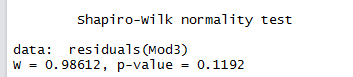
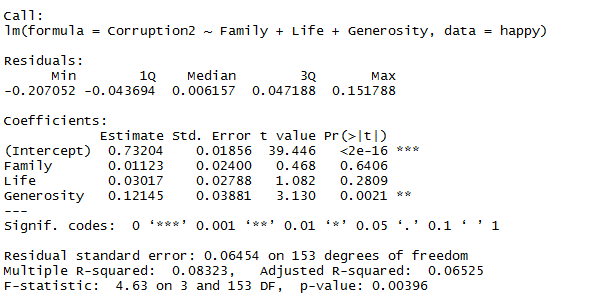
Running the test

* + 
  + Given that the p-value is significant, I can reject the null hypothesis that there is no difference between groups. Thus, I can infer that there is a difference in happiness based on 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**

Test to Use: Multiple Linear Regression (Corruption~Family+Life+Generosity)

Assumptions:

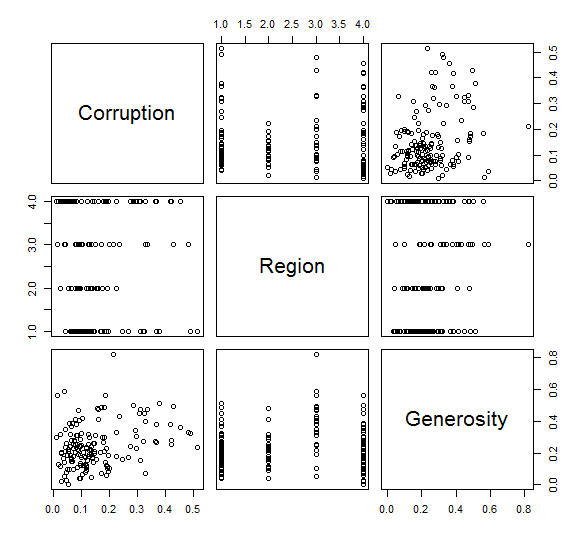
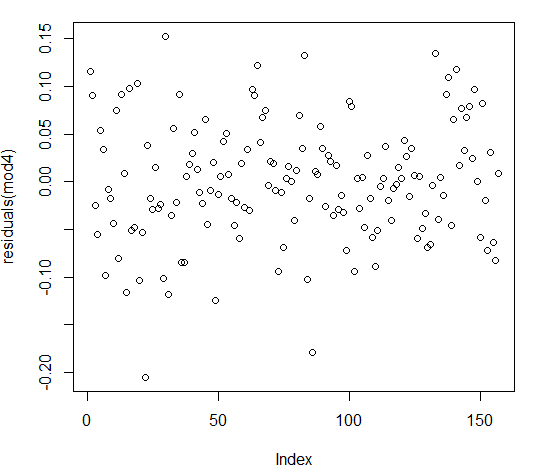
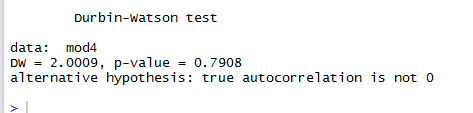
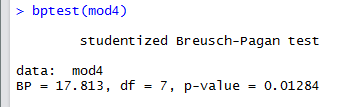
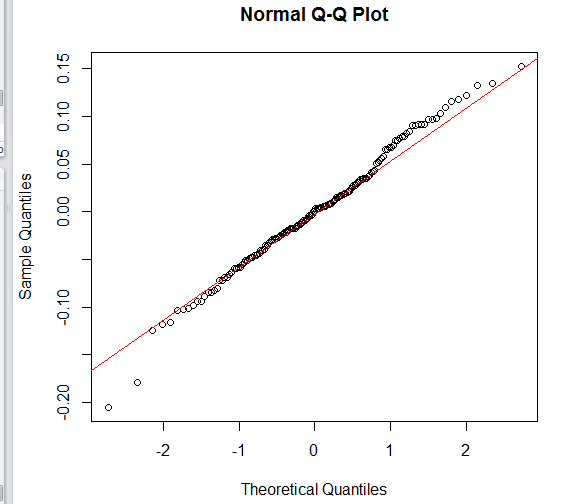
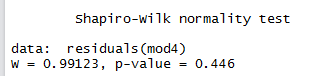
* Linear Relationship: PASSED
  + 
  + 
  + The relationships appear linearly correlated based on the pairs plot above (general positive linear relationship). Furthermore, the correlations between variables reveal that there is a linear relationship between the variables. Thus, the data passes this assumption.
* Independent Errors: FAILED
  + 
  + 
    - Significant p-value, thus reject the null hypothesis that there residuals are not autocorrelated.
* Homoscedasticity: FAILED
  + 
    - Significant p-value, thus reject the null hypothesis that the residuals are homoscedastic.
* Normal Errors: FAILED
  + 
  + 
    - Significant p-value, thus reject the null hypothesis that the population is normally distributed.
* Transform Data
  + mod1 = lm(log(Corruption)~Family+Life+Generosity, data=happy)
    - Re-check assumptions:
      * 
        + Now passes! Insignificant p-value
      * 
        + Now passes! Insignificant p-value
      * 
        + Still fails (significant p-value)
  + Try a different transform:
  + mod2= lm(sqrt(Corruption)~Family+Life+Generosity, data=happy)
    - 
      * Fails (significant p-value)
    - 
      * Fails (significant p-value)
    - 
      * Passes!!
  + Try a different transform:
    - Corruption2= (happy$Corruption)^(1/9)
  + Mod3 = lm(Corruption2~Family+Life+Generosity, data=happy)
    - 
      * Passes!
    - 
      * Failed- However, it is extremely close to passing (since the instructions said that “If your linear model ﬁt is slightly oﬀ even after a transform that’s okay – you can still use it but make it clear what model checking parameters your model still failed”, I will continue on with this model)
    - 
      * Passes!
* Running the Test
  + 
    - For every increase in the ninth root of Corruption, Generosity increases by 0.12145 (this interaction is significant)
    - Family and Life is not significantly influencing Corruption, there should be no significant increase in the ninth root of corruption with an increase in Family or Life.
* Checking Model Fit
  + R^2 is 0.065252, meaning that 6% of the variance in data is explained by my model, which is really low. Not the best model, probably due to all the omitted variables that influence corruption.
  + I checked Residual independence, normality, and homoscedasticity above, and the passed all besides homoscedasticity.

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**

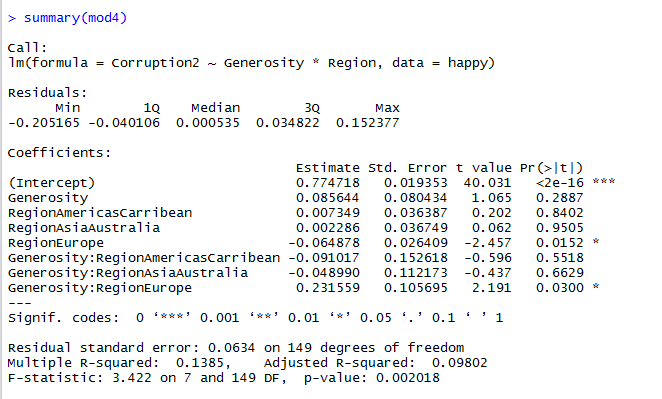
Test to Run: ANCOVA, using Generosity

Mod4 = lm(Corruption2~ Region \* Generosity, data= happy)

Assumptions:

1. Linear Relationship: PASSED (for the continuous variable)
   1. 
   2. The relationships appear linearly correlated based on the pairs plot above (general positive linear relationship). This is, of course, just for the continuous data, since the categorical data would not have a linear relationship.
2. Independent Errors: PASSED
   1. 
   2. 
      1. Not significant p-value, thus fail to reject the null hypothesis that there residuals are not autocorrelated.
3. Homoscedasticity: FAILED
   1. 
      * 1. Failed- However, it is extremely close to passing (since the instructions said that “If your linear model ﬁt is slightly oﬀ even after a transform that’s okay – you can still use it but make it clear what model checking parameters your model still failed”, I will continue on with this model)
4. Normal Errors: PASSED
   1. 
   2. 
      1. Not significant p-value, thus fail to reject the null hypothesis that the population is normally distributed. However, it is so close the insignificant that I am willing to accept the increased chance of Type 1 error.

Running the test:

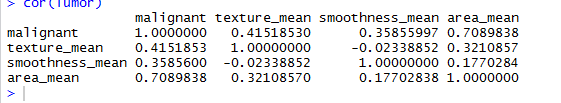
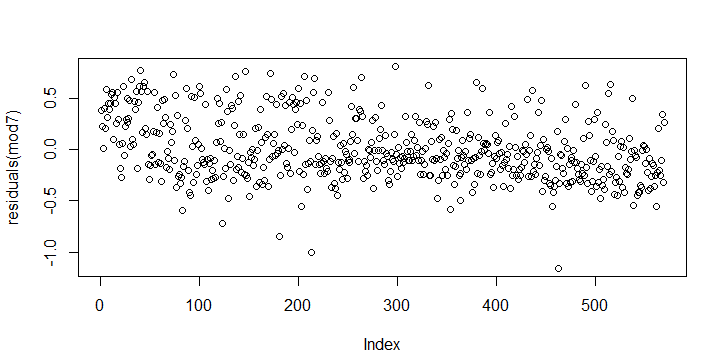
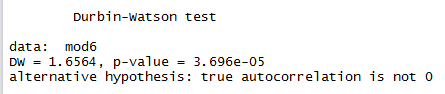
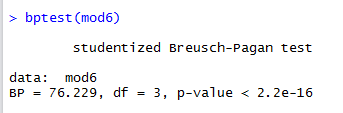
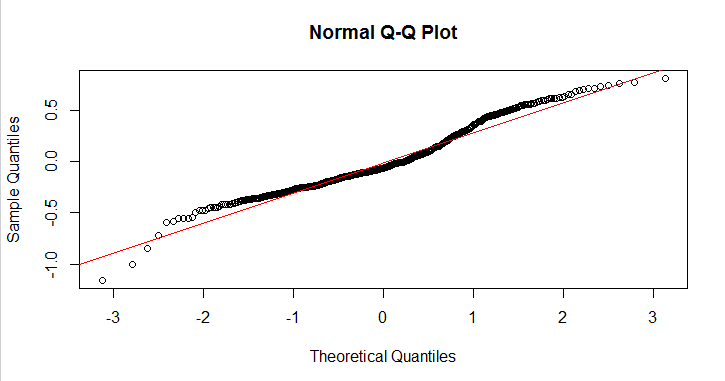
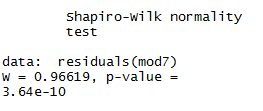
* 
  + Intercept (representing the Region AfricaMideast) is significant. This means that when the effect of Generosity is 0, the effect of Region AfricaMideast on Corruption is 0.774718.
  + The Region Europe is significant, meaning that for every unit of change in Region AfricaMideast, RegionEurope decreases by 0.064878.
  + Finally, the interaction between regionEurope and Generosity is significant. This implies that the slope for the effect of Generosity is significantly different for Region Europe and Region AfricaMideast.
  + In conclusions, the influence of Generosity on Corruption does vary by region. I would interpret this interaction by saying that the influence of Generosity on Corruption varies between Europe and AfricaMiseast.

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**

Test to use: GLM (Binomial Family)

mod6= glm(malignant~texture\_mean+smoothness\_mean+area\_mean, data=cancer, family=binomial(link=”logit))

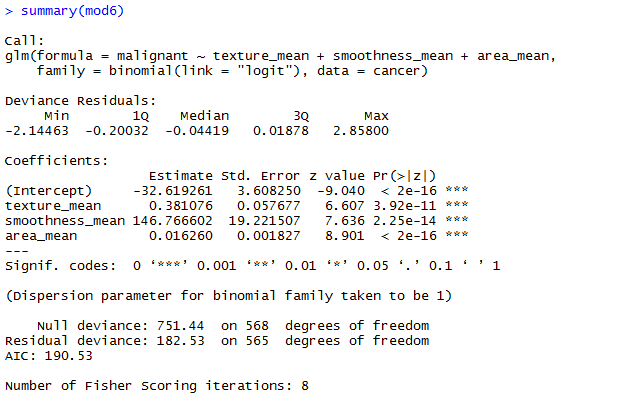
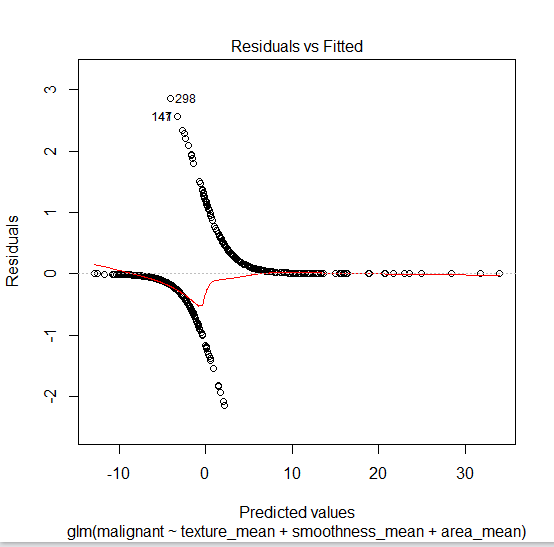
Assumptions:

* Linear Relationship: PASSED
  + 
  + The correlations between variables reveal that there is a linear relationship between the variables. Thus, the data passes this assumption.
* Independent Errors: FAILED
  + 
  + 
    - Significant p-value, thus reject the null hypothesis that there residuals are not autocorrelated.
* Homoscedasticity: FAILED
  + 
    - Significant p-value, thus reject the null hypothesis that the residuals are homoscedastic.
* Normal Errors: FAILED
  + 
  + 
    - Significant p-value, thus reject the null hypothesis that the population is normally distributed.
* Since Homoscedasticity, normality, and independent errors failed, I need to use a GLM. The family would be Binomial (malignant = 1 or 0)

Transform data

* From lm(malignant~texture\_mean+smoothness\_mean+area\_mean, data=cancer)
  + To:
* mod6=glm(malignant~texture\_mean+smoothness\_mean+area\_mean, data=cancer, family=binomial(link='logit'))

Running the test:

* 
  + For every increase in texture\_mean, the log odds of a malignant tumor increases by 0.381076.
  + For every increase in Smoothness\_mean, the log odds of a malignant tumor increase by 146.766602.
  + For every increase in area\_mean, the log odds of a malignant tumor increases by 0.016260
  + Thus, texture, smoothness and area are all significantly associated with if a tumor is malignant or not.
* Check model fit:
  + 
  + And look the deviance. Relatively good model fit (high deviances).

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.**

* Smoothness is most important, followed by texture followed by area.