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=",")

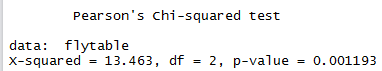
cancer = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/cancer.csv",header=TRUE, sep=",")

**EMBER McCOY – embermcc**

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**
   1. Independent = categorical, dependent = categorical;

**Chi-square test for independence**

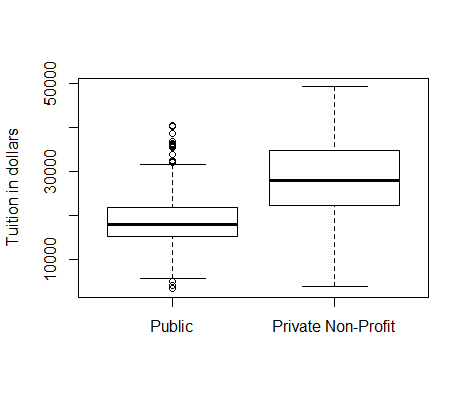
* 1. visual
  2. **Assumptions:**
     1. Independence of observations – yes
     2. No structural zeros – yes
  3. Run test, output:

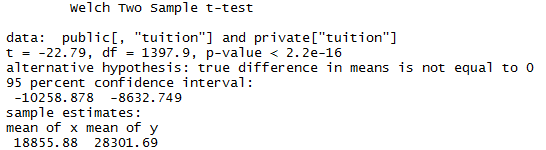


**Results:** P-value = 0.001193 (less than .05), so there is a significant association between gender and whether people think it’ s rude to bring an unruly child on the plane.

* 1. To find which gender tends to think that bringing an unruly child on the plane is more rude, I found the proportion of each gender that found that bringing an unruly child on the plane was “very” rude. Results (below) showed that 47% of males thought that bringing an unruly child on the plane was “very” rude, while only 36% of females thought that bringing an unruly child on the plane was “very” rude.

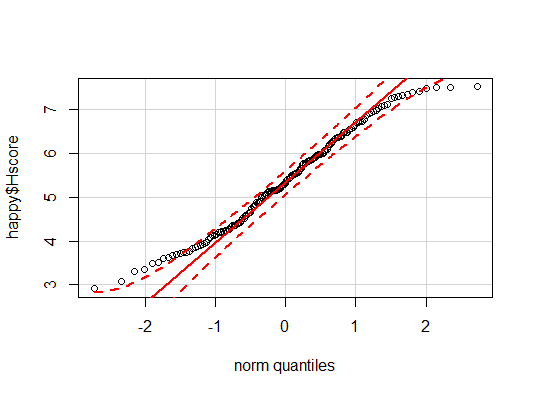


1. Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition? **College**
   1. independent = categorical, dependent = continuous; **Two-Sample** **T-Test**
   2. Visual
   3. **Assumptions:**
      1. Independent, random sample – yes, so unpaired t-test
      2. Continuous dependent variable – yes
      3. Normality of Dependent Variable– I tested for normality of “tuition” using the Shapiro-Wilk normality test, and had a resulting p-value = 3.193e-15 (less than .05), so the distribution is not normal. However, the sample size is >30 (1407 observations), so normality can be assumed with the large sample size.
      4. Equal variances – I used var.test to test equal variance between tuition for public and private non-profit universities. The results were p-value < 2.2e-16 (less than .05), so equal variances cannot be assumed and we must use the **Welch’s Two Sample T-Test**
   4. Run Test, Output:

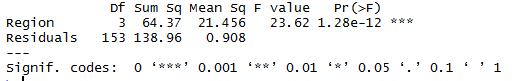


**Results:** P-value < 2.2e-16 (less than .05), so we can reject the null hypothesis and say that there is a significant difference between tuition at public and private non-profit universities. Specifically, the mean tuition for private non-profit universities ($28,301.69) is higher than the mean tuition for public universities ($18,855.88).

1. Is there a significant difference in happiness (Hscore) by region (Region)?**happy**
   1. Independent = categorical, Dependent = continuous; **One-way ANOVA**
   2. Visual
   3. **Assumptions:**
      1. Independent random sample – yes
      2. Normality of Dependent Variable – I tested for normality of Happiness score by using the Shapiro-Wilk normality test. The p-value = 0.01248 (less than .05), so the distribution is not normal. However, the QQPlot looks close to normal and the sample size is 157 (>30) with at least 15 observations in each of the 4 groups, so I am going to assume normality and continue with the test.



* + 1. Equal variances – I used Leven’s test to test equal variances for happiness score between world regions. The results were p-value = 0.5179 (greater than .05), so equal variances can be assumed.
  1. Run test, Output:

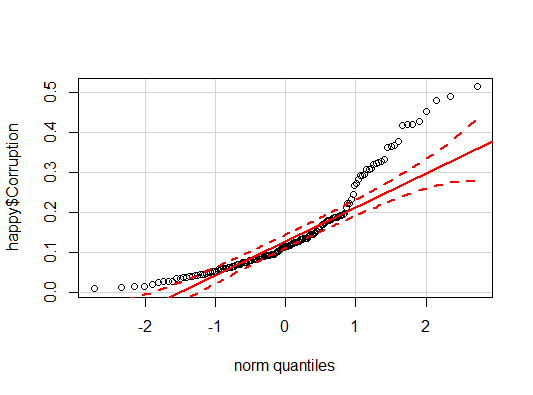
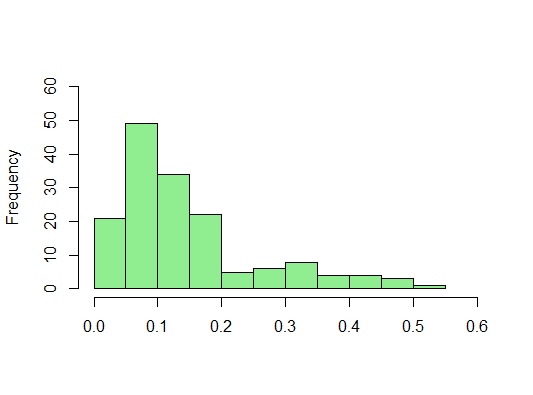


**Results:** P-value = 1.28e-12 (less than .05), therefore there is a significant difference between Region and happiness score. If we wanted to find *which* region is the happiest based on happiness score, we would next run a post-hoc test, such as Tukey HSD.

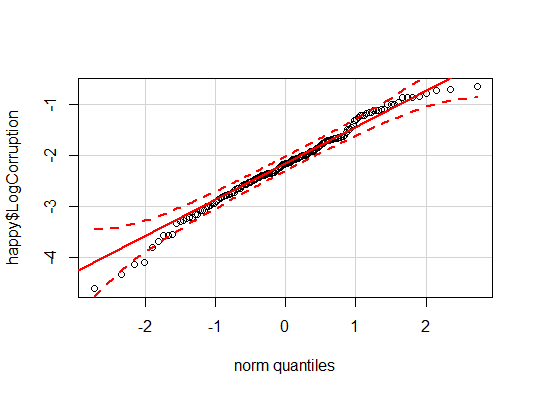
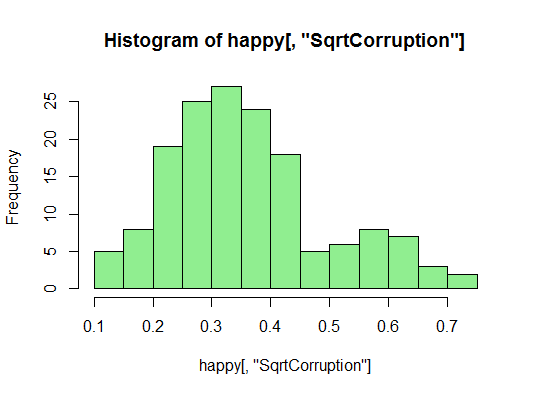
1. What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model. **Happy**
   1. Multiple independent = continuous, dependent = continuous;

**Multiple Linear Regression**

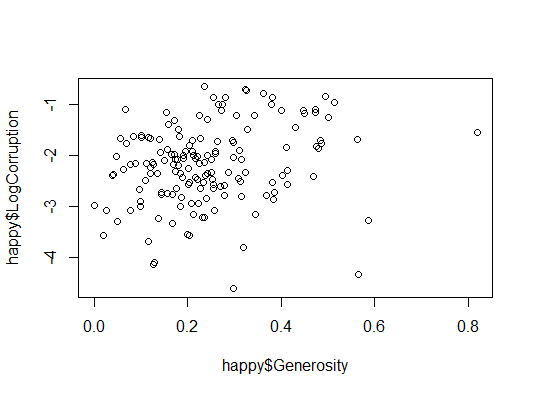
* 1. Three independent variables: GDP, Freedom, Generosity. I chose these variables because they were the least correlated with each other and I didn’t want multi-collinearity issues. All correlations were less than .5 and all VIF values were less than 10, which means I don’t have issues with multi-collinearity and can use these three variables in the model.
     1. GDP to Freedom correlation value = 0.3622828
     2. GDP to Generosity correlation value = -0.02553066
     3. Generosity to Freedom correlation value = 0.3617513
     4. VIF GDP = 1.189712
     5. VIF Freedom = 1.367952
     6. VIF Generosity = 1.189185
  2. **Assumptions:**
     1. FIRST: Hypothesis Testing– Normality of Dependent Variable - Checked histogram & QQPLOT & completed the Shapiro-Wilk Test on Corruption. The results of the Shapiro-Wilk test show that p-value = 5.128e-11 (less than .05), so the distribution is not normal. Nor does it look normal in the QQ plot or histogram, therefore must transform the dependent variable (Corruption).

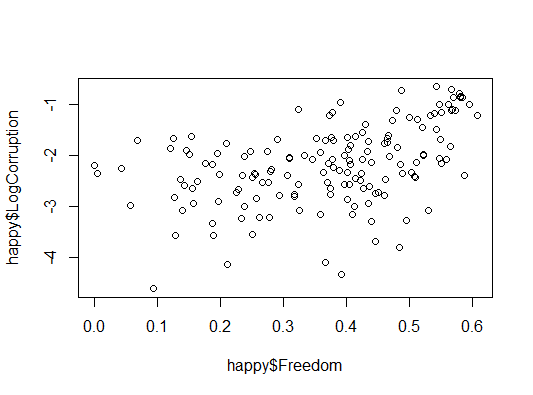
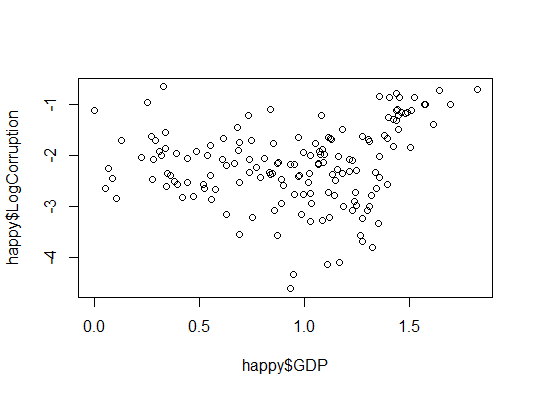


* 1. TRANSFORM:
     1. Log transformation on Corruption. Then ran the QQ plot, histogram, and Shapiro-Wilk test for the new log-transformed Corruption variable. Both the QQ plot and histogram appear to be more normal, and the Shapiro-Wilk test had a resulting p-value of 0.08535 (greater than .05) confirming that the new distribution is normal (barely).

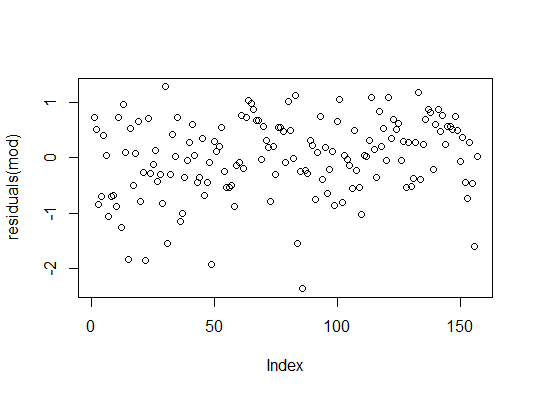
 

* 1. **Assumptions, continued:**
     1. Normality of (Transformed) Dependent Variable – yes, see d.
     2. Linear relationship – assuming yes

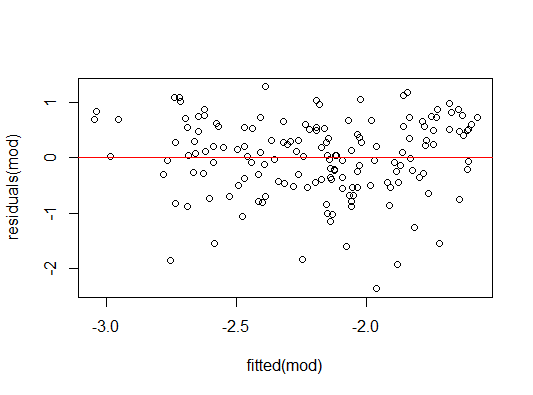




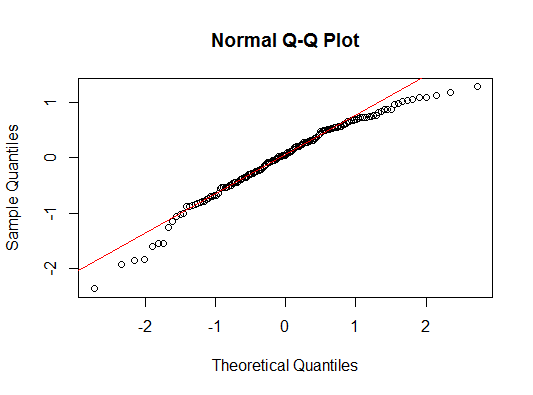
* + 1. Residual independency – I ran the Durbin-Watson statistic test to determine autocorrelation and looked at the plot for the residuals. It does not appear that there is a trend and the DW test resulted in a p-value of 0.185 (greater than .05), which means that there is not an autocorrelation and residuals independency can be assumed.



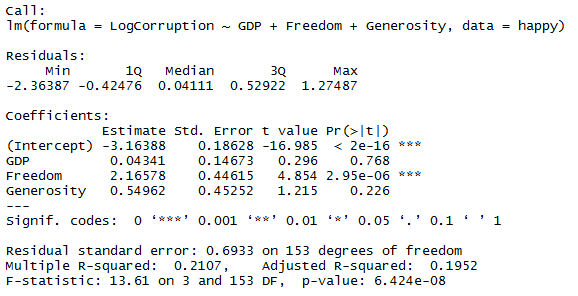
* + 1. Residual homoscedasticity – I ran the studentized Breusch-Pagan test to test residual heteroscedasticity. It appears from the plot that the residuals might not be homoscedastic, because they are more below the line than above, but the BP test resulted in a p-value of 0.06878 (greater than .05), which means that the residuals are NOT heteroscedastic, therefore we can assume residuals are homoscedastic.



* + 1. Residual normality – I ran the Shapiro-Wilk test to test for normality of residuals and looked at the plot. It appears from the plot that residuals are close to normal, but from the Shapiro-Wilk test the p-value = 0.0006382 (less than .05), which means that the residuals are NOT normally distributed. They look close on the QQ plot, so I am going to move on, but you could do a different log transform or a GLM here if you wanted to be completely accurate

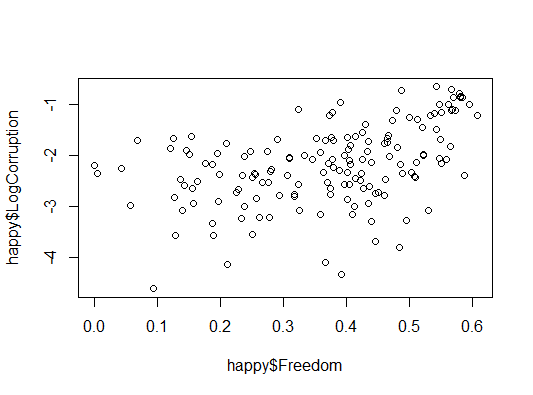


* 1. Run lm, Output:

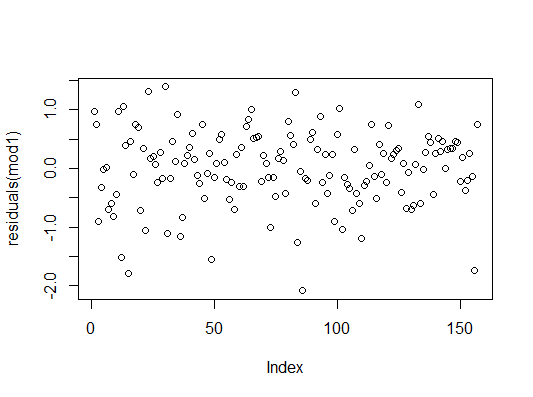


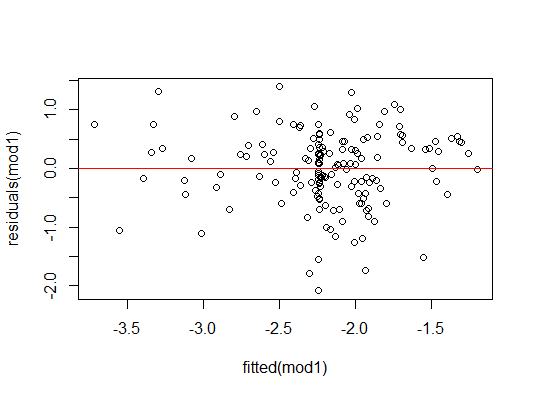
**Results:** The intercept coefficient is significant at the .05 level, therefore we can say if all other factors are 0, (log) corruption would be -3.16388 units. The only factor in the model that are significant at the .05 level is Freedom. This significant coefficient of 2.16578 means that freedom does effect corruption: for every 1-unit increase in Freedom, Log Corruption increases by 2.16578 units. The Adjusted R-squared value is 0.1952, which tells us how much variance can be explained in the model. In this case, the model explains only 19.52% of the variance so I would say this is model is a bad fit, and there are likely other factors that impact corruption in countries.

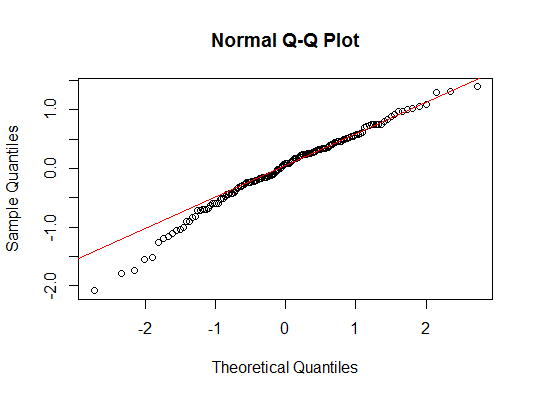
1. Choose one of the continuous independent variables that was significant in the model for Question 4 and interact it with region (Region) to predict corruption (Corruption). This model should only include one continuous independent variable and its interaction with region. Does the influence of your continuous variable on corruption vary by region? If yes, how do you interpret the interaction? **Happy**
   1. Independent=continuous & categorical, Dependent=continuous; **ANCOVA**
   2. **Assumptions**:
      1. Normality of Dependent Variable – from Question #4 I know that the Corruption variable is not normal, but the Log-transformed Corruption variable is, so I will use the Log-transformed Corruption variable again and normality is assumed (see 4d)
      2. Independent random sample – yes
      3. Linear relationship (for continuous variable) – assuming yes



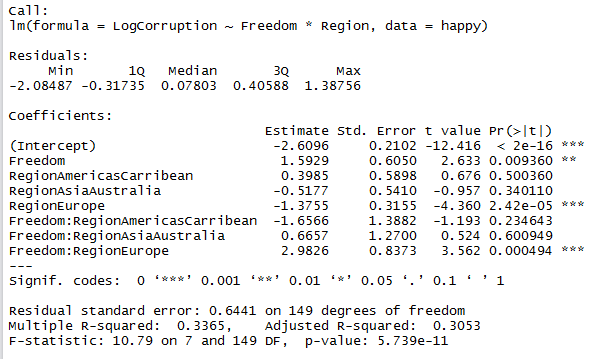
* + 1. Residual independency – I ran the Durbin-Watson statistic test to determine autocorrelation and looked at the plot for the residuals. It does not appear that there is a trend and the DW test resulted in a p-value of 0.7187 (greater than .05), which means that there is not an autocorrelation and residuals independency can be assumed.



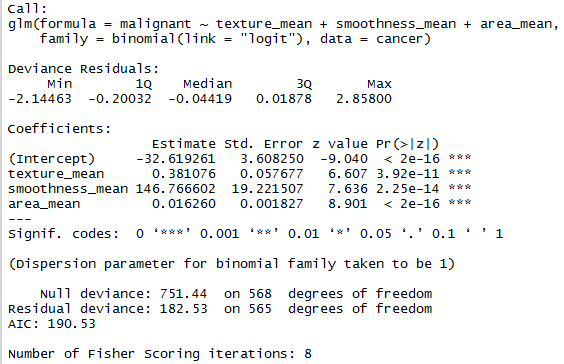
* + 1. Residual homoscedasticity – I ran the studentized Breusch-Pagan test to test residual heteroscedasticity. It appears from the plot that the residuals might not be homoscedastic, but the BP test resulted in a p-value of 0.2501 (greater than .05), which means that the residuals are NOT heteroscedastic, therefore we can assume residuals are homoscedastic.
    2. Residual normality – I ran the Shapiro-Wilk test to test for normality of residuals and looked at the plot. It appears from the plot that residuals are close to normal, but from the Shapiro-Wilk test the p-value = 0.01481 (less than .05), which means that the residuals are NOT normally distributed. They look close on the QQ plot, so I am going to move on, but you could do a different log transform or a GLM here if you wanted to be completely accurate



* 1. Run lm, Output:



**Results:** The influence of Freedom on the log of corruption does vary by region (\*from now on I will say corruption instead of log of corruption). The Freedom coefficient and significance mean that the effect of freedom on corruption is significant at the .05 level for the Africa/Mideast region. For every 1-unit increase in freedom, corruption in the Africa/Mideast region increases by 1.5929 units. The Freedom:RegionEurope coefficient and significance mean that the effect of freedom on corruption is significantly different at the .05 level for Europe than Africa/Mideast. The slope (or effect of freedom on corruption) for Europe is 1.5929 plus 2.9826. The interaction between freedom and the regions Americas/Caribbean and Asia/Australia do not have significant coefficients at the .05 level, which means that the slope (or effect of freedom on corruption) is not significantly different for those regions than for Africa/Mideast. If a graph, I would imagine the lines for Africa/Mideast, Americas/Caribbean, and Asia/Australia are the same, while there is a separate line for the Europe region.

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**
   1. Independent = 3 continuous, Dependent – binominal; **GLM**
   2. Three independent variables: Texture, Smoothness, Area. All correlations were less than .5 and all VIF values were less than 10, which means I don’t have issues with multi-collinearity and can use these three variables.
      1. Texture to Smoothness correlation value = -0.02338852
      2. Texture to Area correlation value = 0.3210857
      3. Smoothness to Area correlation value = 0.1770284
      4. VIF Texture = 1.123269; VIF Smoothness = 1.040059; VIF Area = 1.158975
   3. **Assumptions:**
      1. Normality of Dependent Variable – binomial GLM so N/A
      2. Independent observations – yes
      3. GLM does not assume residual independency, normality, or homoscedasticity, so I am not going to run them, but you still “hope” for that according to Meha’s lecture so you could still run them if you want to be extra sure.
   4. Run GLM, Output:

**Results:**Area, texture, and smoothness all are significant factors associated with whether or not a tumor is malignant (at the .05 level). The significant intercept means that even when all factors are 0, there is still a probability for a tumor to be cancerous. When all other factors are 0, if texture increases by 1 unit, the log odds ratio of the tumor being malignant increases by 0.381076 units; if smoothness increases by 1 unit, the log odds ratio of the tumor being malignant increases by 146.767 units; if area increases by 1 unit, the log odds ratio of the tumor being malignant increases by 0.016260 units.

* 1. Model checking: I ran a model without the factors and compared the AIC. The AIC for the model with these factors was smaller than without the factors, so my model is a good fit comparatively.

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.
   1. In order to do this properly, you would have to standardized all of the variables by creating new variables that standardize all of the different units. Because the units are likely different for smoothness, texture, and area the results of the GLM may look like one of those variables as a much larger effect than the others, when in reality that is just a reflection of the size of unit for that variable.