Sophia Paul

NRE 538

Take home final

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=",")

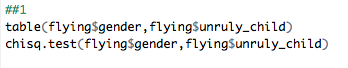
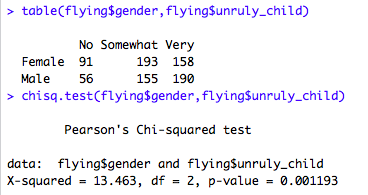
college = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/college.csv",header=TRUE, sep=",")

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**

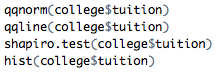
I decided to run a chi square test for independence. The observations are independent (as far as we know), there are no structural zeros, and it appears the cells all have reasonably sized values. The p-value is 0.00119 so there is a difference between men and women in whether they think it is rude to bring an unruly child on a flight. Men are more likely to think that flying with an unruly child is very rude and women are more likely to think it is not rude.

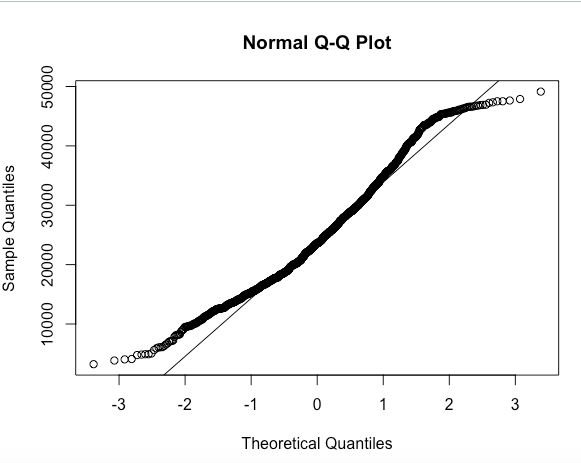


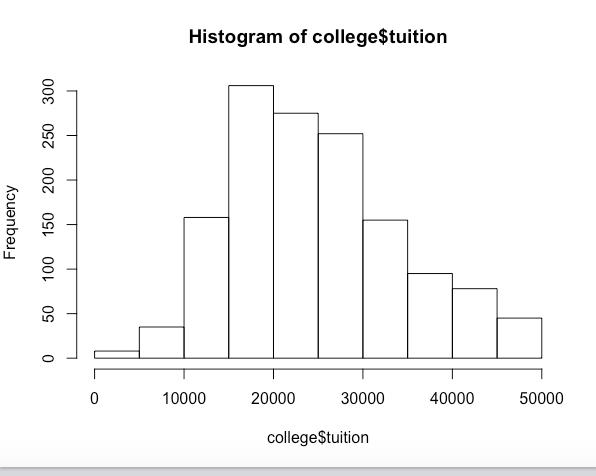
1. Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition? **college**

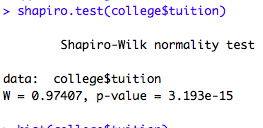
I decided to run a t test. The data are independent and appear continuous. They also seem to be a random sample. I tested for normality, and the tuition data was not normally distributed. A square root transformation helped somewhat and, additionally, the sample size is way over 30 so I can relax that assumption. Next I tested for equal variance (with the transformed data) and the p value was extremely small so variances are not equal so I indicated that in my Welch’s t-test.

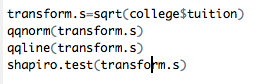
When I ran the t test, the p value was extremely small so I could conclude that there is a significant difference in tuition between public and private nonprofit institutions. Private nonprofit institutions have higher tuition.

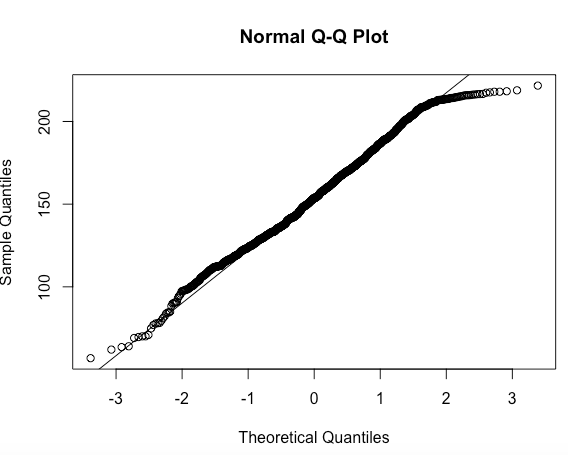


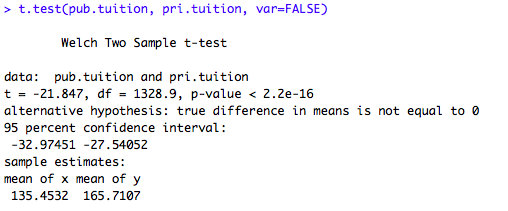
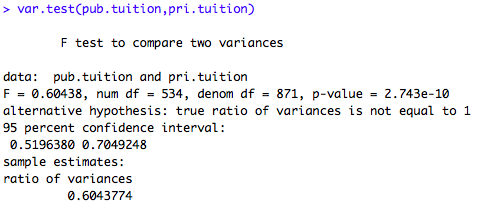
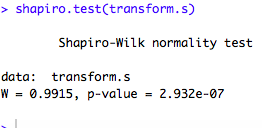
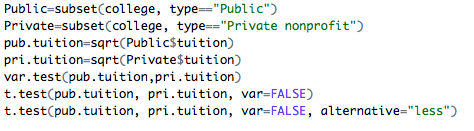


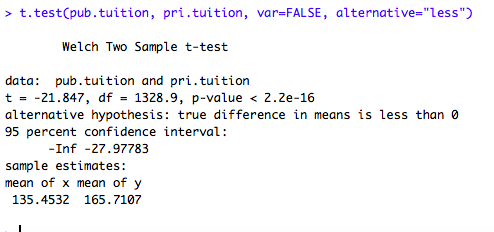








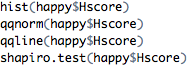


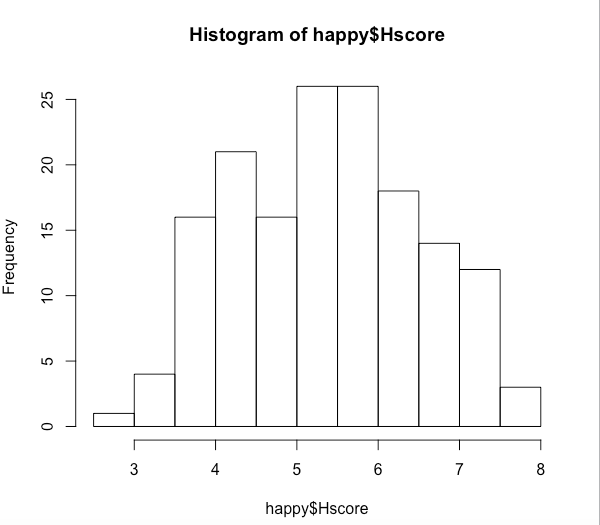


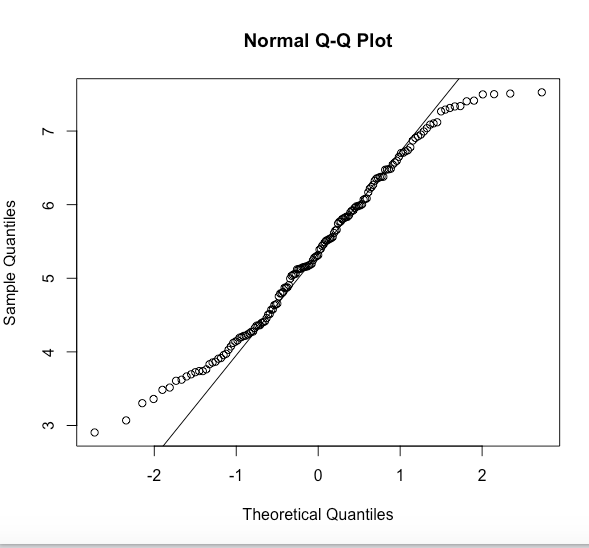
1. Is there a significant difference in happiness (Hscore) by region (Region)? **happy**

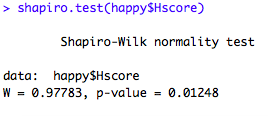
For this test I will run an ANOVA. The samples are independent. I tested for normality of Hscore. Values were not normally distributed, and simple transformations only made them worse. However, given that the sample is fairly large, that assumption can be relaxed anyway. A levene test revealed that the variance was fairly equal.

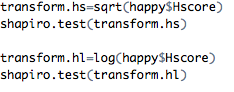
When I ran the ANOVA the p value was less than 0.05 indicating that happiness score does vary significantly by region.

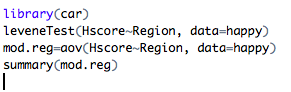
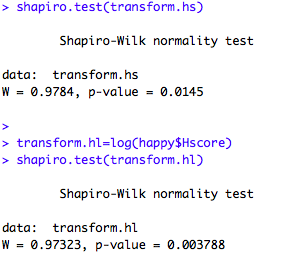


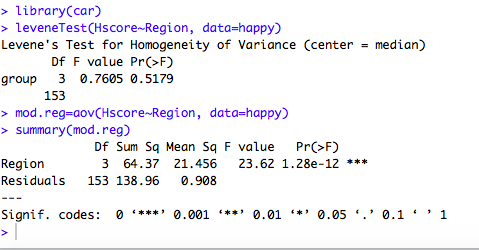












1. What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model. **happy**

I first tested to see if corruption was normally distributed and it was not. However, a log transformation made the corruption value normal enough. I decided to include GDP, freedom, and generosity in my model and tested to see if they were highly correlated with one another and they are not. Additionally all of the VIF scores are between 1 and 2, so they are well below 10 so I am not worried that these two variables are too correlated.

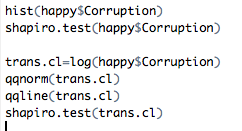
Plotting each of these variables with transformed corruption did not yield nonlinear looking relationships. The relationships did not look that linear either, they mostly appeared random

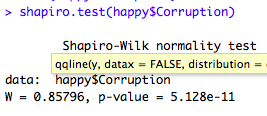
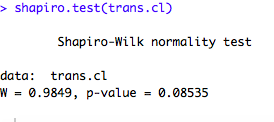
When I ran the model, the intercept was significant (so the transformed corruption is significantly different than zero). GDP and generosity had no significant relationship with corruption so corruption does not change based on those factors. Freedom, however, was significant. Since corruption was transformed it is difficult to say exactly how the relationship operates though as transformed corruption increases by one freedom increases by 0.37. Additionally, the r squared is around .25 indicating that this model only accounts for about 25% of the variance in corruption.

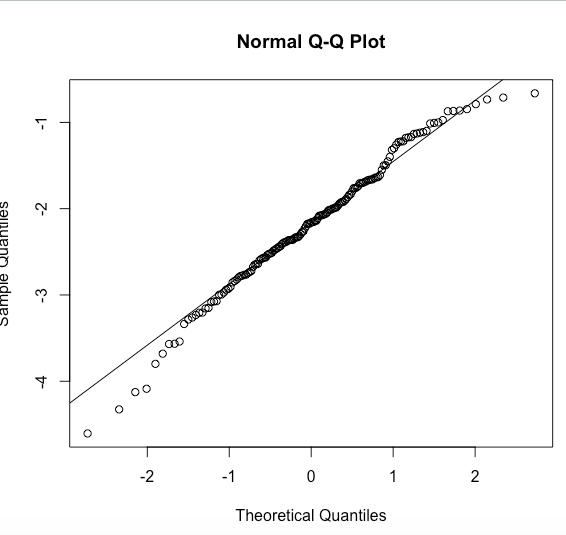
Next I tested the model. I plotted the residuals and they appeared to be scattered randomly. I ran a Durbin Watson test which yielded a p value of 0.1442, suggesting that the residuals are independent and there is no autocorrelation.

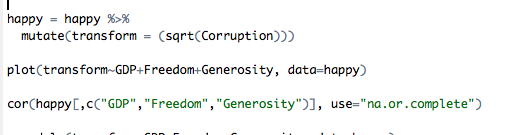
Next I tested to see if the residuals were normal. First I created a qqplot where they appeared fairly normal. A shapiro test confirmed this with a p value of 0.73.

Next I tested for homoscedasticity. On a plot, the residuals to the right appeared slightly more spread out suggesting that the residuals may not be homoscedastic. A bp test yielded a p value of 0.01 suggesting that there is a pattern to the errors. This indicates that while the coefficient is accurate the standard error and p value may not be. A weighted least squares could possibly improve the model.

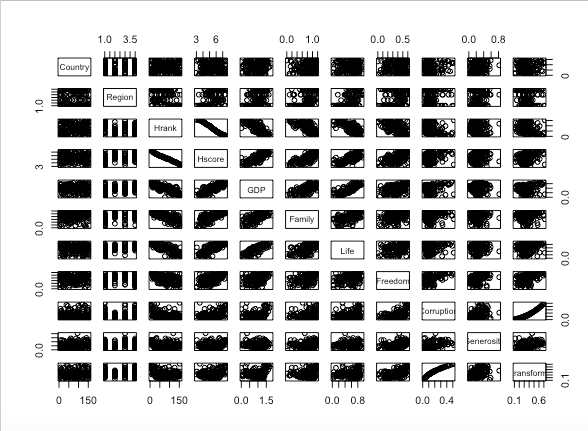


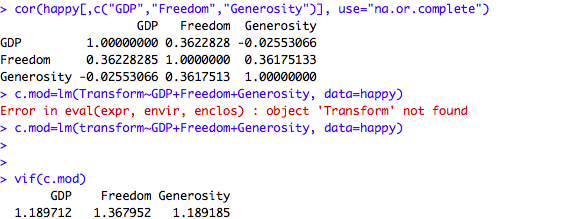


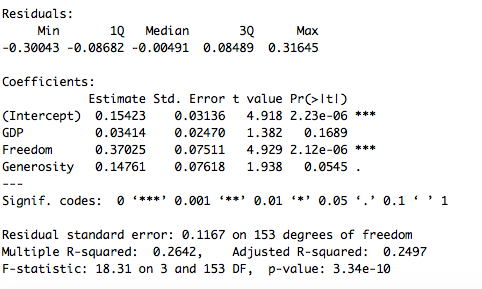


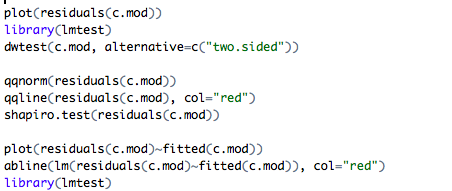


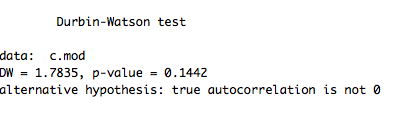
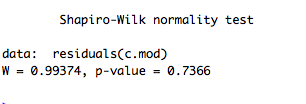
../Desktop/Screen%20Shot%202017-04-20%20at%209.27.29%20PM.png







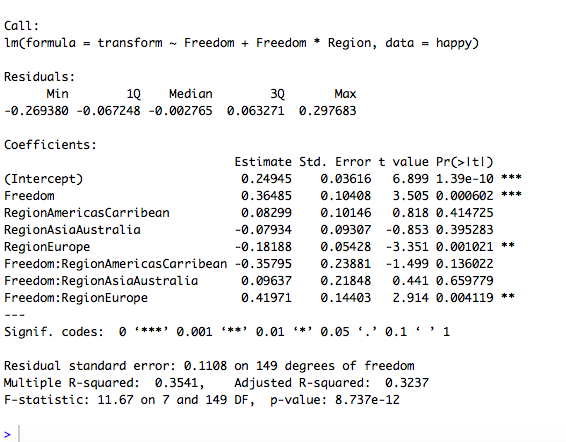


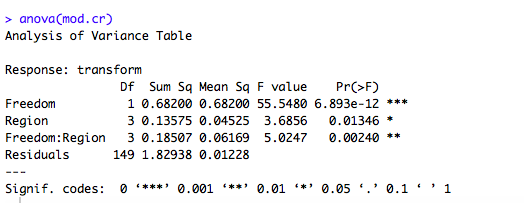


1. Choose one of the continuous independent variables that was significant in the model for Question 4 and interact it with region (Region) to predict corruption (Corruption). This model should only include one continuous independent variable and its interaction with region. Does the influence of your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

I chose to use freedom since it was the only significant variable in my last model. I will, again, use the transformed corruption data. In this model the r squared .35 suggesting it explained 35% of the variance in corruption. Since AfricaMideast was first alphabetically of the regions it is the intercept, when freedom is zero. In this model, freedom is significant and the corruption in Europe is significantly different than the corruption in AfricaMideast, it has a separate intercept. Region overall is not significant.

The interaction term indicates whether the effect of freedom on corruption is different across different regions. It is, and it is a significantly different effect in Europe, but not in other regions. The impact in Europe is less that in other regions.

../Desktop/Screen%20Shot%202017-04-20%20at%208.58.04%20PM.png



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**

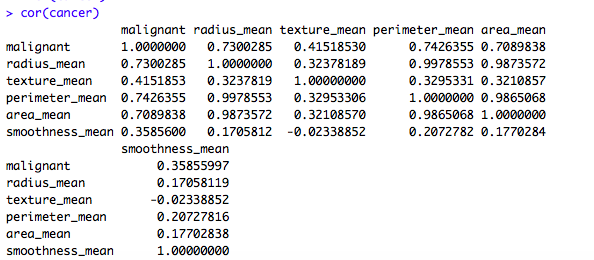
Since malignant is a yes or no variable, I should run a glm with a binomial family. I decided to include radius, texture, and smoothness in my model and a correlation test revealed they were not too highly correlated with each other to be included in a model. A VIF also revealed values all under 3, so I am not worried about them being too highly correlated.

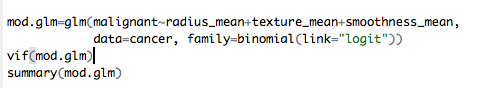
Next, I ran the model. The results indicated that all three factors (radius, texture, and smoothness) were significantly associated with whether or not a tumor was malignant. For every increase in radius the odds of malignancy increase by 1.39. For every increase in texture (however that is quantified) odds of malignancy increase by 0.38. For every increase in smoothness (again, however that is quantified), odds of malignancy increase by 144.67. The intercept is negative and significant, indicating the odds of malignancy when radius, smoothness, and texture are all set to zero (which doesn’t make logical sense). All of these are somewhat transformed due to the GLM, but the important thing to note is that they all do have a significant impact on probability.

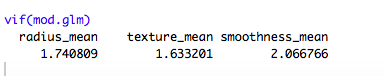
Next I used some of R’s plots to check the model.

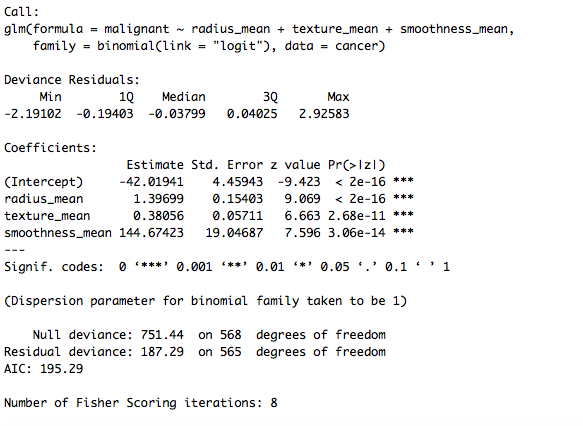
I also created a null model and used anova, Chi test & AIC to compare it to the model I had created. I saw that the model I had created was highly unlikely to produce the same deviance as the null model and including the variables I had included significantly increased the likelihood of the model.

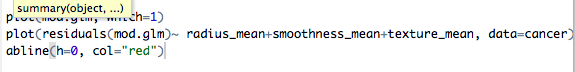
../Desktop/Screen%20Shot%202017-04-20%20at%209.24.57%20PM.png

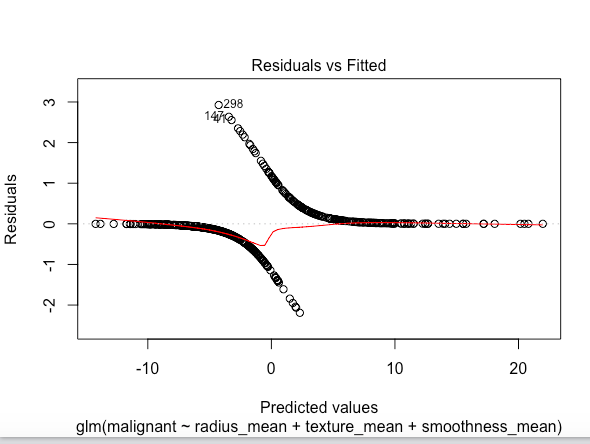


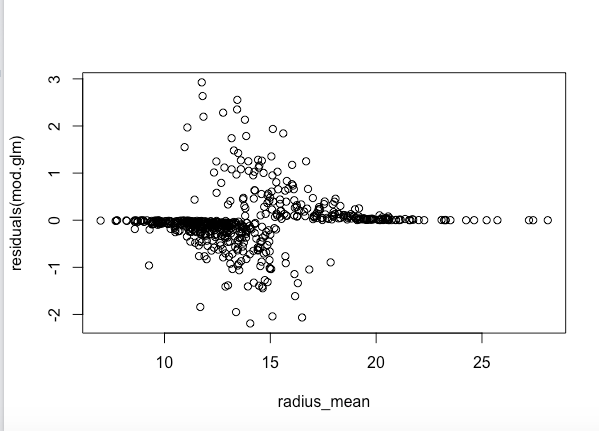


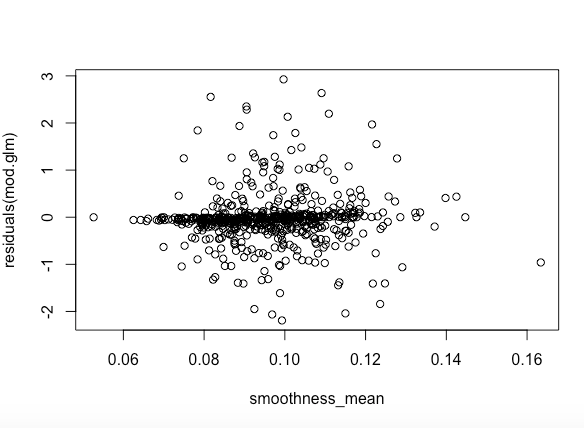


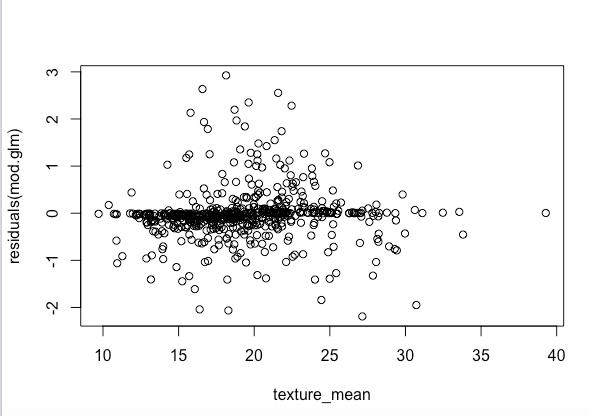


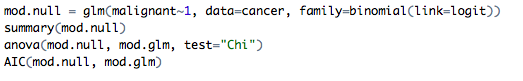


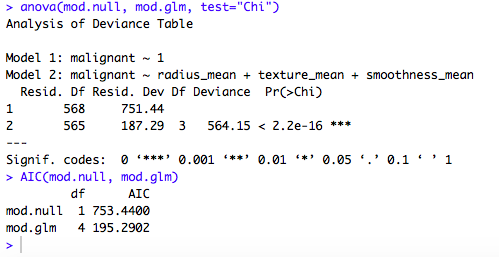












1. BONUS/EXTRA CREDIT: Which independent variables are the most important in explaining whether a breast cancer tumor is malignant or not? Use the same 3 continuous independent variables you chose for question 6. **cancer.**

I created three glms that each only included one of the variables and used a Chi square test in an anova to compare them to the null model. Texture was the only significant factor (meaning that model was the only one significantly different than the null), so it is most important.

