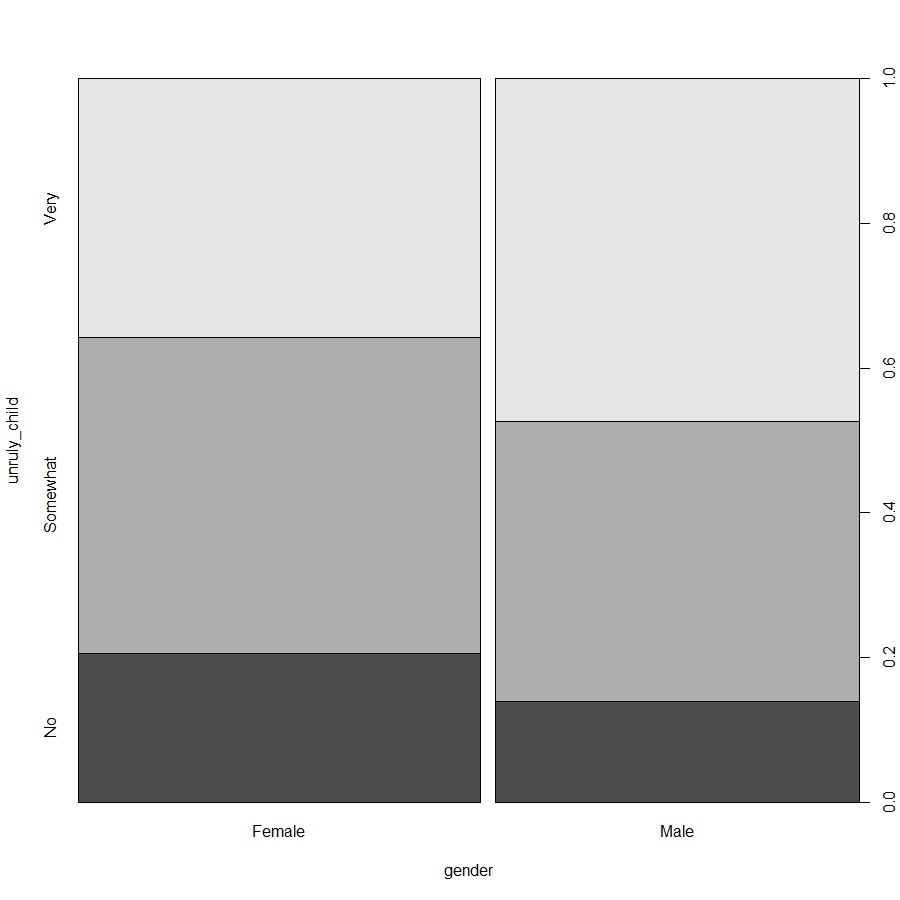
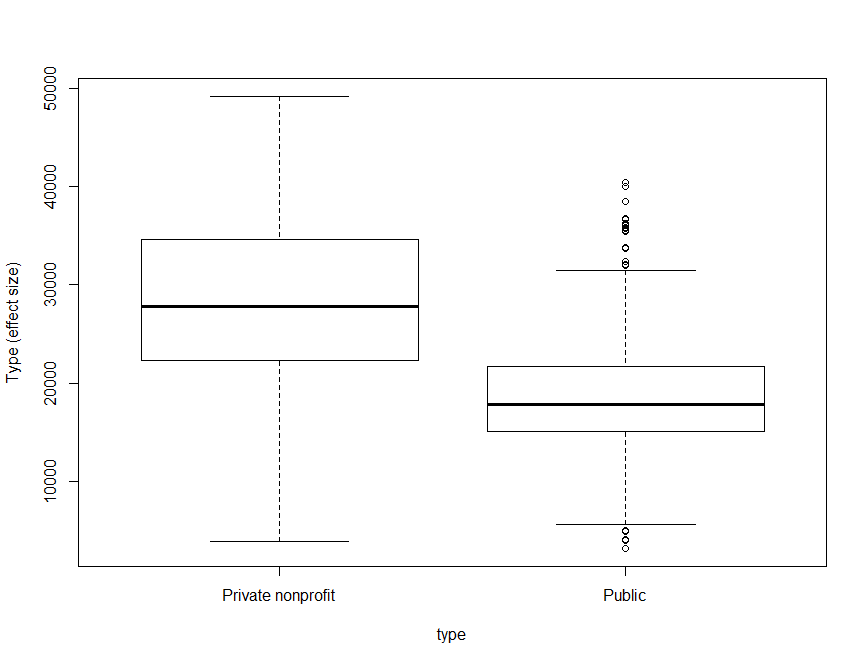
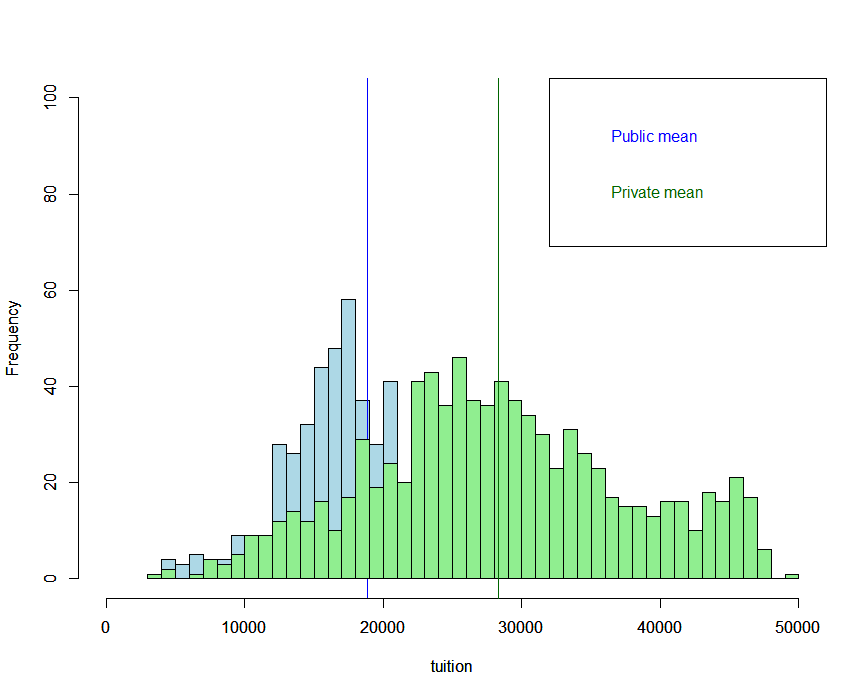
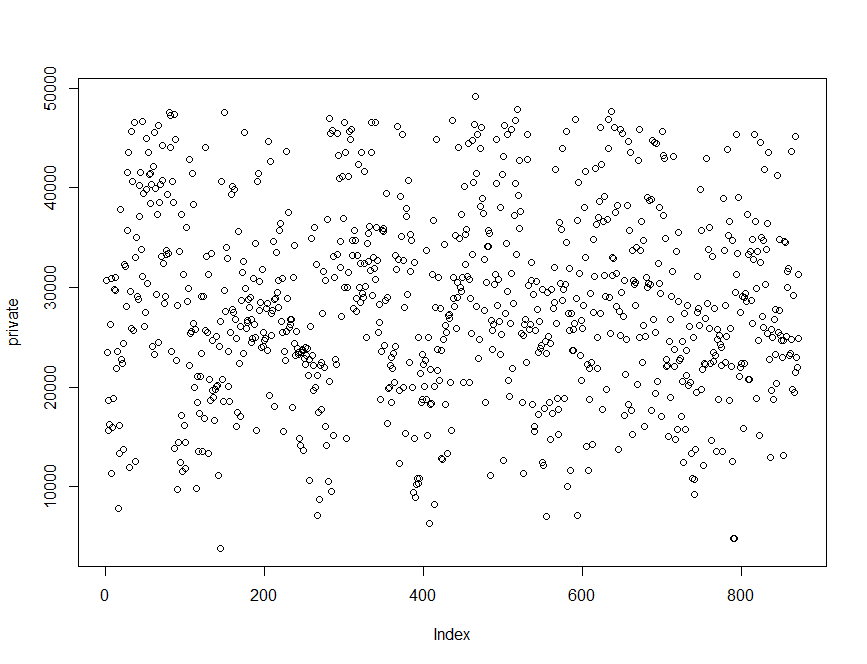
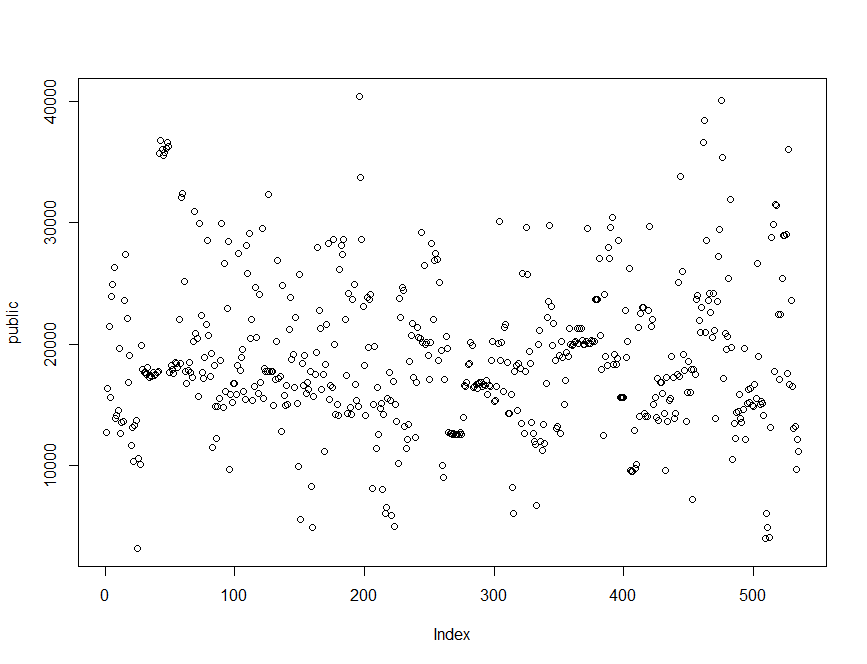
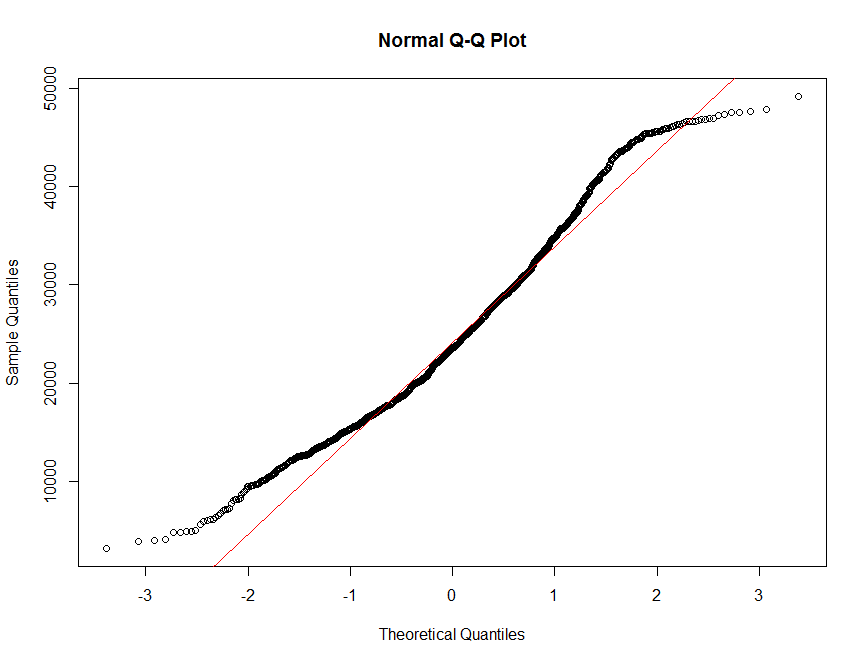
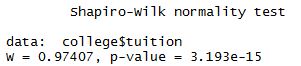
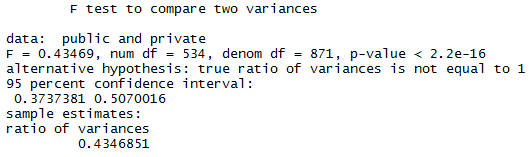
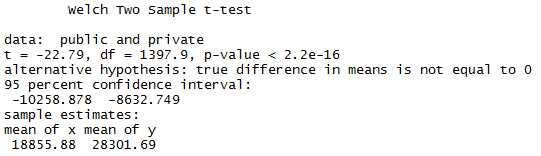
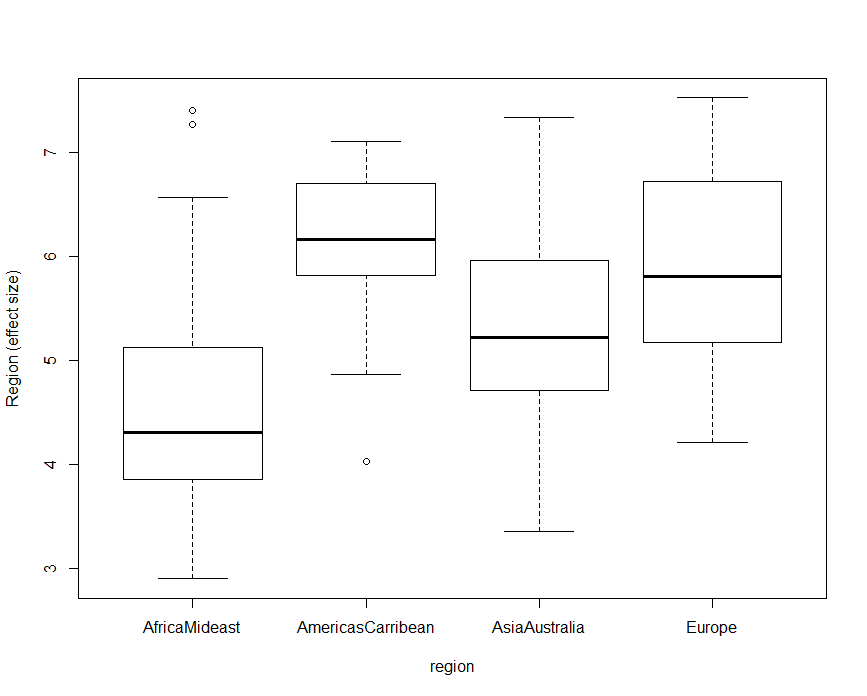
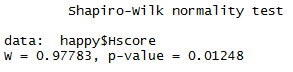
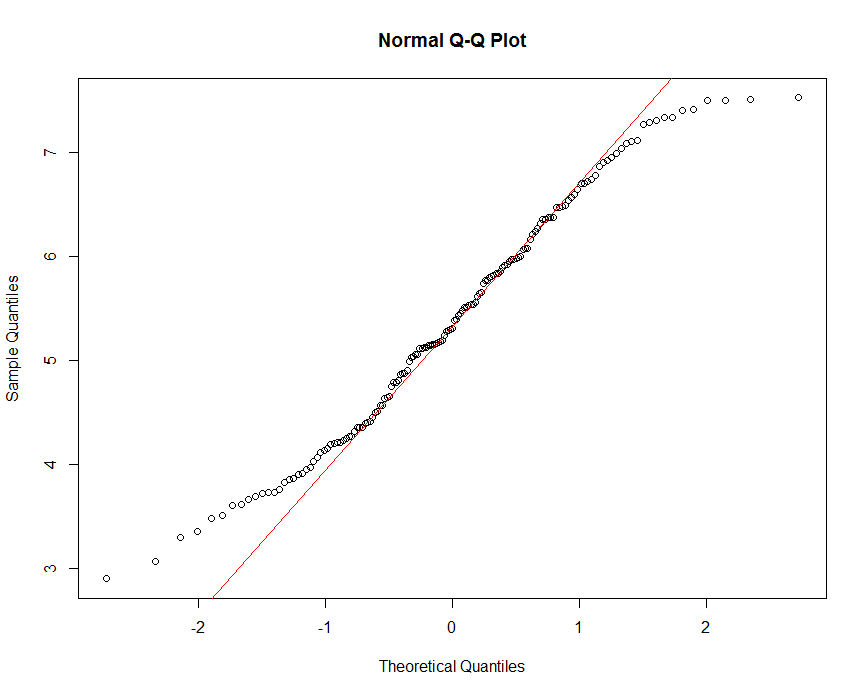
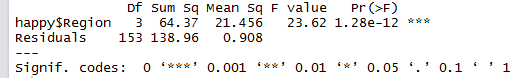
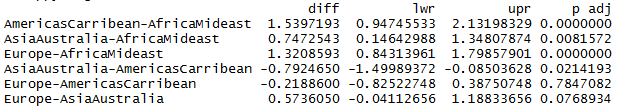
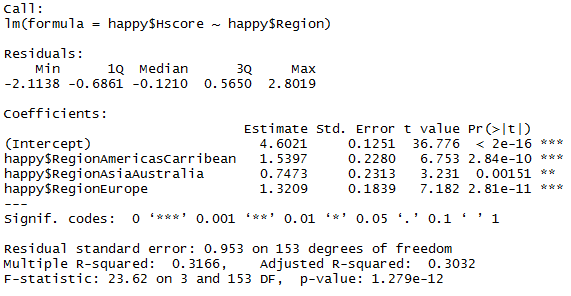
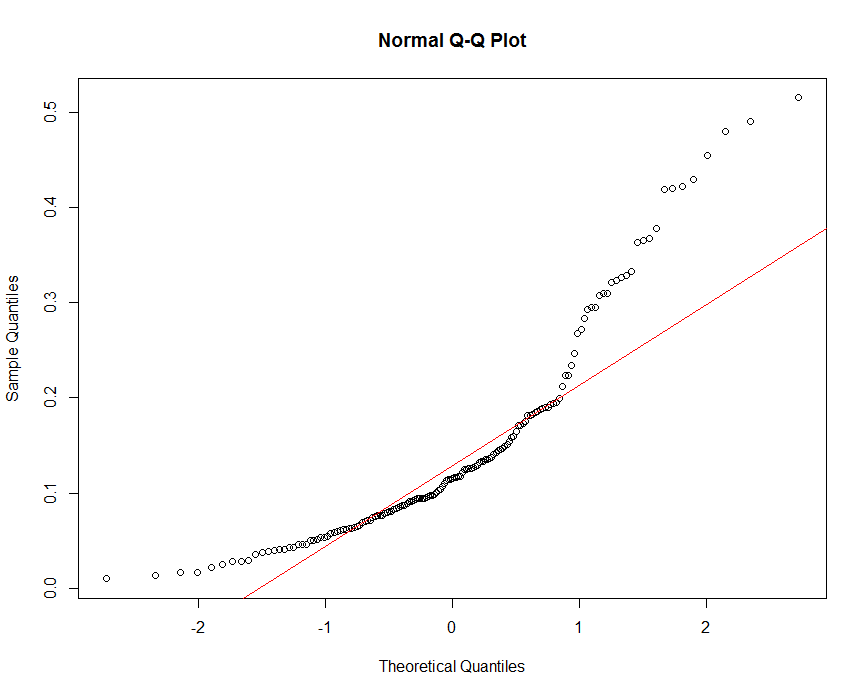
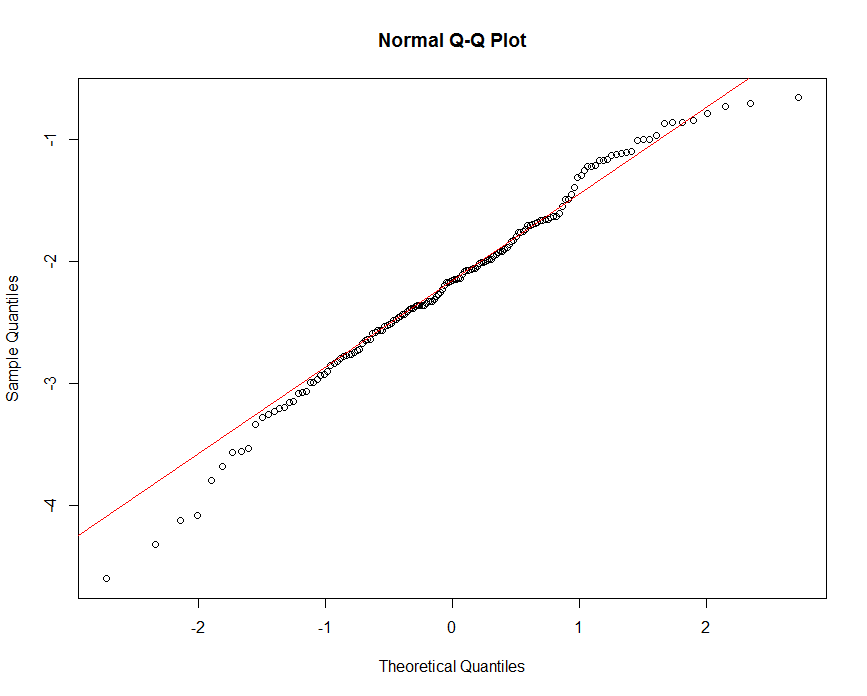
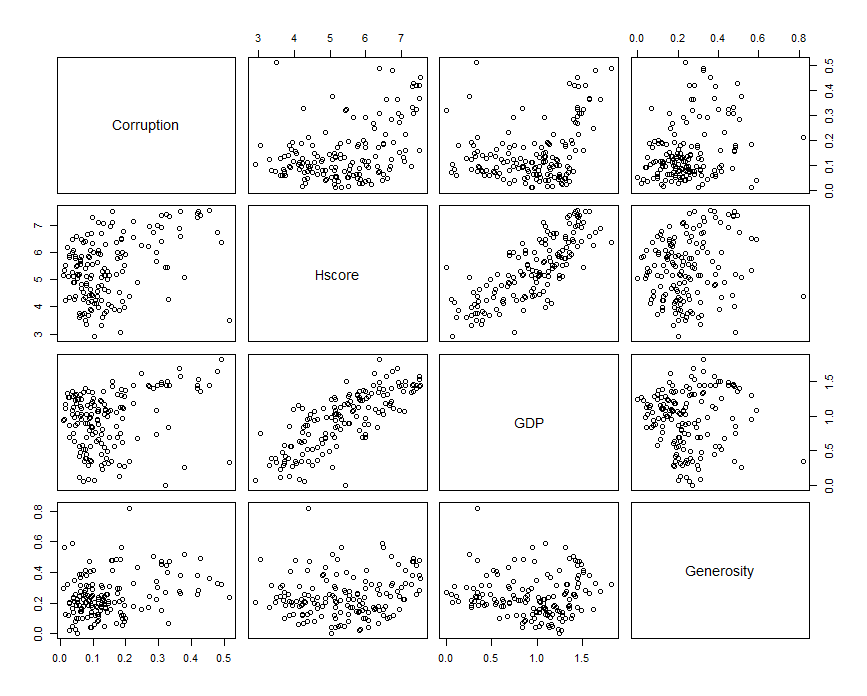
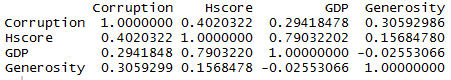
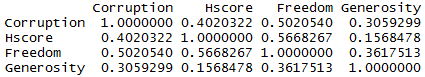
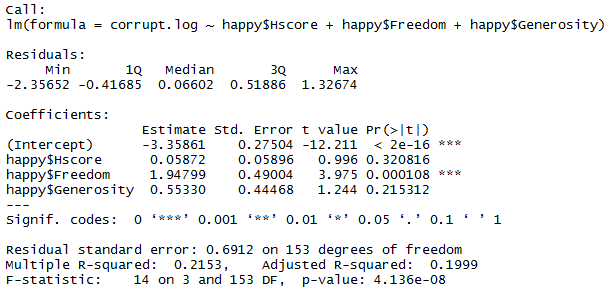
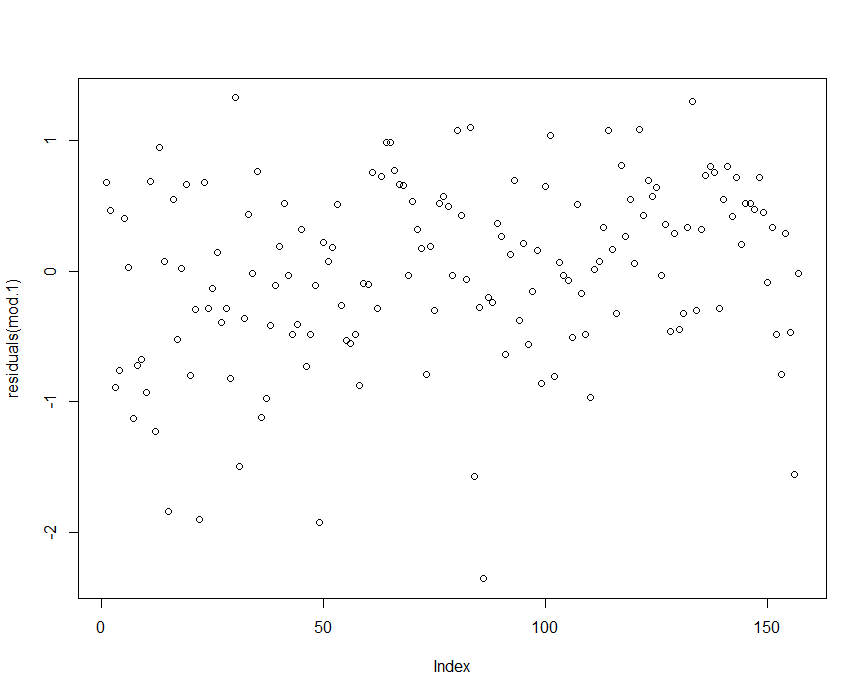
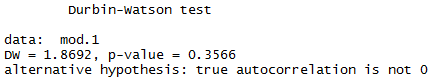
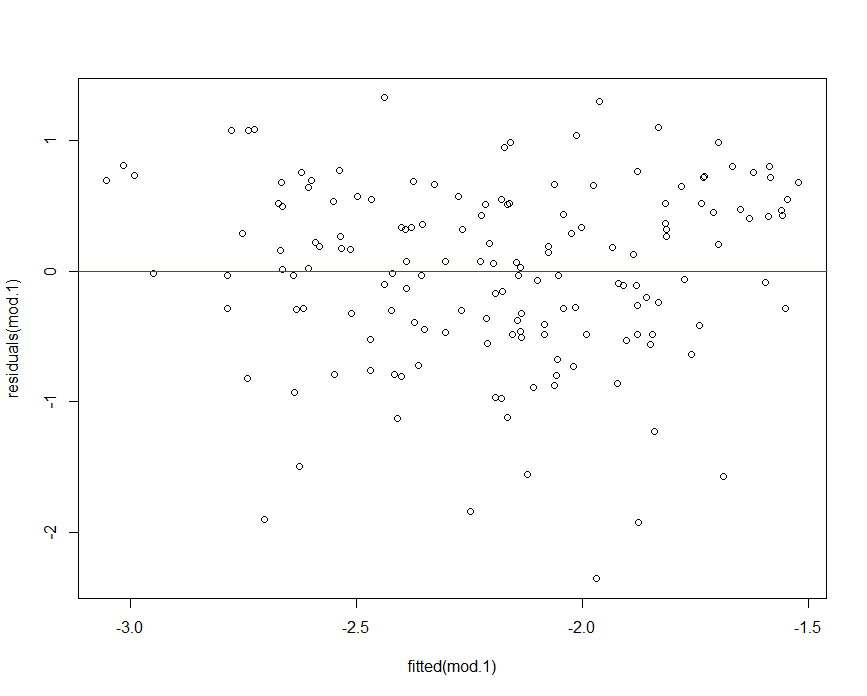
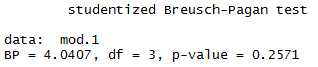
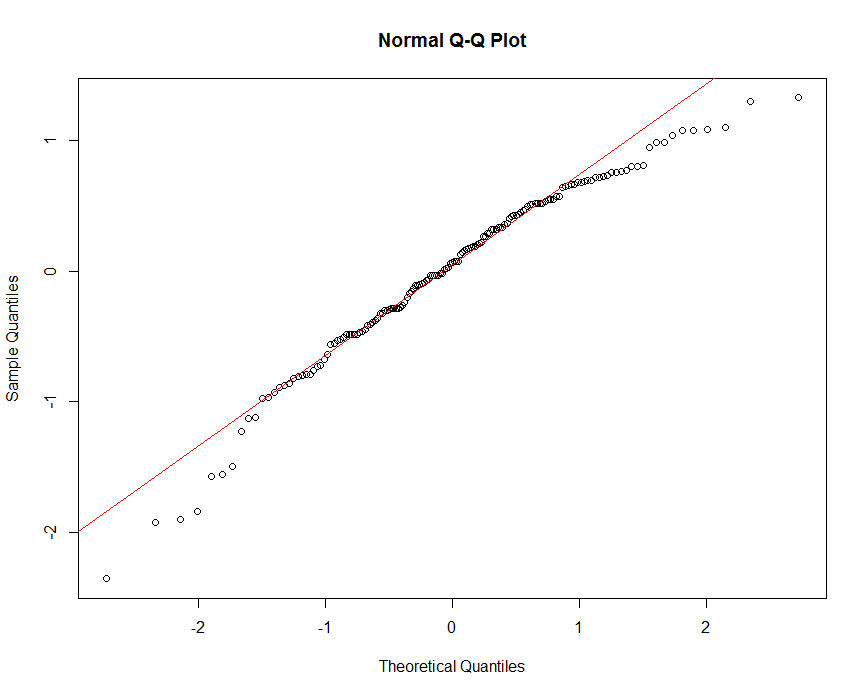
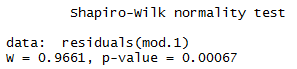
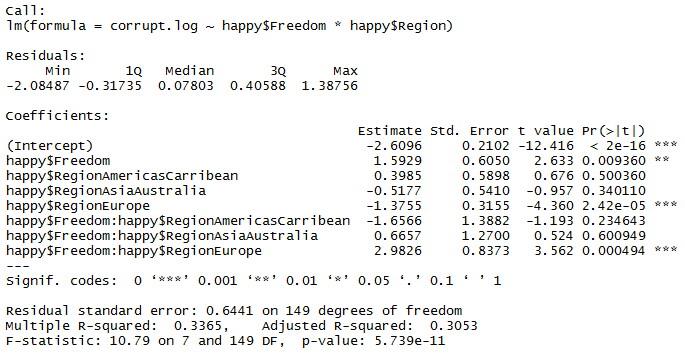
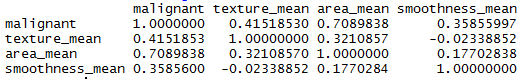
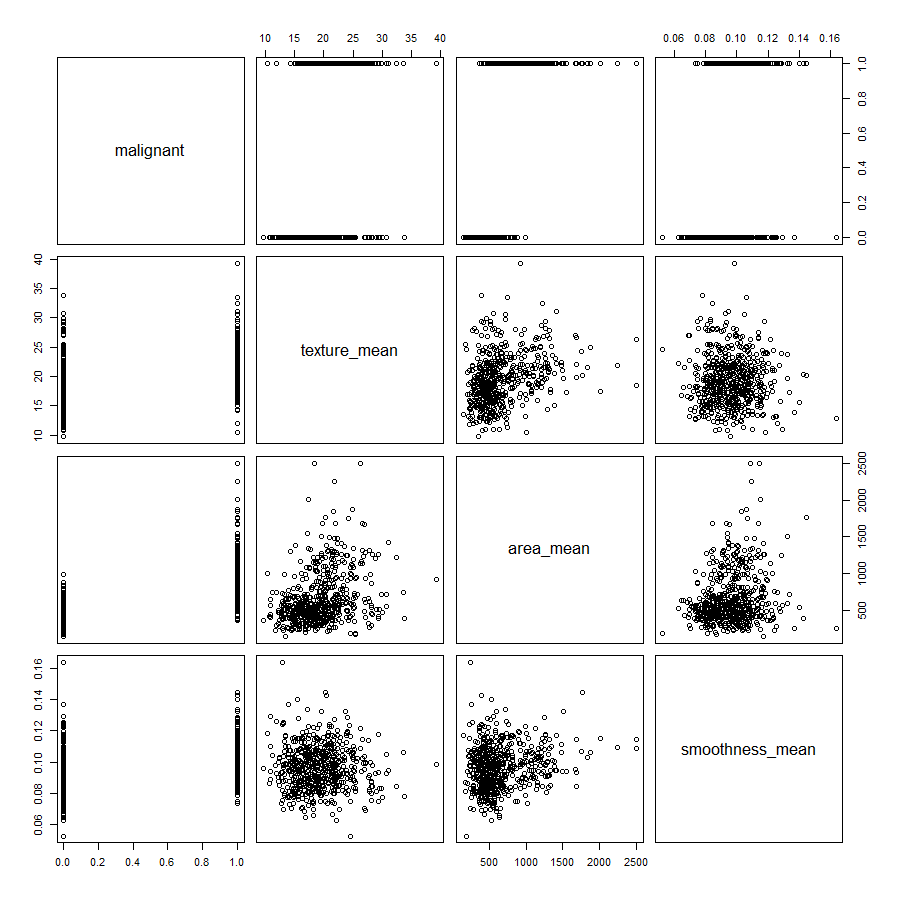
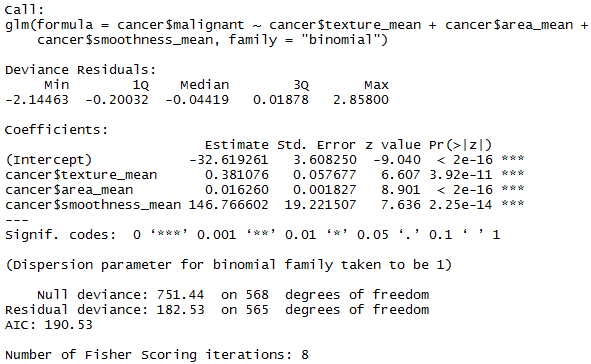
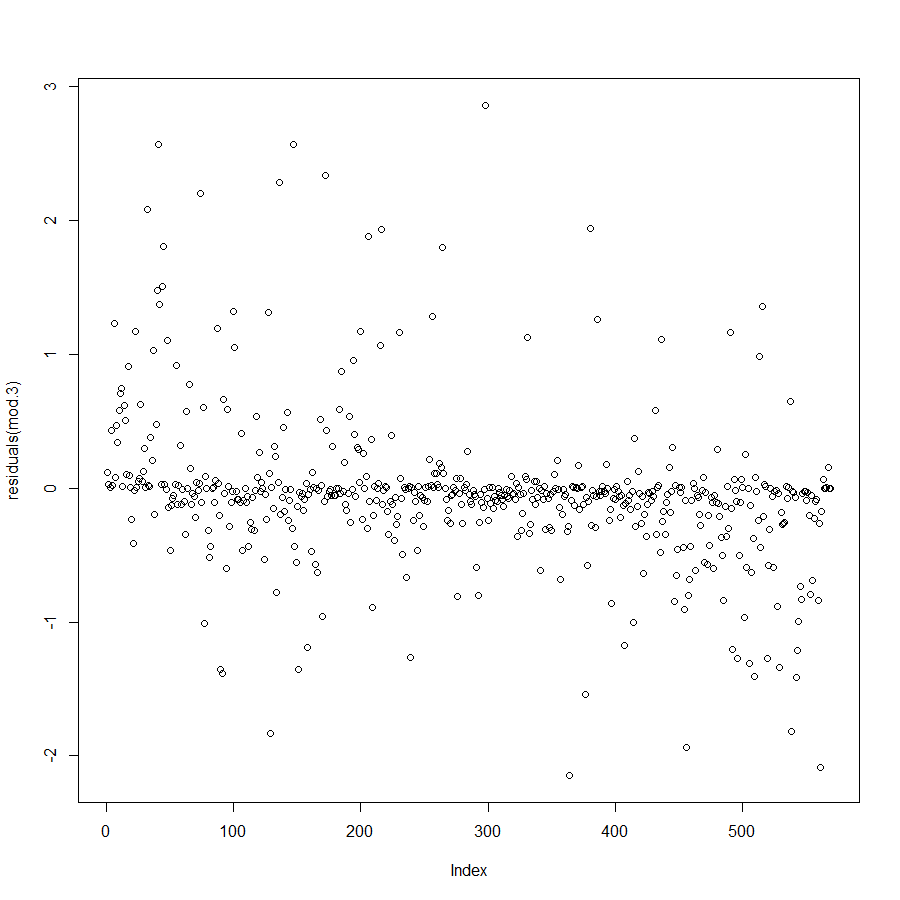
Brittany Szczepanik (szczeb)

NRE 538

4/22/17

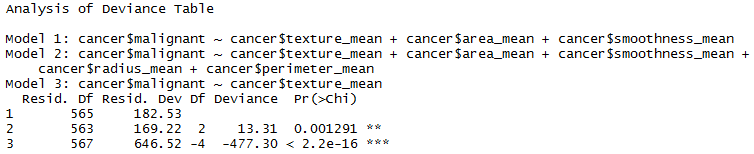
**Final Exam (take-home)**

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. check whether the data meet the requirements/assumptions of the test you plan to run
      1. I was planning to run chi-square test, because we are interested in seeing if two groups (categorical—male vs female) are significantly different from one another in terms of another categorical variable (unruly\_child). However, since the two variables are both factors of different length, I couldn’t get it to work…. Instead, I decided to manipulate the data so that I could turn all of the numbers into integers, and determine if the mean of one sample (male) is significantly different from the mean of the other sample (female). Therefore, I decided to run a two sample, two-tailed t-test.
      2. Visualize the data
         1. In order to get a sense of what we are looking at here, I wanted to plot the data and see what it looked like. From the output below, it appears that females tend to think that it is ruder to bring an unruly child on an airplane. However, we’ll have to test this out to see if this difference is significant.
         2. 
         3. Unfortunately, even after changing all of the information into integers and trying to get an accurate summary, I ran into a ton of errors, and couldn’t quite figure it out in time….Please check my R code so you can see what I attempted to do!
2. Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition? **College**
   1. I plan to run a two sample, two-tailed t-test, because we are dealing with one continuous dependent variable (tuition) in terms of one factor (institution type) with two different levels (public or private).
   2. The first thing I do when running any test is to visualize the data, and try to get a sense of what I would expect from the analysis. From the graphs below, we can see that the two samples appear to be statistically different, particularly from the boxplot, where the means from both private and public do not lie within each other’s confidence intervals, which suggests they are statistically different. Running the test will prove this, though.
   3. 
   4. 
      1. Assumptions
         1. Data are continuous
            1. The values for tuition are continuous, as can be seen by the dataset
         2. Sample randomly selected from the population
            1. We must trust that this data was randomly selected
         3. Observations are independent
            1. Based on the plots of both private and public institutions (see graphs below), it appears that the observations are independent, and I’m assuming that I can trust that the data was collected responsibly.
            2. 
            3. 
         4. Normal distribution
            1. Based on the QQ plot, the data seem to be somewhat normal, but it doesn’t look that great (see below).
            2. 
            3. To test to see whether the data are normally distributed, I also ran a Shapiro-Wilk test (results below). From this test, the data are not normally distributed. The p-value is very low, which means we reject the null hypothesis that the data are normally distributed. However, since the sample size is far greater than 30, we can assume normality based on the central limit theorem.
            4. 
         5. Equal variance (between 2 populations)
            1. To determine if the samples have equal variance, I ran a var.test, which the dataset also failed (see below). The p-value is also very low here, which means that we should reject the null hypothesis that the variances between the two datasets are equal.
            2. 
         6. To correct for the fact that the variances are not equal, we can still run a t-test, but we must specify that the variances are not equal. This would be a Welch’s t-test (which the program runs anyway). The results are below
         7. 
         8. As is evidenced by the Welch’s t-test, the tuition for public and private institutions is significantly different. Since the p-value is far below 0.025 (since we have to split the p-value because it is a two-tailed test), we reject the null hypothesis that the two samples are equal. From the means listed at the bottom of the test, we can see that “x” (public) is far more affordable than private colleges.
3. Is there a significant difference in happiness (Hscore) by region (Region)? **Happy**
   1. For this test, I would run a one-way ANOVA, because we have one continuous dependent variable (Hscore) and several independent, categorical variables that are levels of the same factor (regions).
   2. Again, it is helpful to visualize the data so that we can get a sense of what to expect from running the test. When running an ANOVA, we are testing to see wither or not all the levels in Region are the same. In this case, we expect to reject this null hypothesis, because we can clearly see that the Hscore in AfricaMideast is different from all other regions, and that the AmericasCarribean Hscore is different from AfricaMideast and AsiaAustralia. Running the test should be able to confirm our initial impressions.
   3. 
      1. Assumptions
         1. Populations normally distributed
            1. To test to see if the data are normally distributed, we can both look at the QQ plot, and test for normality using the Shapiro-Wilk normality test. The Shapiro test (below), indicates that we cannot assume normality, because the p-value is less than 0.05, which means we reject the null hypothesis that the distribution is normal. This can also be shown visually in the QQ plot (below), which vaguely follows the red line—except in the beginning and end of the distribution. However, similar to the previous question, we can assume normality because there are more than 30 samples (based on the central limit theorem), and each of the categories has more than 15 samples.
            2. 
            3. 
         2. Independent samples
            1. This one is hard to test, so you need to have some background on the data, which we do, and can assume the observations are independent of one another, and that there is no bias introduced.
         3. Equal variance
            1. To test to see if there is equal variance in the samples, we use a Levene Test (output below). From this test, we see that the p-value is greater than 0.05, so we cannot reject our null hypothesis that the variances are equal, and therefore pass this test.
            2. 
      2. Running the test
      3. 
      4. Since we cannot see which of the regions is significantly different based on this output, we can run a post-hoc test, which is typically the Tukey HSD test. The output of this test is below. As you can see from the output, there is a significant difference between all the regions (as shown by their low adjusted p-values), except between Europe & AmericaCarribean, and Europe & AsiaAustralia (whose p-values are above 0.05—at 0.784 and 0.769, respectively).
      5. 
      6. To get an even better sense of the data, we can run the summary function of an lm instead of just aov, which gives us more information, and allows us to compare the data to one of the regions.
      7. 
      8. From the above output, we can see that all are regions are significantly different from AfricaMideast. We have to note that we are comparing everything to AfricaMideast, because that is the highest alphabetically. Essentially, the intercept is referring to the happiness in AfricaMideast, which is significantly different from 0. All of the other regions are compared against AfricaMideast, and as we can see, are significantly different from the happiness there. It’s just another way to look at the data, but we shouldn’t be surprised that all regions are significantly different from AfricaMideast, because the boxplot shows us that the average for that region is not within the confidence interval of any other region.
4. What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model. **happy**
   1. check whether the data meet the requirements/assumptions of the test you plan to run
      1. For this test, I am choosing to run a linear regression. My dependent variable is corruption, and I am choosing three independent, continuous variables to see if they are associated with a country’s corruption level (Hscore, GDP and Generosity).
      2. Prior to running the test, I wanted to again visualize the data. I looked at a QQ plot of my dependent variable, corruption. From this graph (below), I can see that the data are not normally distributed, which implies that my residuals may also not be normally distributed.
      3. 
      4. To correct for this, I transformed the dependent variable (corruption) using log, since the function looks like exponential growth. After transforming the data, I looked at the QQ plot again, and was pleased with the improvement (see below). Now that the distribution looks more normal, I can proceed with the linear regression, and check the model against my assumption when I’m done.
      5. 
      6. Another aspect of visualizing the data is to see if the data are correlated or not. To see this, we can use the pairs function. We can see from the graph below, that some of the variables appear to be correlated, particularly GDP and Hscore, which seem to have a clear linear pattern. Doing a Pearson correlations test confirms this (data output below the graph), which shows that the GDP and Hscore have a high correlation coefficient of 0.79.
      7. 
      8. 
      9. We typically want to avoid anything higher than 0.5, so I am going to take out GDP and use Freedom instead. It is still highly correlated (higher than 0.5), but not quite as much as the GDP variable, so I’m going to choose to keep this one instead.
      10. 
      11. Running the test
          1. After running the linear regression, I got the following output.
          2. 
          3. From the output, it looks as though the intercept (i.e. corruption), starts off at -3.359. However, this is misleading. Since we took the log of corruption to make the data more normally distributed, we now take the exponent of that, which is 0.03478.
          4. It looks as though Hscore and Generosity do not have a significant impact on corruption. However, Freedom does have a significant impact on corruption. For every unit of freedom increase, the corruption goes up by 1.948.
          5. However, as we can see by the adjusted R-squared valued (0.1999), the model is terrible (only 20% of the variance is being explained by the model), so this would be a good time to check our model against the assumptions necessary for doing linear regressions.
      12. Assumptions
5. Independent errors
   1. To test this assumption, I took a plot of the residuals (below). It appears that the residuals are independent, but to be sure, I ran a dwtest (below the graph). The p-value for that test is higher than 0.05, so we cannot reject the null hypothesis that are errors are independent, which means that we pass this test.
   2. 
   3. 
6. Linear relationship
7. Homoscedasticity
   1. To test whether this dataset is homoscedastistic, we can take another look at the residuals (shown below). This data does appear to be homoscedastistic, which means that there is constant variance of the residuals (i.e. no pattern). Just to be sure, I also ran a bp test (shown below the graph), which confirms what the plot is showing. The p-value is higher than 0.05, so we cannot reject our null hypothesis that the residuals are homoscedastistic, so we pass this test too!
   2. 
   3. 
8. Normal error distribution
   1. We kind of tested this in the beginning (by looking at the actual dataset), but now we need to check the distribution of the residuals to see if they are normally distributed. Our initial analysis should have given us a good indication that they are, but now we can be sure. To visualize this, I plotted a QQ plot of the residuals (below). It doesn’t look great, but it’s not too bad.
   2. 
   3. To check it further, I ran a Shapiro-Wilk test and, unfortunately, our errors are not normally distributed. Unfortunately, the residuals are not normal, as evidenced by the low p-value. Since the p-value is below 0.05, we should reject the null hypothesis that the residuals are normally distributed.
   4. 
   5. Although all assumptions other than normal error distribution check out, the model fit isn’t the best, as evidenced by the very low R^2 value (mentioned previously). There are generally three things we can do to try and improve this model, given the data are not normally distributed: transform the data, resample, or run a GLM. We already did the first one, so we would most likely consider either resampling or running a GLM next, and compare either the R^2 values or the AIC values to determine which model is the best fit. However, given the directions in this test, running either of these appears to be out of scope, so we’ll just leave it as is!
9. 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. Since we are running a linear regression with one continuous variable and one categorical variable, we would run an ANCOVA here, which is kind of the bridge between linear regressions and ANOVAs. Since my only significant variable in question 4 was “Freedom,” I will use that variable to run the analysis with “Region”.
   2. Run the test!
      1. Running the test yields the following output. Following the output, there is a series of explanations for each of the values.
      2. 
      3. Intercept:
         1. From this output, we can see that the intercept is significantly different from 0 based on the very low p-value. The intercept, in this scenario, is “corruption” in the region “AfricaMideast” when the “Freedom” variable is set to 0. Remember that the intercept here is still the log of corruption, so we would need to take the exponent of that to determine the actual value.
      4. Happy$Freedom
         1. According to this test, freedom significantly influences the corruption value in “AfricaMideast.” This particular value is only compared to that region (because it is alphabetically the first one), so we cannot say anything about the other regions within looking at the other outputs.
      5. Happy$Regions
         1. The corruption in “AmericasCarribean” and “AsiaAustralia” is not significantly different from corruption in “AfricaMideast,” which means that the intercepts for these regions are essentially the same (exp -2.6096).
         2. However, the intercept of the corruption in “Europe” would be different, because the coefficient for this is significantly different from the corruption in AfricaMideast.
      6. Happy$Freedom:Regions
         1. For this portion of the graph, we have the interaction terms. Essentially, we are comparing how freedom affects corruption in each of these regions. We can see that the only region where this is significant, is in Europe. Since the AmericasCarribean and AsiaAustralia region does not yield a significant interaction with freedom and corruption when compared to the interaction with freedom and corruption in AfricaMideast, their slopes would essentially be the same as the slope provided by the “happy$Freedom” output, which is exp 1.5929. However, since Europe is significantly different, we can say that the slope increases by exp 2.9826 when compared to the slope of “happy$Freedom.”
10. 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. For this question, I would choose to run a binomial generalized linear regression, because we are testing to see if the effect of three independent variables contribute to breast cancer or not (a binomial question). The three independent variables I’m going to choose are: texture\_mean, area\_mean and smoothness\_mean.
    2. Visualize the data
       1. To get a sense of whether or not these variables are correlated, we can run the pairs and cor functions. The output are below. It appears that area\_mean is strongly correlated with malignancy, so we perhaps should take another independent variable instead.
       2. 
       3. 
       4. To attempt to remove the correlation between malignancy and the independent variable of “area mean”, I tried to use the remaining two variables, “perimeter mean” and “radius mean,” however, both of these are also correlated with malignancy. The good news is that none of these, including area mean, are correlated with each other, so we should still be okay. I’m going to keep my original variable of “area mean” to run the analysis.
    3. Run the test!
    4. 
    5. According to the data output above, all three of the independent variables that I tested are significant, and the intercept is significantly different from 0. Below is a description of each of these outputs. Since we are dealing with linear predictors, the parameters do not necessarily explain the dependent variable (cancer is malignant or not), but they describe the data distribution.
       1. Intercept
          1. The intercept in this case is the probability of cancer when all of the other predictors are at 0. In this case, since the intercept is significant, the value is significantly different from 0. In taking the exp of the value, we see that the odds ratio for cancer to be malignant when all other factors are 0 is very small (6.817654e-15), which makes sense, because that implies that there is no breast at all (i.e. no breast area).
       2. Cancer$texture\_mean
          1. The texture is significant when determining whether cancer is malignant or not. The odds ratio for cancer to not be malignant for every unit increase in texture is 1.463859.
       3. Cancer$area\_mean
          1. Similar to texture, the area is also significant. The odds ratio for cancer to not be malignant is 1.016393 for every unit increase of area
       4. Cancer$smoothness\_mean
          1. Same story here, the smoothness is also significant. The odds ratio for cancer to not be malignant is also 5.494465e+63 for every unit of increase in smoothness.
    6. Checking model fit
       1. Assumptions
          1. The benefit of using a GLM is that you can relax two of the four assumptions: homoscedasticity and normal distribution of errors. We should, however, still check to see if the errors are independent, and if there is a linear relationship
             1. Independent errors/Linear relationship
             2. Below is a plot of the residuals for mod.3 (the glm ran in this question). The residuals are clearly have a linear pattern, which is what we would expect given that the question is has binomial data (they are centered around 0).

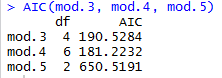


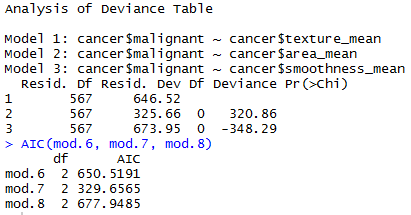
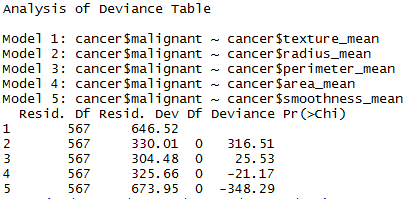
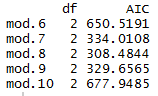
* + 1. There are several ways to check GLMs to see if they are a good fit for capturing patterns in the data: simulation, inspection plots, or model comparisons. We cannot test the variance like we would in typical models without calculating the variance manually be doing simulations. We can only calculate the probability of reproducing data and compare models based on that. Since we do not have an R2 value (to see the variance that is explained by the model), we can instead use ANOVA to compare models based on the chi-squared distribution (not F-distribution, as we did previously).
       1. Model comparison
          1. In order to test this model, I chose to compare it to adding all five of the variables listed (mod.4), and to test it based on only using smoothness (mod.5)

Checking with the chi-squared distribution (below), we can see that our model isn’t too bad when compared to the other two alternatives. We know this because of the low p-value for the second and third models, which means the probability for mod.4 and mod.5 to produce the same deviance as mod.3 is significantly different. The lowest residual deviance is in mod.4 (which means that are residuals are better explained), but mod.3 (the one we used in the output above), isn’t too far off—it’s only a little higher.



To check this further, we can compare using AIC. We want to get a small AIC value, which basically determines the likelihood of predicting the actual impact of the variables in the model on malignancy. Again, we can see that mod.4 (which includes all five of the variables) is slightly better, but less than 10, so it’s not much of an improvement, and we can therefore keep our original model with three instead of five variables.



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.**
   1. To answer this question, I essentially did the same thing I did in the previous question, which was to check the models for how much residual deviance could be explained by various independent variables. In this case, I tested each of the variables independently against malignancy to determine which AIC score was the highest, and to test which model explained the most residual deviance. I’m not sure if this is the best methodology, but it seemed to make the most sense to me. Below is the output from this analysis.
   2. 
   3. From the three independent variables ran in the initial analysis, it appears that the most important variable is area, or mod.7 in this case (based on the low residual deviance score, and the lowest AIC score). This just emphasizes (for me) that the coefficients don’t necessarily mean anything if they aren’t on the same scale. The smoothness coefficient was extremely high in the GLM that we ran previously, which suggests that it makes the largest impact on the malignancy of the tumor; however, we don’t know the scales that are being used here. In order to put all of these items on the same scale (and to truly analyze them in relation to one another), we could normalize these variables based on their standard deviations and measure them that way. If we did that, perhaps we would see that the area variable would have the highest coefficient value.
   4. Just out of curiosity, I also ran all five of the variables to see if the answers would be the same. The output is shown below.
   5. 
   6. 
   7. From the output above (keeping in mind that I changed the model numbers), it appears as though mod.8 (which is the perimeter variable), has the highest likelihood of explaining the most residual deviance (based on the low residual deviance score, and the lowest AIC score), which implies that it is the most significant in terms of determining cancer malignancy. The area variable (which my previous model indicated was the most important), is second most important when including all five variables.