**Final Exam**

**Brooke McWherter**

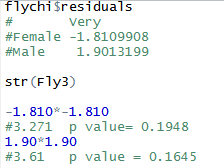
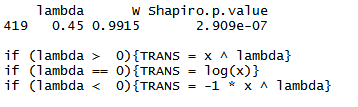
Dataset Please use the following scripts to load in the data from GitHub

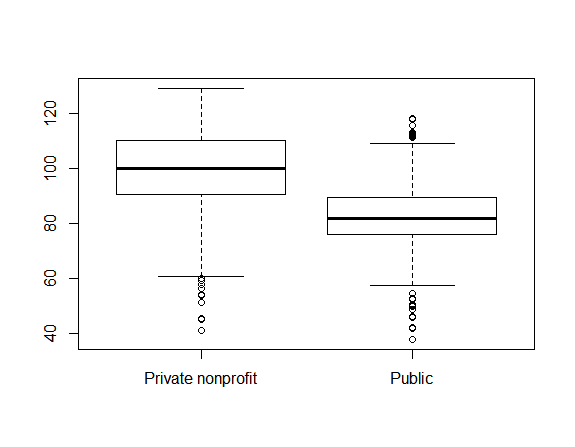
flying = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/flying.csv",header=TRUE, sep=",")

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=",")

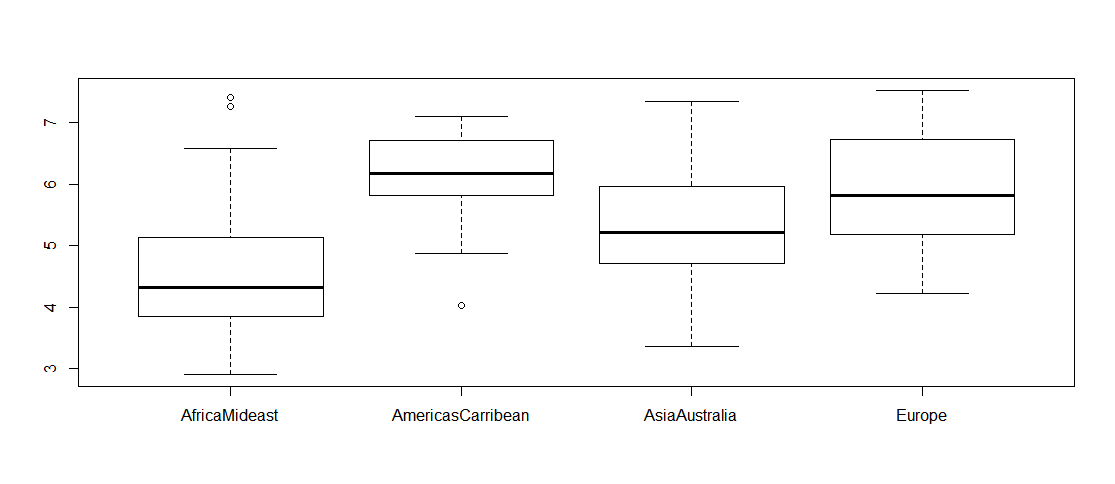
1. Is there a significant association between gender (gender) and whether people think it’s rude to bring an unruly child on the plane (unruly\_child)? If yes, which gender tends to think that bringing an unruly child is more rude? **Flying**
   1. The question asks for test of association between two categorical variables (gender and unruly\_child), as a result the Pearson’s Chi Square Test Independence is needed for the question.
   2. Chi Square Assumptions
      1. Two categorical variables – check
      2. Two or more categories for each variable – check
      3. Independence of observations – relationship between variables will be tested
      4. Relatively large sample size- N= 442 and 401 respectively
   3. Chi-Square Test
      1. A Pearson’s Chi-Square was ran after omitting of Na cells, the test determined that there was an association between gender and the perception of unruly children (p-value =0.001193).
      2. 
   4. Which Gender?
      1. The adjusted residuals were determined through the creation of a cross table, the chi square values were then determined through a squaring of the residuals and the p values were then determined using the residuals and df. A pairwise Normal Independence test was also ran to check results.
      2. 
      3. If we consider the category “Very” to be an indication of a person considering a child ‘more rude’ then based off of my calculations I can’t say that there is any significant relationship between which gender considers a child to be more rude (females p-value = 0.1948, males p-value = 0.1645).
2. Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition? **College**
   1. This scenario looks at a continuous dependent variable (tuition) and a categorical independent variable (type) with two categories (public and private) and looks at the difference between the two therefore a two sample t.test would be best suited.
   2. Assumptions for a t.test are
      1. Bivariate independent variable – check
      2. Continuous dependent variable – check
      3. Sample is randomly selected – I will assume the data was randomly selected
      4. Observations are independent – check
      5. Values are nearly normal or sample size is large enough – Fail
         1. To test for normality, a shapiro wilks test ran, test determined that the x variably was non-normally distributed (p-value=3.193e-15)
         2. Transformations were attempted to fit the assumption; a Tukey’s ladder of Powers transformation was applied through the transform Tukey function. This tests chooses a lambda that maximizes the Shapiro-wilks W statistic or minimizes the Anderson-Darling A statistic. A new lambda value of 0.45, a shapiro wilks was then re-ran and the data fit closer to normal but the data was determined still be non-normally distributed (p-value=2.909e-07).
         3. 
         4. The sample size is sufficiently large 1407 to continue the test.
      6. Equal variance between the two independent variables -Violation
         1. A var.test was ran to determine variance, the primary test for variance resulted in a rejection of the null hypothesis of equal variance (p = 2.2e-16, F=0.43469) after the transformation the data still resulted in the rejection of the null but with an improved fit ( p = 3.678e-09, F-0.62526).
         2. A non-parametric test is needed due to unequal variances
      7. Test – Welch’s Two Sample t-test
         1. The welch’s t-test was ran due to the unequal variances among the two variables.
         2. T.test demonstrated a significant difference between tuition and college type (p = 2.2e-16).
      8. Highest Tuition?
         1. A post hoc Games Howell was ran due to its robustness with unequal variances, a TukeyHSD was also ran to compare results.
         2. Private nonprofit colleges have higher tuition, with public colleges having a lower tuition (difference= -16.45, p value =0.001) this difference in tuition costs is significant.
            1. Visually this is supported in the boxplot representations.



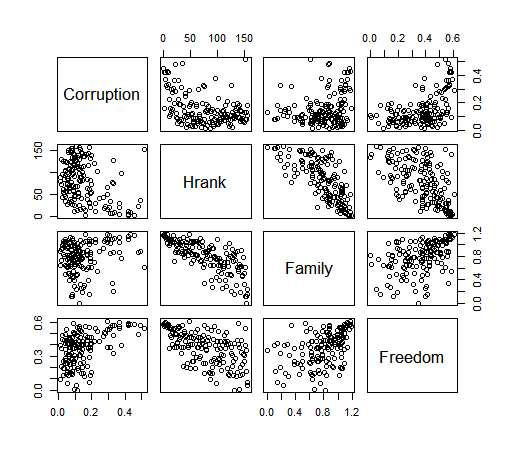
1. Is there a significant difference in happiness (Hscore) by region (Region)? **Happy**
   1. This scenario looks to determine the differences between a numeric/continuous dependent variable (Hscore) and a categorical independent variable (region), based on these factors an ANOVA would be the best fit.
   2. Assumptions
      1. Random sample - check
      2. Observations are independent – we can assume that the different regions do not influence the Hscore in a way measured by this study
         1. Response variable is normal or nearly normal – check

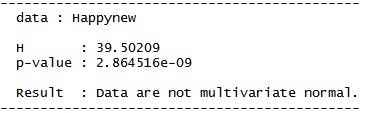
A Shapiro Wilk’s test was ran to test for normality, I failed to reject the null that the data was normally distributed (p. 0.01248)

* + 1. Homogeneity of variances -check
       1. A Levene’s test was ran to check for homogeneity of variance. I failed to reject the null that the group variances are equal (p=0.5179).
  1. Test - ANOVA
     1. An ANOVA was run and I can determine that there is a significant difference in happiness by region (p=1.279e-12). Visually these differences can be observed by looking at the boxplot, with the AmericasCarribian and AfricaMideast as having the largest differences between them over happiness (America has a 1.5397 unit larger, p = 0.0000 – TukeyHSD test).

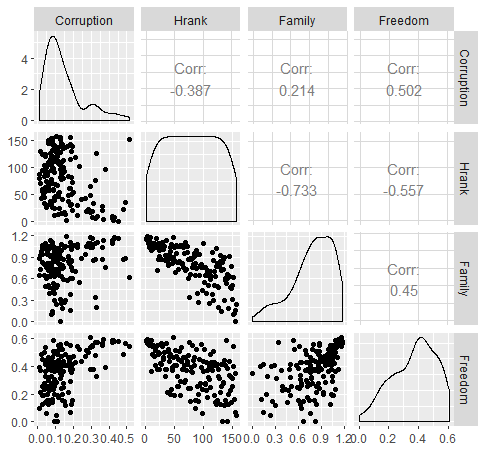


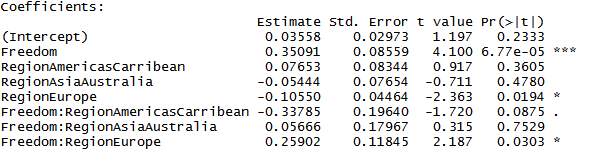
1. What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model. **Happy**
   1. Since this scenario includes both independent and dependent continuous variables a linear regression will be used to determine association.
   2. Assumptions
      1. Linear Relationship - check
         1. Based on visual plots, appears to be a linear relationship

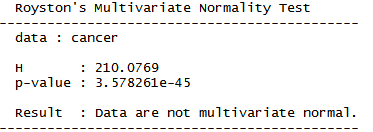


* + 1. Multivariate normality (MVN) – violated – large N
       1. A Royston’s MVN test was carried out, the Roystons test uses the Shapiro-wilk/Shapiro-Francia statistic to test multivariate normality. Based on the test I rejected the null that the data were multivariate normal (H=39.5, p =2.864516e-09).
       2. However, the sample size is large N=157, so I will continue on with the assumption.
       3. 
    2. No or little multicollinearity - check
       1. To evaluate multicollinearity of the model, I calculated the variance inflation factor (VIF) from the lm result. A VIF more than 10 suggests a strong multicollinearity. Based on the results there is little to no multicollinearity in the data (VIF= Hrank=2.514, Family=2.175, Freedom=1.458). Shapiro tests were run individually to confirm.



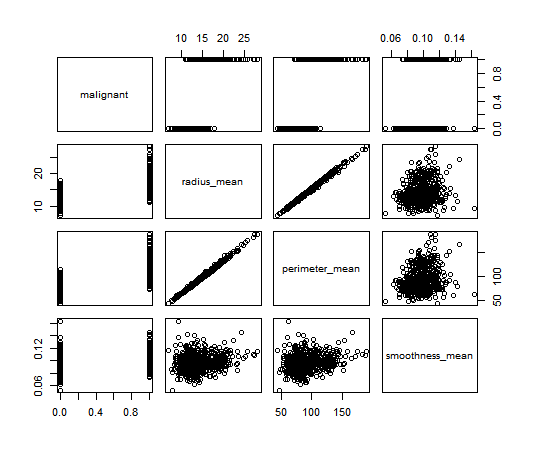
* + 1. No auto-correlation - check
       1. A correlation and pair wise model were used to determine correlation. None of the variables demonstrate any correlation over a 0.5, indicating that there is a moderate positive relationship but no auto correlation.
    2. Homoscedasticity - violated
       1. Bptest was ran to check homoscedasticity. I was able to reject the null that the data was homoscedastic (p=0.002892).
       2. Lack of homoscedasticity will affect standard errors but not model fit.
       3. 
    3. Model Fit
       1. Hrank and Freedom both are significantly associated with corruption. For every increase in corruption, Hrank decreases by a factor of 0.0007058 (p=0.00849), and Freedom increases by a factor of 0.3256979 (p=7.15e07).
       2. A model comparison using ANOVA shows that the best fit excludes Hrank (p=0.008485). With Hrank removed, Freedom remains significant with a p value of (9.34e-10)(r^2=0.2523) however the overall fit is still low.

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. Adding an interaction term with the variable Freedom, did lead to a slightly better model fit (R-squared= 0.3567, but it was not significant P=0.3265)
   2. Based off the results of the linear model, it appears that there is a region that interact with Corruption in a significant way (RegionEurope, p=0.0194). With the interaction of Freedom, Europe remains as having a significant interaction with Corruption, with all things being at 0 as Corruption increases the interaction between freedom and Europe does as well.
   3. I would interpret this as Freedom influencing the interaction between regions and corruption. A secondary model ran without freedom shows a non-significant interaction between all three regions (p= 0.270, 0.726, and 0.591 respectively)
   4. 
2. 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. To determine which factors are closely associated with malignant tumors a multiple linear regression model needs to be run.
   2. The factors selected for this prompt are – radius\_mean, perimeter\_mean, area\_mean.
   3. Assumptions:
      1. Linear Relationship - check
         1. Based on visual plots, appears to be a linear relationship
      2. Multivariate normality (MVN) – violated – large N
         1. A Royston’s MVN test was carried out. Based on the test I rejected the null that the data were multivariate normal (H=210.0769, p =2.864516e-09).
         2. However, the sample size is large N=569, so I will continue on with the assumption.



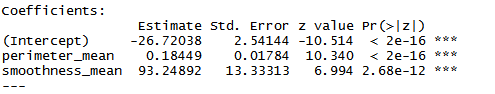
* + 1. No or little multicollinearity - Violated
       1. To evaluate multicollinearity of the model, I calculated the variance inflation factor (VIF) from the lm result. A VIF more than 10 suggests a strong multicollinearity. Based on the results there is a strong collinearity (radius p=343, perimeter p=348, area p =1.5) in two of the three factors.



* + 1. No auto-correlation - check
       1. A correlation and pair wise model were used to determine correlation. Some of the variables demonstrated a stronger correlation trend (0.98) but none of the variables were at or exceeded 1.
       2. 
    2. Homoscedasticity –violated
       1. Bptest was ran to check homoscedasticity. I rejected the null that the data was homoscedastic (p=2.18e-11).



* + 1. Due to the number of violations a Generalized Linear Model was ran
       1. Based on the generalized linear model, all factors have a significant influence on tumor malignancy ( radius p= 0.00264, perimeter p =3.57e-05, and smoothness p=1.10e-05).
       2. Radius has a -3.0795 log odds ratio to that of malignancy, perimeter 0.6388, and smoothness 68.2189.



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. Using a step-wise model comparison followed by an ANOVA, the mode important variable in explaining a breast cancer tumor malignancy is
   2. Based on the ANOVA I determined that model two was the best fit which included the factors perimeter and smoothness.
   3. Using that information I then developed a linear model for each factor and ran an AIC, based on AIC scores perimeter is the most important in explaining whether a breast cancer tumor is malignant or not. (33y vs 715).