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?
   1. Test: Chi-Square Test
   2. Assumptions:
      1. Independent observations: I assume all independent observations based on the fact that each respondent has a unique ID
      2. Randomly selected samples: I assume all samples were randomly selected, though there is no way to verify this for certain as it would depend on the experimental design.
      3. Observations in all cells of contingency table: If we summed the data, we would have observations in all possible cells of the contingency table (i.e. males who answered no, somewhat, and very, and females who answered no, somewhat, and very).
   3. Results:

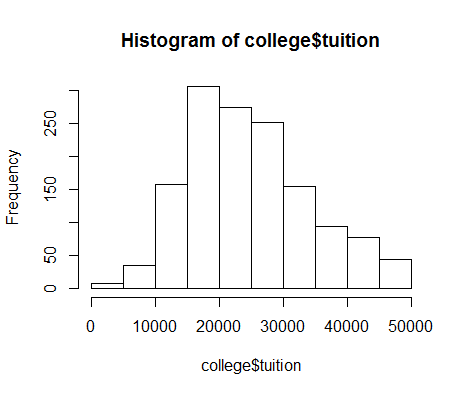
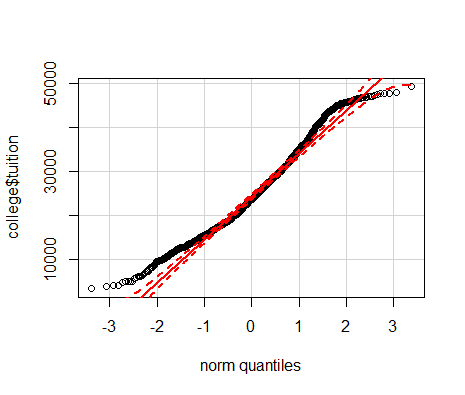
Pearson's Chi-squared test

data: flying$gender and flying$unruly\_child

X-squared = 13.463, df = 2, p-value = 0.001193

* 1. Interpretation: The p-value of the chi-square test is 0.001193, so we can reject the null hypothesis that there is not an association between gender and whether people think it's rude to bring an unruly child on the plane. Therefore, there is a significant association between gender and whether people think it's rude to bring an unruly child on the plane. To determine which gender thought it was more rude to bring an unruly child on the plane, for each gender I summed the number of individuals who answered "somewhat" and "very", and divided it by the total number of individuals. Based on percentages of those surveyed, men think it is more rude (72% for men vs. 66% for women)

1. Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition?
   1. Test: Two sample, one tailed T-test
   2. Assumptions
      1. Independent observations: I assume all independent observations, though there is no way to verify this for certain as it is based on sampling design.
      2. Randomly selected samples: I assume all samples were randomly selected, though there is no way to verify this for certain, as it is based on sampling design.
      3. Normal dependent variable distribution: The p-value from the shapiro test is 3.193e-15, so we can reject the null hypothesis that the distribution of tuition is normal. However, the distribution of tuition looks close to normal from the histogram and qqPlot. Furthermore, we have a large enough sample size to assume normality.

* + 1. Equal variances: The p value from the var.test function is 2.2e-16, which is less than 0.05, so we reject the H0 of this test that the variances are equal (i.e. the variances are not equal). However, this is ok, because we can use Welch's T-test
  1. Results:

Welch Two Sample t-test

data: Private\_NP$tuition and Public$tuition

t = 22.79, df = 1397.9, p-value < 2.2e-16

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

8763.607 Inf

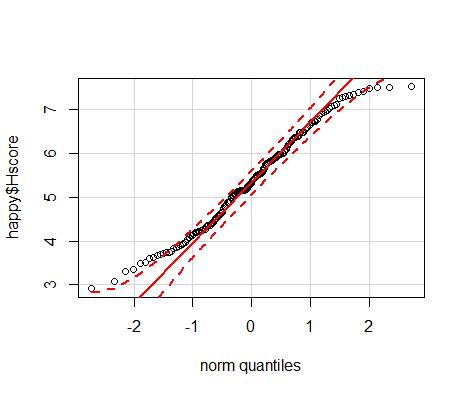
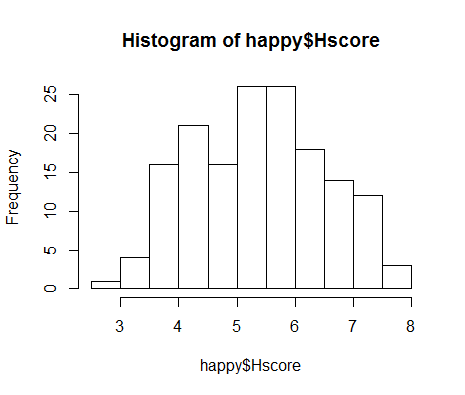
sample estimates:

mean of x mean of y

28301.69 18855.88

* 1. Interpretation: The p-value is 2.2e-16, so we can reject the null hypothesis that the mean of x is less than or equal to the mean of y. Therefore, there is a significant difference in tuition by type of institution. The mean of x is greater than the mean of y, so Private Non-Profit schools have a higher tuition than Public schools.

1. Is there a significant difference in happiness (Hscore) by region (Region)?
   1. Test: ANOVA
   2. Assumptions:
      1. Independent samples: I assume all independent observations, though there is no way to verify this for certain as it is based on sampling design.
      2. Randomly selected samples: I assume all samples were randomly selected, though there is no way to verify this for certain, as it is based on sampling design.
      3. Normal dependent variable distribution: The p-value from the shapiro test is 0.01248, so we can reject the null hypothesis that the distribution of tuition is normal. However, the distribution of Hscore looks close to normal from the histogram and qqPlot. Only the tails deviate significantly from the qqPlot, so I am going to assume normality.



* + 1. Equal variances: The p-value from the levene test is 0.342, which is greater than 0.05, so we fail to reject the null hypothesis of this test that the variances are equal (i.e. the variances are equal)
  1. Results:

Df Sum Sq Mean Sq F value Pr(>F)

Region 3 64.37 21.456 23.62 1.28e-12 \*\*\*

Residuals 153 138.96 0.908

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

* 1. Interpretation: The p-value from the anova summary is 1.28e-12, which is less than 0.05, so we reject the null hypothesis that Hscore is not different based on region. Therefore, happiness differs based on region. If we wanted to identify which regions differed from one another, I would run a post-hoc test.

1. What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model.
   1. Test: Linear Model
   2. Chosen model: Corruption ~ GDP + Generosity + Freedom

These three variables were chosen to avoid collinearity issues based on the correlation and VIF values.

cor(GDP, Generosity) = -0.02553066

cor(GDP, Freedom) = 0.36228285

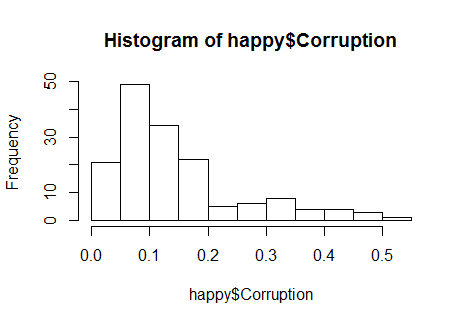
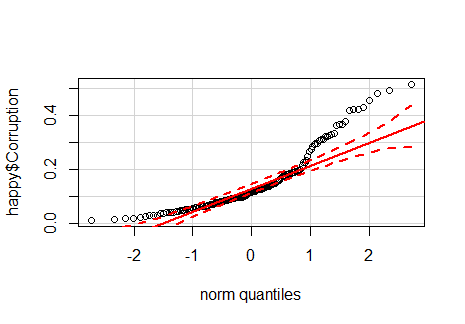
cor(Generosity, Freedom) = 0.3617513

vif.GDP = 1.189712

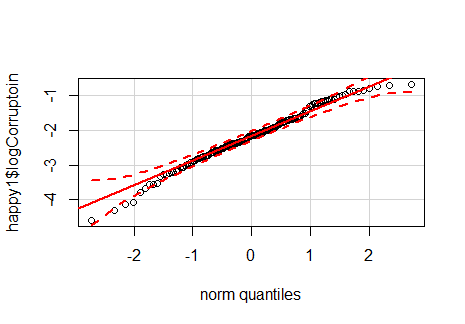
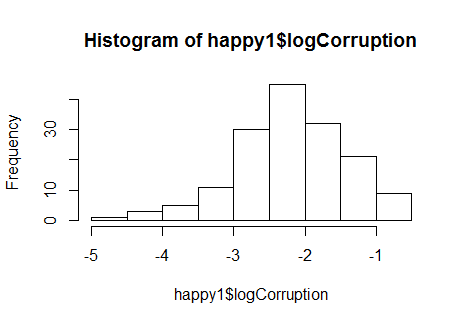
vif.Freedom = 1.367952

vif.Generosity = 1.189185

* 1. Assumptions:
     1. Normal dependent variable distribution: The p-value from the shapiro test is 5.128e-11, so we can reject the null hypothesis that the distribution of corruption is normal. The distribution is also skewed judging from the histogram and qqPlot, so it should be transformed.

After log transforming corruption, the p value from the shapiro test is 0.08535, so we fail to reject the null hypothesis that the distribution of corruption is normal. The distribution of log transformed corruption also looks normal from the histogram and qqPlot.

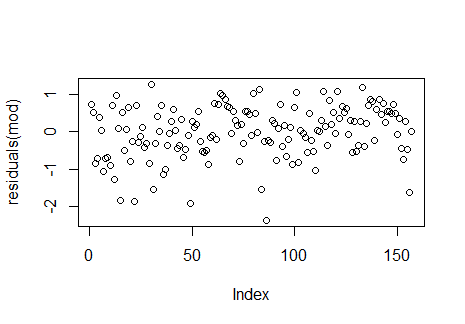


* + 1. Linear relationship: Although the plots of the relationships between the independent variables chosen and logCorruption do not show obvious linear relationships, I assume a linear relationship between the dependent variable and these independent variables.

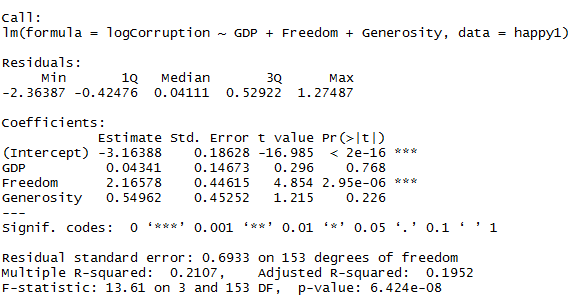
 



* + 1. Residual independence: The p-value from the dwtest is 0.3701, so we fail to reject the null hypothesis of this test that there is not autocorrelation in the residuals. The plot of the residuals also looks to have residual independency. Therefore, the assumption of residual independency is not violated.

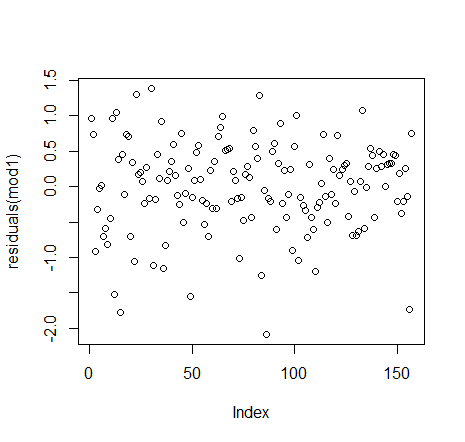


* + 1. Homoscedasticity: the bp test returns a p-value of 0.06878, so we fail to reject the null hypothesis of this test that there is not heteroscedasticity. Heteroscedasticity is also not visible from the plot of the residuals. Therefore, the assumption of residual homoscedasticity is not violated.
    2. Normal errors: The p-value from the shapiro test is 0.0006382, so we reject the null hypothesis that the errors are normally distributed. Therefore, the errors are not normally distributed according to the shapiro test. However, judging from the qqPlot the distribution is close to normal, so I am assuming the distribution of errors is actually normal.
  1. Results:

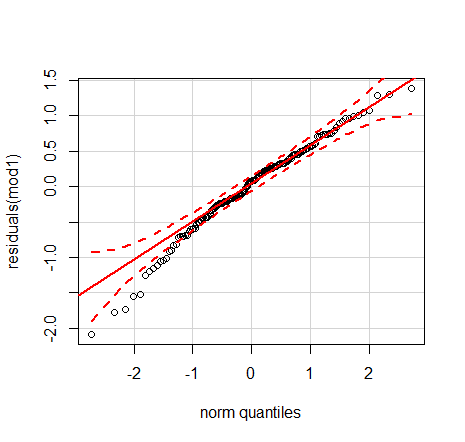


* 1. Interpretation: Based on the p-values, only the intercept and freedom have a significant effect on corruption (based on logCorruption as the dependent variable). The coefficients for GDP and Generosity are not significantly different from zero based on the p-values. Therefore, freedom is the only variable in the model that has a significant effect on Corruption. When all variables in the model are set to 0, the log of Corruption is -3.16388. As freedom increases one unit, the log of corruption increases by 2.16578 units. The model also only explains 19.52% of the variation in log Corruption, so a linear model might not be the best fit for the data.

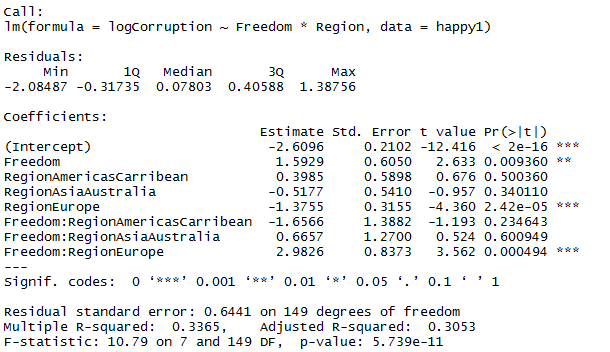
1. Choose one of the continuous independent variables that was significant in the model for Question 4 and interact it with region (Region) to predict corruption (Corruption). This model should only include one continuous independent variable and its interaction with region. Does the influence of your continuous variable on corruption vary by region? If yes, how do you interpret the interaction?
   1. Model: logCorruption ~ Freedom\*Region
   2. Assumptions:
      1. Normal dependent variable distribution: from Number 4 above, I will use log of Corruption so the dependent variable is normally distributed
      2. Linear relationship: from Number 4 above, I am also assuming a linear relationship between the two independent variables and log corruption.
      3. Residual independence: The p-value from the dwtest is 0.5627, so we fail to reject the null hypothesis of this test that there is not autocorrelation in the residuals. The plot of the residuals also looks to have residual independency. Therefore, the assumption of residual independency is not violated.



* + 1. Homoscedasticity: The bp test returns a p-value of 0.2501, so we fail to reject the null hypothesis of this test that there is not heteroscedasticity. Heteroscedasticity is also not visible from the plot of the residuals. Therefore, the assumption of residual homoscedasticity is not violated.
    2. Normal errors: The p-value from the shapiro test is 0.01481, so we reject the null hypothesis that the errors are normally distributed. Therefore, the errors are not normally distributed according to the shapiro test. However, judging from the qqPlot the distribution is close to normal, so I am assuming the distribution of errors is actually normal.



* 1. Results:



* 1. The influence of freedom does vary by region. The coefficient for Freedom means that the effect of freedom on log of corruption is significant for RegionAfricaMideast. Every unit increase in freedom leads to a change of the log of corruption of -2.6096 in RegionAfricaMideast.

The RegionAmericasCarribean and RegionAsiaAustralia variables are not significant, so this means that the intercept for these variables are the same as RegionAfricaMideast. These varibles do not have an effect on log of corruption when freedom is set to 0.

The RegionEurope variable is significant, so this means that the intercept for Europe changes by log(-1.3755) relative to log(-2.9096) when freedom is set to 0.

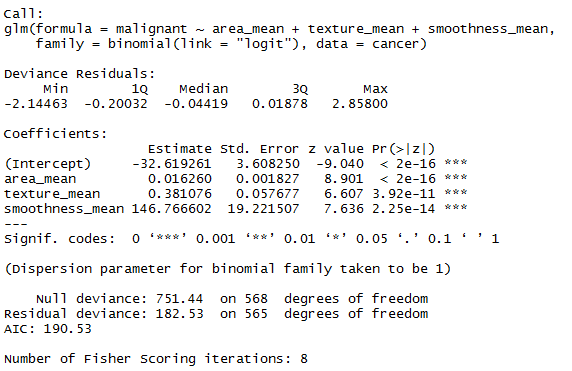
The coefficient for Freedom:Region Europe is significant, and means that the slope for the effect of freedom on log of corruption is significantly different for RegionEurope than any other region. The slope (or effect of freedom on log of corruption) for Europe is log(1.5929) + log(2.9826).

The adjusted R-squared for this model is 0.3053, so the model explains 30.53% of the variation in log corruption.

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.
   1. Model: GLM, binomial distribution

malignant~area\_mean + texture\_mean + smoothness\_mean

* 1. Assumptions:
     1. Independent observations: I assume all independent observations, though there is no way to verify this for certain as it is based on sampling design.
     2. Randomly selected samples: I assume all samples were randomly selected, though there is no way to verify this for certain, as it is based on sampling design.
     3. Linear relationship after transformed dependent variable is assumed.
  2. Results:



* 1. Interpretation: All the variables in the model are associated with whether a breast cancer tumor is malignant or not. When malignant is 0, the log(odds ratio) is equal to -32.619261. If area increases 1 unit, the log(odds ratio) increases by 0.016260. If texture increases 1 unit, the log(odds ratio) increase by 0.381076. If smoothness increases 1 unit, the log (odds ratio) increases by 146.766602. The output shows that when we added the variables, and reduced our degrees of freedom, the model has less deviance than a random model.

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.

Area\_Mean\_Effect = exp(0.016260) = 1.016393

Texture\_Mean\_Effect = exp(0.381076) = 1.463859

Smoothness\_Mean\_Effect = exp(146.766602) = 5.494465e+63

Smoothness is the most important in explaining whether a breast cancer turmor is malignant or not.