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? **flying**

Both gender and unruly child are categorical variables, so I will use chi-square test.

**Hypothesis:**

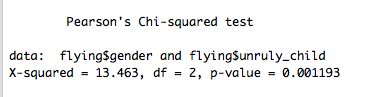
Null hypothesis: The attitude of bring an unruly child to plane is not different between different genders.

Alternative hypothesis: The attitude of bring an unruly child to plane is different between different genders.

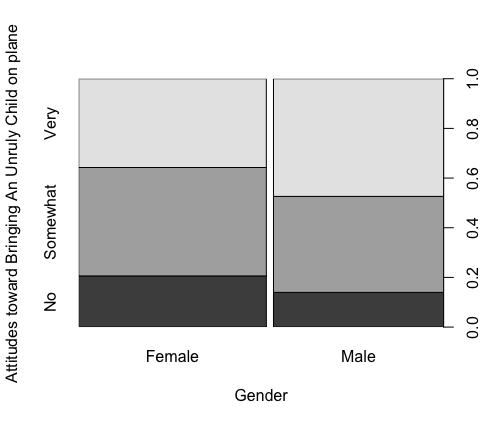
**Assumption Checking:**

1. Samples are randomly selected from the population: I will assume the sample is randomly selected, since it cannot be known from the dataset.
2. Observations are sampled independently: I assume observations are sampled independently, because this cannot be known from just looking at the dataset.
3. No structural zeros: based on the data, there is no structural zeros.

**Test Results:**



The p-value of chi-square test = 0.001193 < 0.05, so the null hypothesis should be rejected. The attitude of bring an unruly child to plane is significantly different between male and female.



The plot shows, fewer males think take an unruly child to plane is not rude, so males tend to think taking an unruly child to plane is rude.

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

Because type is categorical variable, there are two groups in “type”, and the tuition is a continuous variable, so I want to use a two sample t-test.

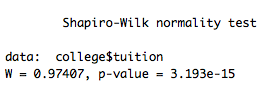
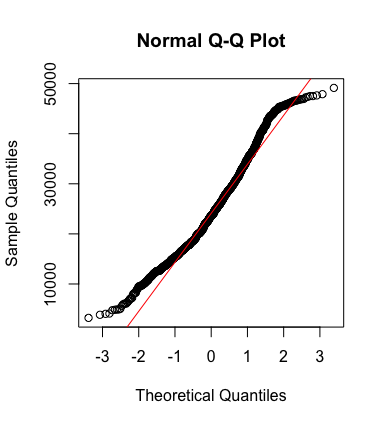
**Hypothesis:**

Null hypothesis: the tuition has no difference between public and nonprofit private colleges.

Alternative hypothesis: the tuition has significant difference between public and private nonprofit collegess.

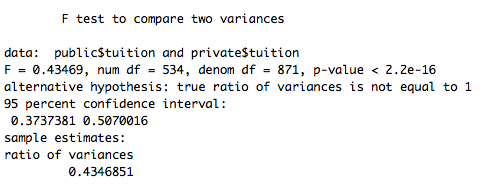
**Assumptions and check of two sample t-test:**

1. Observations should be independent: I assume each observation is independent.
2. Samples should be randomly selected from population: I assume samples are randomly selected from population.
3. Samples should be normally distributed or the sample size should be larger than 30:



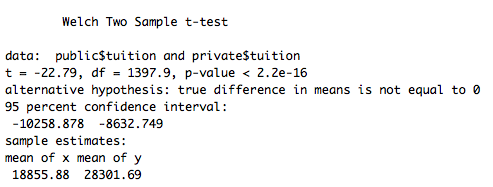
Although the sample is not normally distributed, but the sample size of tuition is larger than 30.

1. Variance are equal:

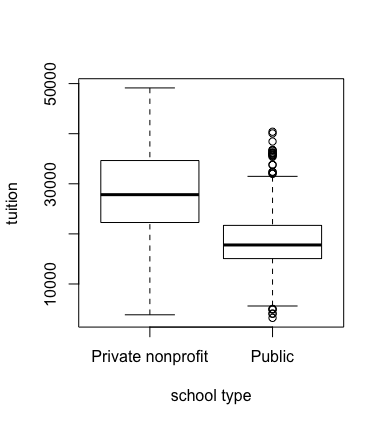


p-value<2.2e-16, so two variances are significantly different from each other.

**Results of Welch two sample t-test:**

****

Since the p-value is smaller than 2.2e-16 (<0.05), so we reject the null hypothesis. Thus, the tuition of public colleges is significantly different from non-profit private colleges.



The mean of public is 18855.88. The mean of private is 28301.69.

The boxplot shows the tuition of private nonprofit colleges is significantly higher than public colleges.

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

Happiness is continuous, but region is categorical and more than two groups, so I want to use a one-way ANOVA

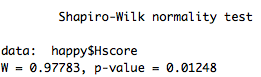
**Hypothesis:**

Null hypothesis: happiness has no significant difference among different regions.

Alternative hypothesis: happiness is significantly different among different regions.

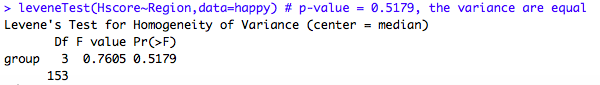
**Assumptions and Checking:**

1. Sample is normally distributed: shapiro test can be used to check normality



p-value is 0.01248, smaller than 0.05, so sample is not normally distributed.

1. Variance between different groups is same: this can be tested by Levene’s test.

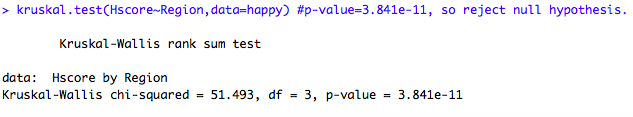


Since the p-value is equal to 0.5170>0.05, so the variance of different groups is equal to each other.

1. Samples are independent: I assume each observation is dependent with each other

**Results of test:**

since the variance is equal, but data is not normally distributed, I use Kruskal-Wallis test.



p-value is 3.841e-11 <0.05, so at least one region’s happiness is significantly different with other regions.

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

Corruption, GDP, Family, and Freedom are all continuous variables, so I will use muti-linear regression.

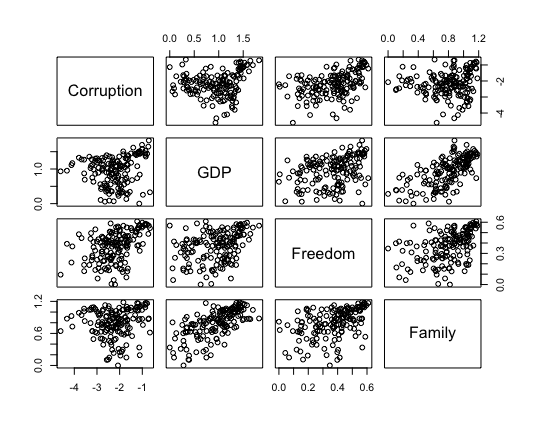
**Hypothesis:**

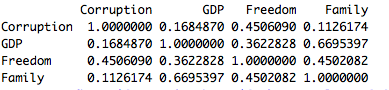
Null Hypothesis: Family, Freedom, and GDP have no effect on Corruption

Alternative Hypothesis: Family, Freedom and GDP have significant effect on Corruption.

**Assumptions and Checking**:

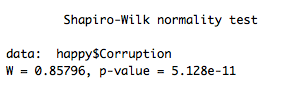
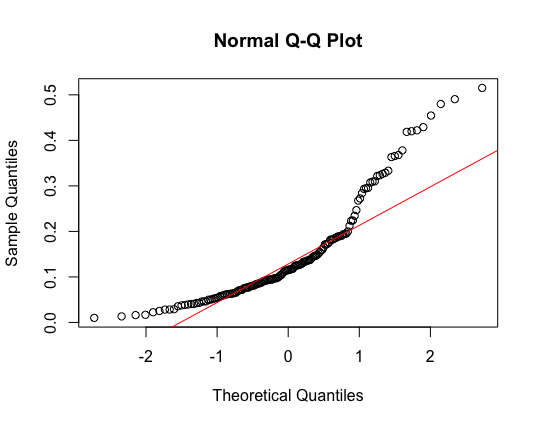
1. Samples are randomly selected from population: I assume data is selected randomly.
2. Observations are sampled independently: I also assume observations are sampled independently in this question.
3. There is a linear relationship between dependent variable and independent variables:





From the result of correlation, we can assume Corruption has linear relationship with each of these three independent variables.

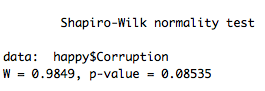
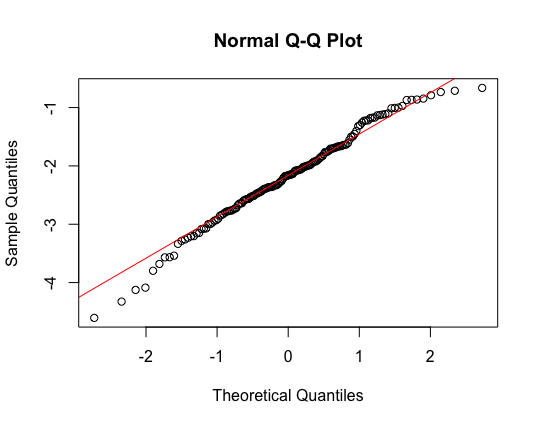
1. Dependent variable is normally distributed



Both the result of shapiro test and QQ plot show Corruption is not normally distributed.

**Transform data for linear regression:**

1. Log transform data and check the normality of corruption



The QQ plot and shapiro test of log transformed dependent variable is normally distributed.

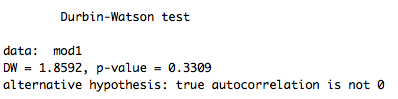
**Test Multicollinearity between independent variables:**

**Screen%20Shot%202017-04-23%20at%2013.57.13.png**

Since all VIFs are smaller than 10, so there is no multicollinearity between different independent variables.

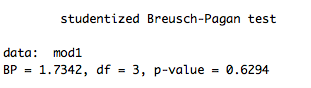
**Build model and check for the first time:**

1. Independency of residuals:



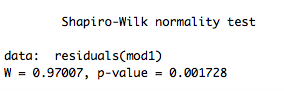
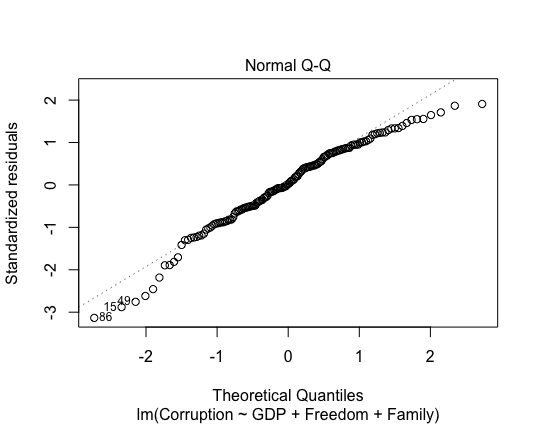
The result of dw test shows residuals are independent

1. Homoscedasticity of residuals:



The result of bp test shows residuals are homoscedastic

1. Normality of residuals:

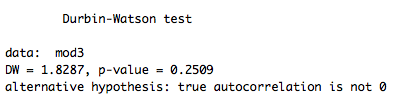


The QQ plot and shapiro test results show residual is not normally distributed.

**Transformation and check model:**

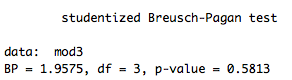
After *two* times of removing outliers, I get another linear model.

1. Independency of residuals:



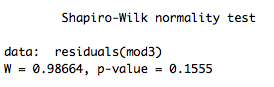
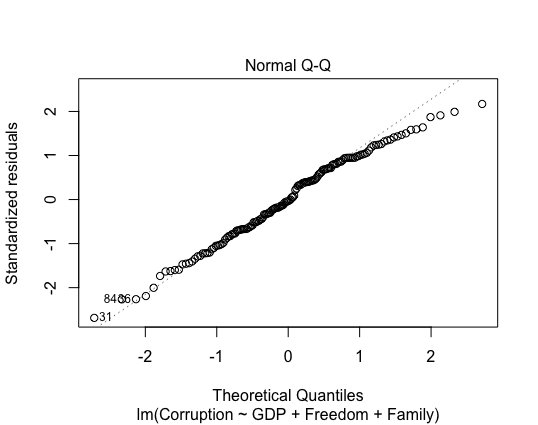
The result of dw test shows residuals are independent because p-value is 0.2509 > 0.05.

1. Homoscedasticity of residuals



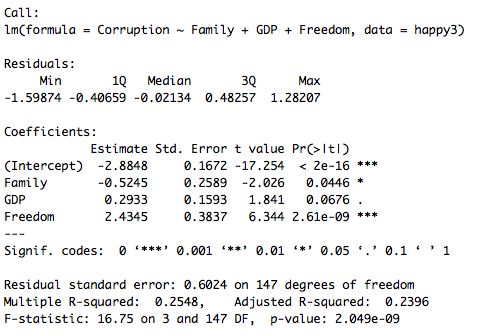
The result of bp test shows residuals are homoscedastic because p-value is 0.5813>0.05.

1. Normality of residuals



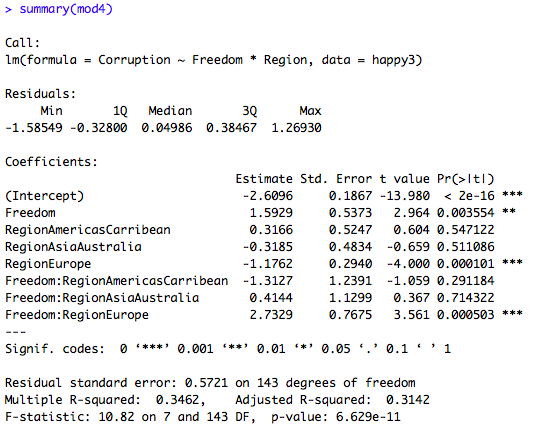
The QQ plot and result of shapiro test shows now the residuals are normally distributed (p-value of shapiro test is larger than 0.05).

**Run the model:**



The result shows, both freedom and family have a significant effect on the log transformed corruption. Freedom has a positive effect on log corruption, but family has a negative effect on log corruption. When both family, GDP and Freedom are zero, log corruption is the intercept value: -2.8848. When Family increases one unit, the log corruption decreases 0.5245 units. When Freedom increases one unit, the log corruption increases 2.4345 units. The adjusted R-square is 0.2396, this means this model can explain 23.96% of 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? **Happy**

****

When the region is AfricaMideast and Freedom is zero, the log transformed corruption is the intercept value: -2.6096, and this value is significantly different from 0. When the region is AfricaMidest, the log corruption will increase 1.5929 units, if freedom increases 1 unit. This increasing is significant (p-value=0.003554<0.05). When the freedom is zero, and we change location to Europe, the log corruption (intercept of Europe) is -1.1762 lower than AfricaMidest. This difference is significant. So the intercept of Europe is -2.6096-1.1762=-3.7858.

In Europe, when freedom increases one unit, the log corruption of Europe will increase 2.7329+1.5929=4.3258 units, and this increasing is significant. Changing region to other areas like AmericasCarribean, and AsiaAustralia will not cause significant change in both intercept and slope.

The adjusted R square is 0.3142, this means the model can explain 31.42% of data.

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**

Because the dependent variable is binary (the malign column includes just 0, and 1) so I consider to use logistic model (GLM) instead of a normal linear model.

**Hypothesis:**

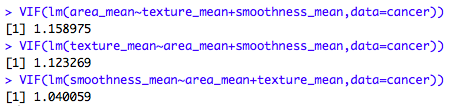
Null hypothesis: area, texture, and smoothness are not associated with whether a breast cancer tumor is malignant or not.

Alternative hypothesis: area, texture and smoothness are associated with whether a breast cancer tumor is malignant or not.

**Assumptions and Checking**:

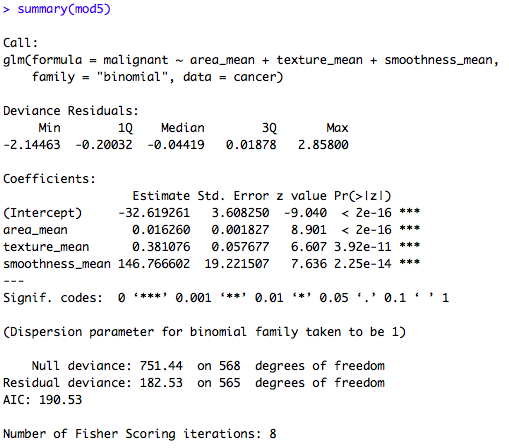
1. Observations are independent: I assume each observation is independent with each other.
2. Samples are randomly selected from population: I assume the random selection for my test.
3. Dependent variable meets the distribution of binomial distribution: since in data there is just 0 and 1, so I think the dependent variable meets the binomial distribution.

**Multicollinearity test:**



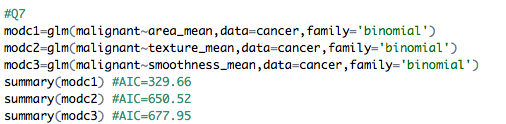
In all multicollinearity tests, the VIF values are smaller than 10, so we can think there is no multicollinearity among different independent variables.

**Generalized linear model:**



The result means, when area, texture, and smoothness are controlled or equal to zero, the malignant should be -32.619261. area\_mean, texture\_mean, and smoothness\_mean can all have significant effect on malignant. When area\_mean increases 1 unit, malignant will increase 0.016160 units; when texture\_mean increases 1 unit, malignant will increase 0.381076 units; when smoothness\_mean increases 1 unit, malignant will increase 146.766602 units.

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.**



when I run glm for three independent variables separately, the summary will show AIC of these three models. Because the AIC of modc1 is the lowest, so the model that relates malignant and area\_mean loss least information, so area\_mean, among my three independent variables, is most important to explain whether a breast cancer tumor is malignant or not.