Prakruthi Shetty

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

Dependent Variable: Gender (Categorical)

Independent Variable: Unruly\_Child (Categorical)

Test: Chi Square test

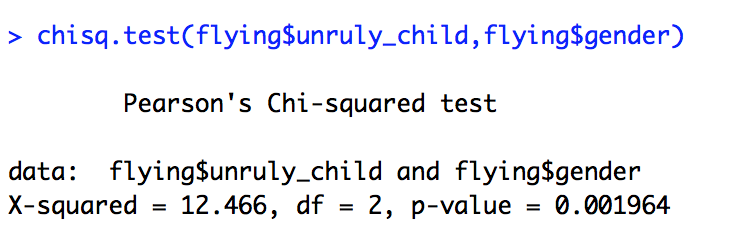
Assumptions:

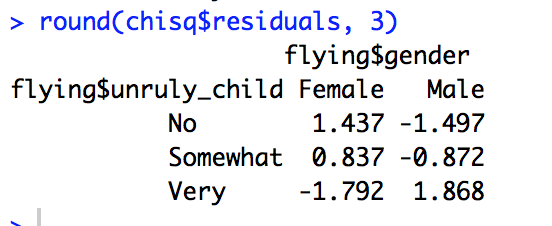
1. Observations are independent [ASSUMED TRUE]
2. Sample size is greater than is large (>30)

Hypothesis

Null : There is no association between gender and if an individual believes its rude to bring an unruly child on the plane.

Alternate: There is an association between gender and if an individual believes its rude to bring an unruly child on the plane

Test result:

Since the p-value is less than 0.05, we can reject the null hypothesis, thus there exists a significant association between gender and if an individual believes its rude to bring an unruly child on the plane.

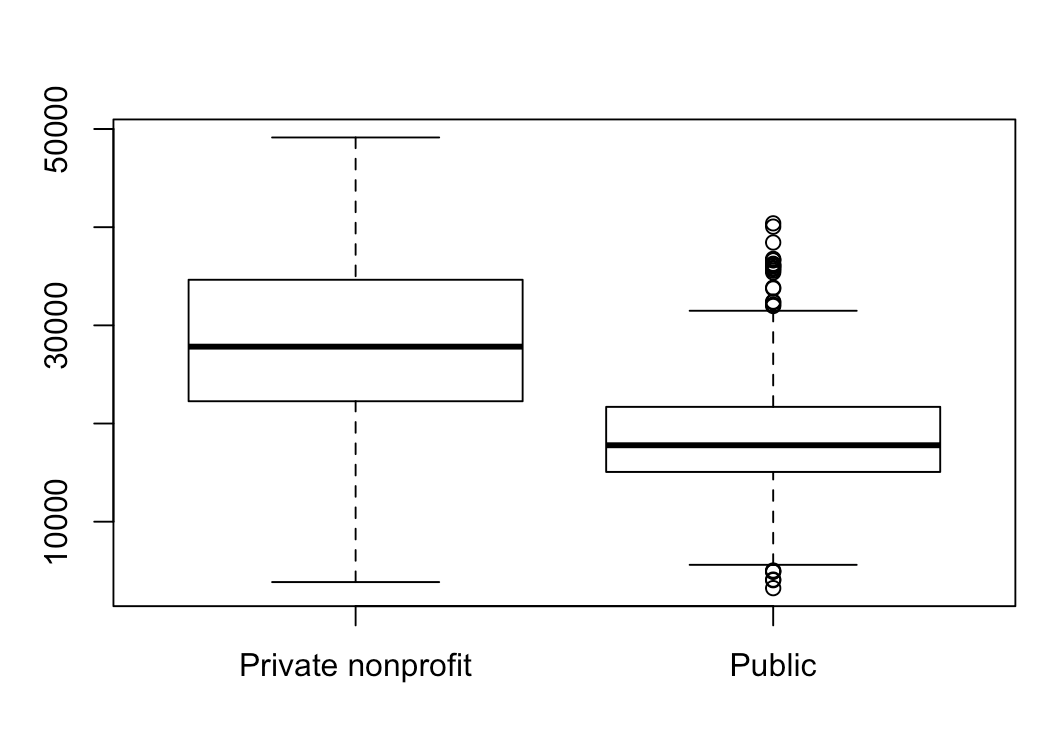
Males tend to think that brining an unruly child is more rude than Females.

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

Dependent Variable: Type of institution (Categorical)

Independent Variable: Tuition (Continuos)

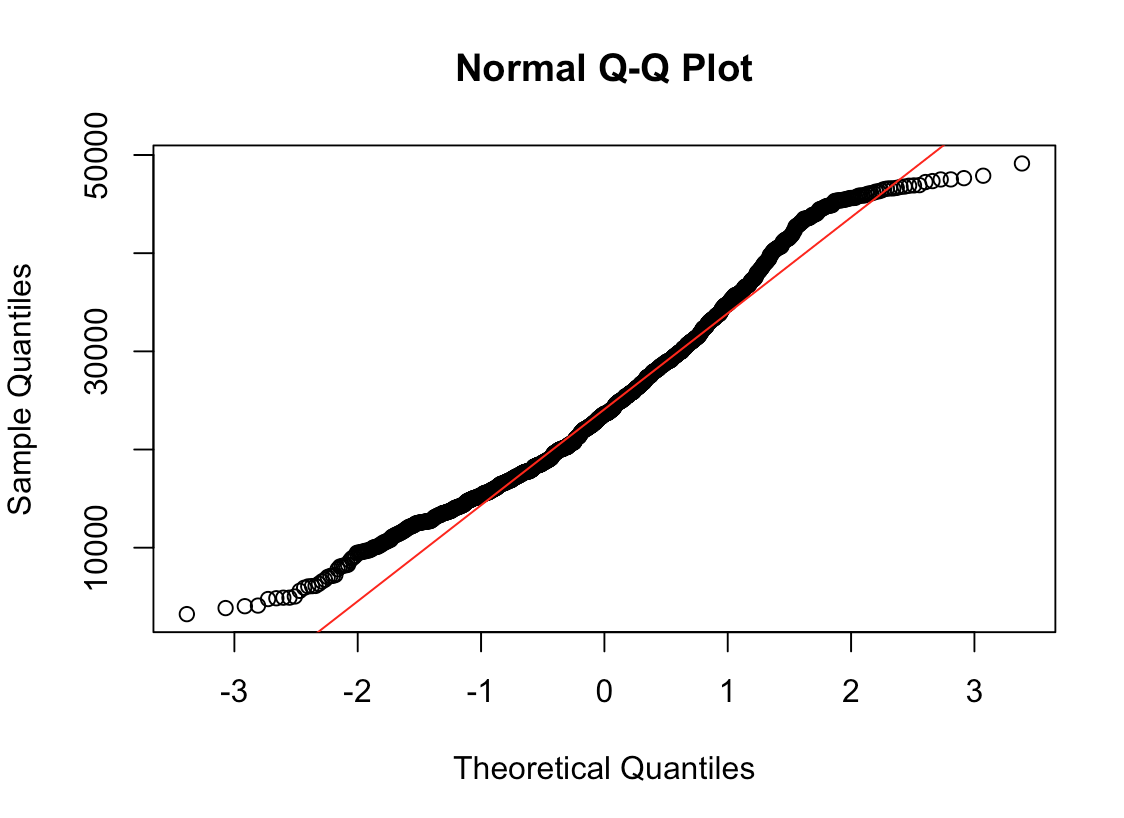
Test: T-Test

Plot: Box plot

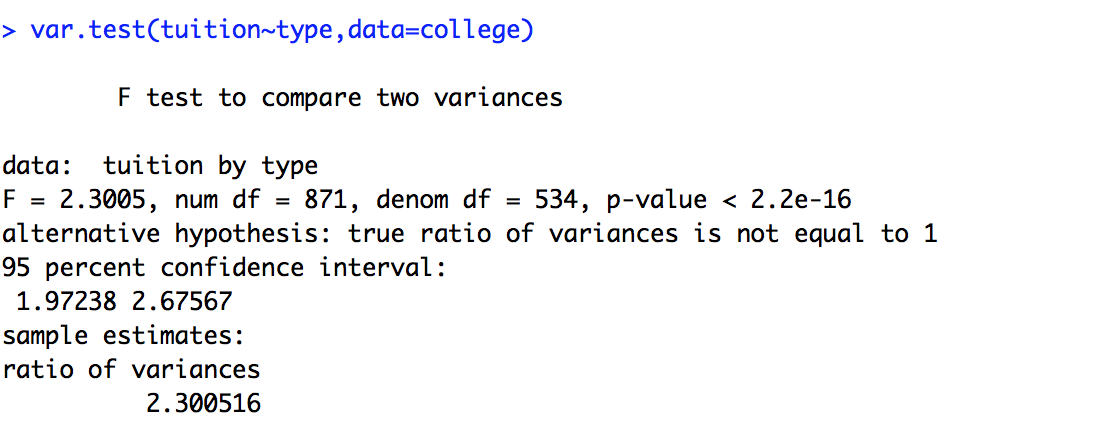
Assumptions:

1. Normality - TRUE

More than 30 observations.

Also can be assumed to be normal by looking at the qqplot.

1. Samples have equal variance - FALSE

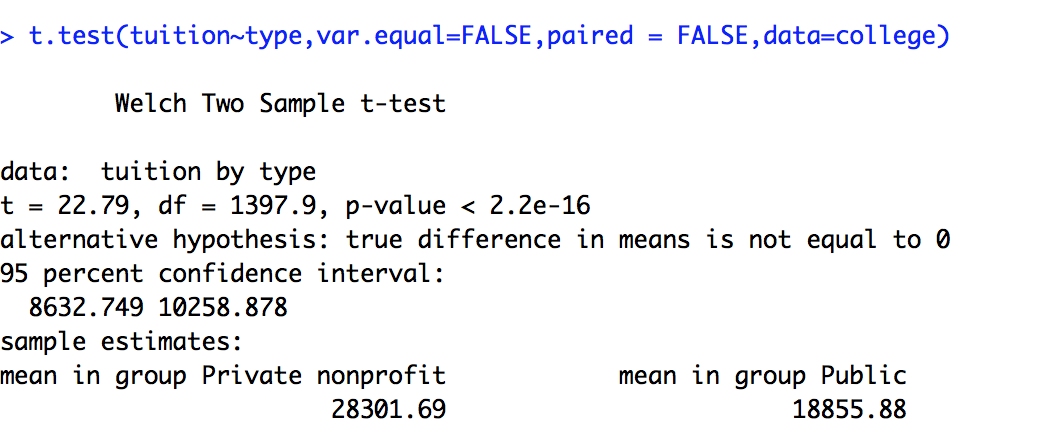
Thus using the Welch’s t-test

3. Independent observations - Assumed TRUE

Hypothesis

Null : There is no significant difference in tuition by type of institution.

Alternate: There is significant difference in tuition by type of institution.

Test result:

Since p:2.2e-16 < 0.05, we reject the null hypothesis.

Thus there exists significant difference in tuition by type of institution.

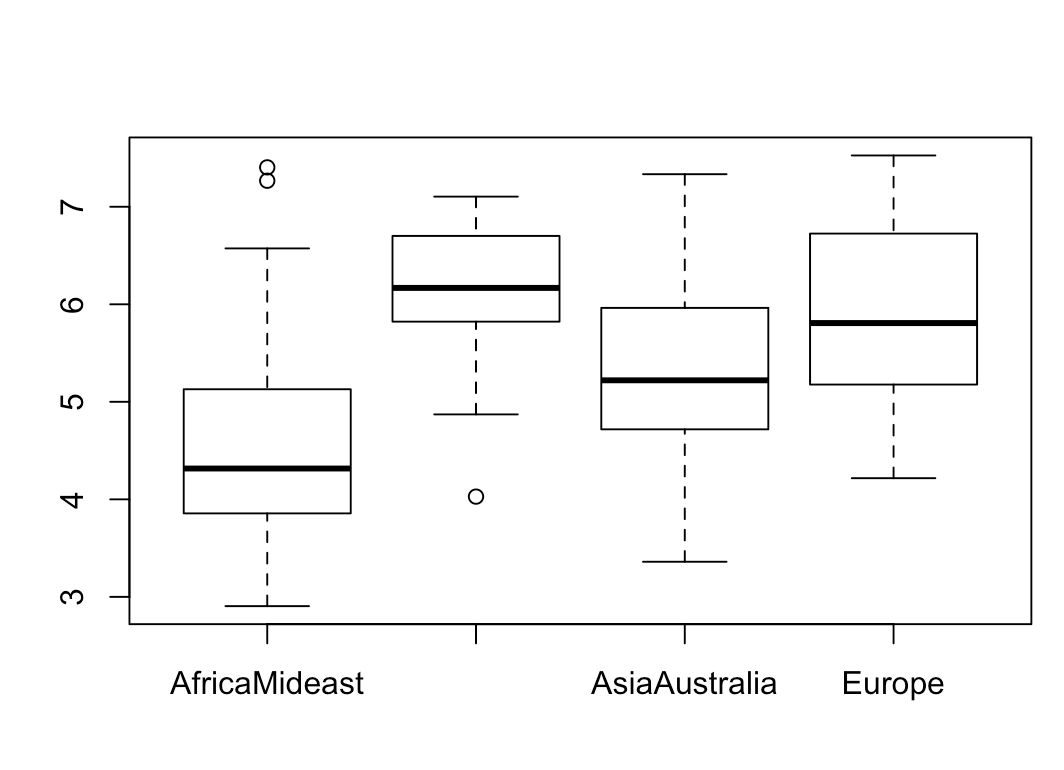
Private non profit institutes have higher tuition as indicated by the higher tuition mean of 28301.69 compared to Public institutes with mean tuition of 18855.88.

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

Dependent Variable: Region (Categorical)

Independent Variable: Hscore (Continuos)

Test: Anova

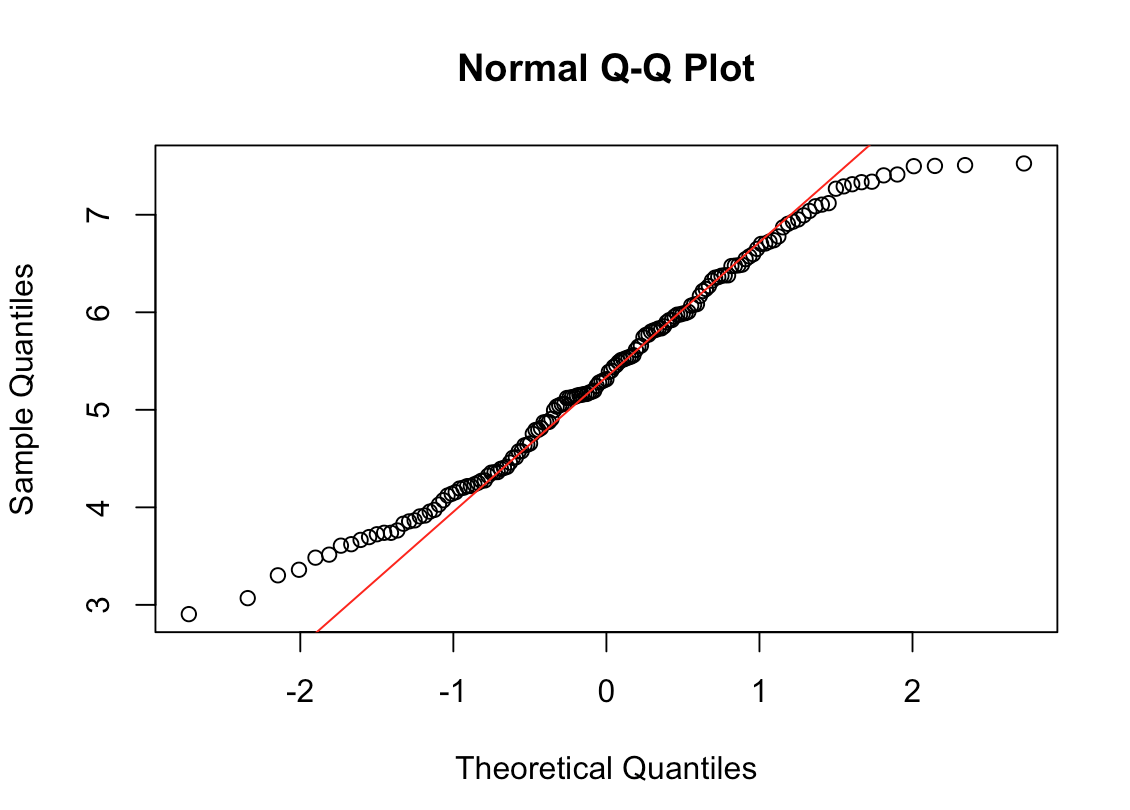
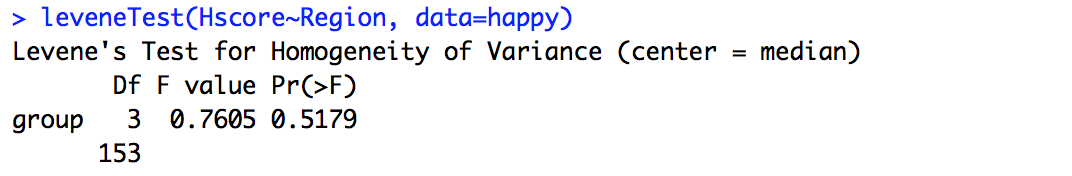
Plot: Box plot

Assumptions:

1. Normality - TRUE

More than 30 observations.

Also can be assumed to be normal by looking at the qqplot.

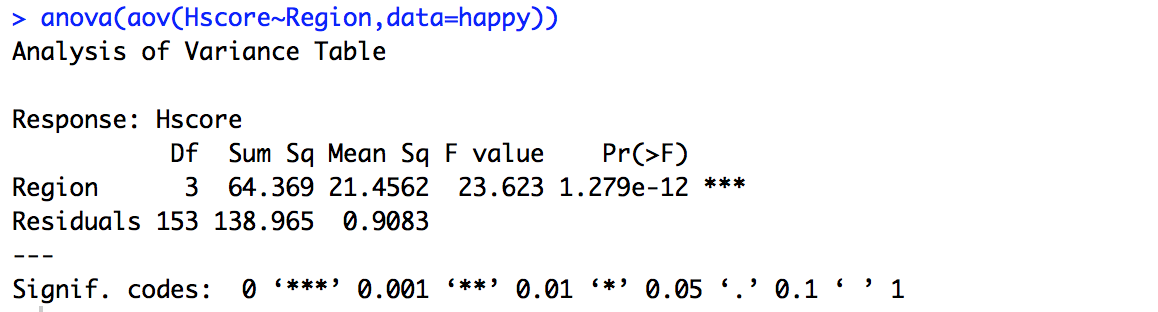
2. Samples have equal variance - TRUE

3. Independent observations - Assumed True

Hypothesis

Null : Hscore doesn't differ depending on the region.

Alternate: Hscore differs depending on the region.

Test result:

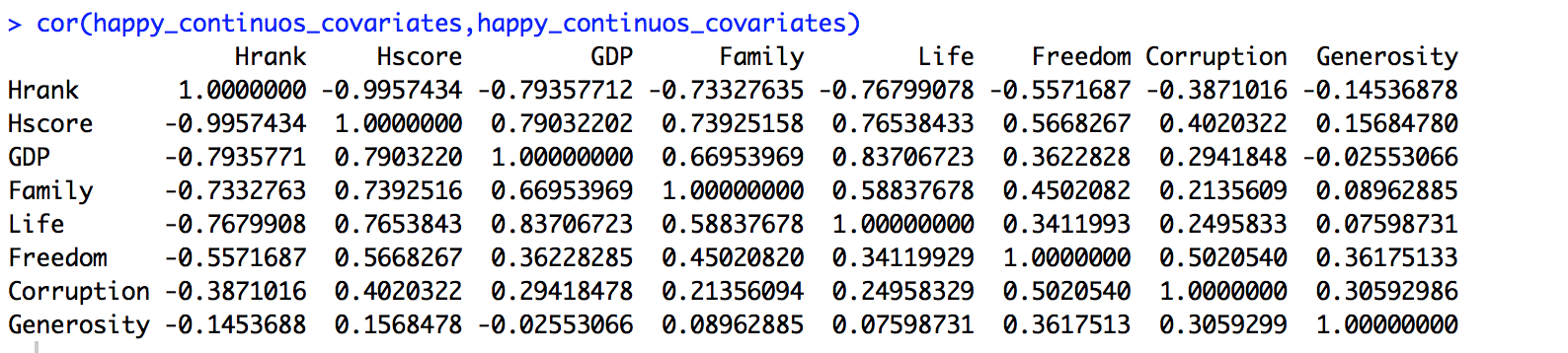
P-value: 1.279e-12 < 0.05, thus we reject the null hypothesis and hence there is a significant difference in happiness score by region.

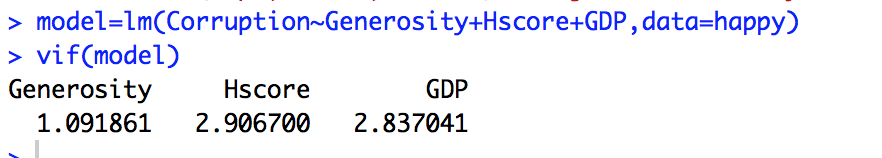
1. What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model. happy

Dependent Variable: Corruption Levels (Continuos)

Independent Variable: Hscore, Generosity, GDP (Continuos)

Test: Multiple Linear regression

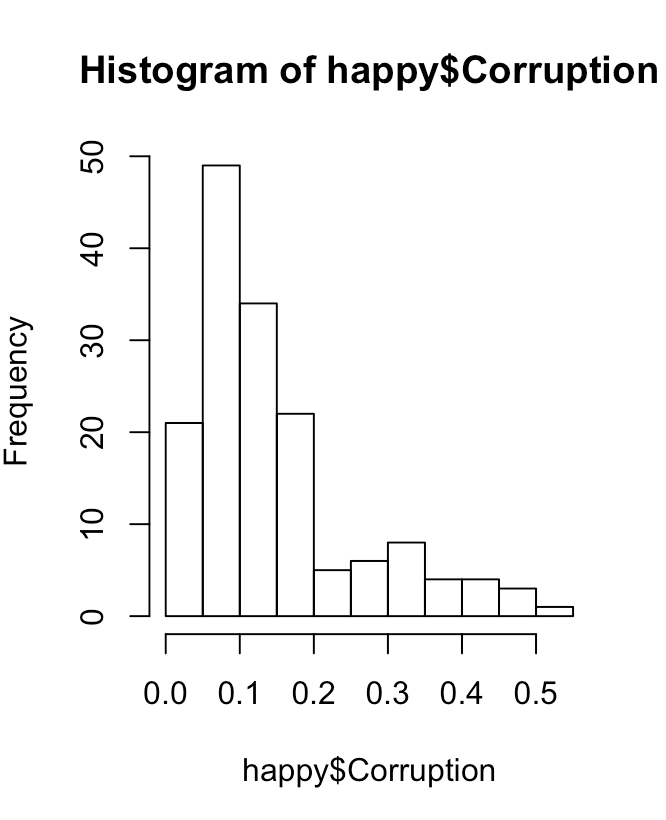
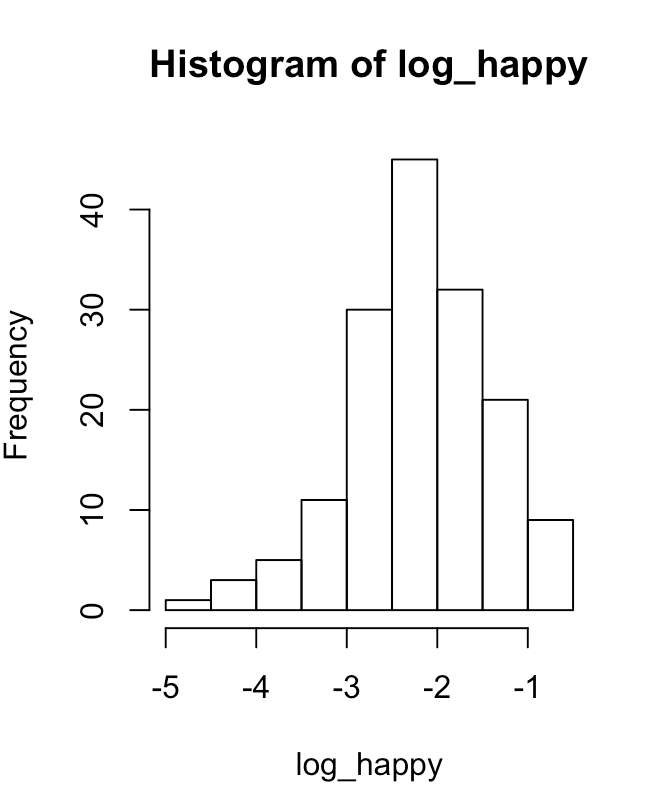
Correlation:

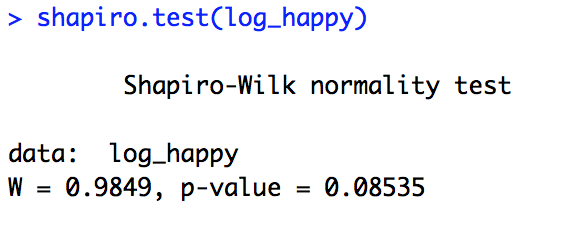
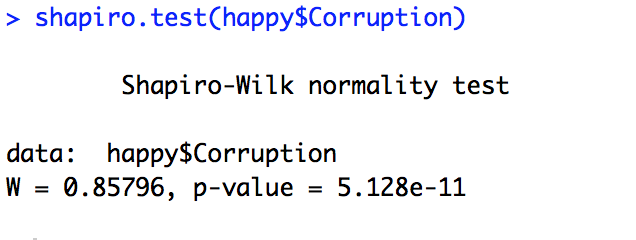
Covariates chosen based on lower correlational value: Generosity, GDP, Hscore.

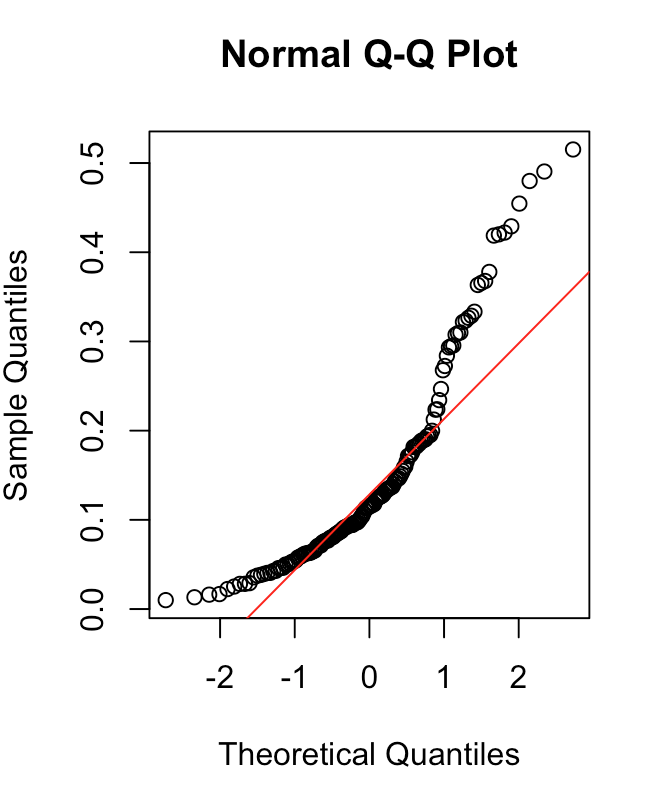
Since VIF value for all 3 variable is less than 5, the collinearity between the variables doesn't affect the model.

Assumptions:

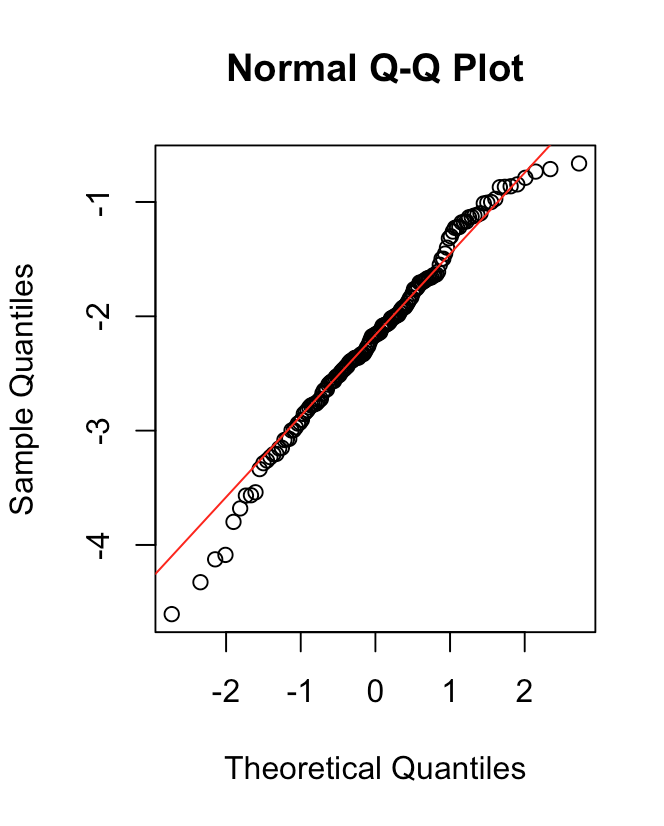
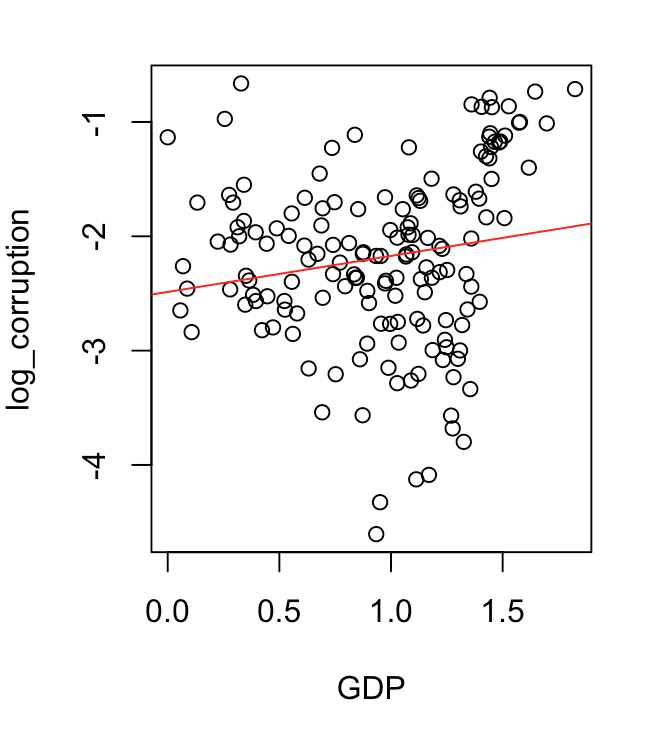
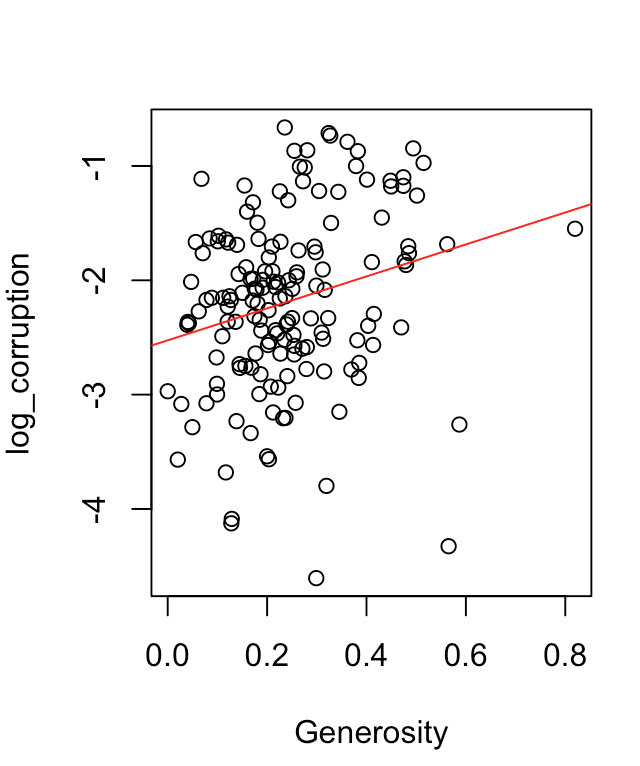
Since the y-variable(Corruption) was not normally distributed, transformed the y variable using log transformation.

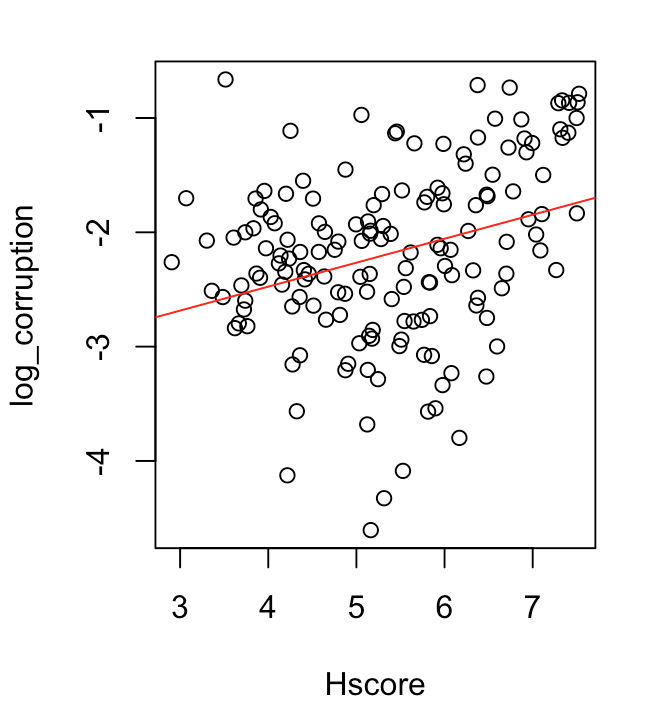
Before Transformation After log transformation



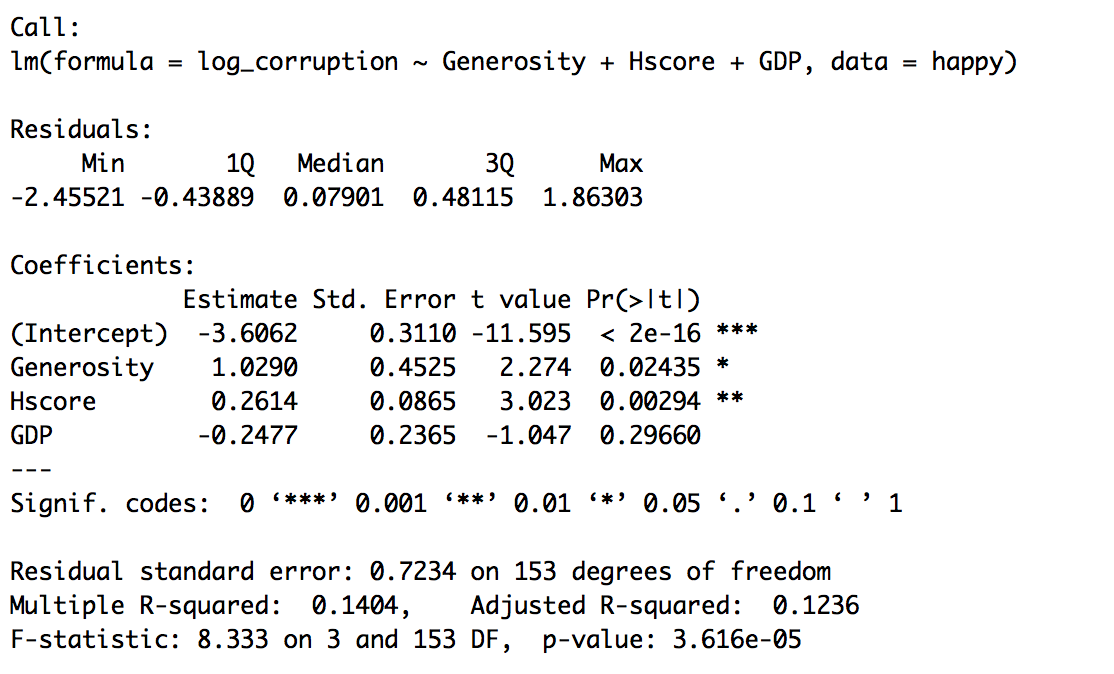


Linear relationship: After log transformation

Corruption vs GDP  Corruption vs Generosity

Corruption vs Hscore

All the above plots shows that the every independent variable is linearly related to the dependent variable. The correlation between dependent and independent variable is not high. Also the residuals seem to be homostedastic.

Test result:

The value ‘coefficient estimate of the intercept’, shows that the value of corruption is exp(-36062)=0.02716 [ i.e.log(y) = -3.6062] when Generosity=GDP=Hscore=0.

This is a significant value indicated by the p value and significant codes.

The value ‘coefficient estimate of the Generosity', shows that the value of Corruption increases by exp(1.0290) = 2.7983 [i.e.log(y) = 1.0290] with increase in Generosity value by 1 unit. This is a significant value indicated by the p value and significant codes.

The value ‘coefficient estimate of the Hscore’, shows that the value of Corruption increases by exp(0.2614)=1.2988 [i.e. log(y) = 0.2614] with increase in Hscore value by 1 unit. This is a significant value indicated by the p value and significant codes.

Since value ‘coefficient estimate of the GDP’ is not a significant value as indicated by the p value and significant codes, there is no effect of GDP on Corruption.

The coefficient standard error for each variable shows average amount by which the coefficient of estimates vary from the actual average value of dependent variable.

The coefficient t-value indicate how many standard deviations is the coefficient estimate away from 0.

**Thus factors that are significantly associated with a country’s corruption levels are Generosity and Hscore.**

Model Check:

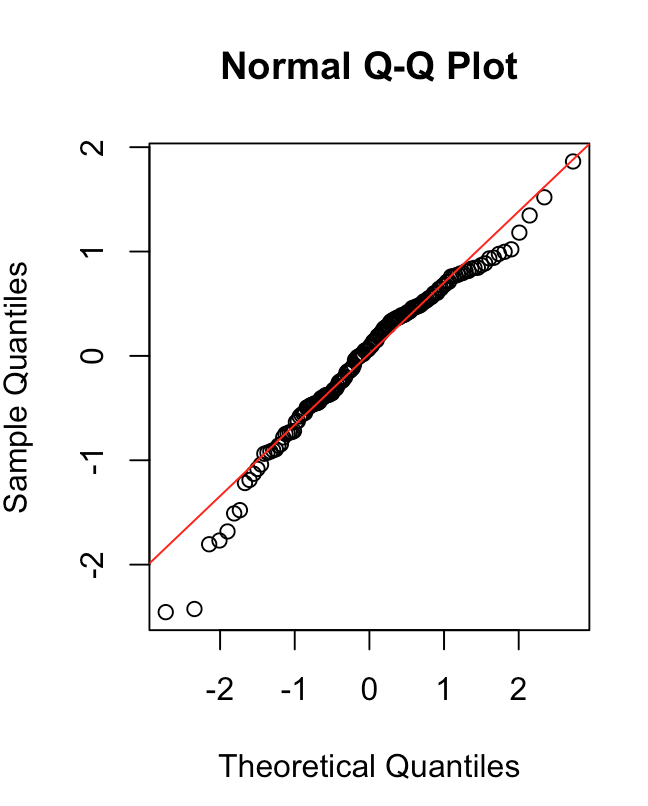
1. Residual homostedastic - TRUE

As seen by the bptest of the residuals, p: 0.05496, indicates that the residuals are homostedastic.

1. Residual independence - TRUE

As seen by the dwtest of the residuals, p-value: 0.286 > 0.05, since p-value is greater than 0.05, the null hypothesis is true and there doesn’t exists autocorrelation and the residuals are independent.

1. Normality of the residuals - TRUE

As shown by the qqplot below.

4. The f-statistic is 8.33 and p-value is 3.616e-05, indicating that we should reject the null hypothesis that the variables Generosity, GDP, Hscore collectively have no effect on Corruption

5. R-square value being 0.1404, indicating that the model explains only 14.4% of variance exhibited by the data.

Although the residuals are homostedastic, normally distributed and independent , this model not be the best fit as the R-square value is very low and explains only 14.4% of variance exhibited by 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? happy

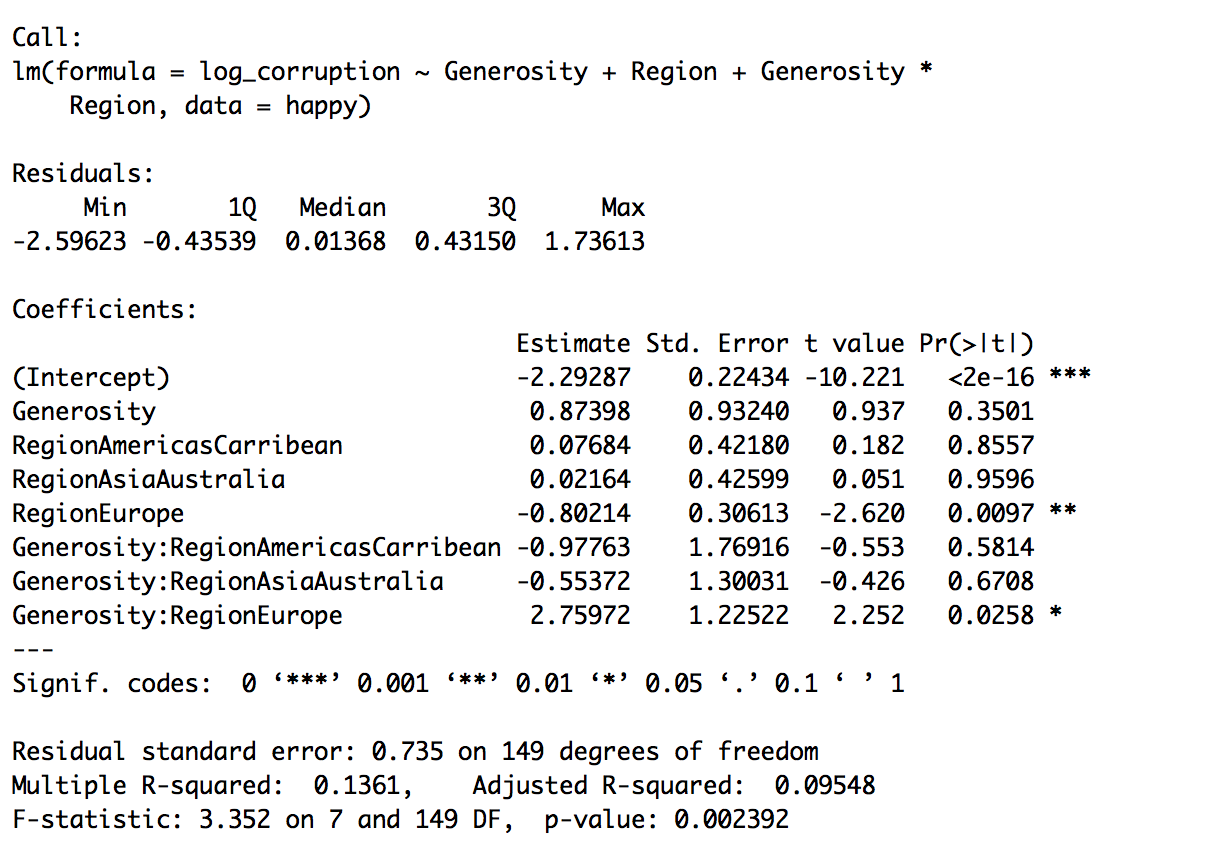
Dependent variable: Corruption(Continuos)

Independent variable: Generosity(Continuos), Region(Categorical)

Test: Ancova

Assumptions:

1. Since Corruption is not normally distributed, we log transform the y variable which follows normal distribution. (in detail in the above problem)
2. Linear regression assumptions of normality of dependent variable, correlation between variables and linear relationship between continuos variable are met as shows for the above problem.

Test Results

The significant effects that we see from the above model are :

1. When Generosity is zero, there is a significant effect on Corruption for the region AfricaMideast. The intercept value exp(-2.29287) = 0.10097 indicates the corruption value for Region AfricaMideast when Generosity is zero.
2. Another region that has significant effect on corruption is Europe. When generosity is zero, this region, as shown by its coefficient value has a corruption value of exp(-0.80214)=0.44836.

From above two observations that are significant there is an effect on corruption across region when generosity is =0.

1. The coefficient of Generosity:RegionEurope is significant and this means that the slope for the effect of Generosity on Corruption is significantly different for Europe than other regions. The slope (or effect of Generosity on Corruption) for Europe is exp(0.87398)=2.3964 plus exp(2.75972)=15.7954

Model Fit

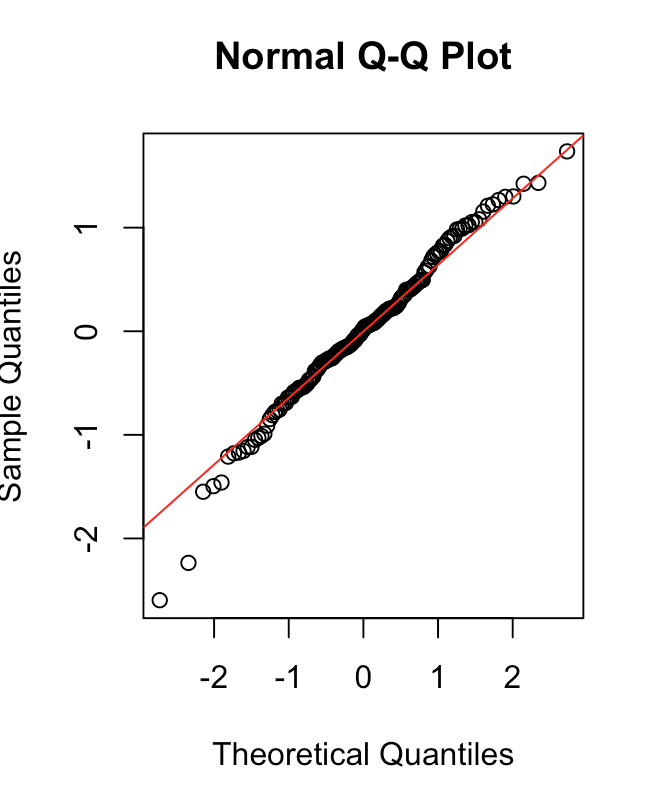
1. Residual homostedastic - FALSE

As seen by the bptest of the residuals, p:0.01836<0.05, indicates that the residuals are not homostedastic.

1. Residual independence - TRUE

As seen by the dwtest of the residuals, p-value: 0.9724 > 0.05, since p-value is greater than 0.05, the null hypothesis is true and there doesn’t exists autocorrelation and the residuals are independent.

1. Normality of the residuals - TRUE

As shown by the qqplot below.

4. The f-statistic is 3.352 and p-value is 0.002, indicating that we should reject the null hypothesis that the variables Generosity, Region and the interaction between Region&Generosity have no effect on Corruption

5. R-square value being 0.1361, indicating that the model explains only 13.61% of variance exhibited by the data.

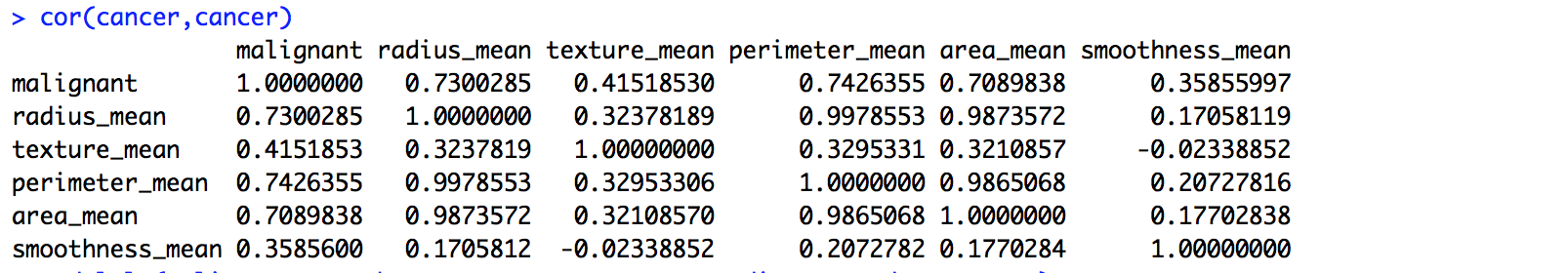
This model might not be a great fit in explaining corruption as the R2 value is very low and residuals are heterostedastic.

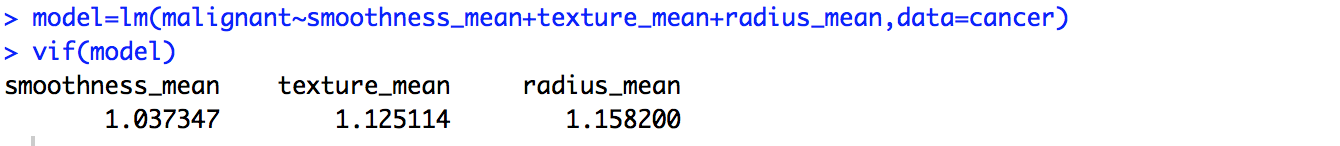
1. Which factors are significantly associated with whether a breast cancer tumor is malignant or not? Choose three continuous independent variables to include in your model. cancer

Dependent Variable: Malignant Levels (Continuos)

Independent Variable: smoothness\_mean, texture\_mean, radius\_mean (Continuos)

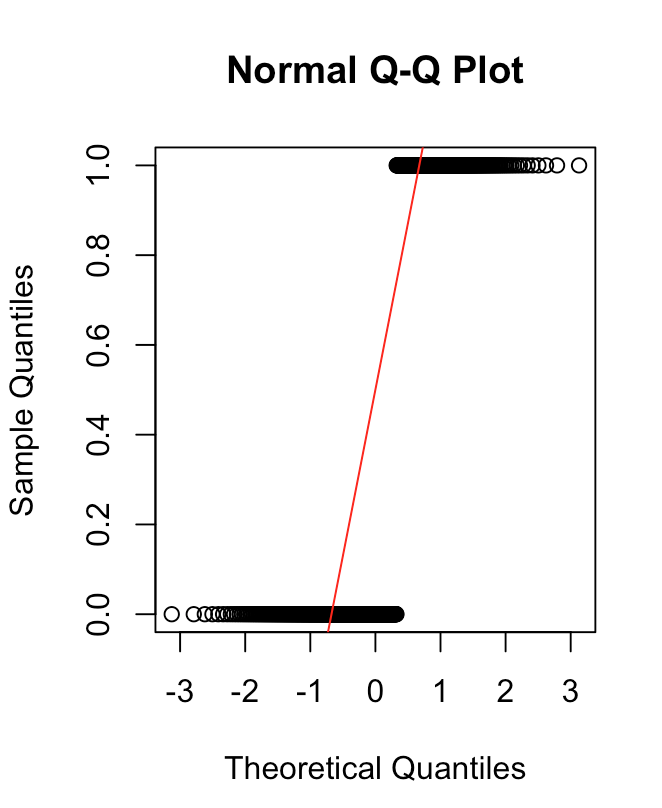
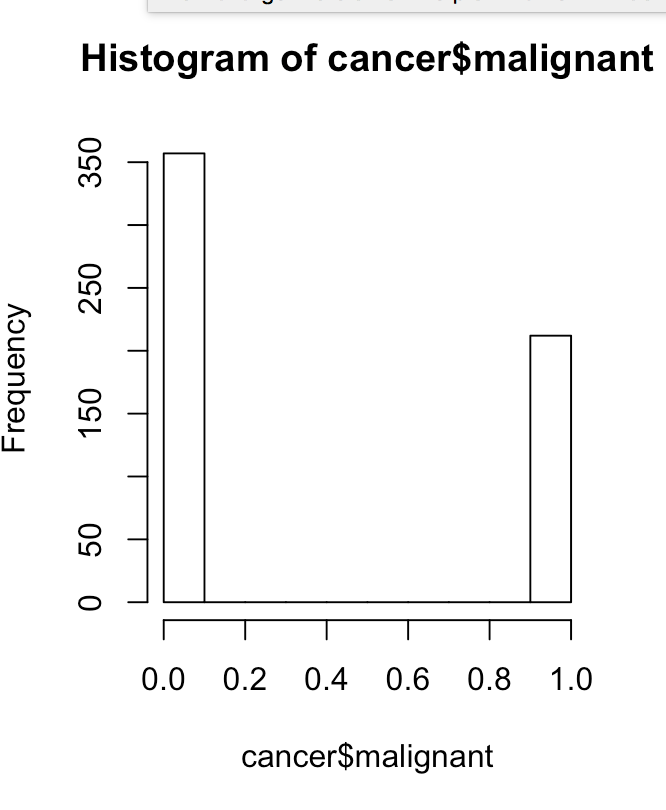
Test: Multiple Linear regression (Generalized)

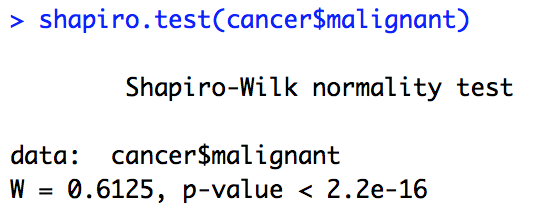
Correlation:

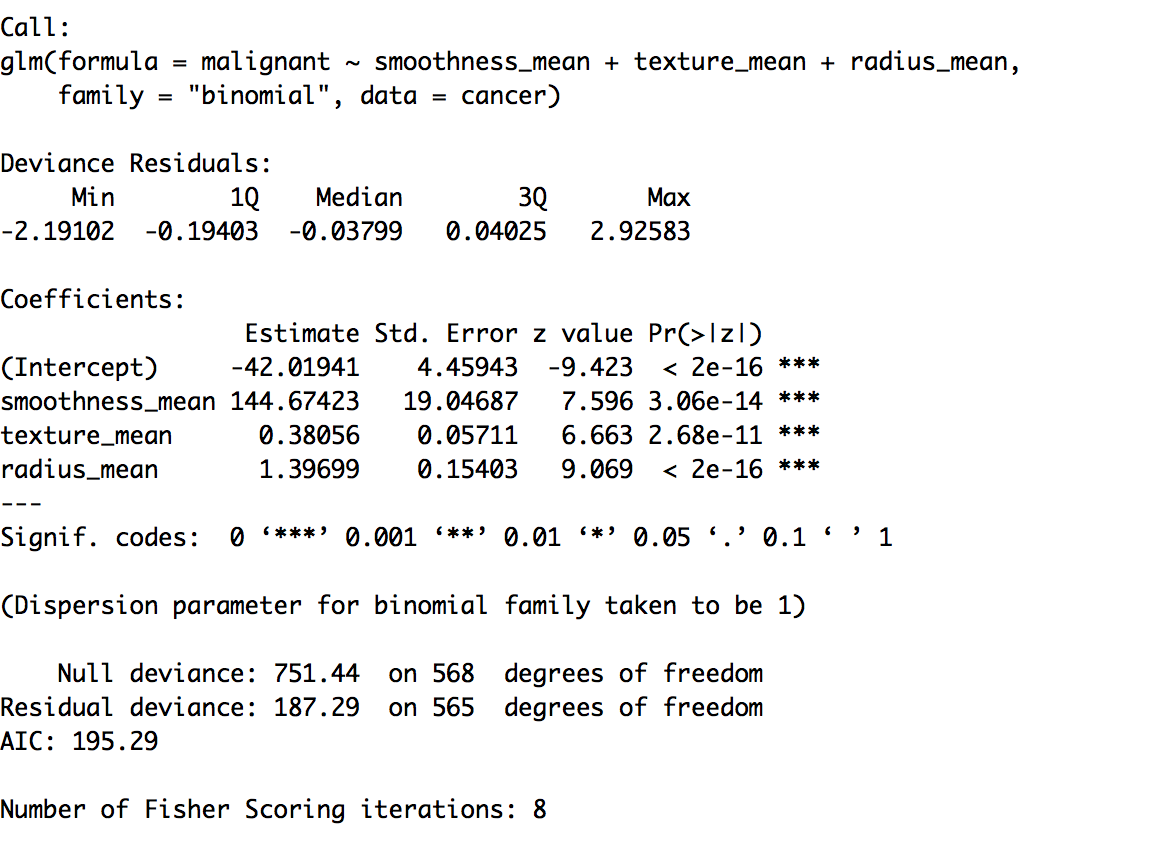
Covariates chosen based on lower correlational value: smoothness\_mean, texture\_mean, radius\_mean.

Since VIF value for all 3 variable is less than 5, the collinearity between the variables doesn't affect the model.

Assumptions: Check for normality of y variable(malignant)

Since the y-variable(Corruption) was not normally distributed and since the histogram and qqplot looks like a binomial distribution, as shown below., using GLM with binomial distribution and using logic link function.



Test Result:

All the independent variable, i.e. smoothness\_mean, texture\_mean, radius\_mean have a significant effect on the dependent variable ‘malignant’.

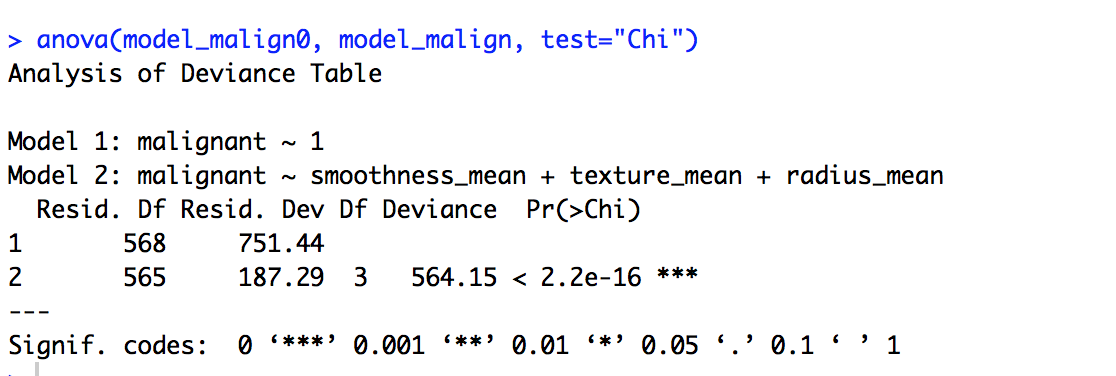
The intercept indicates that the log odds of a tumor being malignant is -42.01941 when all 3 variables smoothness\_mean, texture\_mean, radius\_mean is 0. This is a significant effect.

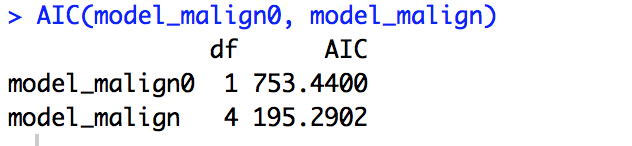
The coefficient estimate for smoothness\_mean indicates that for an increase in one unit of smoothness\_mean the log odds of a tumor being malignant is 144.674. This is a significant effect.

The coefficient estimate for texture\_mean indicates that for an increase in one unit of texture\_mean the log odds of a tumor being malignant is 0.38056. This is a significant effect.

The coefficient estimate for radius\_mean indicates that for an increase in one unit of radius\_mean the log odds of a tumor being malignant is 1.39699. This is a significant effect.

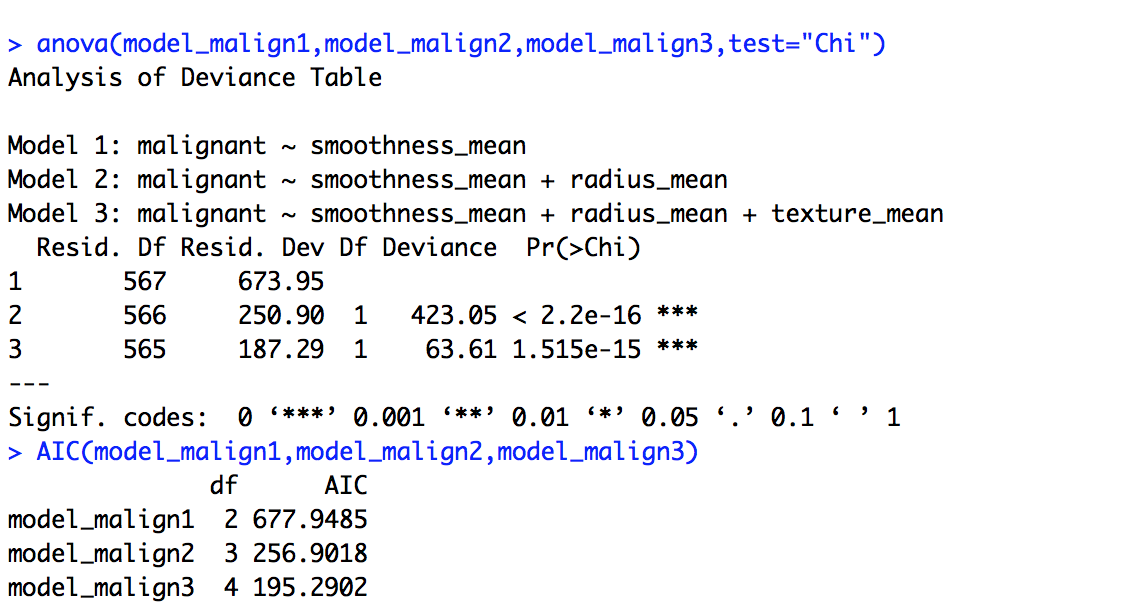
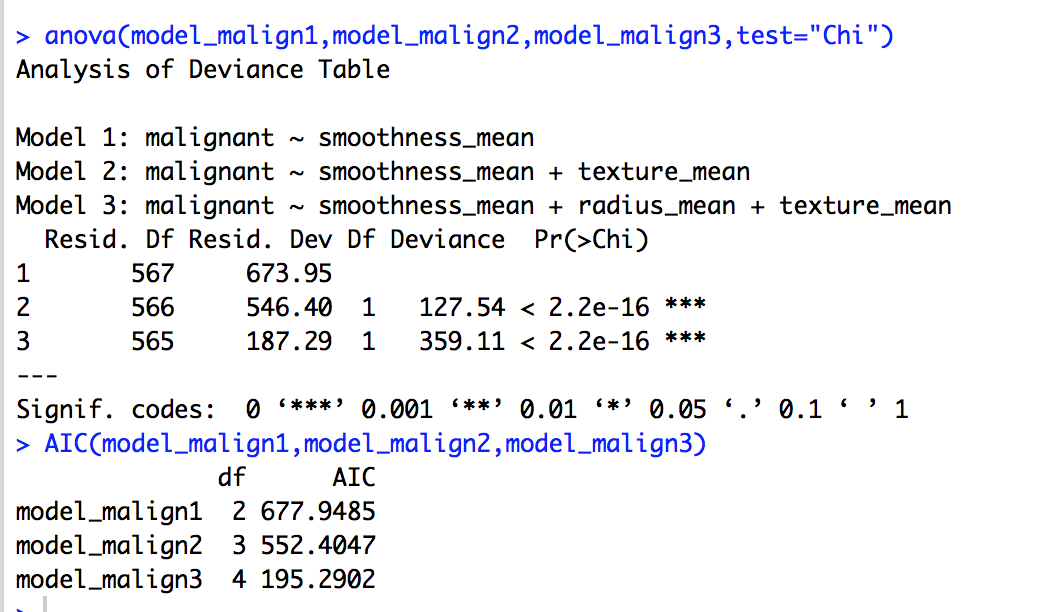
Thus all 3 factors: smoothness\_mean, texture\_mean, radius\_mean are significantly associated with whether a breast cancer tumor is malignant or not.

Model Fit: Good fit



From the above results the model chosen seems to be a good fit, as we see that including smoothness\_mean, texture\_mean, radius\_mean as independent variables (model\_malign) could increase the likelihood of the model and this increase is significant as indicated by the p-value. Also the AIC test indicates model\_malign to have low AIC score 195.2902, way less compared to model\_malign 0.

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.

radius\_mean is the most important variable among smoothness\_mean, texture\_mean, radius\_mean.

From the above tests of different model comparisons, including radius\_mean as an independent variable increases the likelihood of the model and this increase is significant as indicated by the p-value of anova comparison of models.

Also the residual deviance of the model decreases by a large number when the radius\_mean is included indicating that this variable is most important in explaining whether a breast cancer tumor is malignant or not.

This can also be observed by comparing the AIC models as the AIC value decreases by large number when the radius\_mean in included in the model, thus making that model a better fit when radius\_mean is included in explaining whether a breast cancer tumor is malignant or not.