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**

Since both independent (gender) and dependent (unruly\_child) variables are categorical, chi-square test for independence will be used to analyze the association.

**Hypothesis:**

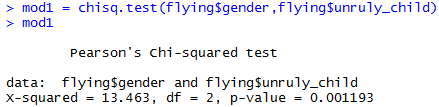
Null hypothesis: Whether people think it is rude to bring an unruly child on the plane is not different between males and females.

Alternative hypothesis: Whether people think it is rude to bring an unruly child on the plane is different between males and females.

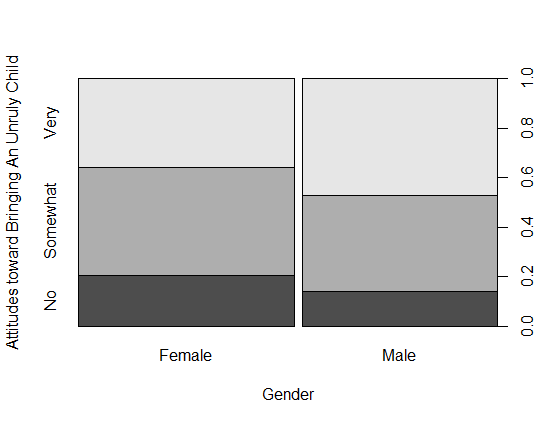
**Assumption Checking:**

1. Samples are randomly selected from the population: This is unlikely to test; however, we assume that samples from the dataset are selected randomly.
2. Each observation is sampled independently: This is unlikely to test; however, samples are collected individually, so we assume independence of observations.
3. No structural zeros: Data are factors.

**Test Results:**



The p-value of chi-square test = 0.001193 < 0.05, so we reject the null hypothesis. Therefore, whether people think it is rude to bring an unruly child on the plane is significantly different between males and females.



Based on the plot, males have higher percentage than females who think bringing an unruly child on the plane is very or somewhat rude. Therefore, males tend to think that bringing an unruly child on the plane is more rude.

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

Since the independent variable is categorical (Public & Private nonprofit) and the dependent variable is continuous (tuition), t-test will be used to analyze the difference.

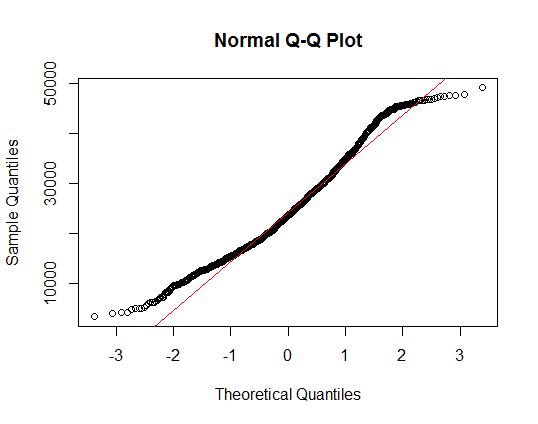
**Hypothesis:**

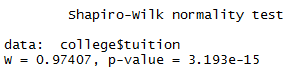
Null hypothesis: The tuition fee is not different between public and private institutions.

Alternative hypothesis: The tuition fee is different between public and private institutions.

**Assumption Checking:**

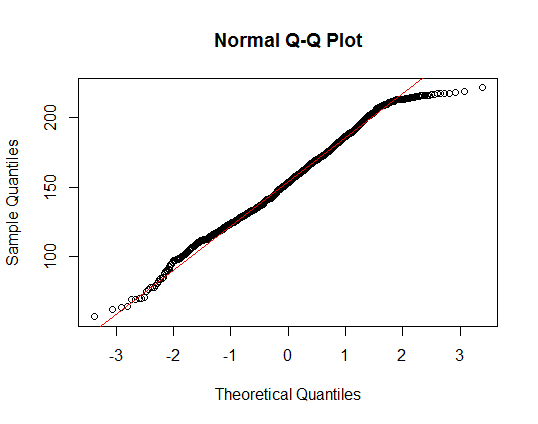
1. Samples are randomly selected from the population: This is unlikely to test; however, we assume that samples from the dataset are selected randomly.
2. Each observation is sampled independently: This is unlikely to test; however, samples are collected by institutions separately, so we assume independence of observations.
3. Samples are normally distributed:

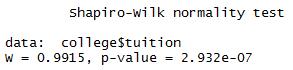




Based on the QQ plot and shapiro test (p-value = 3.193e-15 < 0.05), the dependent variable is not normally distributed.

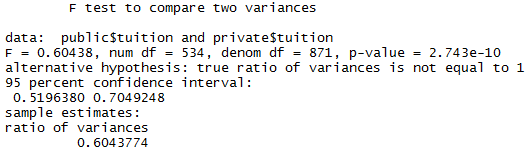
We transform the dependent variable by *square root*.





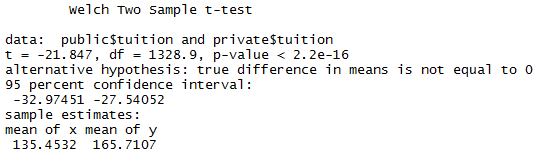
Based on the QQ plot and shapiro test of the *sqrt* transformed dependent variable, the p-value of shapiro test = 2.932e-07 < 0.05. The transformed dependent variable is still not normally distributed, but it is far better than the original one (p-value = 3.193e-15). In addition, since we have 1,407 observations (> 30), we can relax the assumption and assume that the dependent variable is normally distributed. Therefore, we will use the *sqrt* transformed dependent variable to run the following test.

1. Samples have equal variance between groups:



The p-value of variance test between public and private institutions = 2.743e-10 < 0.05, so we do not meet the equal variance assumption. Therefore, we will use Welch’s t-test.

**Test Results:**



The p-value of Welch’s t-test = 2.2e-16 < 0.05, so we reject the null hypothesis. Therefore, the tuition fee is significantly different between public and private institutions.



Based on the mean of *sqrt* transformed tuition in both institution types, private institutions have higher tuition fee than public institutions.

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

Since the independent variable of four levels is categorical (AfricaMideast, AmericasCarribean, AsiaAustralia & Europe) and the dependent variable is continuous (Hscore), ANOVA will be used to analyze the difference.

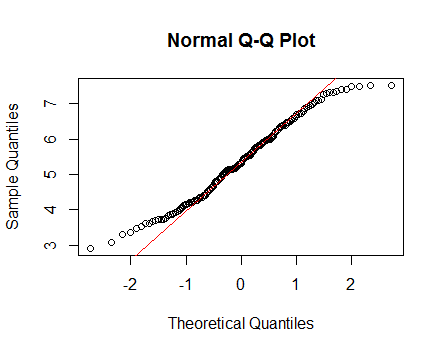
**Hypothesis:**

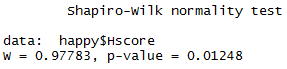
Null hypothesis: The level of happiness is not different between regions.

Alternative hypothesis: The level of happiness is different between regions.

**Assumption Checking:**

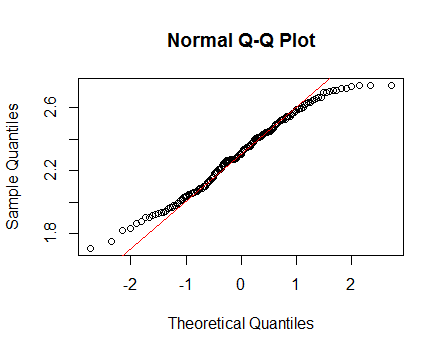
1. Samples are randomly selected from the population: This is unlikely to test; however, we assume that samples from the dataset are selected randomly.
2. Each observation is sampled independently: This is unlikely to test; however, samples are collected by countries separately, so we assume independence of observations.
3. Samples are normally distributed:

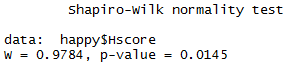




Based on the QQ plot and shapiro test (p-value = 0.01245 < 0.05), the dependent variable is not normally distributed.

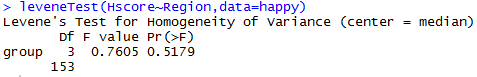
We transform the dependent variable by *square root*.





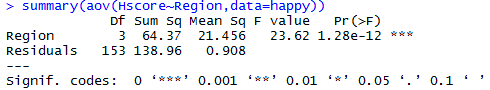
Based on the QQ plot and shapiro test of the *sqrt* transformed dependent variable, the p-value of shapiro test = 0.0145 < 0.05. The transformed dependent variable is still not normally distributed, and it is not better than the original one (p-value = 0.01245). Since we have 157 observations (> 30) and each level of regions has over 15 observations (AfricaMideast: 58, AmericasCarribean: 25, AsiaAustralia: 24 & Europe: 50), we can relax the assumption and assume that the dependent variable is normally distributed. Therefore, we will use the original form of dependent variable to run the following test.

1. Samples have equal variance between groups:



The p-value of Levene test = 0.5179 > 0.05, so we met the equal variance assumption.

**Test Results:**



The p-value of ANOVA = 1.28e-12 < 0.05, so we reject the null hypothesis. Therefore, the level of happiness is different between 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**

Since the dependent and three independent variables are continuous, multiple linear regression will be used to analyze the association. Three covariates that will be included in the multiple linear model as independent variables are GDP, Freedom, and Generosity.

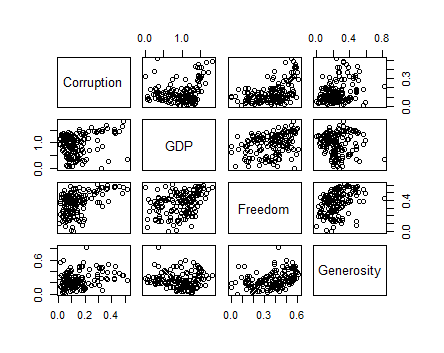
**Hypothesis:**

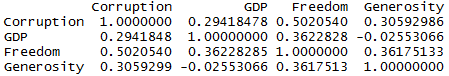
Null hypothesis: GDP, freedom, and generosity are not associated with a country’s corruption levels.

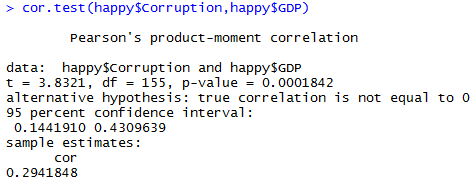
Alternative hypothesis: Either GDP, freedom, or generosity is associated with a country’s corruption levels.

**Assumption Checking:**

1. Samples are randomly selected from the population: This is unlikely to test; however, we assume that samples from the dataset are selected randomly.
2. Each observation is sampled independently: This is unlikely to test; however, samples are collected by countries separately, so we assume independence of observations.
3. There is a linear relationship between independent and dependent variables:

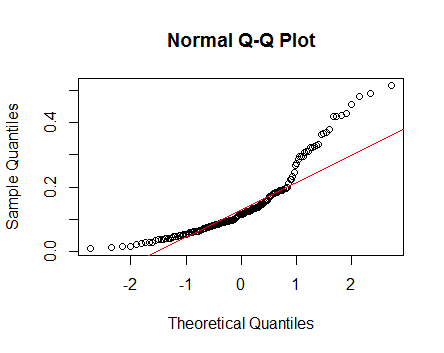


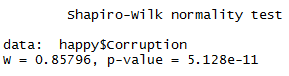




Based on the correlation plot, correlation coefficients of Corruption with each independent variable, and the correlation test of the smallest correlation coefficient (0.2941848 between Corruption and GDP) (p-value = 0.0001842 < 0.05), there is a linear relationship between Corruption and each independent variable.

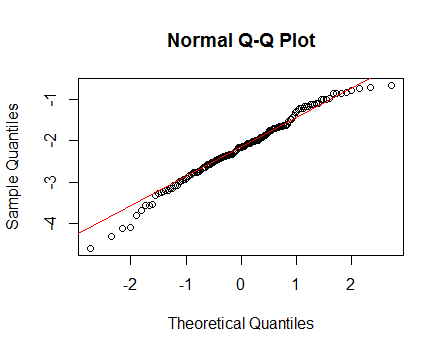
1. Samples are normally distributed (Normality of the dependent variable):

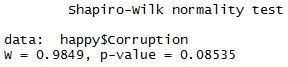




Based on the QQ plot and shapiro test (p-value = 5.128e-11 < 0.05), the dependent variable is not normally distributed.

We transform the dependent variable by *log*.

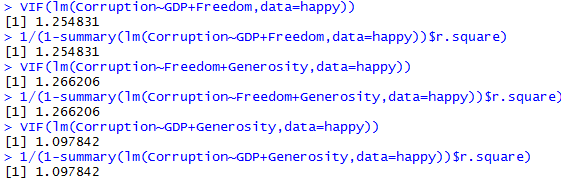




Based on the QQ plot and shapiro test of the *log* transformed dependent variable, the p-value of shapiro test = 0.08535 > 0.05. The transformed dependent variable became normally distributed. Therefore, we will use the *log* transformed dependent variable to run the following test.

1. There is no multi-collinearity between independent variables:

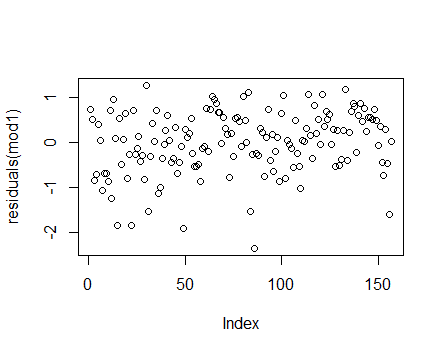


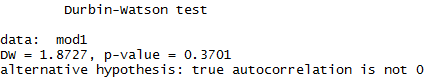


All the correlation coefficients between the independent variables are < 0.5, and VIF between two of the independent variables = 1.254831, 1.266206, and 1.097842 (< 10), so there is no multi-collinearity between the independent variables.

*Run the multiple linear model*

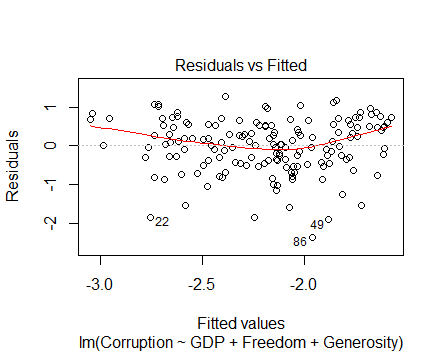
1. Residual independency:

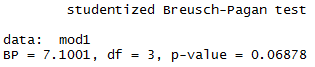




Based on the residual plot and dw test (p-value = 0.3701 > 0.05), there is no autocorrelation between the residuals. Therefore, the residuals are independent.

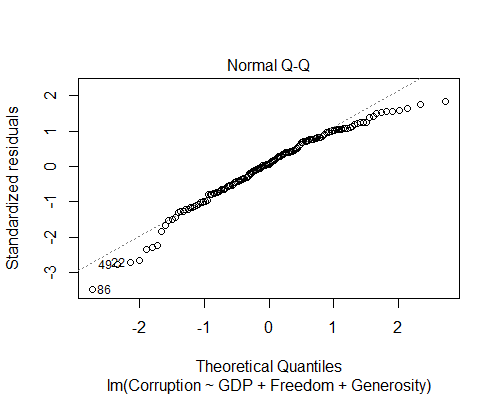
1. Residual homoscedasticity:

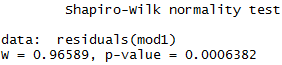




Based on the residual plot and bp test (p-value = 0.06878 > 0.05), the residuals are homoscedatic.

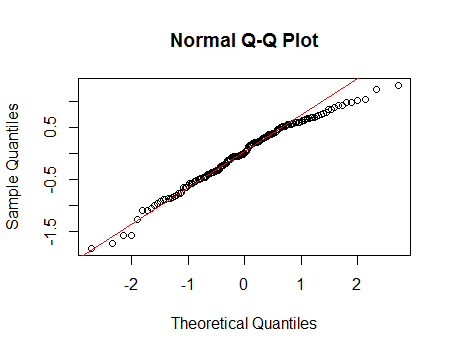
1. Residual normality:

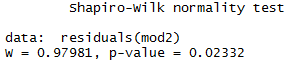




Based on the QQ plot and shapiro test (p-value = 0.0006382 < 0.05), the residuals are not normally distributed.

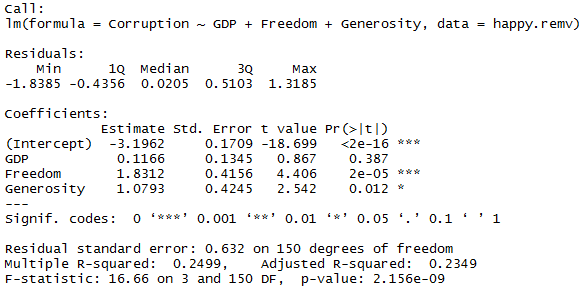
We remove the outliers based on the QQ plot.





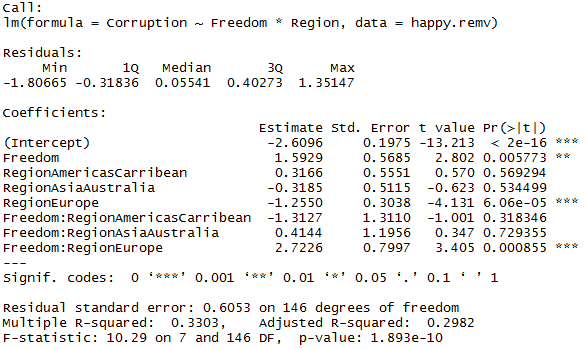
Based on the QQ plot and shapiro test (p-value = 0.02332 > 0.05), the residuals became normally distributed. Therefore, we will use the outliers-removed dataset to run the multiple linear regression.

**Test Results:**



When all the independent variables are controlled, the *log* transformed Corruption level is -3.1962 (intercept). Based on the p-values of coefficients, Freedom and Generosity have significant effects on the Corruption level, while GDP does not. When increasing one unit of Freedom (GDP and Generosity are controlled), the *log* transformed Corruption level will increase 1.8312 (slope of Freedom). When increasing one unit of Generosity (GDP and Freedom are controlled), the *log* transformed Corruption level will increase 1.0793 (slope of Generosity). And the adjusted R-square is 0.2349, which means around 24% of the dependent variable (Corruption level) are explained by the model with GDP, Freedom, and Generosity as independent variables.

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 controlled, the *log* transformed Corruption level is -2.6096 (intercept). Based on the p-values of coefficients, only Freedom, Europe region, and interaction between Freedom and Europe region have significant effect on the Corruption level, while others do not. When increasing one unit of Freedom, the *log* transformed Corruption level of AfricaMideast will increase 1.5929 (slope of Freedom @ AfricaMideast). When there is no interaction and Freedom is controlled, changing the region from AfricaMideast to Europe will decrease the *log* transformed Corruption level by 1.255. Looking at the interaction between Freedom and Europe region, the effect of Freedom depends on the regions (AfricaMideast or Europe), which means the effect of Freedom on Corruption level differ from AfricaMideast to Europe. When changing from AfricaMideast to Europe, the effect of Freedom on the *log* transformed Corruption level will increase 2.7226. That is to say, when the region is Europe, increasing one unit of Freedom, the *log* transformed Corruption level of Europe will increase 1.5929 + 2.7226 = 4.3155 (slope of Freedom @ Europe). And the adjusted R-square is 0.2982, which means around 30% of the dependent variable (Corruption level) are explained by the model with Freedom and Region as independent variables.

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**

Since whether a breast cancer tumor is malignant or not is a binary data as the dependent variable, and three independent variables are continuous, binomial family of generalized linear regression will be used to analyze the association. Three covariates that will be included in the generalized linear model as independent variables are radius, texture, and smoothness.

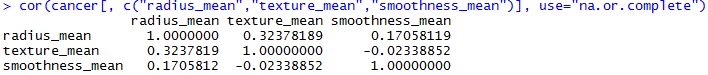
**Hypothesis:**

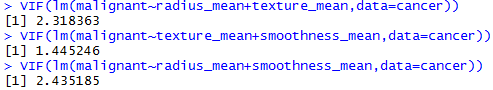
Null hypothesis: Radius, texture, and smoothness are not associated whether a breast cancer tumor is malignant or not.

Alternative hypothesis: Either radius, texture, and smoothness is associated whether a breast cancer tumor is malignant or not.

**Assumption Checking:**

1. Samples are randomly selected from the population: This is unlikely to test; however, we assume that samples from the dataset are selected randomly.
2. Each observation is sampled independently: This is unlikely to test; however, samples are collected by randomly selected patients, so we assume independence of observations.
3. There is no multi-collinearity between independent variables:

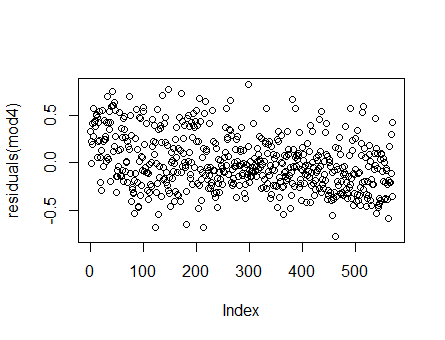




All the correlation coefficients between the independent variables are < 0.5, and VIF between two of the independent variables = 2.318363, 1.445246, and 2.435185 (< 10), so there is no multi-collinearity between the independent variables.

*Run the multiple linear model*

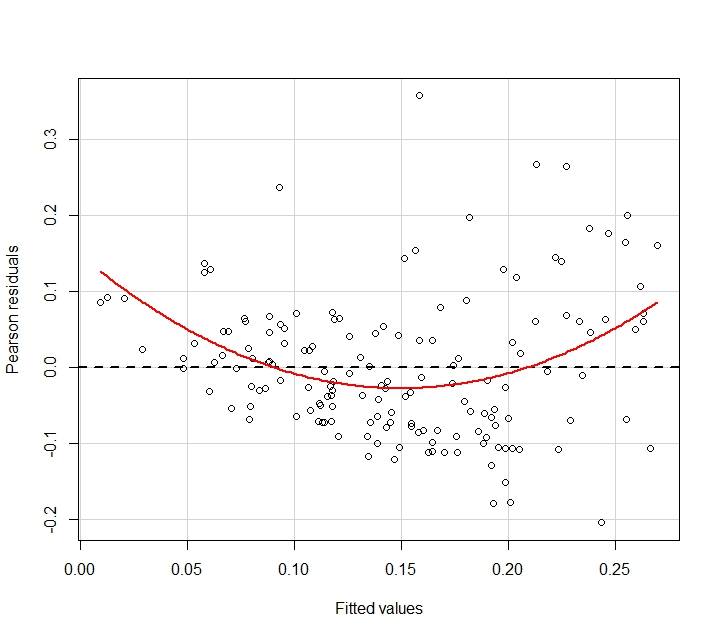
1. Residual independency:



Based on the residual plot, there is slight autocorrelation between the residuals. Since we don’t know it is time, space or other autocorrelation from the dataset, we still run the generalized linear regression.

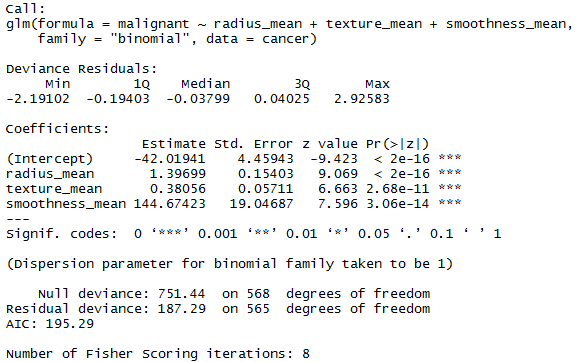
*Run generalized linear model*

1. There is a linear relationship between independent and transformed dependent variables:



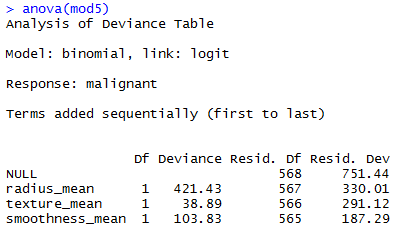
Based on the plot, there is a slight linear relationship between independent and transformed dependent variables.

**Test Results:**



Based on the p-values of coefficients, the three independent variables (radius, texture, and smoothness) are significant associated with whether a breast cancer tumor is malignant or not. When all the independent variables are controlled, the *log odds ratio* of whether a breast cancer tumor is malignant or not is -42.01941 (intercept). When increasing one unit of radius\_mean (texture\_mean and smoothness\_mean are controlled), the *log odds ratio* of whether a breast cancer tumor is malignant or not will increase 1.39699 (coefficent of radius\_mean). When increasing one unit of texture\_mean (radius\_mean and smoothness\_mean are controlled), the *log odds ratio* of whether a breast cancer tumor is malignant or not will increase 0.38056 (coefficent of texture\_mean). When increasing one unit of smoothness\_mean (radius\_mean and texture\_mean are controlled), the *log odds ratio* of whether a breast cancer tumor is malignant or not will increase 144.67423 (coefficent of smoothness\_mean).

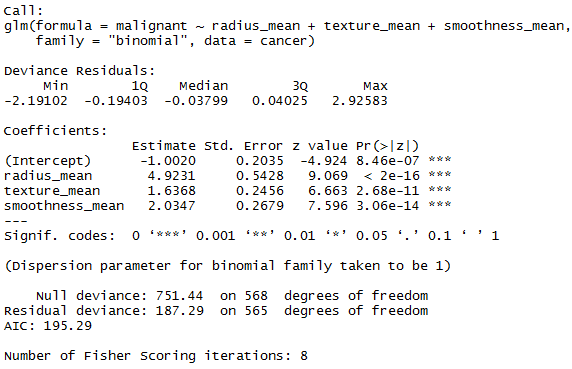
1. BONUS/EXTRA CREDIT: Which independent variables are the most important in explaining whether a breast cancer tumor is malignant or not? Use the same 3 continuous independent variables you chose for question 6. **cancer.**



To compare the relative importance of the independent variables, we use anova to see the likelihood of each independent variable to regenerate the observed data. The smaller deviance, the more likelihood to generate the same observations based on the model. Texture has the smallest deviance compared to radius and smoothness; therefore, texture is the most important variable in explaining whether a breast cancer tumor is malignant or not.

To compare the relative increase of the independent variables, we standardize the scale of each independent variable through subtracting by the mean of each variable and divided by the standard deviation of each variable.

**Test Results of Standardized Independent Variables:**



Since all the independent variables are on the same scale, each unit of radius has the largest influence (increasing log odds ratio by 4.9231) on whether a breast cancer tumor is malignant or not. Therefore, radius is the variable that increases the log odds ratio the most by each unit.