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**

# Q1)

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

head(flying)

# Chi-square test will be used in this question

# assumptions checking

# 1) independent observations

# we cannot check for independent sampling in R, we assume it follows the assumption

# 2) must have observations in each of the cells in the contingency table

# each cell is filled and the rows with NA in either "gender" or "unruly\_child" will be excluded

# 3) random sample from large population

# we cannot check for sampling process in R, we assume it follows the assumption

# transformation

# no transformation is required for this question

# run the test

library(MASS)

q1tbl = table(flying$gender,flying$unruly\_child)

q1tbl

chisq.test(q1tbl)

# results interpretation

# p-value is 0.00119 which is significant, it means there is association between geneder and

# whether people think it's rude o bring an unruly child on plane.

# model fit checking

# model fitness check is not applicable here

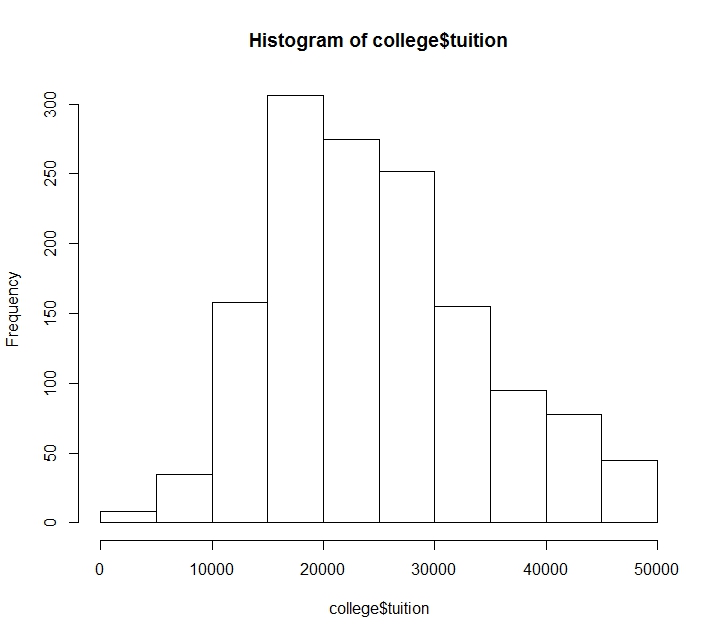
1. Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition? **college**

# Q2)

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

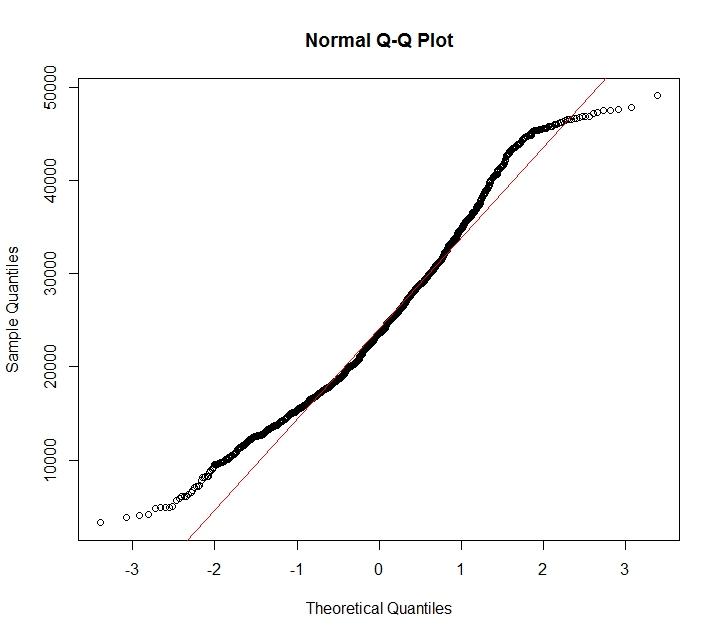
head(college)

hist(college$tuition)



qqnorm(college$tuition)

qqline(college$tuition,col="red")



# from just eyeballing the proxy graphs, tuition is not normally distributed

# two-tailed two samples t-test will be used in this question

# assumptions checking

publicshool<-subset(college,type=="Public")

is.na(publicshool)

privateschool<-subset(college,type=="Private nonprofit")

is.na(privateschool)

# 1) equal variances

var.test(publicshool[,'tuition'],privateschool[,'tuition'])

# p-value is 2.2e-16 which is significant, it means variances of tuition from public and private schools are not equal, it violates the assumption

# 2) normality test

shapiro.test(college$tuition)

# p-value is 3.193e-15 which is significant, it means the tuition data is not normally distributed, it violates the assumption

# 3) independent sampling

# we cannot check for independent sampling in R, we assume it follows the assumption

# transformation

# since N>30 in this case, i will assume it to be normally distributed according to CLT and use Welch T-test to deal with non-equal variance issue

# run the test

t.test(publicshool[,'tuition'],privateschool[,'tuition'],paired = FALSE)

# results interpretation

# p-value is 2.2e-16 whih is significant, it means that there is a significant difference for the tuition between public school and private non profit

# model fit checking

# model fitness check is not applicable here

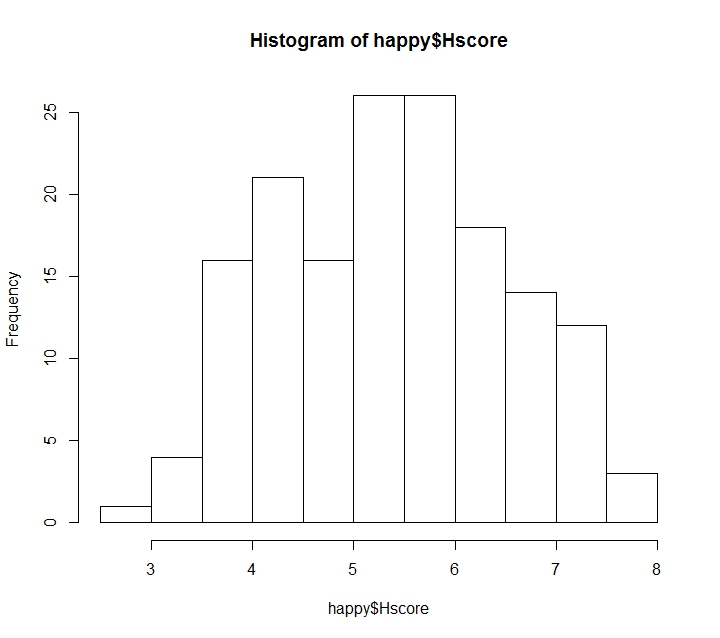
1. Is there a significant difference in happiness (Hscore) by region (Region)? **happy**

# Q3)

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

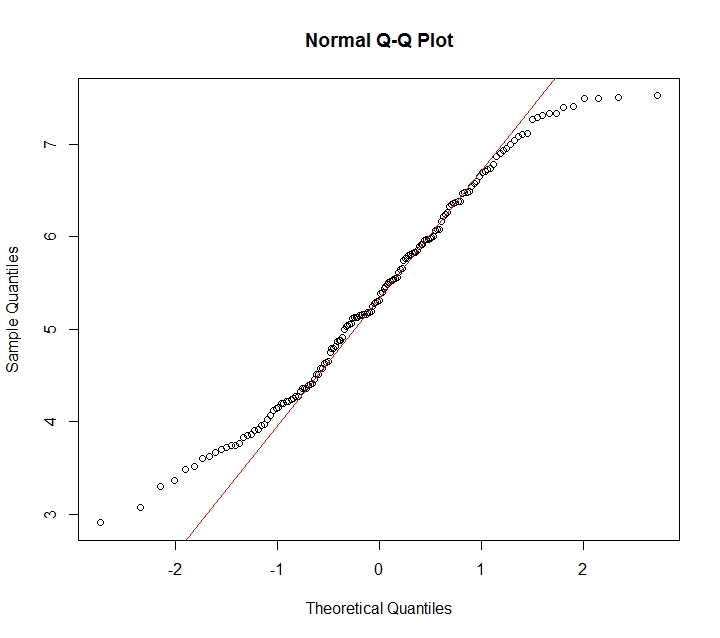
head(happy)

hist(happy$Hscore)



qqnorm(happy$Hscore)

qqline(happy$Hscore, col="red")



# from just eyeballing the proxy graphs, Hscore is not normally distributed

# One-way ANOVA will be used in this question

# assumptions checking

# 1) equal variances

# subset and stack the data

AsiaAustralia<-subset(happy, Region=="AsiaAustralia")

Europe<-subset(happy, Region=="Europe")

AmericasCarribean<-subset(happy, Region=="AmericasCarribean")

AfricaMideast<-subset(happy, Region=="AfricaMideast")

country.happy=stack(list(AsiaAustralia=AsiaAustralia$Hscore, Europe=Europe$Hscore, AmericasCarribean=AmericasCarribean$Hscore,AfricaMideast=AfricaMideast$Hscore))

library(car)

leveneTest(values~ind, country.happy)

# p-value is 0.5179 which is not significant, it means that the variances among 4 regions in Hscore are equal and follows the assumption

# 2) normality test

shapiro.test(happy$Hscore)

# p-value is 0.01248 which is significant, it means that the Hscore data is not normally distributed and violates the assumption

# 3) independent sampling

# we cannot check for independent sampling in R, we assume it follows the assumption

# transformation

# since N>30 in this case and each of the region has >15 data, i will assume it to be normally distributed according to CLT

# run the test

countryHscore=aov(values~ind, data=country.happy)

# results interpretation

summary(countryHscore)

# p-value is 1.28e-12 which is significant, it means that there is a significant differnece in Hscore among different regions

# model fit checking

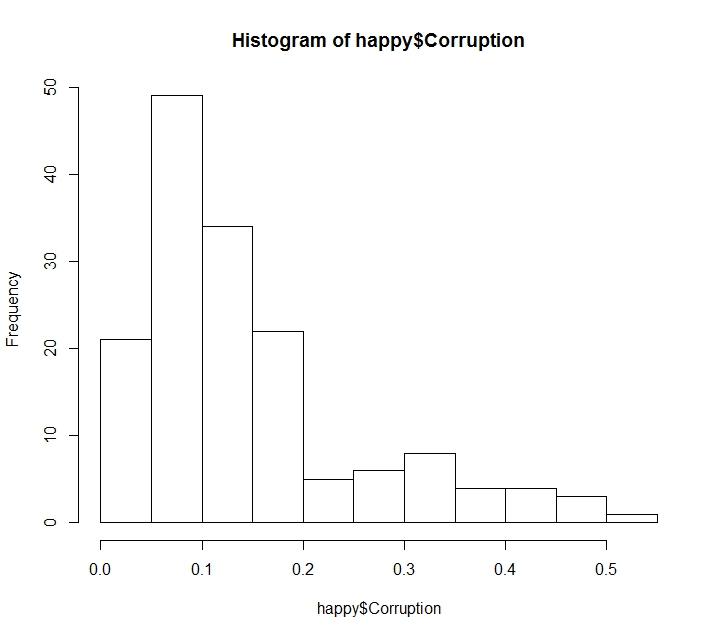
# model fitness check is not applicable here

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

# Q4)

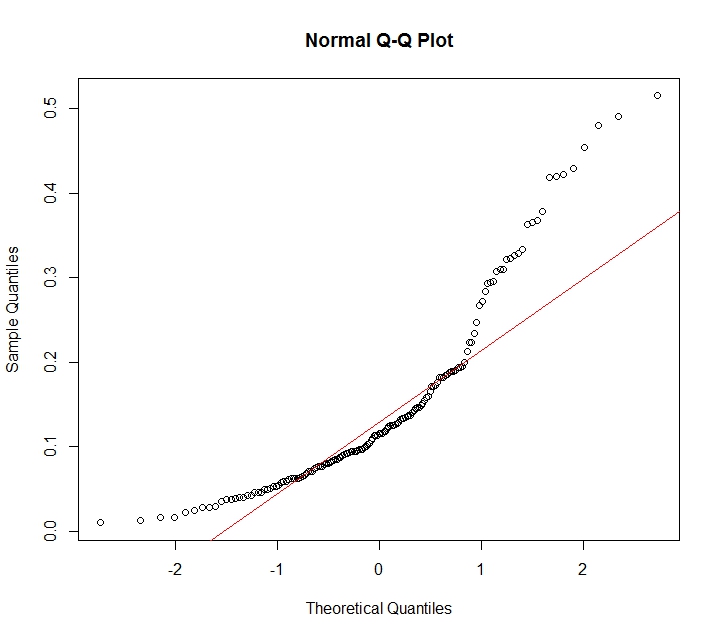
head(happy)

hist(happy$Corruption)



qqnorm(happy$Corruption)

qqline(happy$Corruption, col="red")



# from the plot on dependent variable as a proxy, it is not normally distributed and transformation later is needed

# multiple linear regression will be used

q4lm=lm(Corruption~Hscore+Freedom+Generosity, data=happy)

summary(q4lm)

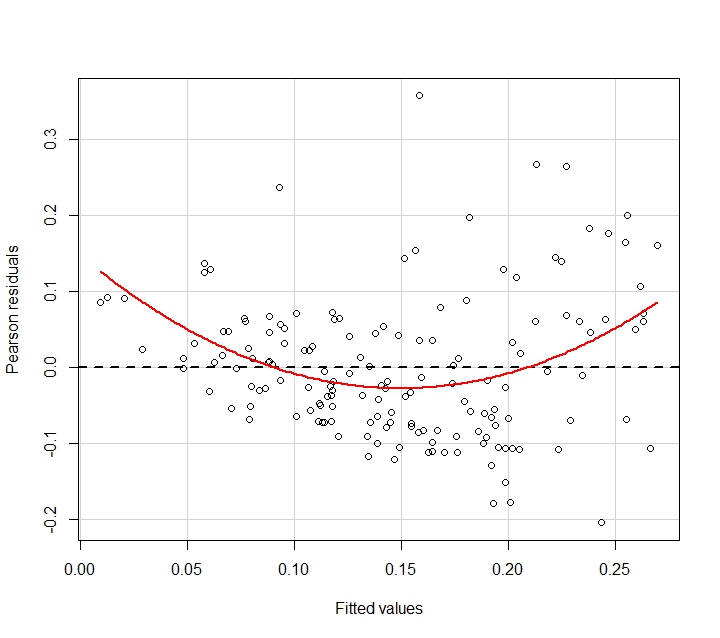
vif(q4cubicrootlm)

# from the vif result, the three values are around 1.5 which means the chosen variables are not highly correlated

# assumptions checking

# 1) linear relationship

residualPlot(q4lm)



# from the residual plot, the red line deviates greatly from the central horizontal line, it violates the assumption of linearity

# 2) normality test on dependent variable (as a proxy)

shapiro.test(happy$Corruption)

# p-value is 5.128e-11 which is significant, it means that corruption data is not normally distributed and violates the assumption

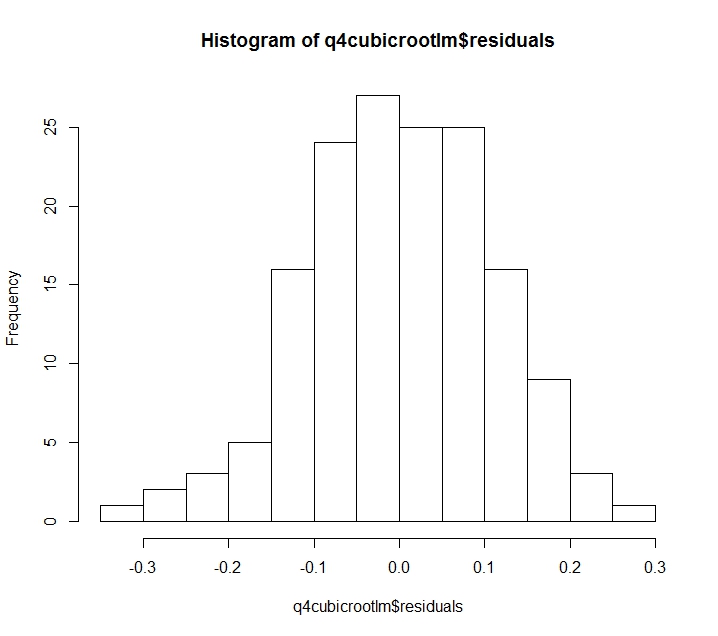
# 3) independent observations

# we cannot check for independent observations in R, we assume it follows the assumption

# transformation

q4cubicrootlm=lm((Corruption)^(1/3)~Hscore+Freedom+Generosity, data=happy)

hist(q4cubicrootlm$residuals)



# from the histogram, it looks more normal after taking cubic root

shapiro.test(residuals(q4cubicrootlm))

# the p-value from shapiro test is 0.876 which is not significant, it means they are normally distributed now and we are good to go

# run the test

q4cubicrootlm=lm((Corruption)^(1/3)~Hscore+Freedom+Generosity, data=happy)

# results interpretation

summary(q4cubicrootlm)

# from the summary table, Corruption^(1/3)=0.278+0.0143Hscore+0.310Freedom+0.117Genorosity

# since i took cubic root for transformation, i need to backtransform the estimates

0.277794^(3) #0.0214

0.014308^(3) #2.93e-06

0.310309^(3) #0.0299

0.117062^(3) #0.00160

# after taking cubic on those estimates, Corruption=0.0214+2.93e-6Hscore+0.0299Freedom+0.00160Genorosity

# from the 1st line, p-value is 1.68e-09 which is significant, when Hscore, Freedom, and Generosity are set to zero, Corruption = 0.0214

# from the 2nd line, p-value is 0.125 which is not significant, holding Freedom and Generosity zero, it has no effect to Corruption

# from the 3rd line, p-value is 9.08e-05 which is significant, holding Hscore and Generosity zero, a unit increase in Freedom will lead to 0.0299increase in Corruption

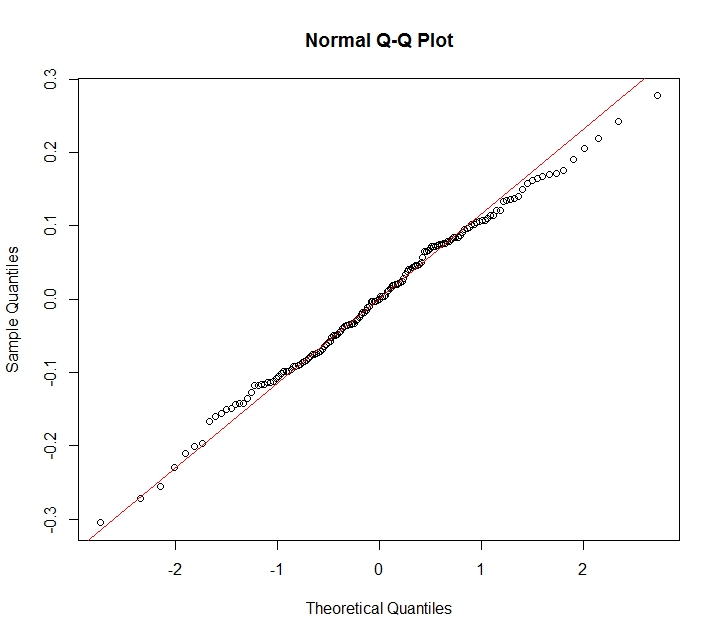
# from the 4th line, p-value is 0.0967 which is not significant, holding Hscore and Freedom zero, it has no effect to Corruption

# model fit checking

# from the summary table, multiple R-squared is 0.255 which means that 25.5% of variances are explained by the model

# 1) residual normality

qqnorm(residuals(q4cubicrootlm))

qqline(residuals(q4cubicrootlm),col="red")

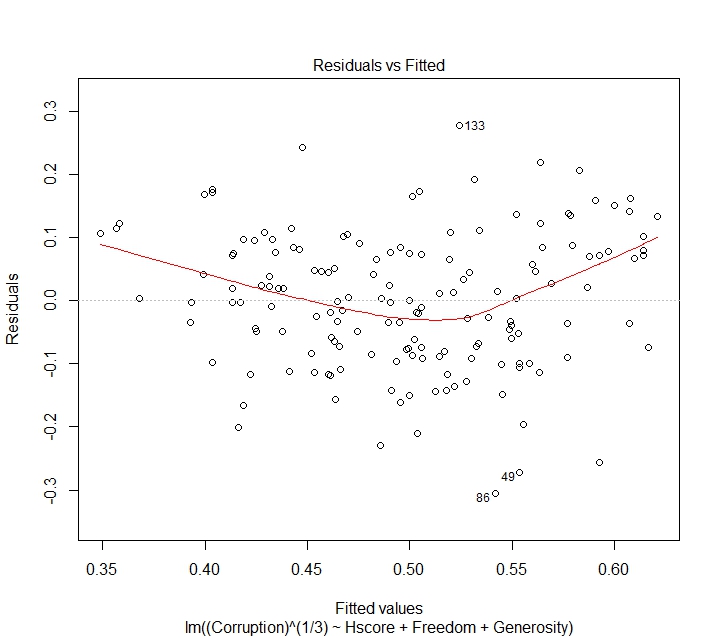
# from just eyeballing, the residuals seem to be normally distributed

shapiro.test(residuals(q4cubicrootlm))

# from the shapiro test, p-value=0.876 which is not significant meaning they are normally distributed and agree with the qqplot result

# 2) residual homoscedasticity

plot(q4cubicrootlm, which=c(1))



# from just eyeballing the residual plot, the result seems to be heteroscedastic

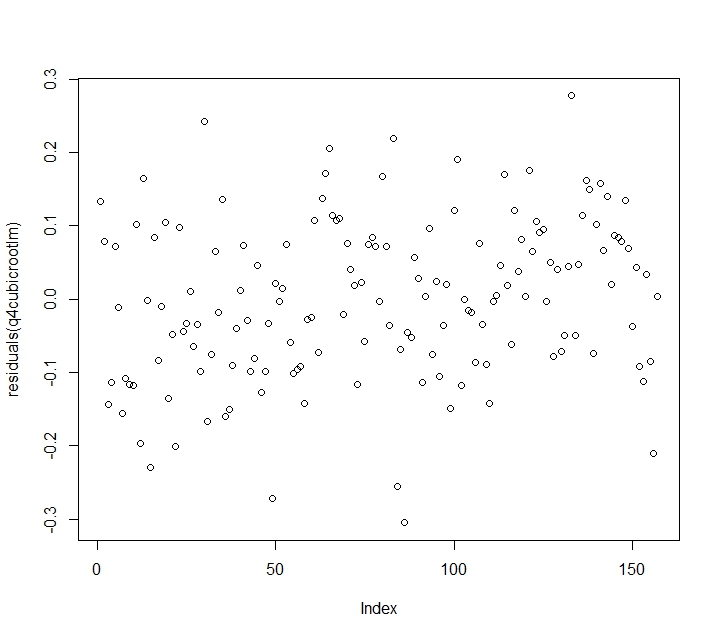
library(lmtest)

bptest(q4cubicrootlm)

# from the bptest, p-value=0.136 which is not significant, it means that the residuals are homoscedastic

# 3) residual interdependency

plot(residuals(q4cubicrootlm))



# from just eyeballing the residual plot, it has no obvious pattern

dwtest(q4cubicrootlm, alternative=c("two.sided"))

# from the dwtest, p-value=0.1684 which is not significant, it means that the residuals are independent

# overall, this is a good model to explain Corruption

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**

# Q5)

head(happy)

# from Q4, the significant variable will be Freedom

# from Q4 plot on Corruption, it is not normally distributed and transformation later is needed

# multiple linear regression will be used in this case

q5lm=lm(Corruption~Freedom\*Region, data=happy)

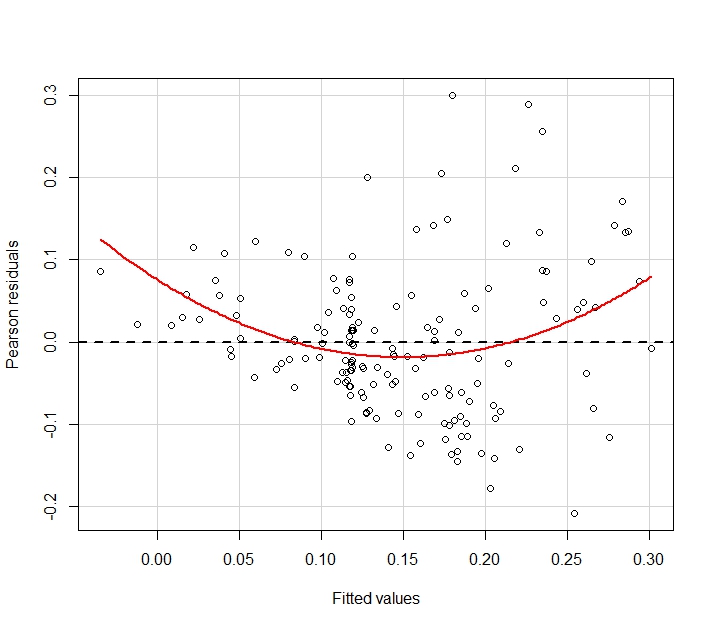
vif(q5cubicrootlm)

# from the adjusted VIF result, (GVIF^(2\*Df), Freedom, Region, and interaction term are 1.7, 3.3 and 3.6 respectively, which means they are not highly correlated

# assumptions checking

# 1) linear relationship

residualPlot(q5lm)



# from the residual plot, the red line deviates greatly from the central horizontal line, it violates the assumption of linearity

# 2) normality test on depdenent variable (as a proxy)

shapiro.test(happy$Corruption)

# p-value is 5.128e-11 which is significant, it means thats corruption data is not normally distributed

# 3) independent observations

# we cannot check for independent observations in R, we assume it follows the assumption

# transformation

q5cubicrootlm=lm((Corruption)^(1/3)~Freedom\*Region, data=happy)

hist(q5cubicrootlm$residuals)

# from the histogram, it looks more normal after taking cubic root

shapiro.test(residuals(q5cubicrootlm))

# the p-value from shapiro test is 0.8519 which is not significant, it means they are normally distributed now and we are good to go

# run the test

q5cubicrootlm=lm((Corruption)^(1/3)~Freedom\*Region, data=happy)

# results interpretation

summary(q5cubicrootlm)

# from the summary table, since i took cubic root before, i have to backtransform all the estimates by taking cubic power

(0.40617)^(3) #0.0670

(0.31375)^(3) #0.0309

(0.07311)^(3) #0.000391

(-0.07734)^(3) #-0.000463

(-0.18682)^(3) #-0.00652

(-0.31201)^(3) #-0.0304

(0.09658)^(3) #0.000901

(0.42202)^(3) #0.0752

# from the 1st line, p-value is 2e-16 which is significant, it means that the holding Freedom zero, Corruption in AfricaMideast will be 0.0670

# from the 2nd line, p-value is 0.00147 which is significant, it means that a unit increase in Freedom will lead to 0.0309 increase in Corruption in AfricaMideast

# from the 3rd line, p-value is 0.440 which is not significant, it means that AmericaCarribean has no effect on the change in intercept

# from the 4th line, p-value is 0.373 which is not significant, it means that AsiaAustralia has no effect on the change in intercept

# from the 5th line, p-value is 0.000301 which is significant, it means that the intercept of Europe is 0.00652 lower than the intercept of AfricaMideast, that is 0.0605

# from the 6th line, p-value is 0.162 which is not significant, it means that there is no significant interaction between Freedom and AmericasCarribean

# from the 7th line, p-value is 0.635 which is not significant, it means that there is no significant interaction between Freedom and AsiaAustralia

# from the 8th line, p-value is 0.00197 which is significant, it means that in Europe when Freedom increase by 1 unit, Corruption will increase by 0.0309+0.0752 = 0.106 unit

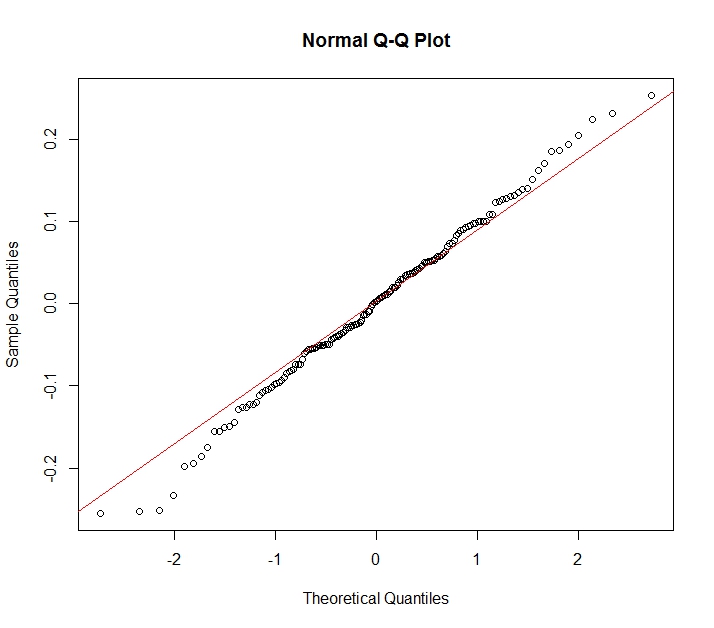
# therefore, there is a difference in influence by Freedom in AfricaMideast and Europe towards Corruption

# model fit checking

# from the summary table, multiple R-squared is 0.350 which means that 35% of variances are explained by the model

# 1) residual normality

qqnorm(residuals(q5cubicrootlm))

qqline(residuals(q5cubicrootlm),col="red")

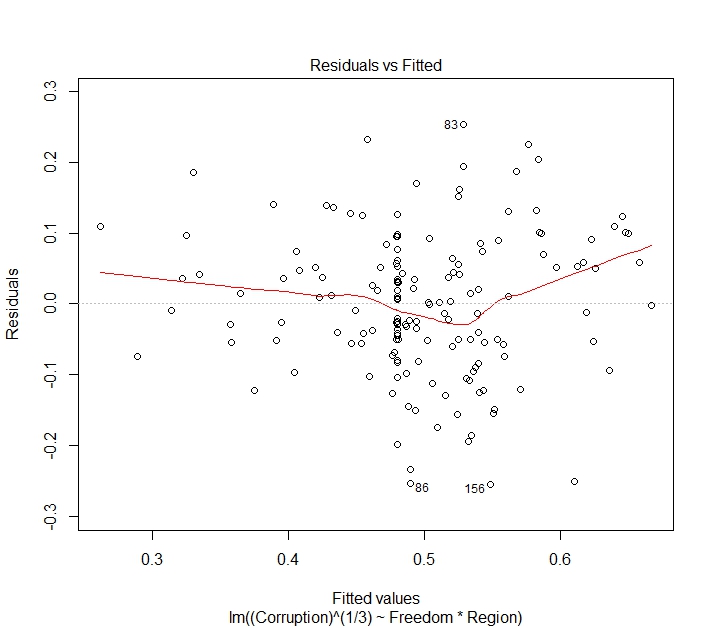
# from just eyeballing, the residuals seem to be normally distributed

shapiro.test(residuals(q5cubicrootlm))

# from the shapiro test, p-value=0.852 which is not significant and agree with the qqplot result

# 2) residual homoscedasticity

plot(q5cubicrootlm, which=c(1))



# from just eyeballing the residual plot, the result seems to to be homoscedastic

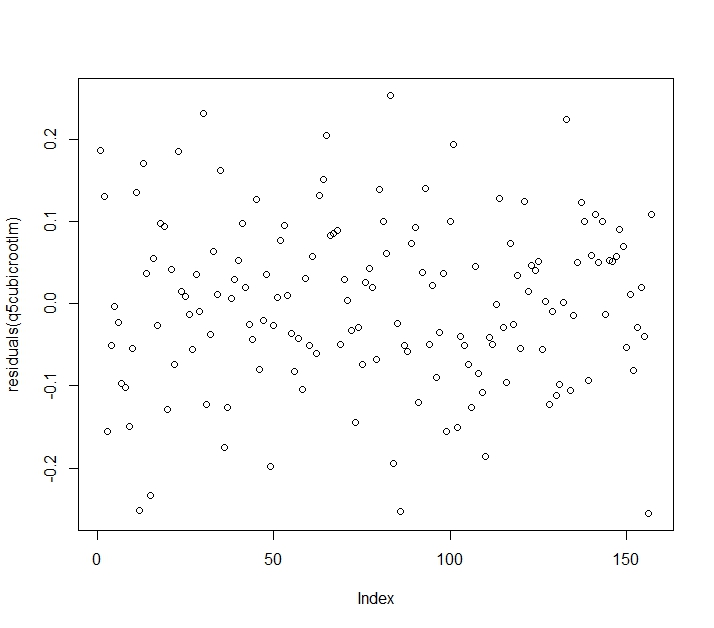
library(lmtest)

bptest(q5cubicrootlm)

# from the bptest, p-value=0.0554 which is not significant, it means that the residuals are homoscedastic

# 3) residual interdependency

plot(residuals(q5cubicrootlm))



# from just eyeballing the residual plot, it has no obvious pattern

dwtest(q5cubicrootlm, alternative=c("two.sided"))

# from the dwtest, p-value=0.846 which is not significant, it means that the residuals are indepedent

# overall, this is a good model to explain Corruption

AIC(q4cubicrootlm,q5cubicrootlm)

# from the AIC score, model in q5 has a lower score which means it is a better model compared to q4

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**

# Q6)

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

head(cancer)

# GLM(Binomial) will be used in this question

library(plyr)

library(reshape2)

library(magrittr)

library(RCurl)

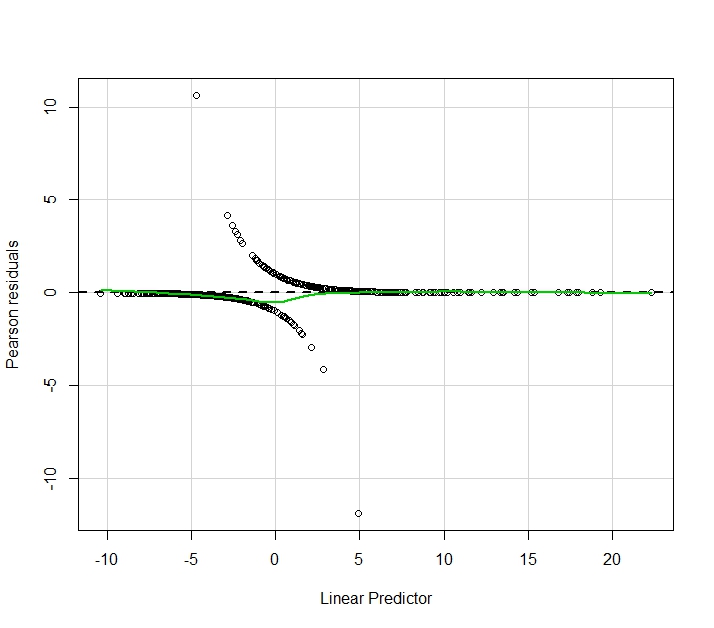
library(ggplot2)

q6glm=glm(malignant~radius\_mean+texture\_mean+perimeter\_mean,data=cancer, family=binomial(link=logit))

# assumptions checking

# 1) linear relationship between transformed dependent variable and independent variable

residualPlot(q6glm)



# from the residual plot, the greenline is very close to the central horizontal line which means it follows the assumption for linearity

# 2) independent observations

# we cannot check for independent observations in R, we assume it follows the assumption

# transformation

# the glm function is ready to run

# run the test

q6glm=glm(malignant~radius\_mean+texture\_mean+perimeter\_mean,data=cancer, family=binomial(link=logit))

# results interpretation

summary(q6glm)

# odd ratio = P(malignant)/(1-P(malignant))

# from the summary table, log(odd ratio)=-18.4-6.03radius\_mean+0.243texture\_mean+1.068perimeter\_mean

# since i used logit transformation, i will have to backtransform the estimates to interpret the real effect

exp(-18.32552)/(1+exp(-18.32552)) #1.10e-08

exp(-6.02623)/(1+exp(-6.02623)) #0.00241

exp(0.243)/(1+exp(0.243)) #0.560

exp(1.068)/(1+exp(1.068)) #0.744

# from the 1st line, the p-value is 2e-16 which is significant, it means that holding radius, texture, and perimeter zero, the predicted probability of malignant is 1.10e-8

# from the 2nd line, the p-value is 1.46e-09 which is significant, it means that holding texture and perimeter zero, the probability of malignant tumor when experiencing average radius is 0.00241

# from the 3rd line, the p-value is 9.21e-08 which is significant, it means that holding radius and perimeter zero, the probability of malignant tumor when experiencing average texture is 0.560

# from the 4th line, the p-value is 5.63e-12 which is significant, it means that holding radius and texture zero, the probability of malignant tumor when experiencing average perimeter is 0.744

# therefore, radius, texture, and perimeter of the tumor are all significantly affecting whether the tumor is malignant or not

# model fit checking

# from the summary table, residual deviance is 218.90 which is 532.54 lower than the Null deviance.

# It means that the model has a much higher likelihood to reproduce original data

# model comparison is used to check the fitness

q6glm1=glm(malignant~radius\_mean+texture\_mean, data=cancer, family=binomial(link=logit))

q6glm2=glm(malignant~radius\_mean, data=cancer, family=binomial(link=logit))

q6glm3=glm(malignant~1, data=cancer, family=binomial(link=logit))

anova(q6glm,q6glm1,q6glm2,q6glm3, test="Chi")

# from the deviance table, adding each of the three independent variables into the function will increase the likelihood of the model and

# the increase are all significant as the 3 p-values are all significant

AIC(q6glm,q6glm1,q6glm2,q6glm3)

# from the AIC score table, it agree with the deviance table that adding each of those three independent variables into the function will decrease the score,

# which means becomes a better model as it will has a higher likelihood to predict the original data

# therefore, the original model that includes radius, texture, and perimeter is the best model to predict whether the tumor is malignant or not

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.**

# Q7)

# to compare the variable importance, varImp function is used

install.packages("caret")

library(caret)

q7importance<-varImp(q6glm, scale=FALSE)

q7importance

# from the results of relative importance, perimeter\_mean scored 6.89, followed by 6.05 from radius\_mean and 5.34 from texture\_mean

# therefore, perimeter is the most important variable in explaining whether a breast cancer tumor is malignant or not