MA575 Project

library(readxl)  
US <- read\_excel("US oil price.xlsx")  
US

## # A tibble: 418 × 2  
## Date `US oil price (Dollars per Barrel)`  
## <dttm> <dbl>  
## 1 1987-05-15 00:00:00 19.4  
## 2 1987-06-15 00:00:00 20.1  
## 3 1987-07-15 00:00:00 21.3  
## 4 1987-08-15 00:00:00 20.3  
## 5 1987-09-15 00:00:00 19.5  
## 6 1987-10-15 00:00:00 19.9  
## 7 1987-11-15 00:00:00 18.8  
## 8 1987-12-15 00:00:00 17.3  
## 9 1988-01-15 00:00:00 17.1  
## 10 1988-02-15 00:00:00 16.8  
## # … with 408 more rows

library(readxl)  
Europe<- read\_excel("Europe oil price.xlsx")  
Europe

## # A tibble: 418 × 2  
## Date `Europe oil price (Dollars per Barrel)`  
## <dttm> <dbl>  
## 1 1987-05-15 00:00:00 18.6  
## 2 1987-06-15 00:00:00 18.9  
## 3 1987-07-15 00:00:00 19.9  
## 4 1987-08-15 00:00:00 19.0  
## 5 1987-09-15 00:00:00 18.3  
## 6 1987-10-15 00:00:00 18.8  
## 7 1987-11-15 00:00:00 17.8  
## 8 1987-12-15 00:00:00 17.0  
## 9 1988-01-15 00:00:00 16.8  
## 10 1988-02-15 00:00:00 15.7  
## # … with 408 more rows

Part1 A. Hypothesis test for mean Population: Daily U.S crude oil price from 1978 to 2022.

Alpha: 0.1 (so that both tail was 5%)

Data set: Monthly U.S crude oil price from 1978 to 2022.( I choose every 15th day in a month to obtain the monthly U.S crude oil price) <https://www.eia.gov/dnav/pet/hist/RWTCD.htm>

Claim: I believe the average U.S crude oil price per barrel is 84 dollars(I thought the average petroleum price is about 2.5 dollars per gallon from 1978 to 2022. And by deducting the government tax from it which is 50 cents, then the average price is about 2 dollars per gallon. As a barrel is 42 gallons, so I estimate the average U.S crude oil price per barrel is about 84 dollars)

whether or not the sample data could be biased? I take the oil price on every 15th day of a month to conduct the data set while the oil price on 15th day of a month may not fully represent the oil price of that month. So it may be biased.

H0: mu = 84 Ha: mu != 84 mu represents for the mean of monthly US oil price from May 1987 to Feb 2022.

#part1 A  
US\_oil\_price <- US$`US oil price (Dollars per Barrel)`  
Europe\_oil\_price <- Europe$`Europe oil price (Dollars per Barrel)`  
  
alpha <-0.10  
n=length(US\_oil\_price)  
#v=df  
v=n-1  
sx = sd(US\_oil\_price)  
xbar=mean(US\_oil\_price)  
tcrit<- qt(alpha/2, df=n-1, lower.tail=FALSE)  
  
SE<- sx/sqrt(n)  
  
#margin of error  
eps<- tcrit\*SE  
  
#claimed value of the mean  
mu0<-84  
  
#confidence interval  
CIL <- xbar-eps  
CIU <- xbar+eps  
  
#Test Statistics  
tstat <- (xbar-mu0)/SE  
  
#p-value  
pval <- 2\*pt(-abs(tstat),df=n-1,lower.tail = TRUE)  
  
#data frame summary  
metric\_name=c("CI.lower","CI.upper","Claimed value",  
 "Test Statistics", "T critical value",  
 "p-value","alpha")  
  
metric\_val=c(xbar-eps, xbar+eps, mu0,  
 tstat, tcrit, pval, alpha)  
  
  
datasummary <- data.frame(metric\_name,metric\_val)  
datasummary

## metric\_name metric\_val  
## 1 CI.lower 4.369280e+01  
## 2 CI.upper 4.832701e+01  
## 3 Claimed value 8.400000e+01  
## 4 Test Statistics -2.702826e+01  
## 5 T critical value 1.648516e+00  
## 6 p-value 1.063910e-93  
## 7 alpha 1.000000e-01

From the result, we know that the p-value is very small and it is way smaller than the alpha value 0.1. So we reject the null hypothesis and conclude that the mean of monthly U.S crude oil price is not equal to 84 dollars per barrel. This is not correspondent to the claim I made before.

Part1 B. Hypothesis test for sigma Population: Daily U.S crude oil price from 1978 to 2022.

Alpha: 0.1 (so that both tail was 5%)

Data set: Monthly U.S crude oil price from 1978 to 2022.( I choose every 15th day in a month to obtain the monthly U.S crude oil price)

Claim:I believe the standard deviation of crude oil price of the U.S per barrel is 20 dollars.(I thought intuitively that the standard deviation of crude oil price is about 20 dollars.)

H0: sigma = 20 Ha: sigma != 20

#Part1 B  
sig0 <- 20  
#chi-square  
cstat <- ((n-1)\*sx^2/sig0^2)  
cstat

## [1] 860.9085

ccriticL <- qchisq(alpha/2,df = v,lower.tail = T)  
ccriticU <- qchisq(alpha/2,df = v,lower.tail = F)  
  
C\_pval <- 2\*pchisq(cstat, df = v, lower.tail = F)  
  
metric\_name=c("CI.lower","CI.upper","Claimed value",  
 "Test Statistics",  
 "p-value","alpha")  
  
metric\_val=c(ccriticL, ccriticU, sig0,  
 cstat, C\_pval, alpha)  
  
  
datasummary <- data.frame(metric\_name,metric\_val)  
datasummary

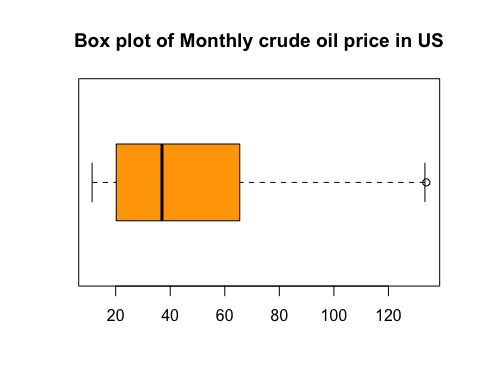
## metric\_name metric\_val  
## 1 CI.lower 3.706621e+02  
## 2 CI.upper 4.656114e+02  
## 3 Claimed value 2.000000e+01  
## 4 Test Statistics 8.609085e+02  
## 5 p-value 9.077894e-33  
## 6 alpha 1.000000e-01

In the result, we can see that the test statistics is not within the confidence interval range with alpha = 0.1. So we reject the null hypothesis and conclude that the standard deviation of monthly U.S crude oil price is not equal to 20 dollars per barrel. My claim is not correct.

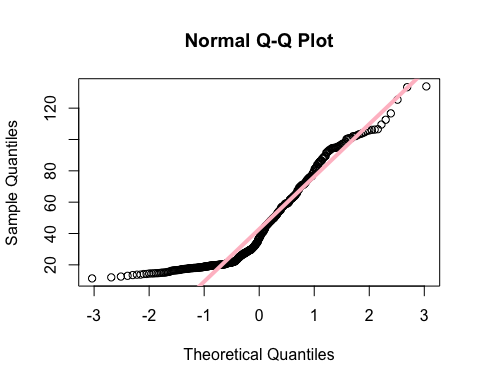
Part1 C.Normal QQ plots for both data sets

Claim: I believe the crude oil price of the U.S per barrel is normally distributed I believe the crude oil price of Europe per barrel is normally distributed

#Part1 C  
#US NQQ plots  
boxplot(US\_oil\_price,horizontal = T, col="orange",  
 main = "Box plot of Monthly crude oil price in US")



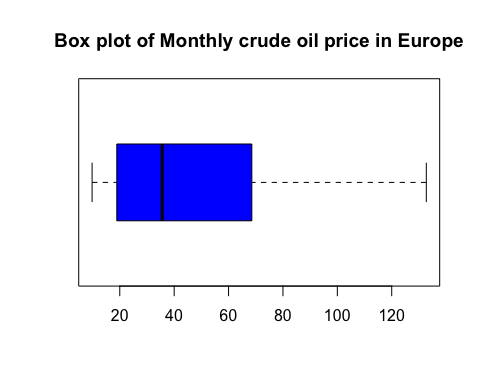
Q\_US <- qqnorm(US\_oil\_price)  
  
qqline(US\_oil\_price, col = "pink",lwd = 4)



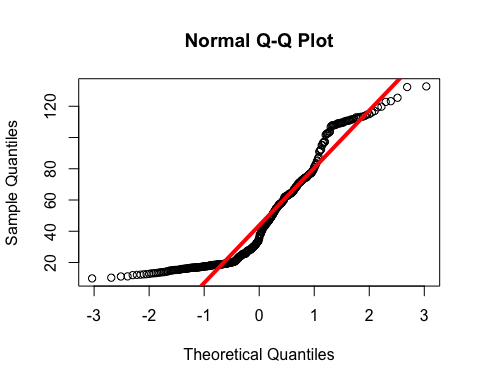
rnqq\_US <- cor(Q\_US$x,Q\_US$y)  
rnqq\_US

## [1] 0.943429

#Europe NQQ plots  
boxplot(Europe\_oil\_price,horizontal = T, col="blue",  
 main = "Box plot of Monthly crude oil price in Europe")



Q\_Europe <- qqnorm(Europe\_oil\_price)  
  
qqline(Europe\_oil\_price,col = "red", lwd = 4)



rnqq\_Europe <- cor(Q\_Europe$x, Q\_Europe$y)  
rnqq\_Europe

## [1] 0.9358836

From both data set’s QQ plot,since the correlation for both QQ plot is 0.94, we can see that data points are approximately fall along the line. Thus, we can conclude that both monthly U.S crude oil price and monthly Europe crude oil price is normally distributed which is correspondent to what we assumed in our claim stated before.

Part1 D. Hypothesis test for mean of two data sets

Claim:I believe there is not a significant difference in the average crude oil price of the U.S compared to the average crude oil price of Europe.

H0: mu1 = mu2 Ha: mu1 != mu2

#Part1 D  
  
xbar1 <- mean(US\_oil\_price)  
  
xbar2 <- mean(Europe\_oil\_price)  
  
med1 <- median(US\_oil\_price)  
  
med2 <- median(Europe\_oil\_price)  
   
sd1 <- sd(US\_oil\_price)  
  
sd2 <- sd(Europe\_oil\_price)  
  
n1 <- length(US\_oil\_price)  
  
n2 <- length(Europe\_oil\_price)  
  
alpha\_mean <- 0.1  
  
SxD <- Europe\_oil\_price - US\_oil\_price  
  
xbarD <- mean(SxD)  
  
sdD <- sd(SxD)  
  
n <- n1  
  
SE\_mean <- sdD/sqrt(n)  
  
tcrit <- qt(alpha\_mean/2,df = n-1, lower.tail = F)  
  
meps1 <- tcrit \* SE\_mean  
  
tstat\_mean <- (xbarD - 0)/SE\_mean  
  
pval\_mean <- 2\*pt(-abs(tstat\_mean),df = n-1)  
  
metric\_name = c("CI.lower","CI.upper","Test Statistics",  
 "T-critical value", "p-value","alpha")  
  
metric\_val\_mean <- c(xbarD-meps1,xbarD+meps1, tstat\_mean,tcrit, pval\_mean,alpha\_mean)  
  
datasummary <- data.frame(metric\_name,metric\_val\_mean)  
datasummary

## metric\_name metric\_val\_mean  
## 1 CI.lower 8.611096e-01  
## 2 CI.upper 1.721570e+00  
## 3 Test Statistics 4.948036e+00  
## 4 T-critical value 1.648516e+00  
## 5 p-value 1.089624e-06  
## 6 alpha 1.000000e-01

From the result, the test statistics does not fall into the confidence interval region and the p value is very small. So we reject the null hypothesis and conclude that there is significant difference between the mean of monthly crude oil price in U.S compared to the mean of monthly crude oil price in Europe which is not correspondent to our claim before.

Part1 E. Homoscedasticity

Claim: I believe that the population of the U.S and Europe are homoscedastic to one another in terms of their crude oil price

H0:sigma1^2 = sigma2^2 Ha:sigma1^2 != sigma2^2

#Part1 E  
alpha\_homo <- 0.1  
  
fstat <- sd1^2/sd2^2  
  
fcritL <- qf(alpha\_homo/2, df1 = n1 -1, df2 = n2-1, lower.tail = T)  
fcritU <- qf(alpha\_homo/2, df1 = n1 -1, df2 = n2-1, lower.tail = F)  
  
fstatL <- min(fstat, 1/fstat)  
fstatR <- max(fstat,1/fstat)  
  
pval\_homo <- pf(fstatL, df1 = n1-1, df2 = n2-1, lower.tail = T ) + pf(fstatR, df1 = n1-1, df2 = n2-1, lower.tail = F)  
  
metric\_name\_homo <- c("CI\_lower","CI\_Upper", "FstatL","FstatU","FcritL","FcritU","pval","alpha")  
  
metric\_val\_homo <- c(fcritL\*fstat, fcritU\*fstat, fstatL, fstatR, fcritL, fcritU, pval\_homo, alpha\_homo)  
  
data\_summary\_homo <- data.frame(metric\_name\_homo, metric\_val\_homo)  
data\_summary\_homo

## metric\_name\_homo metric\_val\_homo  
## 1 CI\_lower 0.67909648  
## 2 CI\_Upper 0.93760259  
## 3 FstatL 0.79794900  
## 4 FstatU 1.25321292  
## 5 FcritL 0.85105248  
## 6 FcritU 1.17501567  
## 7 pval 0.02140559  
## 8 alpha 0.10000000

From the result above, we can see the pvalue is 0.021 which is slightly smaller than the alpha value 0.1. So we reject the hypothesis test and conclude that the population of the U.S and Europe are not homoscedastic to one another in terms of their monthly crude oil price.

Part2 A Simple Linear regression

Response Variable: Average Retail Gasoline Price in California from 1999-2022 (Weekly) Explanatory Variable: Average US Crude Oil Price from 1999-2022 (Weekly)

Claim: I believe the average retail gasoline price in california is linearly related to average us crude oil price since the gasoline is produced from crude oil.

library(readxl)  
average\_retail\_price\_in\_Cali <- read\_excel("~/Desktop/MA575/MA575 Project/MA575 project/average retail price in Cali.xlsx")

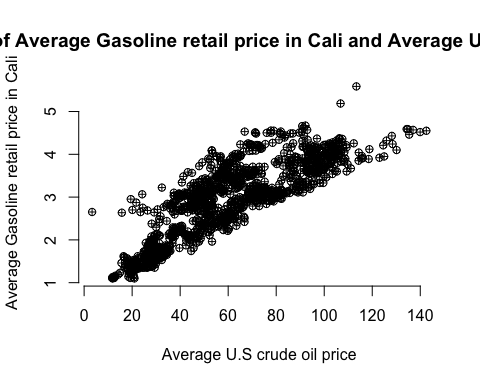
## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting numeric in H1209 / R1209C8: got a date

average\_retail\_price\_in\_Cali

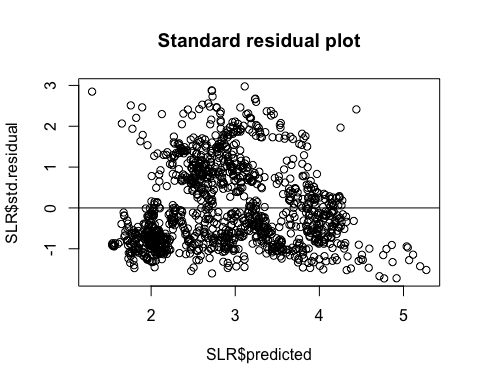
## # A tibble: 1,208 × 8  
## `Date/cali` `Average Retail pr…` `Date/US crude` `Average US Cr…`  
## <dttm> <dbl> <dttm> <dbl>  
## 1 1999-01-04 00:00:00 1.13 1999-01-01 00:00:00 11.8  
## 2 1999-01-11 00:00:00 1.13 1999-01-08 00:00:00 12.7  
## 3 1999-01-18 00:00:00 1.12 1999-01-15 00:00:00 12.6  
## 4 1999-01-25 00:00:00 1.12 1999-01-22 00:00:00 12.3  
## 5 1999-02-01 00:00:00 1.11 1999-01-29 00:00:00 12.5  
## 6 1999-02-08 00:00:00 1.11 1999-02-05 00:00:00 12.2  
## 7 1999-02-15 00:00:00 1.1 1999-02-12 00:00:00 11.8  
## 8 1999-02-22 00:00:00 1.11 1999-02-19 00:00:00 11.7  
## 9 1999-03-01 00:00:00 1.12 1999-02-26 00:00:00 12.4  
## 10 1999-03-08 00:00:00 1.16 1999-03-05 00:00:00 12.9  
## # … with 1,198 more rows, and 4 more variables: `Date/Europe crude` <dttm>,  
## # `Average Europe Crude oil price` <dbl>, `Refinery Costs and Profits` <dbl>,  
## # `Distribution Costs, Marketing Costs, and Profits` <dbl>

source("nemolm2.R")

x <- average\_retail\_price\_in\_Cali$`Average US Crude oil price`  
y <- average\_retail\_price\_in\_Cali$`Average Retail prices`  
#scatter plot  
plot(x, y, main = "Scatter plot of Average Gasoline retail price in Cali and Average U.S crude oil price",  
 xlab = "Average U.S crude oil price", ylab = "Average Gasoline retail price in Cali",  
 pch = 10, frame = FALSE)



#standard residuals plot  
SLR <- nemolm2(y,x)  
plot(SLR$predicted,SLR$std.residual,main = "Standard residual plot")  
abline(0,0)



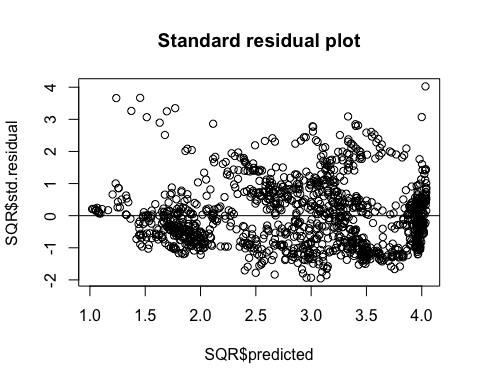
#table for the model parameters, standard errors, Rsquare adjusted, Rsquare and pvalue  
metric\_name\_SLR <- c("Slope","Intercept","SE of Slope","SE of Intercept","R square adjusted","R square","p-value")  
metric\_val\_SLR <- c(SLR$betahat[2],SLR$betahat[1],SLR$SEbetahat[2],SLR$SEbetahat[1],SLR$r2adj,SLR$r2,SLR$pval)  
data\_summary\_SLR <- data.frame(metric\_name\_SLR, metric\_val\_SLR)  
data\_summary\_SLR

## metric\_name\_SLR metric\_val\_SLR  
## 1 Slope 0.0285413936  
## 2 Intercept 1.2034570441  
## 3 SE of Slope 0.0005161666  
## 4 SE of Intercept 0.0336538770  
## 5 R square adjusted 0.7169011652  
## 6 R square 0.7171357127  
## 7 p-value 0.0000000000

Part2 A. Simple linear regression As showed in the data summary, the model is Average Gasoline retail price in Cali = 0.0285\* Average US crude oil price + 1.2035. Both R square and R square adjusted is 0.717 which indicates the model have strong positive relationship. The p value equals to 0 suggests that the null hypothesis is rejected and the model is statistically significant. By using this SLR model, we can predicted the average gasoline retail price in Cali by using the Average US crude oil price. The model means for every dollar increase in Average US crude oil price, there will be 0.0285 dollar increase in average gasoline retail price in cali.

Part2 B. Simple Quadratic Regression Claim: I believe the average gasoline retail price in cali is not quadratically related to the average Europe crude oil price since the gasoline price in cali seems no relationship with the average Europe crude oil price.

X <- average\_retail\_price\_in\_Cali$`Average Europe Crude oil price`  
SQR <- nemolm2(y,cbind(X,X^2))  
#standard residuals plot  
plot(SQR$predicted,SQR$std.residual,main = "Standard residual plot")  
abline(0,0)



#table for the model parameters, standard errors, Rsquare adjusted, Rsquare and pvalue  
metric\_name\_SQR <- c("Slope of X","Slope of X^2","Intercept","SE of Slope of X","SE of Slope of X^2","SE of Intercept","R square adjusted","R square","p-value")  
metric\_val\_SQR <- c(SQR$betahat[2],SQR$betahat[3],SQR$betahat[1],SQR$SEbetahat[2],SQR$SEbetahat[3],SQR$SEbetahat[1],SQR$r2adj,SQR$r2,SLR$pval)  
data\_summary\_SQR <- data.frame(metric\_name\_SQR, metric\_val\_SQR)  
data\_summary\_SQR

## metric\_name\_SQR metric\_val\_SQR  
## 1 Slope of X 5.529750e-02  
## 2 Slope of X^2 -2.155292e-04  
## 3 Intercept 4.943513e-01  
## 4 SE of Slope of X 1.682693e-03  
## 5 SE of Slope of X^2 1.205619e-05  
## 6 SE of Intercept 5.131327e-02  
## 7 R square adjusted 8.128104e-01  
## 8 R square 8.131205e-01  
## 9 p-value 0.000000e+00

As showed in the data summary, the model is Average Gasoline retail price in Cali = 0.0553\* Average Europe crude oil price - 0.000216\*Average Europe crude oil price^2 + 0.494. Both R square and R square adjusted is around 0.813 which indicates the model have strong positive relationship. The p value equals to 0 suggests that the null hypothesis is rejected and the model is statistically significant. This is to say our claim is incorrect. By using this SQR model, we can predicted the average gasoline retail price in Cali by using the Average Europe crude oil price. Since the r square is smaller, we prefer using the SQR model over the SLR model.

Part3 Multiple Linear Regression

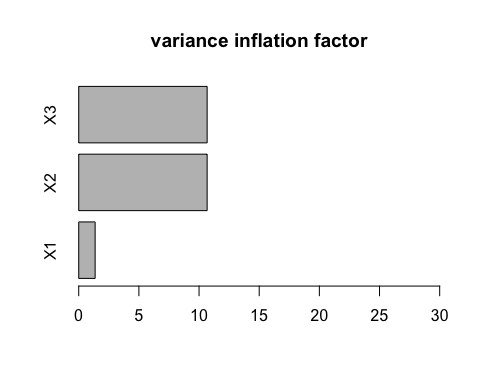
Response Variable: Average Retail Gasoline Price in California from 1999-2022 (Weekly) X1: Average US Crude Oil Price from 1999-2022 (Weekly) X2: Refinery Costs and Profits in Cali (weekly) X3: Distribution Costs, Marketing Costs, and Profits in Cali(weekly)

Claim: I believe all of the variables are good predictors of response variable

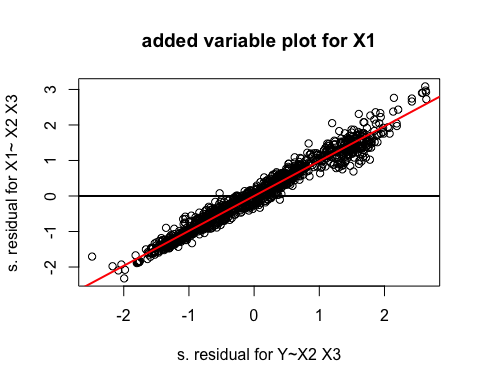
X1 <- average\_retail\_price\_in\_Cali$`Average US Crude oil price`  
X2 <- average\_retail\_price\_in\_Cali$`Refinery Costs and Profits`  
X3 <- average\_retail\_price\_in\_Cali$`Distribution Costs, Marketing Costs, and Profits`  
MLR <- nemolm2(y,cbind(X1,X2,X3))  
#ANOVA table  
metric\_name\_ANOVA <- c("SST","SSE","SSM","MST","MSE","MSM","Fstat","p-value")  
metric\_val\_ANOVA <- c(MLR$sst,MLR$sse,MLR$ssm,MLR$mst,MLR$mse,MLR$msm,MLR$Fstat,MLR$pval)  
data\_summary\_ANOVA <- data.frame(metric\_name\_ANOVA, metric\_val\_ANOVA)  
data\_summary\_ANOVA

## metric\_name\_ANOVA metric\_val\_ANOVA  
## 1 SST 9.680640e+02  
## 2 SSE 2.456747e+01  
## 3 SSM 9.434966e+02  
## 4 MST 8.020414e-01  
## 5 MSE 2.040487e-02  
## 6 MSM 3.144989e+02  
## 7 Fstat 1.541293e+04  
## 8 p-value 0.000000e+00

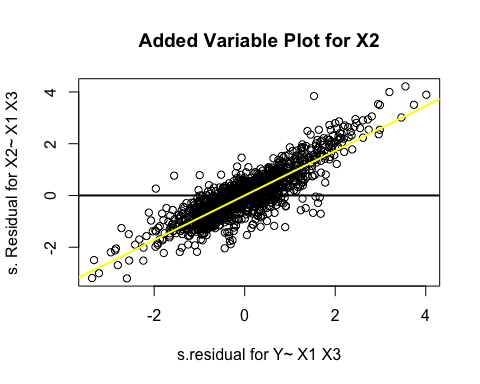
#Variance Inflation factors  
MYv12 <- nemolm2(y, cbind(X1, X2))  
MYv13 <- nemolm2(y, cbind(X1, X3))  
MYv23 <- nemolm2(y, cbind(X2, X3))  
M1v23 <- nemolm2(X1, cbind(X2, X3))  
M2v13 <- nemolm2(X2, cbind(X1, X3))  
M3v12 <- nemolm2(X3, cbind(X1,X2))  
  
vif1 <- 1/(1-MYv23$r2)  
vif2 <- 1/(1-MYv12$r2)  
vif3 <- 1/(1-MYv12$r2)  
  
vif <- c(vif1, vif2, vif3)  
  
barplot(vif, horiz=T,  
 main='variance inflation factor',  
 names.arg=c('X1','X2','X3'),  
 xlim=c(0,30))



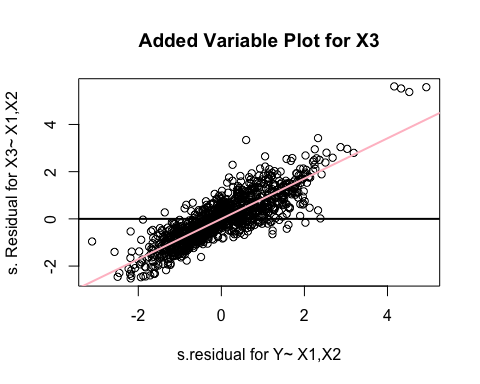
#Added variable plot  
#AVP for X1  
plot(MYv23$std.residual,M1v23$std.residual,  
 main = "added variable plot for X1",  
 xlab = "s. residual for Y~X2 X3",  
 ylab = "s. residual for X1~ X2 X3")  
abline(0,0,lwd=2)  
  
m1 <- cor(MYv23$std.residual,M1v23$std.residual)\*sd(M1v23$std.residual)/sd(MYv23$std.residual)  
abline(mean(MYv23$std.residual)-m1\*mean(M1v23$std.residual),m1,  
 col="red",lwd=2)



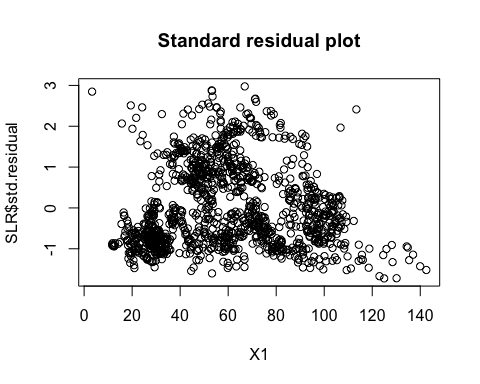
#AVP for X2  
plot(MYv13$std.residual, M2v13$std.residual,  
 main='Added Variable Plot for X2',  
 xlab='s.residual for Y~ X1 X3', ylab='s. Residual for X2~ X1 X3')  
abline(0,0,lwd=2)  
  
m2<-cor(MYv13$std.residual, M2v13$std.residual)\*sd(M2v13$std.residual/sd(MYv13$std.residual))  
abline(mean(MYv13$std.residual)-m2\*mean(M2v13$std.residual),m2,  
 col="yellow",lwd=2)



#AVP for X3  
plot(MYv12$std.residual, M3v12$std.residual,  
 main='Added Variable Plot for X3',  
 xlab='s.residual for Y~ X1,X2', ylab='s. Residual for X3~ X1,X2')  
abline(0,0,lwd=2)  
  
m3<-cor(MYv12$std.residual, M3v12$std.residual)\*sd(M3v12$std.residual/sd(MYv12$std.residual))  
abline(mean(MYv12$std.residual)-m3\*mean(M3v12$std.residual),m3,  
 col="pink",lwd=2)



#standard residuals plot  
plot(X1,SLR$std.residual,main = "Standard residual plot")



#correlation matrix between Y and all three variables  
cor(cbind(y,X1,X2,X3))

## y X1 X2 X3  
## y 1.0000000 0.84683866 0.37797718 0.4823931  
## X1 0.8468387 1.00000000 -0.06613842 0.0659103  
## X2 0.3779772 -0.06613842 1.00000000 0.4383531  
## X3 0.4823931 0.06591030 0.43835315 1.0000000

From the added variable plots for each variable, we can see that all the variables have a significant slope. So we conclude that all of the variables are significant in the model. Considering the variance inflation factors with a value-5 threshold, we can see that X1 is good for the threshold while X2 X3 are larger than 5. So we conclude that X2 X3 have multicolinearity. By looking at the standardized residuals plot, we can see a not random trend which indicates non-constant variance.

The model we conducted in MLR: Average Retail Gasoline Price in California = 0.2294 + 0.0286*Average US Crude Oil Price + 1.0536*Refinery Costs and Profits in Cali + 1.4618\*Distribution Costs, Marketing Costs, and Profits in Cali

This means a dollar increase in Average US crude oil price can lead to 0.0286 dollar increase in Average retail gasoline price, a dollar increase in refinery costs and profits in california can lead to 1.0536 dollar increase in Average retail gasoline price, a dollar increase in Distribution Costs, Marketing Costs, and Profits in California can lead to 1.4618 dollar increase in Average retail gasoline price.

For example, if we have Average US Crude Oil Price = 14dollars, Refinery Costs and Profits in Cali= 0.4 dollar,and Distribution Costs, Marketing Costs, and Profits in Cali = 0.3 dollar, then the average retail gasoline price in california = 0.2294 + 0.0286*2 + 0.4*1.0536 + 1.4618\*0.3 = 0.42144 +0.43854 +0.4004 +0.2294 = 1.49 dollar