# Assignment4

## Question1

library(tidyverse)

kungsan <- read\_csv('http://stats.apiolaza.net/data/kungsan\_full.csv')

set.seed(17053777)

## Question2

my\_kungsan <- kungsan %>%

sample\_n(540) %>%

mutate(weight2 = weight ^ 2)

## Question 3

# install.packages("GGally")

library(GGally)

my\_kungsan %>% ggpairs(columns = c('weight', 'weight2', 'sex', 'height'))

A picture containing graphical user interface

Description automatically generated

# The plot show there is a strong positive correlation relationship between weight and height and the correlation is 0.94. And the relationship between weight2 and height is in a strong positive correlation too since the correlation is 0.868. But for the relationship between height and sex, there is no correlation here because height is different in different sex, so we cannot see the correlation here. Instead, there is a strong positive correlation between weight and weight2 with correlation 0.980.

## Question 4

# install.packages('car')

library(car)

m1 <- lm(height ~ weight, data = my\_kungsan)

summary(m1)

A screenshot of a computer

Description automatically generated with low confidence

m2 <- lm(height ~ weight + weight2, data = my\_kungsan)

summary(m2)

vif(m2)

A screenshot of a computer

Description automatically generated with low confidence

m3 <- lm(height ~ weight + weight2 + sex, data=my\_kungsan)

summary(m3)

vif(m3)

A screenshot of a computer

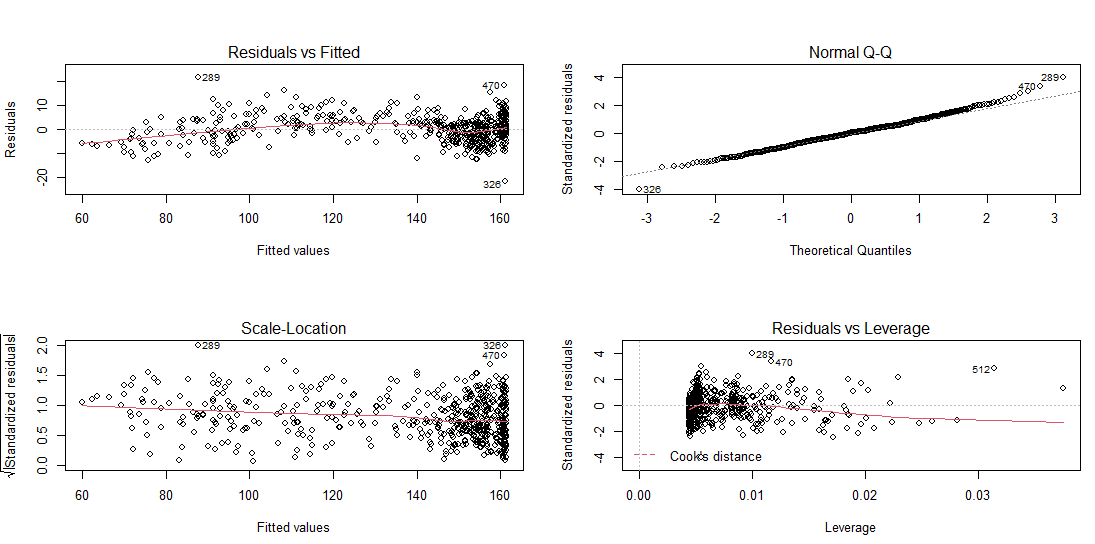
Description automatically generated with low confidence

# The variation inflation for m2 and m3 are greater than 10. Due to the rule of thumb, if VIF is greater than 10, then these predictors are too specific which can cause a wrong correlation coefficient and makes collinearity a problem.

## Question5

par(mfrow = c(2, 2))

plot(m3)



# From Residuals vs Fitted, we observe that there is deviation from linearity.

# From Normal Q-Q: we observe that after the theoretical qualities are greater than 2, the observation points are further way from the predicted line. The residuals increase as the quality increases.

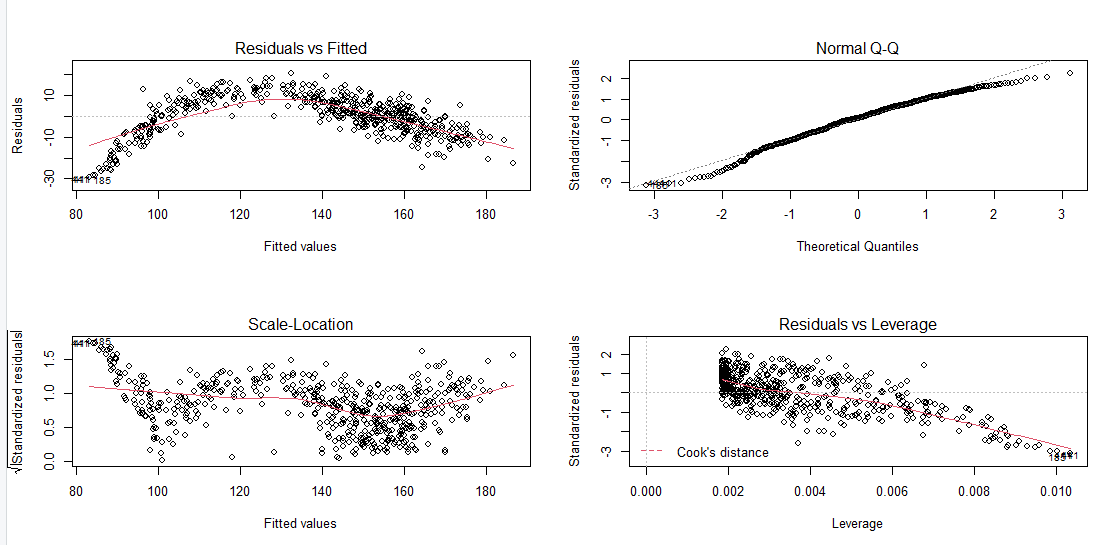
# From Scale-Location, we observe the red line is approximately horizontal, but the variance is decreasing which does not meet the variance assumptions

# From Residuals vs Leverage, we can see there is no strong inflation observation

# Overall, the model meets the assumptions well.

par(mfrow = c(2, 2))

plot(m1)



# From Residuals vs Fitted, we observe that there are non-linear patterns.

# From Normal Q-Q: we observe that after the theoretical qualities are greater than 2 and the theoretical qualities are less than -2, the observation points are further way from the predicted line. The residuals increase as the quality increases.

# From Scale-Location, we observe the red line is not that horizontal, and the pattern is not clear.

# From Residuals vs Leverage, we observe that there are some points which are close to the red dash line, which could be considered as outliers.

## Question6

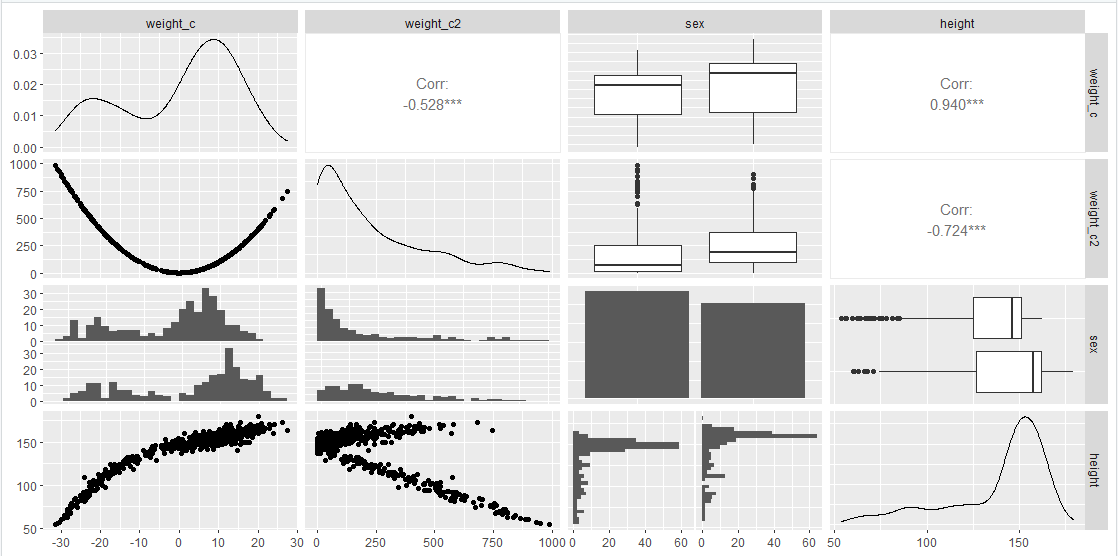
# install.packages('farver')

my\_kungsan <- my\_kungsan %>%

mutate(weight\_c = weight - mean(weight),

weight\_c2 = weight\_c ^ 2)

my\_kungsan %>% ggpairs(columns = c('weight\_c', 'weight\_c2', 'sex', 'height'))



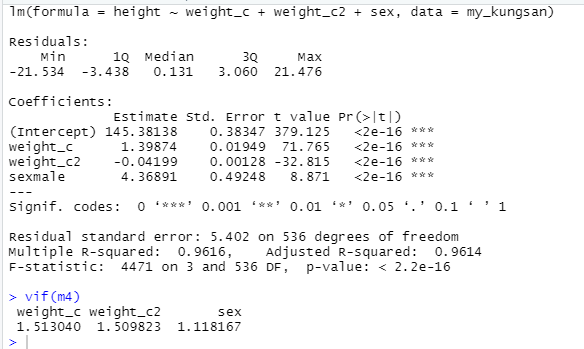
# The difference is the correlation between weight\_c and weight\_c2, it decreases from 0.980 in the graph in questions 2 to -0.528 in this graph.

## Question7

m4 <- lm(height ~ weight\_c + weight\_c2 + sex, data = my\_kungsan)

summary(m4)

vif(m4)

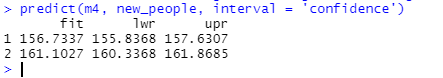


## Question8

new\_people <- tibble(weight = c(50, 50), sex = c('female', 'male')) %>%

mutate(weight\_c = weight - 36, weight\_c2 = weight\_c ^2)

predict(m4, new\_people, interval = 'confidence')



## Question9

wine <- read\_csv('http://stats.apiolaza.net/data/white\_wines.csv')

my\_wine <- wine %>% sample\_n(4800)

my\_wine

w1 <- lm(quality~fix\_acid+vol\_acid+

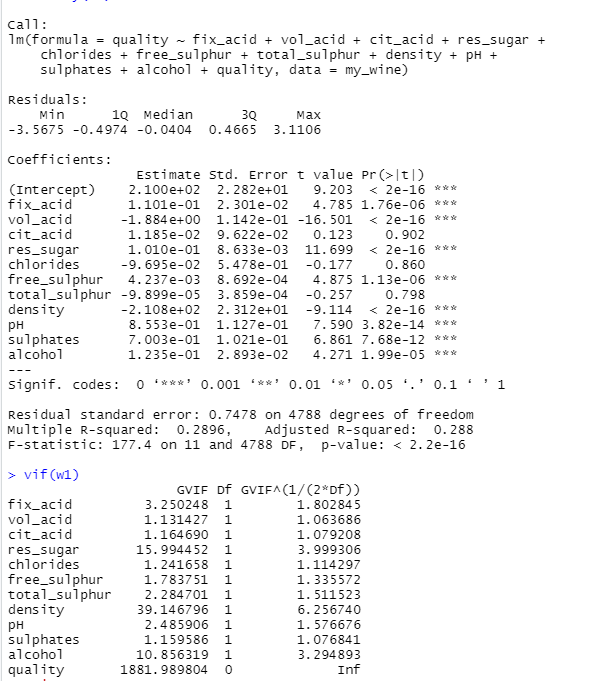
cit\_acid+res\_sugar+chlorides+free\_sulphur+

total\_sulphur+density+pH+sulphates+

alcohol+quality , data=my\_wine)

summary(w1)

vif(w1)



**#** adj-R2 is 0.288

# residual standard error: 0.7478 on 4788 degrees of freedom

# VIF for all slopes is shown above.

## Question10

library(leaps)

par(mfrow = c(1, 1))

all\_mods <- regsubsets(quality ~ ., data = my\_wine)

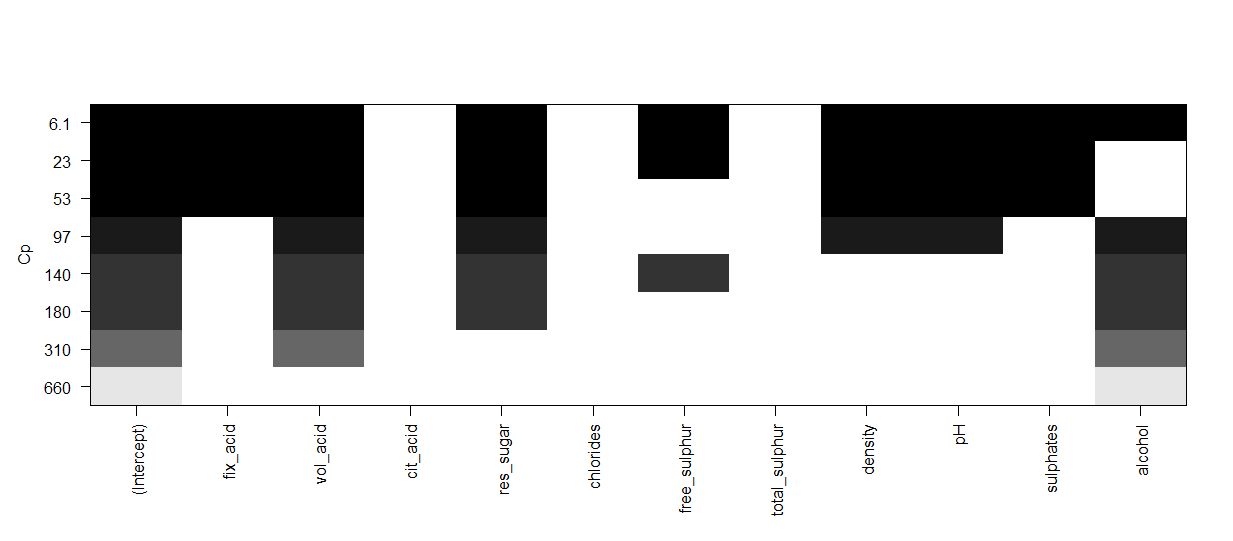
plot(all\_mods, scale = 'Cp')

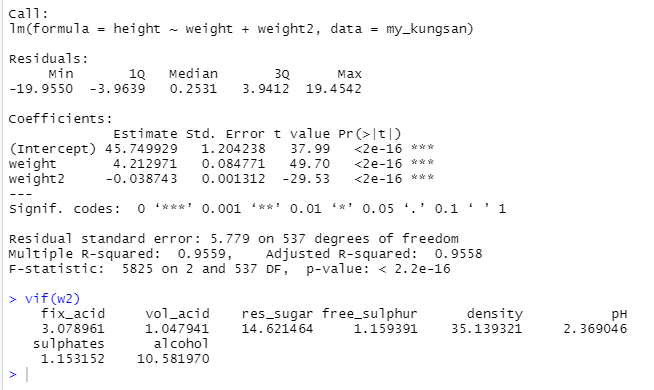
# As we know, the lower the Cp, the better the model. We can observe that cit\_acid, chlorides, total\_sulphur are the variables that need t be dropped. So, the predictors which need to be contained in the best model is fix\_acid, vol\_acid, res\_sugar, free\_sulphur, density, pH, sulphates, alcohol.

w2 <- lm(quality ~ (fix\_acid + vol\_acid + res\_sugar + free\_sulphur + density + pH + sulphates + alcohol), data = my\_wine)

summary(m2)

vif(w2)





# adj-R2 = 0.9558

# residual standard error = 5.779 on 537 degrees f freedom

# VIF is shown above.

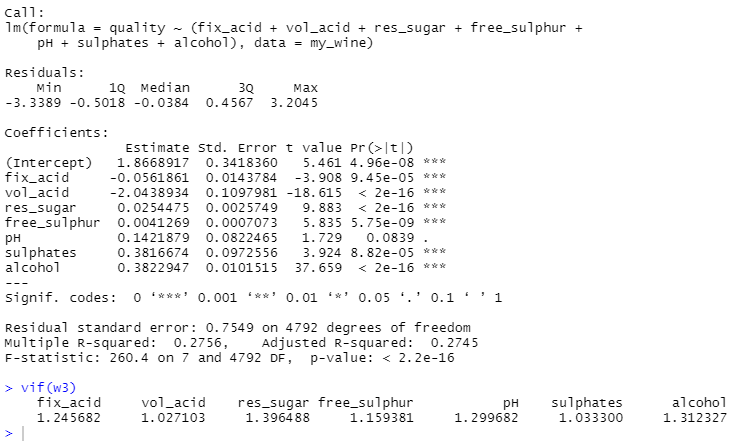
## Question11

# from the previous vif, we can see desity has the highest density.

w3 <- lm(quality ~ (fix\_acid + vol\_acid + res\_sugar + free\_sulphur + pH + sulphates + alcohol), data = my\_wine)

summary(w3)

vif(w3)



# adj-R2 = 0.2745

# residual standard error = 0.7549 on 4792 degrees f freedom

# VIF is shown above.

# From w2 to w3, we can see the adjusted R-squared decreases significantly which means predictors improve the model by less than expected by chance. And the VIF for all factors in w3 are lower than VIF in w2, which means there is no likely to occur collinearity a problem. So we can say the model w3 is better.