Final Project

Aziz Alshalfan & Aman Goswami

15 December, 2020

# Importing Required Libraries  
library(car)  
library(broom)  
library(readxl)  
library(dplyr)  
library(faraway)  
library(ggplot2)  
library(lubridate)  
library(leaps)

# Question 1

The file **stockdata.csv** uploaded to the blackboard is a dataset that contains the price of a stock in the last 100 days as the response and the following as predictors:

* **vol**: Volatility of the stock.
* **cap.to.gdp**: The ratio of the market cap to GDP.
* **q.ratio**: The ratio of market cap to net worth.
* **gaap**: Shiller Cape Index.
* **avg.allocation**: Average investor equity allocation of the stock.

Fit a model to explain price in terms of the predictors. Perform regression diagnostics to answer the following questions. Display any plots that are relevant and explain your reasoning. Suggest possible improvements if there are any.

# Importing Stocks Data  
  
stocks <- read.csv(file = "stockdata.csv",  
 header = TRUE)  
  
head(stocks)

X cap.to.gdp q.ratio gaap trailing.pe avg.allocation price vol  
1 1 0.06936092 0.4829763 0.9729927 0.5996109 0.001411161 1.142843 0.92  
2 2 0.81777520 0.8410553 0.2172974 0.8306384 0.276647039 1.161696 0.22  
3 3 0.94262173 0.4551459 0.5205087 0.7396654 0.479843133 1.164927 0.02  
4 4 0.26938188 0.8605173 0.8279428 0.5644036 0.510342157 1.164664 0.37  
5 5 0.16934812 0.6761024 0.9641110 0.3643197 0.903116010 1.165892 0.93  
6 6 0.03389562 0.7275831 0.6318801 0.9902194 0.155371382 1.153048 0.42

## (a)

Fit a model to explain **price** in terms of the predictors. Which variables are important, can any of the variables be removed? Please use F-test to justify.

stocks\_form <- "price ~ cap.to.gdp + q.ratio + gaap + trailing.pe + avg.allocation + vol"  
  
stocks\_model <- lm(formula = stocks\_form,  
 data = stocks)  
  
summary(stocks\_model)

Call:  
lm(formula = stocks\_form, data = stocks)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.0067787 -0.0015687 0.0002342 0.0019888 0.0075661   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.1087154 0.0012722 871.527 <2e-16 \*\*\*  
cap.to.gdp 0.0209002 0.0010535 19.839 <2e-16 \*\*\*  
q.ratio 0.0181111 0.0010414 17.391 <2e-16 \*\*\*  
gaap 0.0163251 0.0009298 17.557 <2e-16 \*\*\*  
trailing.pe 0.0143780 0.0009750 14.747 <2e-16 \*\*\*  
avg.allocation 0.0225869 0.0009978 22.637 <2e-16 \*\*\*  
vol -0.0005667 0.0009918 -0.571 0.569   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.002758 on 93 degrees of freedom  
Multiple R-squared: 0.9535, Adjusted R-squared: 0.9505   
F-statistic: 318 on 6 and 93 DF, p-value: < 2.2e-16

The variable **vol** turns out to be insignificant in the model.

Therefore, we can test the individual effect of the variable **vol** in the model. For that, we can use F-test.

**Hypothesis for Partial F-test**:

The model with **vol** is not significantly better than the model without **vol**.

The model with **vol** is significantly better than the model without **vol**.

stocks\_form1 <- "price ~ cap.to.gdp + q.ratio + gaap + trailing.pe + avg.allocation"  
  
stocks\_model1 <- lm(formula = stocks\_form1,  
 data = stocks)  
  
anova(stocks\_model1, stocks\_model)

Analysis of Variance Table  
  
Model 1: price ~ cap.to.gdp + q.ratio + gaap + trailing.pe + avg.allocation  
Model 2: price ~ cap.to.gdp + q.ratio + gaap + trailing.pe + avg.allocation +   
 vol  
 Res.Df RSS Df Sum of Sq F Pr(>F)  
1 94 0.00070986   
2 93 0.00070738 1 2.4832e-06 0.3265 0.5691

Since the p-value > 0.05, we do not reject and conclude that the model with **vol** is not significantly better than the model without **vol**.

Therefore, the final model is:

summary(stocks\_model1)

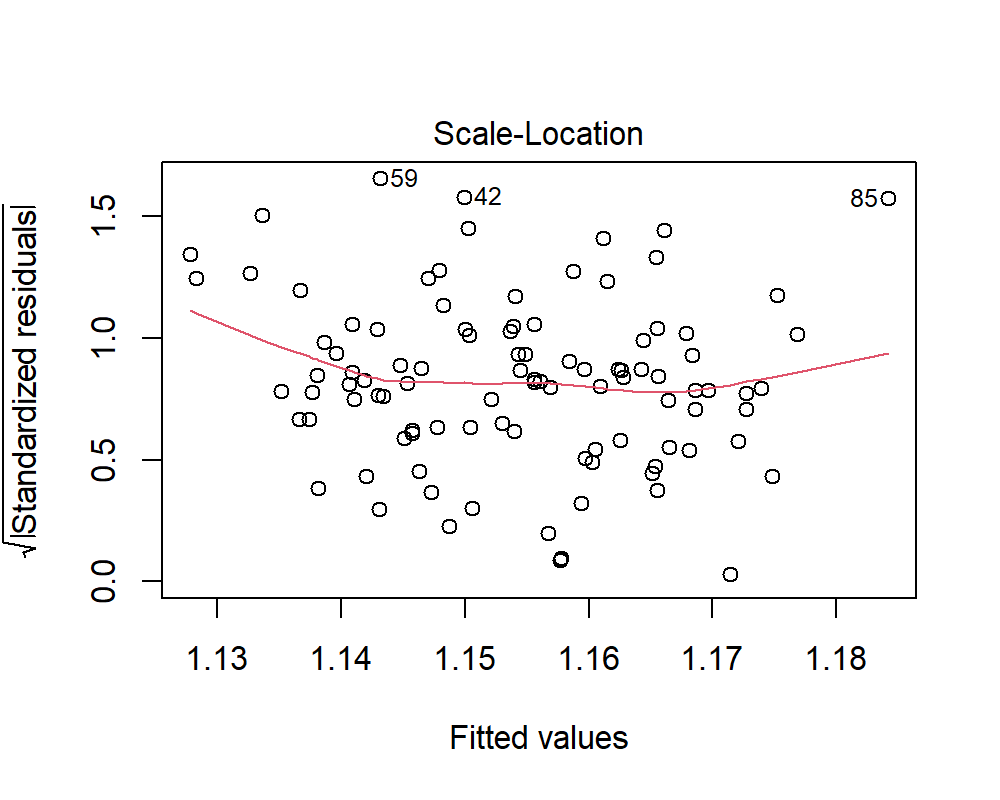
Call:  
lm(formula = stocks\_form1, data = stocks)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.0066732 -0.0015245 0.0003056 0.0019045 0.0073869   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.1083616 0.0011074 1000.89 <2e-16 \*\*\*  
cap.to.gdp 0.0210170 0.0010297 20.41 <2e-16 \*\*\*  
q.ratio 0.0181196 0.0010375 17.46 <2e-16 \*\*\*  
gaap 0.0162510 0.0009174 17.71 <2e-16 \*\*\*  
trailing.pe 0.0144476 0.0009639 14.99 <2e-16 \*\*\*  
avg.allocation 0.0226051 0.0009937 22.75 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.002748 on 94 degrees of freedom  
Multiple R-squared: 0.9534, Adjusted R-squared: 0.9509   
F-statistic: 384.3 on 5 and 94 DF, p-value: < 2.2e-16

Here, all the variables are significant in the model.

## (b)

Check the constant variance assumption of the errors.

plot(x = stocks\_model1,  
 which = 3,  
 sub.caption = NA)



There is no pattern in the above residual plot. The residuals are equally spread over the entire range. Therefore, the assumption of constant variance of the errors is not violated.

## (c)

Check the independentness of the errors assumption.

**Hypothesis**:

There is no correlation among the residuals.

The residuals are auto-correlated.

durbinWatsonTest(stocks\_model1)

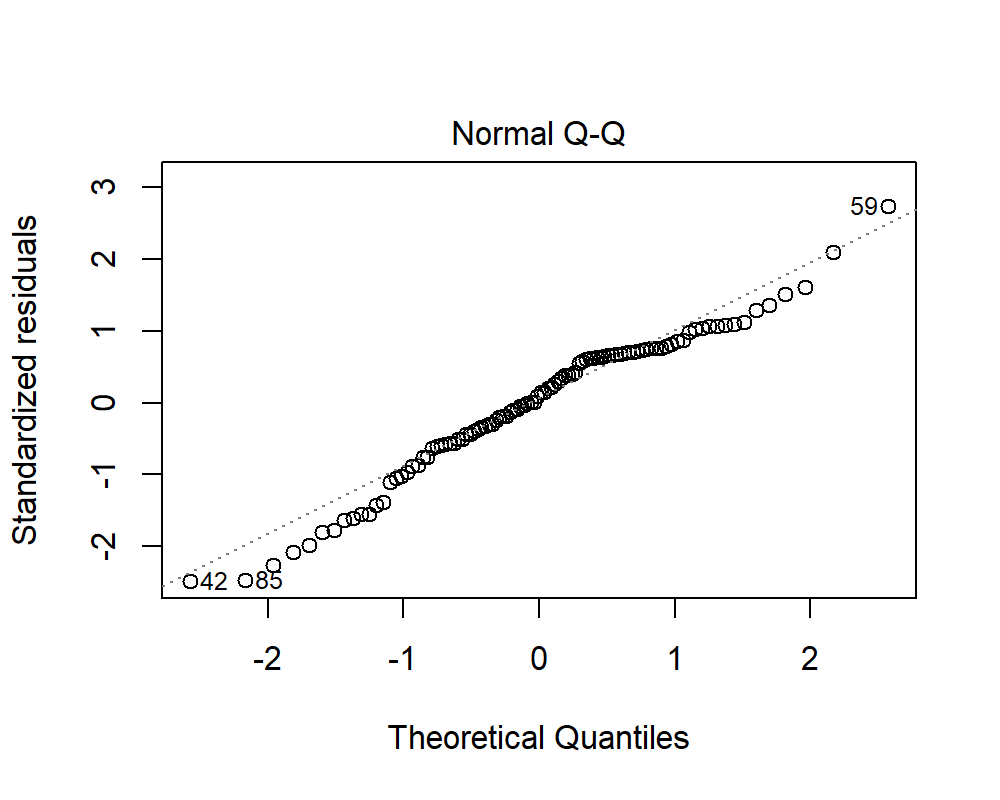
lag Autocorrelation D-W Statistic p-value  
 1 0.07656504 1.839522 0.398  
 Alternative hypothesis: rho != 0

Since, p-value > 0.05, therefore we do not reject and conclude that there is no correlation among the residuals.

## (d)

Check the normality assumption.

plot(x = stocks\_model1,  
 which = 2,  
 sub.caption = NA)



**Hypothesis**:

The residuals are normally distributed.

The residuals are not normally distributed.

shapiro.test(x = residuals(stocks\_model1))

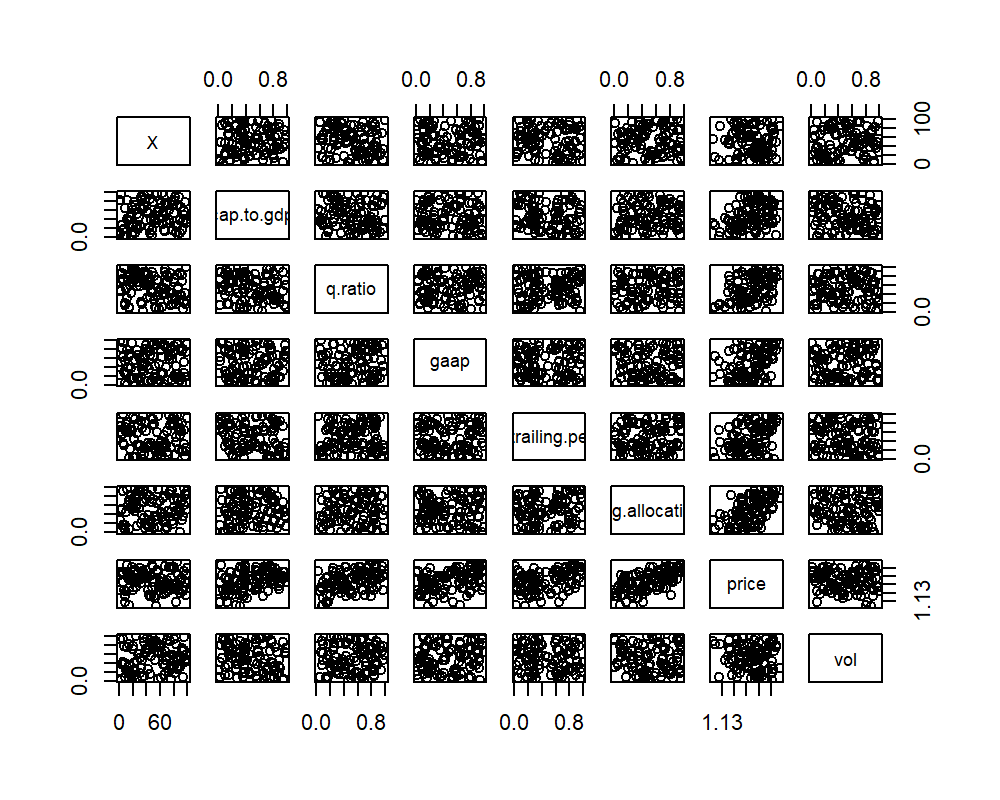
Shapiro-Wilk normality test  
  
data: residuals(stocks\_model1)  
W = 0.97164, p-value = 0.02955

For the Shapiro Wilk Test, since the p-value < 0.05, we reject and conclude that the residuals are not normally distributed. The same can be observed from the QQ-Plot of the residuals. The points deviate from the normal line at tails.

## (e)

Is non-linearity a problem?

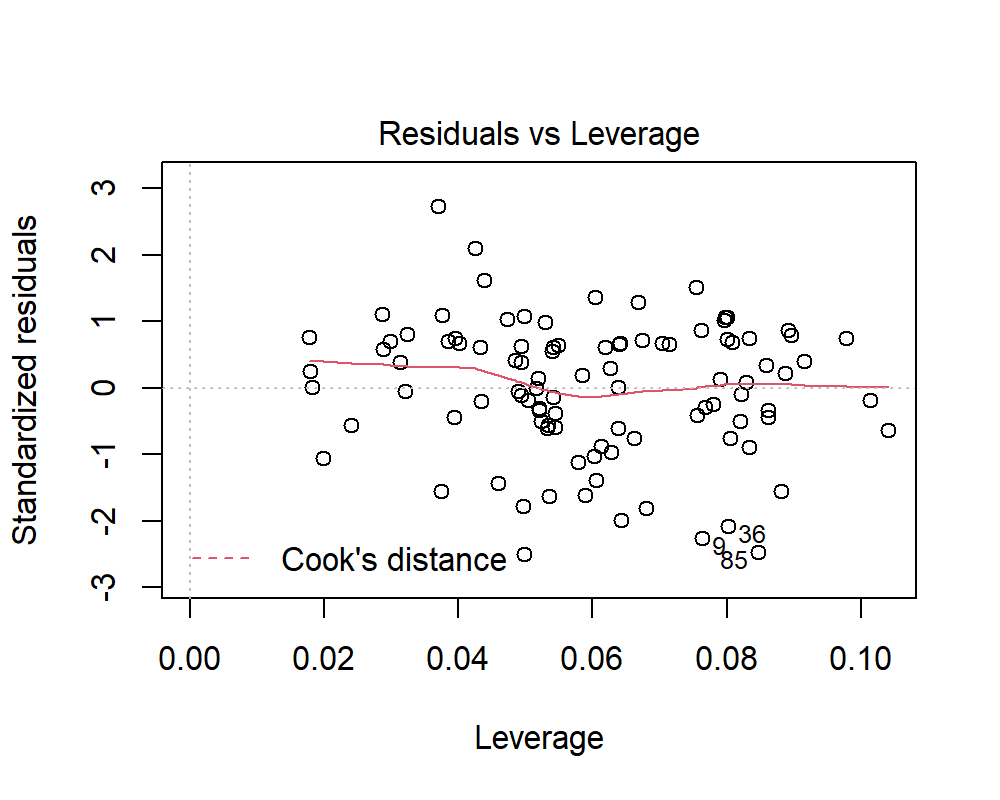
pairs(stocks)



## (f)

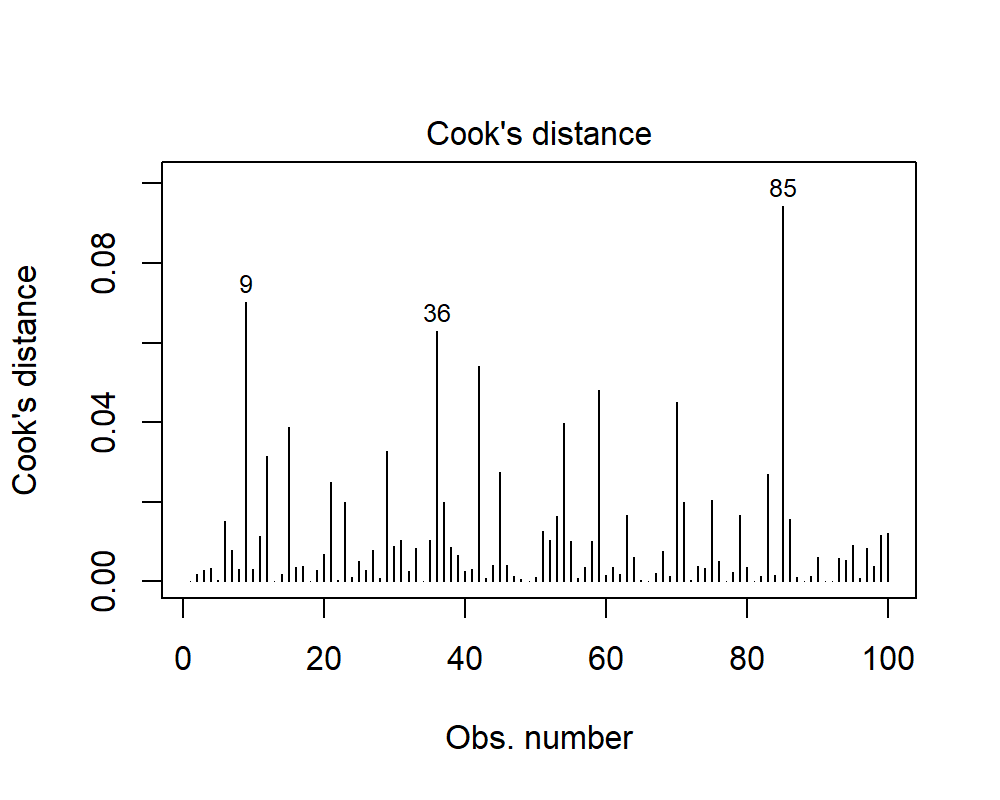
Check for outliers, compute and plot Cook’s Distance

plot(x = stocks\_model1,  
 which = 5,  
 sub.caption = NA)



The observations 9, 36 and 85 are potential outliers in the data.

plot(x = stocks\_model1,  
 which = 4,  
 sub.caption = NA)



## (g)

Check for influential points.

augment(stocks\_model1) %>%   
 data.frame() %>%   
 top\_n(wt = .cooksd,  
 n = 3)

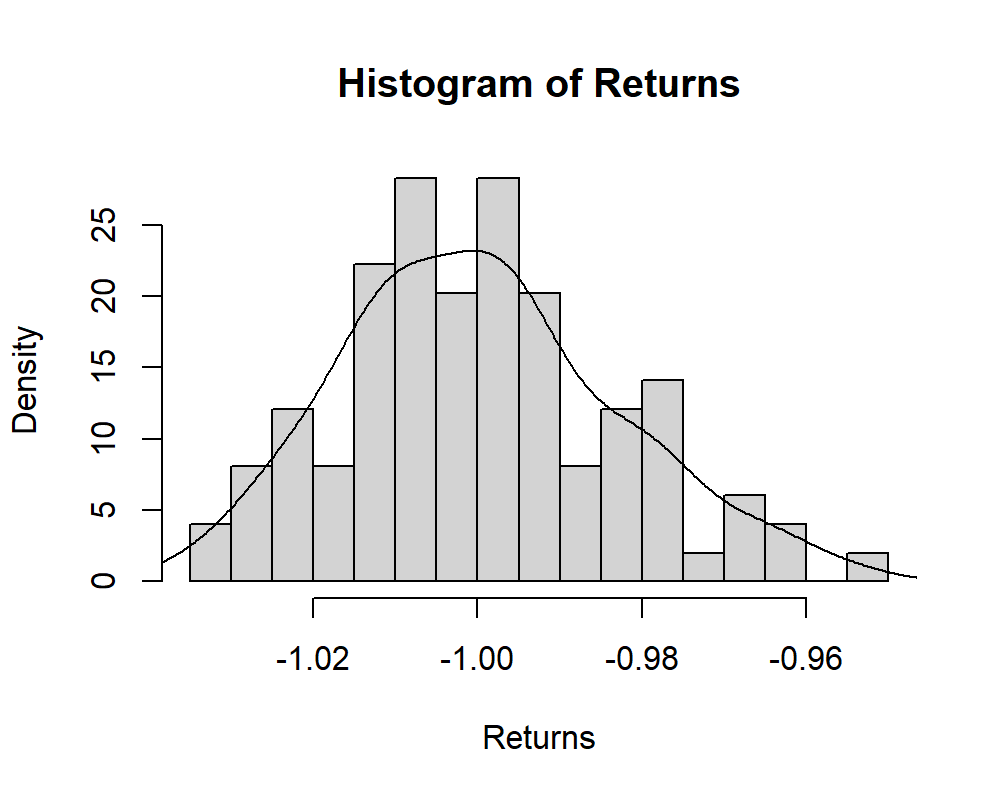
price cap.to.gdp q.ratio gaap trailing.pe avg.allocation .fitted  
1 1.127683 0.02287774 0.1466545 0.5426289 0.6547472 0.1711072 1.133645  
2 1.160614 0.94108754 0.8894984 0.3154241 0.1179111 0.6639421 1.166096  
3 1.177698 0.71841339 0.8152898 0.9223454 0.9728031 0.7486021 1.184199  
 .resid .std.resid .hat .sigma .cooksd  
1 -0.005962750 -2.257768 0.07639043 0.002686820 0.07026806  
2 -0.005481926 -2.079967 0.08017026 0.002698449 0.06284450  
3 -0.006500792 -2.472631 0.08469051 0.002671417 0.09428313

## (h)

The return at time is defined as:

where p is the price data for day . Are the returns normally distributed? Please justify your answer using QQ-Plots and Normality Tests.

returns <- ((stocks$price - lag(x = stocks$price)) - 1)[-1]  
  
hist(x = returns,  
 breaks = 20,  
 probability = TRUE,  
 main = "Histogram of Returns",  
 xlab = "Returns")  
  
lines(density(returns))



**Hypothesis**:

The returns are normally distributed.

The returns are not normally distributed.

shapiro.test(x = returns)

Shapiro-Wilk normality test  
  
data: returns  
W = 0.98581, p-value = 0.3695

Since the p-value > 0.05, we do not reject and conclude that the returns are normally distributed. Also, the histogram of returns follows normal distribution.

# Question 2

Repeat the same from (a) to (h) on the **cheddar** dataset (except part (h)) form the book by fitting a model with **taste** as the response and the other three variables as predictors. Answer the questions posed in the first question.

data(cheddar)  
  
head(cheddar)

taste Acetic H2S Lactic  
1 12.3 4.543 3.135 0.86  
2 20.9 5.159 5.043 1.53  
3 39.0 5.366 5.438 1.57  
4 47.9 5.759 7.496 1.81  
5 5.6 4.663 3.807 0.99  
6 25.9 5.697 7.601 1.09

## (a)

Fit a model to explain **taste** in terms of the predictors. Which variables are important, can any of the variables be removed? Please use F-test to justify.

taste\_form <- "taste ~ ."  
  
taste\_model <- lm(formula = taste\_form,  
 data = cheddar)  
  
summary(taste\_model)

Call:  
lm(formula = taste\_form, data = cheddar)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-17.390 -6.612 -1.009 4.908 25.449   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -28.8768 19.7354 -1.463 0.15540   
Acetic 0.3277 4.4598 0.073 0.94198   
H2S 3.9118 1.2484 3.133 0.00425 \*\*  
Lactic 19.6705 8.6291 2.280 0.03108 \*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 10.13 on 26 degrees of freedom  
Multiple R-squared: 0.6518, Adjusted R-squared: 0.6116   
F-statistic: 16.22 on 3 and 26 DF, p-value: 3.81e-06

The variable **Acetic** turns out to be insignificant in the model.

Therefore, we can test the individual effect of the variable **Acetic** in the model. For that, we can use F-test.

**Hypothesis for Partial F-test**:

The model with **Acetic** is not significantly better than the model without **Acetic**.

The model with **Acetic** is significantly better than the model without **Acetic**.

taste\_form1 <- "taste ~ H2S + Lactic"  
  
taste\_model1 <- lm(formula = taste\_form1,  
 data = cheddar)  
  
anova(taste\_model1, taste\_model)

Analysis of Variance Table  
  
Model 1: taste ~ H2S + Lactic  
Model 2: taste ~ Acetic + H2S + Lactic  
 Res.Df RSS Df Sum of Sq F Pr(>F)  
1 27 2669.0   
2 26 2668.4 1 0.55427 0.0054 0.942

Since the p-value > 0.05, we do not reject and conclude that the model with **Acetic** is not significantly better than the model without **Acetic**.

Now, we’ll also test for the significance of the variable **Lactic** in the model.

**Hypothesis for Partial F-test**:

The model with **Lactic** is not significantly better than the model without **Lactic**.

The model with **Lactic** is significantly better than the model without **Lactic**.

taste\_form2 <- "taste ~ H2S"  
  
taste\_model2 <- lm(formula = taste\_form2,  
 data = cheddar)  
  
anova(taste\_model2, taste\_model1)

Analysis of Variance Table  
  
Model 1: taste ~ H2S  
Model 2: taste ~ H2S + Lactic  
 Res.Df RSS Df Sum of Sq F Pr(>F)   
1 28 3286.1   
2 27 2669.0 1 617.18 6.2435 0.01885 \*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Here, since the p-value < 0.05, we reject and conclude that the model with **Lactic** is better than the model without **Lactic**.

Therefore, the final model is:

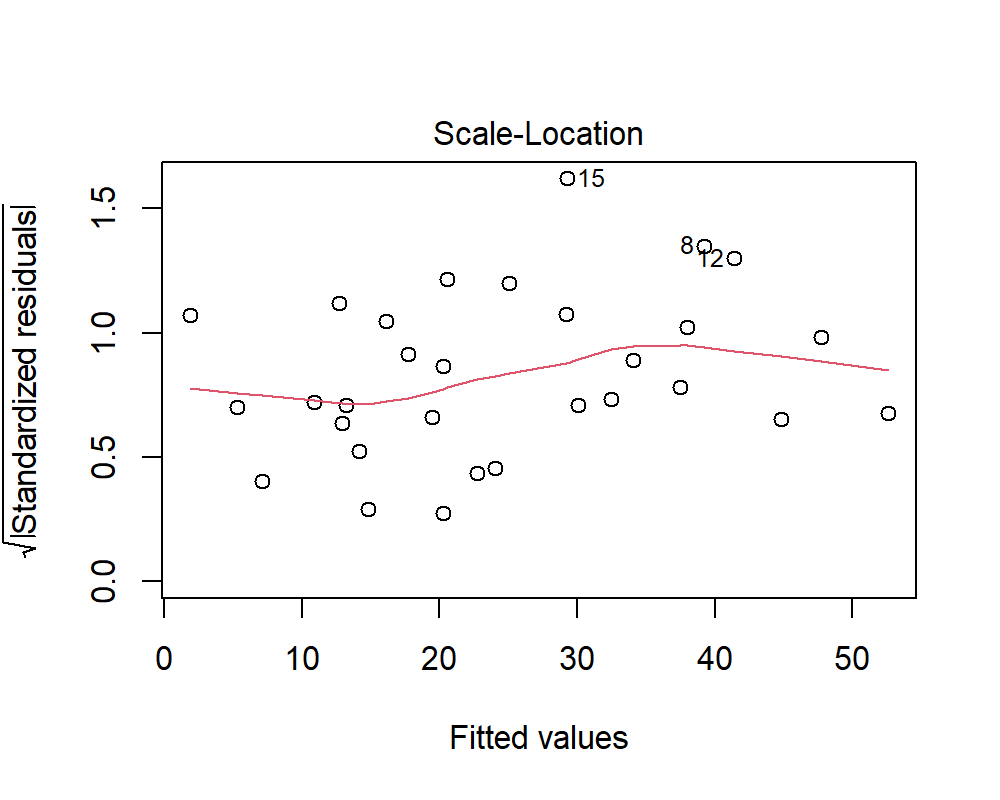
summary(taste\_model1)

Call:  
lm(formula = taste\_form1, data = cheddar)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-17.343 -6.530 -1.164 4.844 25.618   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -27.592 8.982 -3.072 0.00481 \*\*  
H2S 3.946 1.136 3.475 0.00174 \*\*  
Lactic 19.887 7.959 2.499 0.01885 \*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 9.942 on 27 degrees of freedom  
Multiple R-squared: 0.6517, Adjusted R-squared: 0.6259   
F-statistic: 25.26 on 2 and 27 DF, p-value: 6.551e-07

## (b)

Check the constant variance assumption of the errors.

plot(x = taste\_model1,  
 which = 3,  
 sub.caption = NA)



There is no pattern in the above residual plot. The residuals are equally spread over the entire range. Therefore, the assumption of constant variance of the errors is not violated.

## (c)

Check the independentness of the errors assumption.

**Hypothesis**:

There is no correlation among the residuals.

The residuals are auto-correlated.

durbinWatsonTest(taste\_model1)

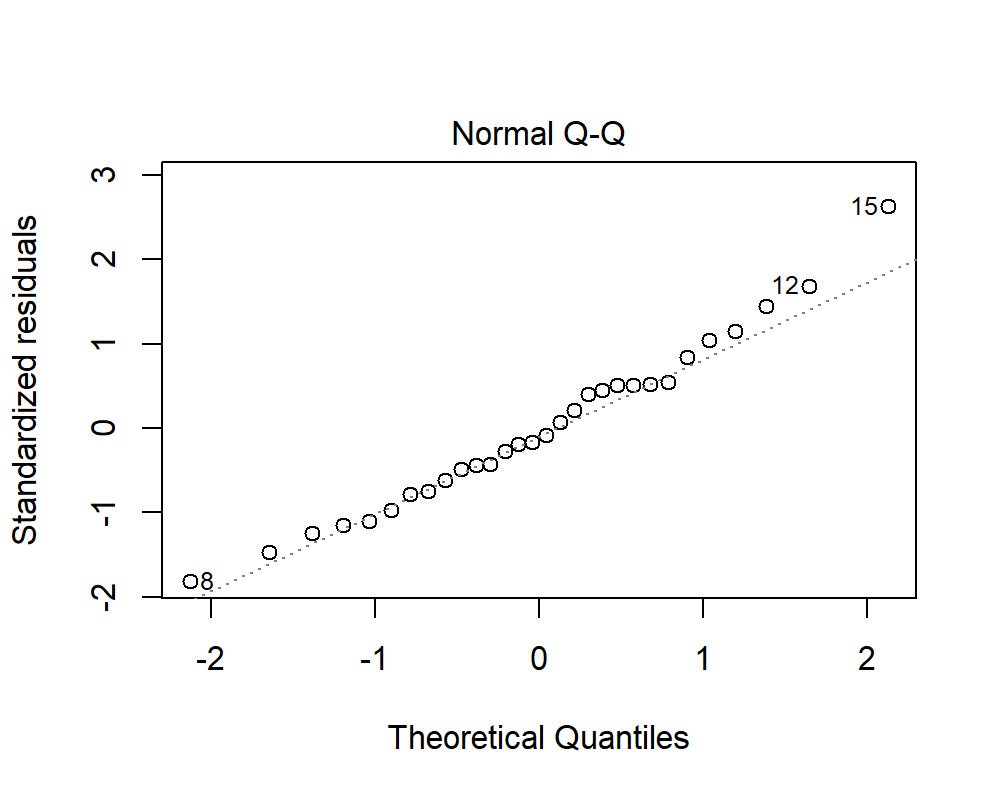
lag Autocorrelation D-W Statistic p-value  
 1 0.167847 1.581086 0.234  
 Alternative hypothesis: rho != 0

Since, p-value > 0.05, therefore we do not reject and conclude that there is no correlation among the residuals.

## (d)

Check the normality assumption.

plot(x = taste\_model1,  
 which = 2,  
 sub.caption = NA)



**Hypothesis**:

The residuals are normally distributed.

The residuals are not normally distributed.

shapiro.test(x = residuals(taste\_model1))

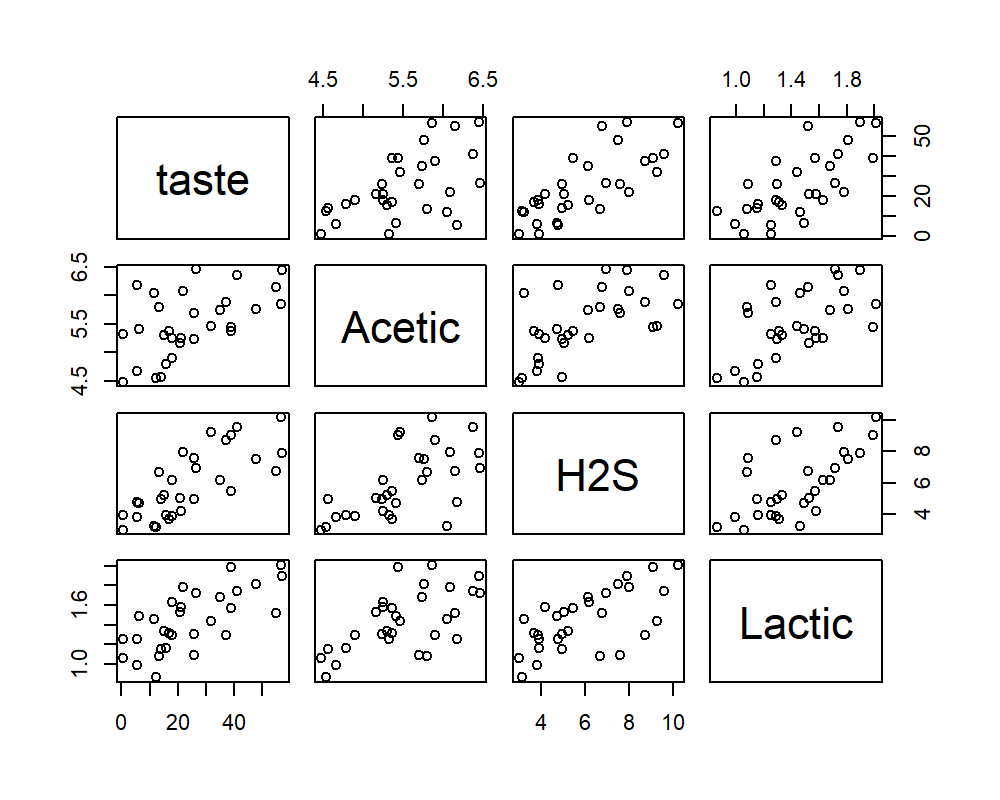
Shapiro-Wilk normality test  
  
data: residuals(taste\_model1)  
W = 0.97945, p-value = 0.8107

For the Shapiro Wilk Test, since the p-value > 0.05, we do not reject and conclude that the residuals are normally distributed. The same can be observed from the QQ-Plot of the residuals.

## (e)

Is non-linearity a problem?

pairs(cheddar)

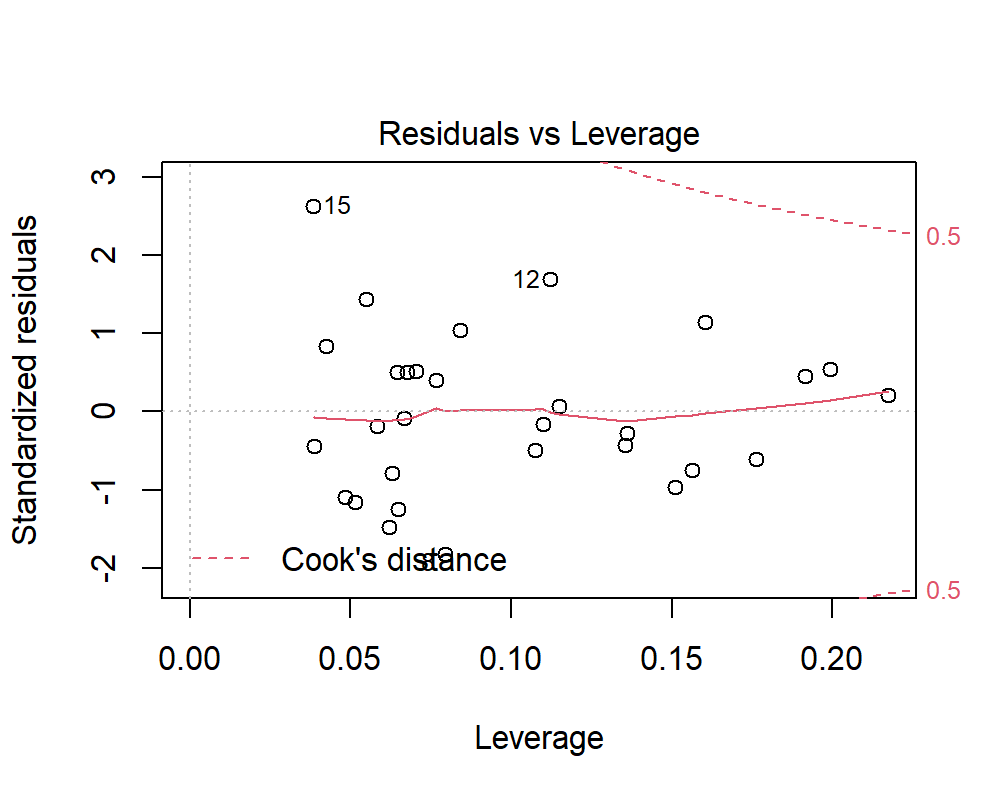


There is no issue of non-linearity in the data. Clear positive linear relationships can be observed. Also, the point to note here is the problem of multi-collinearity in the data as the predictors are also exhibit linear relationships.

## (f)

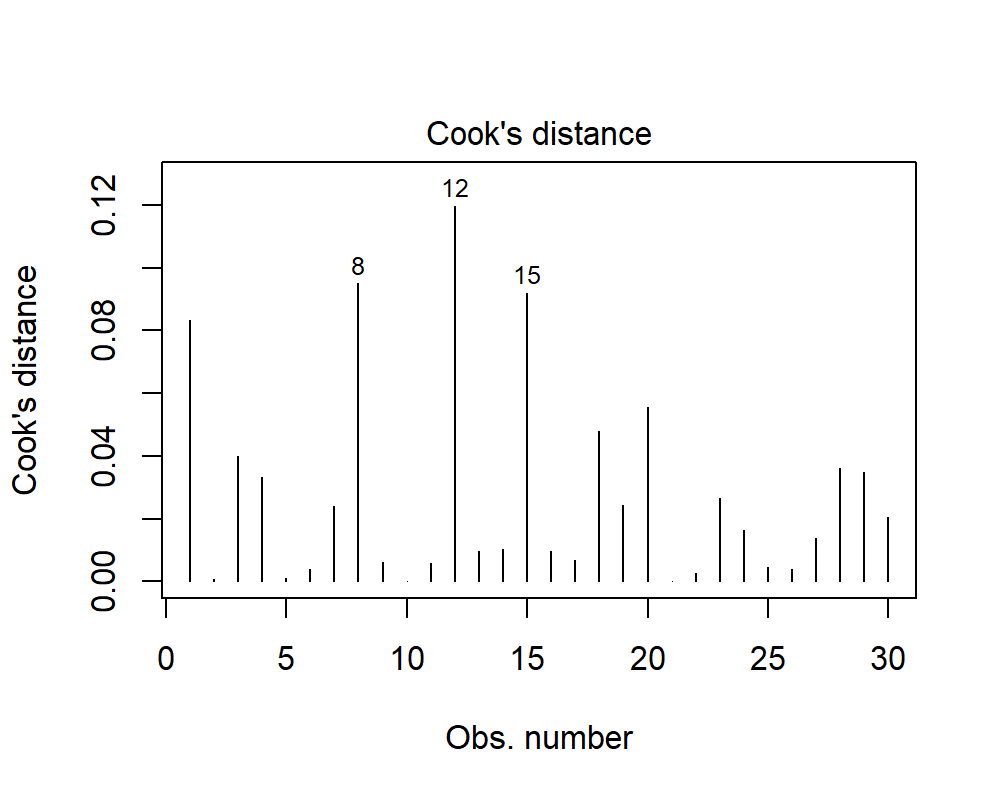
Check for outliers, compute and plot Cook’s Distance

plot(x = taste\_model1,  
 which = 5,  
 sub.caption = NA)



The observations 15, 12, and 30 can be considered as outliers.

plot(x = taste\_model1,  
 which = 4,  
 sub.caption = NA)



We can observe that the observations discussed are the ones having higher value of Cook’s Distance.

## (g)

Check for influential points.

augment(taste\_model1) %>%  
 data.frame() %>%  
 top\_n(wt = .cooksd,  
 n = 3)

taste H2S Lactic .fitted .resid .std.resid .hat .sigma  
1 21.9 7.966 1.78 39.24337 -17.34337 -1.817987 0.07932505 9.491406  
2 57.2 7.908 1.90 41.40096 15.79904 1.686304 0.11200441 9.583383  
3 54.9 6.752 1.52 29.28193 25.61807 2.627527 0.03834518 8.740954  
 .cooksd  
1 0.09492147  
2 0.11955693  
3 0.09176239

# Question 3

The problem is to discover relation between US new house construction starts data (**HOUST**) and macro economic indicators: **GDP**, **CPI**, Population (**POP**). Please download the relevant data from **house.zip** from blackboard. The description for this data can be found in <https://fred.stlouisfed.org/>.

houst <- read\_xls(path = "HOUST.xls",  
 range = "A12:B172",  
 col\_names = c("DATE", "HOUST"))  
  
head(houst)

# A tibble: 6 x 2  
 DATE HOUST  
 <dttm> <dbl>  
1 1975-10-01 00:00:00 297.  
2 1976-01-01 00:00:00 281.  
3 1976-04-01 00:00:00 439.  
4 1976-07-01 00:00:00 434.  
5 1976-10-01 00:00:00 383.  
6 1977-01-01 00:00:00 367.

gdp <- read\_xls(path = "GDP.xls",  
 range = "A21:B183",  
 col\_names = c("DATE", "GDP"))  
  
head(gdp)

# A tibble: 6 x 2  
 DATE GDP  
 <dttm> <dbl>  
1 1976-01-01 00:00:00 58.6  
2 1976-04-01 00:00:00 32.4  
3 1976-07-01 00:00:00 33.6  
4 1976-10-01 00:00:00 47.9  
5 1977-01-01 00:00:00 54.1  
6 1977-04-01 00:00:00 67.7

cpi <- read\_xls(path = "CPI.xls",  
 range = "A57:B217",  
 col\_names = c("DATE", "CPI"))  
  
head(cpi)

# A tibble: 6 x 2  
 DATE CPI  
 <dttm> <dbl>  
1 1976-01-01 00:00:00 0.633  
2 1976-04-01 00:00:00 0.5   
3 1976-07-01 00:00:00 0.9   
4 1976-10-01 00:00:00 0.833  
5 1977-01-01 00:00:00 1.07   
6 1977-04-01 00:00:00 1.03

pop <- read\_xls(path = "POP.xls",  
 range = "A12:B171",  
 col\_names = c("DATE", "POP"))  
  
head(pop)

# A tibble: 6 x 2  
 DATE POP  
 <dttm> <dbl>  
1 1976-01-01 00:00:00 462  
2 1976-04-01 00:00:00 562  
3 1976-07-01 00:00:00 579  
4 1976-10-01 00:00:00 510  
5 1977-01-01 00:00:00 529  
6 1977-04-01 00:00:00 617

## (a)

Data preparation: combine all data into an R dataframe object, and construct dummy or factor variable for 4 quarters. First model is *HOUST ~ GDP + CPI + QUARTER*.

house <- merge(x = houst, y = pop, by = "DATE")  
house <- merge(x = house, y = gdp, by = "DATE")  
house <- merge(x = house, y = cpi, by = "DATE")  
  
house <- house %>%   
 mutate(QUARTER = factor(quarter(x = DATE)))  
  
house\_model1 <- lm(formula = "HOUST ~ GDP + CPI + QUARTER",  
 data = house)  
  
summary(house\_model1)

Call:  
lm(formula = "HOUST ~ GDP + CPI + QUARTER", data = house)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-266.68 -69.19 12.62 71.28 217.21   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 271.0686 20.9148 12.961 < 2e-16 \*\*\*  
GDP 0.2208 0.1207 1.829 0.069328 .   
CPI 1.8468 9.8302 0.188 0.851224   
QUARTER2 105.3363 23.5381 4.475 1.48e-05 \*\*\*  
QUARTER3 88.2852 23.4548 3.764 0.000237 \*\*\*  
QUARTER4 30.4973 23.4202 1.302 0.194801   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 104.4 on 154 degrees of freedom  
Multiple R-squared: 0.1782, Adjusted R-squared: 0.1515   
F-statistic: 6.677 on 5 and 154 DF, p-value: 1.173e-05

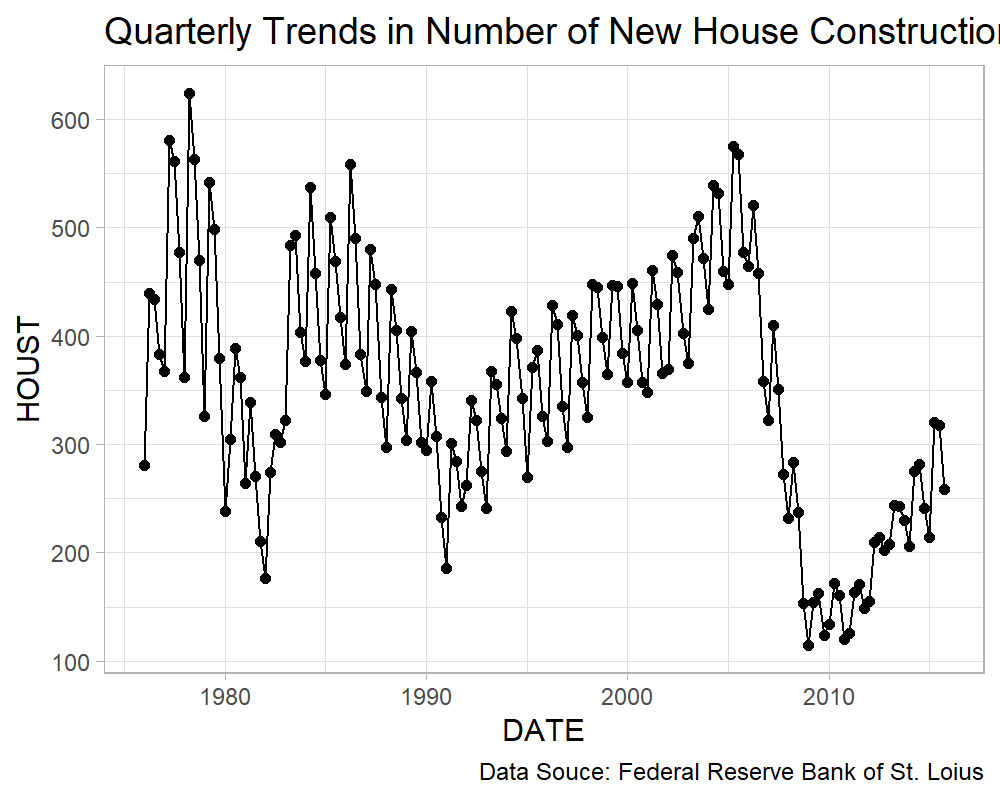
## (b)

Do you think the data needs some cleaning? If so, clean the data.

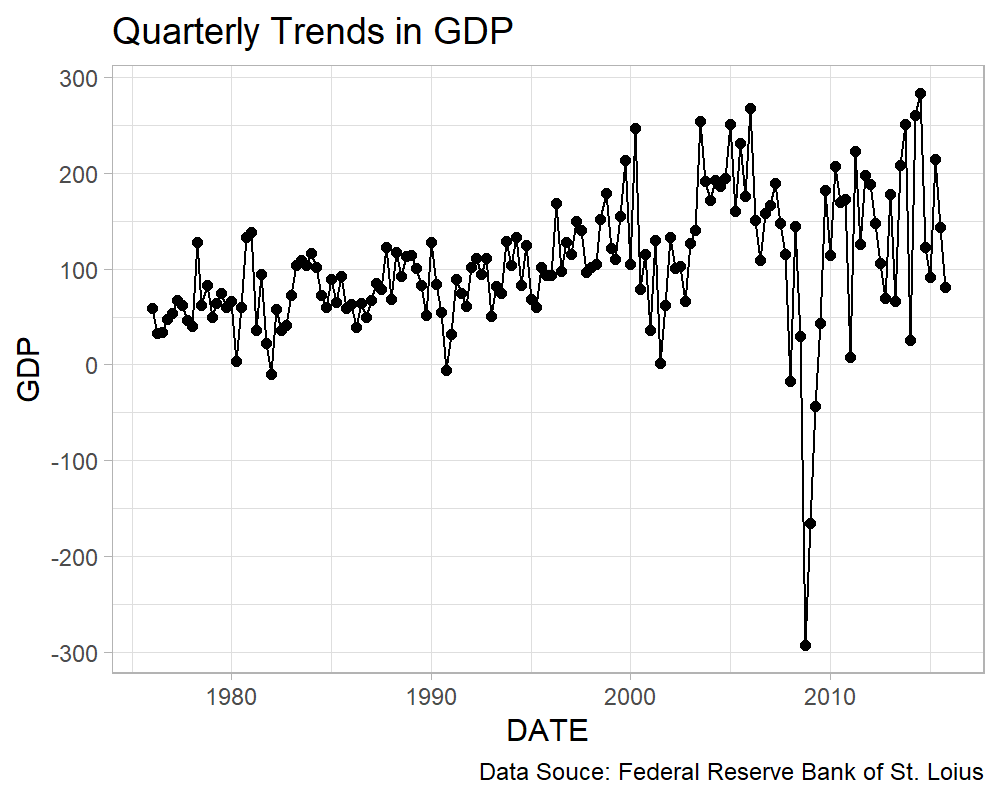
## (c)

Is there a seasonal effect you observe in the data? Show necessary steps and explanation. This is an open-ended question and you are free to use any tool that you find appropriate.

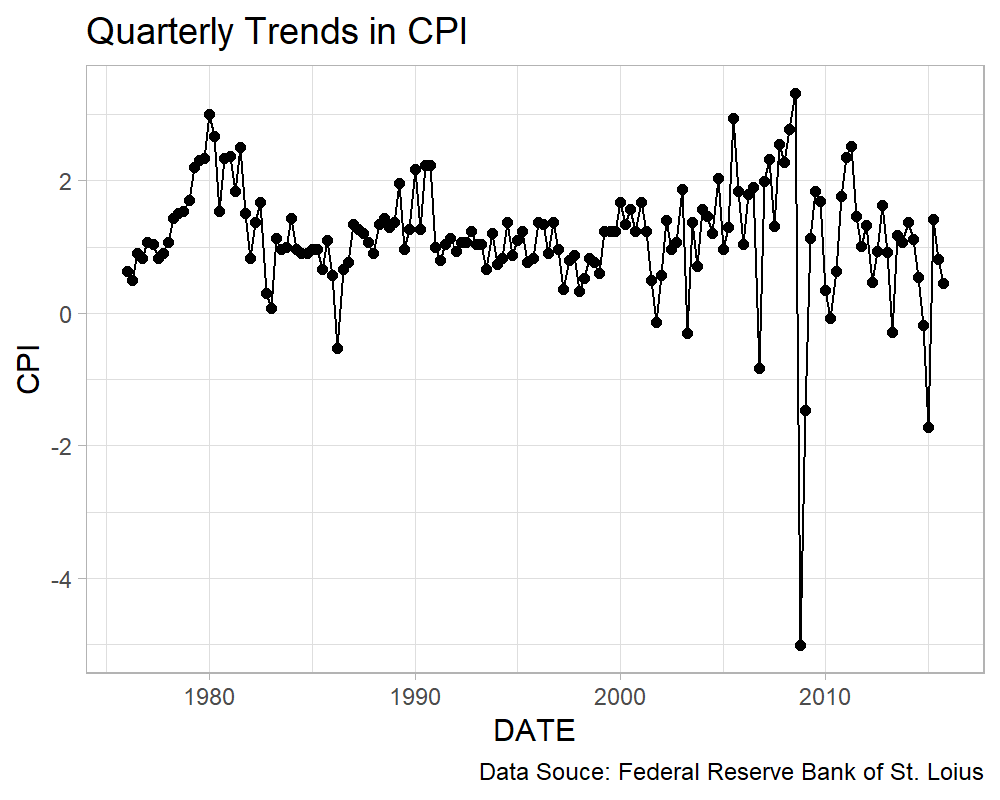
ggplot(house,  
 aes(x = DATE,  
 y = HOUST)) +  
 geom\_point() +  
 geom\_line() +  
 theme\_light() +  
 labs(title = "Quarterly Trends in Number of New House Constructions",  
 caption = "Data Souce: Federal Reserve Bank of St. Loius")



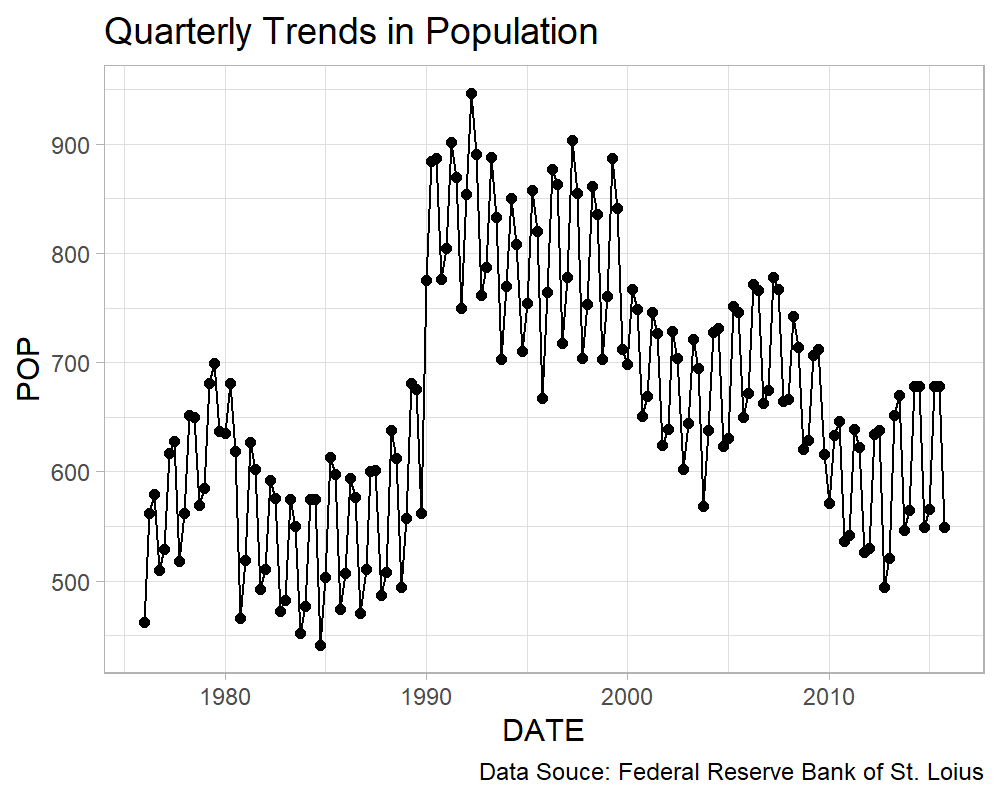
ggplot(house,  
 aes(x = DATE,  
 y = GDP)) +  
 geom\_point() +  
 geom\_line() +  
 theme\_light() +  
 labs(title = "Quarterly Trends in GDP",  
 caption = "Data Souce: Federal Reserve Bank of St. Loius")



ggplot(house,  
 aes(x = DATE,  
 y = CPI)) +  
 geom\_point() +  
 geom\_line() +  
 theme\_light() +  
 labs(title = "Quarterly Trends in CPI",  
 caption = "Data Souce: Federal Reserve Bank of St. Loius")



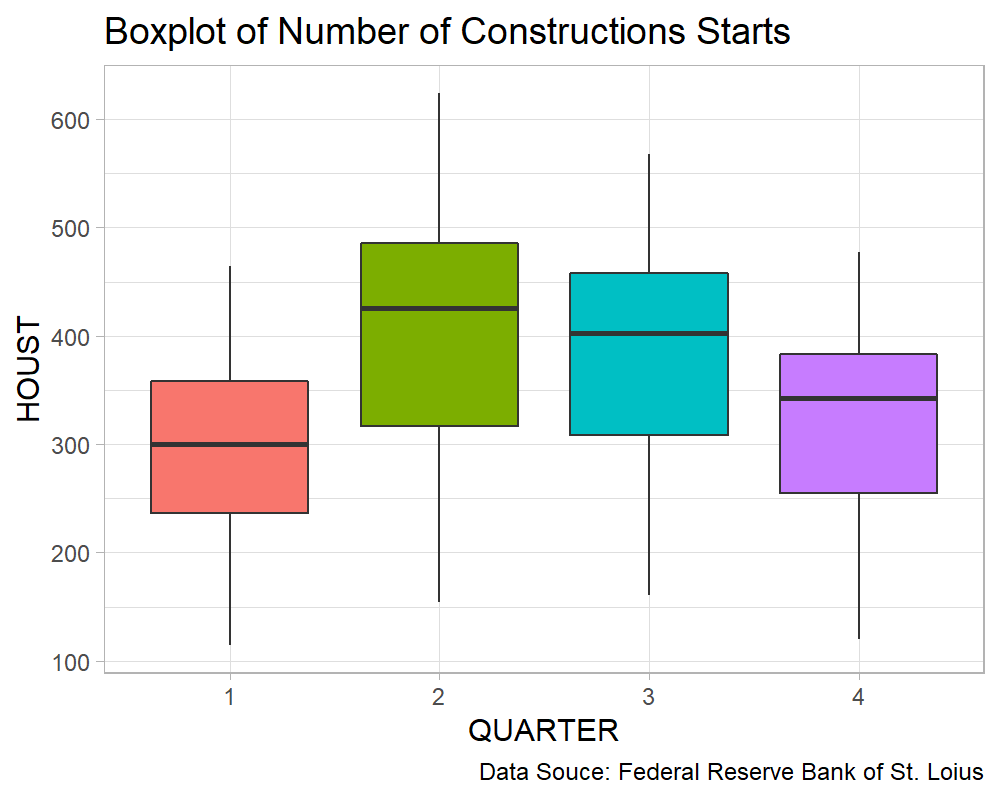
ggplot(house,  
 aes(x = DATE,  
 y = POP)) +  
 geom\_point() +  
 geom\_line() +  
 theme\_light() +  
 labs(title = "Quarterly Trends in Population",  
 caption = "Data Souce: Federal Reserve Bank of St. Loius")



## (d)

Do a pair-wise comparison for different quarters. Which quarter do you think is the best one to buy a house? Show necessary steps and explanation. Use any statistical test or tool that you think is appropriate, this is an open-ended question and there is no one way of answering this question.

ggplot(house,  
 aes(x = QUARTER,  
 y = HOUST,  
 fill = QUARTER)) +  
 geom\_boxplot() +  
 theme\_light() +  
 theme(legend.position = "none") +  
 labs(title = "Boxplot of Number of Constructions Starts",  
 caption = "Data Souce: Federal Reserve Bank of St. Loius")



From the above plot, since the number of construction starts is maximum in the 2nd quarter, it will be recommended for a person not looking for ready-to-move investment option to buy in the 2nd quarter of the year as the prices will be lower considering the number of options available.

We can use ANOVA Test to check whether there is a significant difference between the number of construction starts between various quarters of the year.

**Hypothesis**

There is no significant difference between the number of construction starts between any quarters of the year.

There is a significant difference between the number of construction starts between at least any two quarters of the year.

anova\_model <- aov(formula = HOUST ~ QUARTER,  
 data = house)  
  
summary(anova\_model)

Df Sum Sq Mean Sq F value Pr(>F)   
QUARTER 3 320981 106994 9.696 6.63e-06 \*\*\*  
Residuals 156 1721498 11035   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Since, the p-value < 0.05, we reject and conclude that there is a significant difference between the number of constructions starts between at least any two quarters of the year.

To check which quarters exhibit the difference in the number of construction starts, we will use the pairwise comparison test.

**Hypothesis**

There is no significant difference between the number of construction starts in the pair.

There is a significant difference between the number of construction starts in the pair.

TukeyHSD(anova\_model, conf.level = 0.95)

Tukey multiple comparisons of means  
 95% family-wise confidence level  
  
Fit: aov(formula = HOUST ~ QUARTER, data = house)  
  
$QUARTER  
 diff lwr upr p adj  
2-1 111.2400 50.2389 172.241102 0.0000287  
3-1 92.3625 31.3614 153.363602 0.0007236  
4-1 32.5150 -28.4861 93.516102 0.5111122  
3-2 -18.8775 -79.8786 42.123602 0.8526299  
4-2 -78.7250 -139.7261 -17.723898 0.0055150  
4-3 -59.8475 -120.8486 1.153602 0.0566456

We will only consider the comparison with the 2nd quarter since that’s our point of interest.

From the above table of multiple comparison, we can conclude that, the there is a significant difference in the number of construction starts in the 2nd quarter differ from the 1st and 4th quarter but doesn’t significantly differ from the 3rd quarter.

## (e)

Add population to the first model, do the steps (b) and (c) again.

house\_model2 <- lm(formula = "HOUST ~ GDP + CPI + POP + QUARTER",  
 data = house)  
  
summary(house\_model2)

Call:  
lm(formula = "HOUST ~ GDP + CPI + POP + QUARTER", data = house)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-271.25 -64.11 15.78 70.93 213.90   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 299.00784 53.40285 5.599 9.73e-08 \*\*\*  
GDP 0.22720 0.12149 1.870 0.06337 .   
CPI 1.88789 9.85210 0.192 0.84829   
POP -0.04569 0.08032 -0.569 0.57031   
QUARTER2 109.61624 24.76083 4.427 1.81e-05 \*\*\*  
QUARTER3 91.91961 24.35938 3.773 0.00023 \*\*\*  
QUARTER4 28.98273 23.62236 1.227 0.22174   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 104.6 on 153 degrees of freedom  
Multiple R-squared: 0.1799, Adjusted R-squared: 0.1477   
F-statistic: 5.594 on 6 and 153 DF, p-value: 2.852e-05

# Question 4

Read the **train.csv** and **test.csv** files in R which contains training and test data containing information on ten thousand customers. The aim here is to predict which customer will default on their credit card debt. These datasets contains the following information/variables:

* **default** : A factor with levels **No** and **Yes** indicating whether the customer defaulted on their debt.
* **student** : A factor with levels **No** and **Yes** indicating whether the customer is a student.
* **balance** : The average balance that the customer has remaining on their credit card after making their monthly payment.
* **income** : Income of customer.

## (a)

def\_train <- read.csv(file = "train-default.csv")  
  
head(def\_train)

X default student balance income  
1 1 No No 729.5265 44361.625  
2 3 No No 1073.5492 31767.139  
3 6 No Yes 919.5885 7491.559  
4 7 No No 825.5133 24905.227  
5 9 No No 1161.0579 37468.529  
6 10 No No 0.0000 29275.268

def\_test <- read.csv(file = "test-default.csv")  
  
head(def\_test)

X default student balance income  
1 2 No Yes 817.1804 12106.13  
2 4 No No 529.2506 35704.49  
3 5 No No 785.6559 38463.50  
4 8 No Yes 808.6675 17600.45  
5 11 No Yes 0.0000 21871.07  
6 13 No No 237.0451 28251.70

def\_train <- def\_train %>%   
 mutate(default = ifelse(test = default == "Yes",  
 yes = 1,  
 no = 0),  
 student = ifelse(test = student == "Yes",  
 yes = 1,  
 no = 0))  
  
default\_model1 <- glm(formula = "default ~ balance + income",  
 data = def\_train,  
 family = "binomial")  
  
summary(default\_model1)

Call:  
glm(formula = "default ~ balance + income", family = "binomial",   
 data = def\_train)  
  
Deviance Residuals:   
 Min 1Q Median 3Q Max   
-2.4579 -0.1367 -0.0517 -0.0183 3.3748   
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) -1.173e+01 5.768e-01 -20.340 < 2e-16 \*\*\*  
balance 5.787e-03 3.028e-04 19.115 < 2e-16 \*\*\*  
income 1.719e-05 6.466e-06 2.658 0.00785 \*\*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
 Null deviance: 1723.03 on 6046 degrees of freedom  
Residual deviance: 901.62 on 6044 degrees of freedom  
AIC: 907.62  
  
Number of Fisher Scoring iterations: 8

## (b)

AIC(default\_model1)

[1] 907.6183

## (c)

Give an interpretation of regression coefficients (in words).

**Interpretation**:

coeffs <- round(coefficients(default\_model1), digits = 4)  
  
exp(coeffs[[1]]); exp(coeffs[[2]]); exp(coeffs[[3]])

[1] 8.034225e-06

[1] 1.005817

[1] 1

* For a unit increase in balance, the odds of defaulting credit card debt increases by **1.0058**
* For a unit increase in income, the odds of defaulting credit card debt increases by **1**

## (d)

Form the confusion matrix over the test data. What percentage of the time, are your predictions correct?

def\_test <- def\_test %>%   
 mutate(default\_pred = predict(object = default\_model1,  
 newdata = def\_test,  
 type = "response"),  
 default\_pred = ifelse(test = round(default\_pred) >= 0.5,  
 yes = "Yes",  
 no = "No"))  
  
table(x = def\_test$default,  
 y = def\_test$default\_pred)

y  
x No Yes  
 No 3805 10  
 Yes 99 39

mean(def\_test$default == def\_test$default\_pred) \* 100

[1] 97.2426

The model is able to predict **97.2426%** of the cases in the test data correctly.

## (e)

In your model, what is the estimated probability of default for a student with a credit card balance of $2,000 and an income of $40,000?

What is the probability of the default for a non-student with the same credit card balance and income to default?

coefficients(default\_model1)

(Intercept) balance income   
-1.173178e+01 5.787483e-03 1.718741e-05

coeffs <- coefficients(default\_model1)  
  
prob <- 1/(1 + exp(-(coeffs[[1]] + coeffs[[2]] \* 2000 + coeffs[[3]] \* 40000)))  
  
prob

[1] 0.6296417

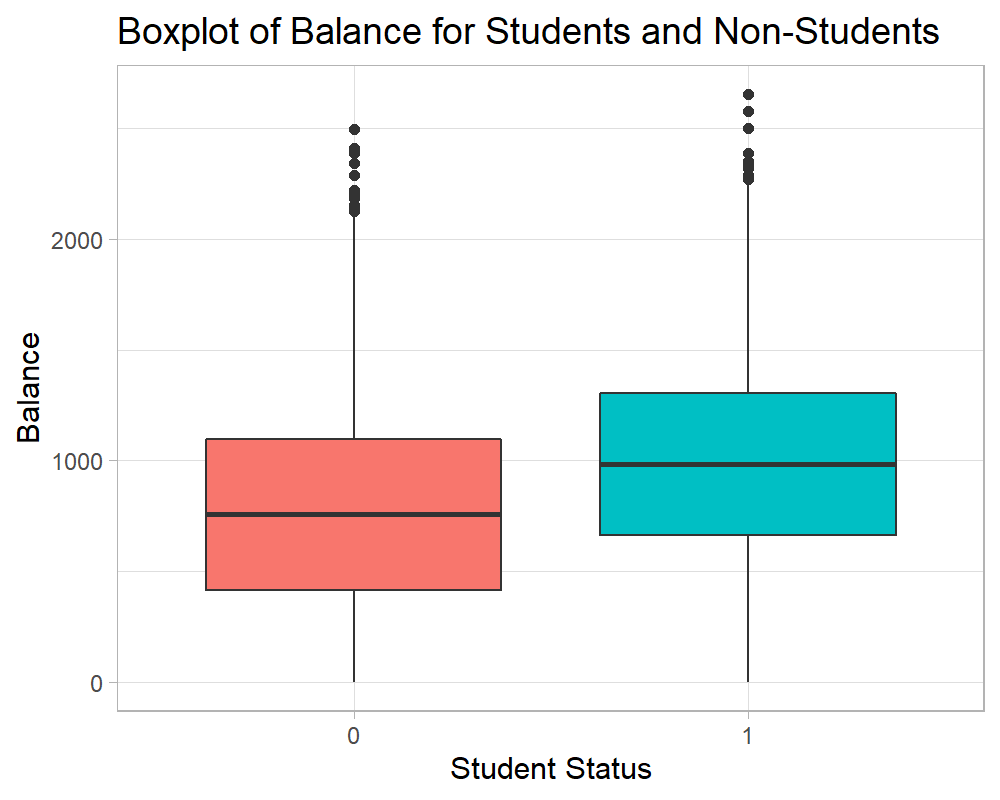
The estimated probability of default for a student with a credit card balance of $2,000 and an income of $40,000 is **0.6296**.

Currently, the model does not take into account the effect of the status of the person being a student or non-student. Therefore, the probability of a non-student with the same credit card balance and income to default is the same i.e. **0.6296**.

## (f)

Are the variables **students** and **balance** correlated? If yes, why do you think this is the case? If no, please explain.

ggplot(def\_train,  
 aes(x = factor(student),  
 y = balance,  
 group = student,  
 fill = factor(student))) +  
 geom\_boxplot() +  
 theme\_light() +  
 theme(legend.position = "none") +  
 labs(title = "Boxplot of Balance for Students and Non-Students",  
 x = "Student Status",  
 y = "Balance")



Here, we can observe that there is a minimal difference between the average balance of students and non-students.

One odd thing to note here is that the average balance for students is more than that of non-students. This may be due to the factor that some of the students included in this data correspond to highly skilled working professionals having a high income that are pursuing further education to enhance their career.

cor(x = def\_train$student,  
 y = def\_train$balance)

[1] 0.2024676

The correlation between the variables **students** and **balance** correlated is **0.2025**. The correlation is very low. Also, the same inference made from the above chart.

## (g)

Now, let’s add the binary variable **student** to the model. Fit a logistic regression model of the form **default ~ balance + income + student**, in other words, regress the variable **default** to all the other predictor with logistic regression.

def\_train <- def\_train %>%   
 mutate(student = factor(student))  
  
default\_model2 <- glm(formula = "default ~ balance + income + student + 0",  
 data = def\_train,  
 family = "binomial")  
  
summary(default\_model2)

Call:  
glm(formula = "default ~ balance + income + student + 0", family = "binomial",   
 data = def\_train)  
  
Deviance Residuals:   
 Min 1Q Median 3Q Max   
-2.4556 -0.1344 -0.0499 -0.0174 3.4155   
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
balance 5.907e-03 3.102e-04 19.040 <2e-16 \*\*\*  
income -5.013e-06 1.079e-05 -0.465 0.642   
student0 -1.091e+01 6.481e-01 -16.830 <2e-16 \*\*\*  
student1 -1.172e+01 5.825e-01 -20.114 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
 Null deviance: 8382.92 on 6047 degrees of freedom  
Residual deviance: 895.02 on 6043 degrees of freedom  
AIC: 903.02  
  
Number of Fisher Scoring iterations: 8

## (h)

Does the data say that it is more likely for a student to default compared to a non-student for different values of income level? Please comment.

coeffs <- round(x = coefficients(default\_model2),  
 digits = 4)  
  
coeffs

balance income student0 student1   
 0.0059 0.0000 -10.9070 -11.7165

The odds for a non-student being a defaulter is **-10.907** while the odds for a student being a defaulter is **-11.7165**. Here, the odds of defaulting the credit card debt is less for students as compared to non-students.

Therefore, the data says that it is more likely for a non-student to default their credit card debt compared to a student for different values of income level.

# Question 5

These days, there are a lot of discussions about what should the healthcare system look like in US. For a scientific discussion, one should need to have a model of demand in the healthcare system. In this question, we will work on the dataset **dvisit** which is about modeling the demand for doctor visits in terms of explanatory variables such as **age**, **income**, **existence** of health insurance and others. To load this dataset, we type in the commands:

install.packages("faraway")  
library(faraway)  
data(dvisits)

which downloads the library **faraway**, loads this library and then pulls up the dataset **dvisits** from this library. The information about this dataset from the Faraway package can be found at the following document: <https://cran.r-project.org/web/packages/faraway/faraway.pdf>

The following exercise is about fitting a model to data and checking diagnostics of it, making sure that our model is right.

**Hints**:  
In class, we provide solutions in R to a similar problem but for a different dataset. I will also give many hints in the class for doing the homework, we will go over the homework questions together.

## (a)

Using the **dvisits** dataset, fit a model with the **hospdays** as the response and other variables as potential predictors. Make sure that variables in your model are significant. Note that there is no single perfect model for this dataset, you can do your best for the fit. We can accept any model as long as your variables are reasonably significant and you can justify the variables in your model in words about why/how they should be predictive. We will accept all the models as long as they do not have any serious flaws in them, so feel free to be creative and do not be afraid about playing with variables. Perform regression disgnostics on this model to answer the following questions. Display any plots that are relevant.

library(faraway)  
data(dvisits)  
  
head(dvisits)

sex age agesq income levyplus freepoor freerepa illness actdays hscore  
1 1 0.19 0.0361 0.55 1 0 0 1 4 1  
2 1 0.19 0.0361 0.45 1 0 0 1 2 1  
3 0 0.19 0.0361 0.90 0 0 0 3 0 0  
4 0 0.19 0.0361 0.15 0 0 0 1 0 0  
5 0 0.19 0.0361 0.45 0 0 0 2 5 1  
6 1 0.19 0.0361 0.35 0 0 0 5 1 9  
 chcond1 chcond2 doctorco nondocco hospadmi hospdays medicine prescrib  
1 0 0 1 0 0 0 1 1  
2 0 0 1 0 0 0 2 1  
3 0 0 1 0 1 4 2 1  
4 0 0 1 0 0 0 0 0  
5 1 0 1 0 0 0 3 1  
6 1 0 1 0 0 0 1 1  
 nonpresc  
1 0  
2 1  
3 1  
4 0  
5 2  
6 0

First, we’ll try to fit the model including all the variables as predictors.

vis\_form <- "hospdays ~ ."  
  
vis\_model <- lm(formula = vis\_form,  
 data = dvisits)  
  
summary(vis\_model)

Call:  
lm(formula = vis\_form, data = dvisits)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-25.794 -0.881 -0.080 0.270 74.496   
  
Coefficients: (1 not defined because of singularities)  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.435267 0.531699 2.699 0.00697 \*\*   
sex -0.385078 0.160614 -2.398 0.01654 \*   
age -8.624152 3.019310 -2.856 0.00430 \*\*   
agesq 10.752913 3.383752 3.178 0.00149 \*\*   
income -0.143137 0.243207 -0.589 0.55620   
levyplus -0.003811 0.182961 -0.021 0.98338   
freepoor -0.249185 0.385497 -0.646 0.51805   
freerepa 0.593503 0.280839 2.113 0.03462 \*   
illness -0.022598 0.064293 -0.351 0.72523   
actdays 0.082811 0.029459 2.811 0.00496 \*\*   
hscore 0.042208 0.038294 1.102 0.27041   
chcond1 0.211276 0.176071 1.200 0.23021   
chcond2 0.881456 0.269974 3.265 0.00110 \*\*   
doctorco -0.174721 0.104492 -1.672 0.09457 .   
nondocco 0.480668 0.078761 6.103 1.12e-09 \*\*\*  
hospadmi 5.530524 0.153613 36.003 < 2e-16 \*\*\*  
medicine -0.059448 0.105265 -0.565 0.57227   
prescrib 0.152538 0.118373 1.289 0.19759   
nonpresc NA NA NA NA   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 5.243 on 5172 degrees of freedom  
Multiple R-squared: 0.2686, Adjusted R-squared: 0.2662   
F-statistic: 111.7 on 17 and 5172 DF, p-value: < 2.2e-16

We can observe that there are many insignificant variables in the model. Rather than testing the significance of individual variables one-by-one we’ll use step-wise regression.

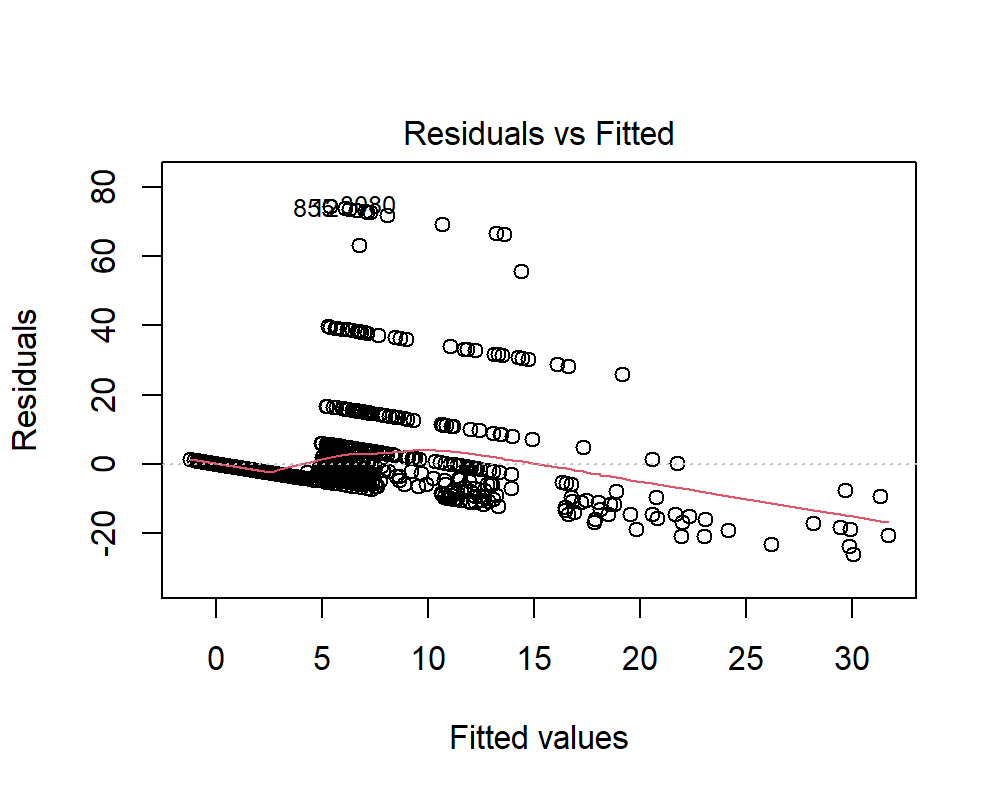
# Intercept Model  
int\_model <- lm(formula = "hospdays ~ 1",  
 data = dvisits)  
  
vis\_step <- step(object = int\_model,  
 scope = list(lower = int\_model,  
 upper = formula(vis\_model)),  
 direction = "forward",  
 data = dvisits,  
 trace = FALSE)  
  
summary(vis\_step)

Call:  
lm(formula = hospdays ~ hospadmi + nondocco + freerepa + chcond2 +   
 actdays + agesq + age + sex + prescrib + doctorco, data = dvisits)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-26.029 -0.868 -0.045 0.258 74.601   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.39221 0.52026 2.676 0.007475 \*\*   
hospadmi 5.54557 0.15318 36.204 < 2e-16 \*\*\*  
nondocco 0.48512 0.07866 6.167 7.47e-10 \*\*\*  
freerepa 0.65017 0.23129 2.811 0.004956 \*\*   
chcond2 0.76962 0.24331 3.163 0.001570 \*\*   
actdays 0.08659 0.02896 2.990 0.002800 \*\*   
agesq 11.08011 3.15707 3.510 0.000453 \*\*\*  
age -8.80575 2.81302 -3.130 0.001756 \*\*   
sex -0.36007 0.15497 -2.323 0.020197 \*   
prescrib 0.11124 0.06267 1.775 0.075933 .   
doctorco -0.16824 0.10404 -1.617 0.105933   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 5.241 on 5179 degrees of freedom  
Multiple R-squared: 0.2681, Adjusted R-squared: 0.2667   
F-statistic: 189.7 on 10 and 5179 DF, p-value: < 2.2e-16

## (b)

Why is your model a good/reasonable model? Check the constant variance assumption for the errors.

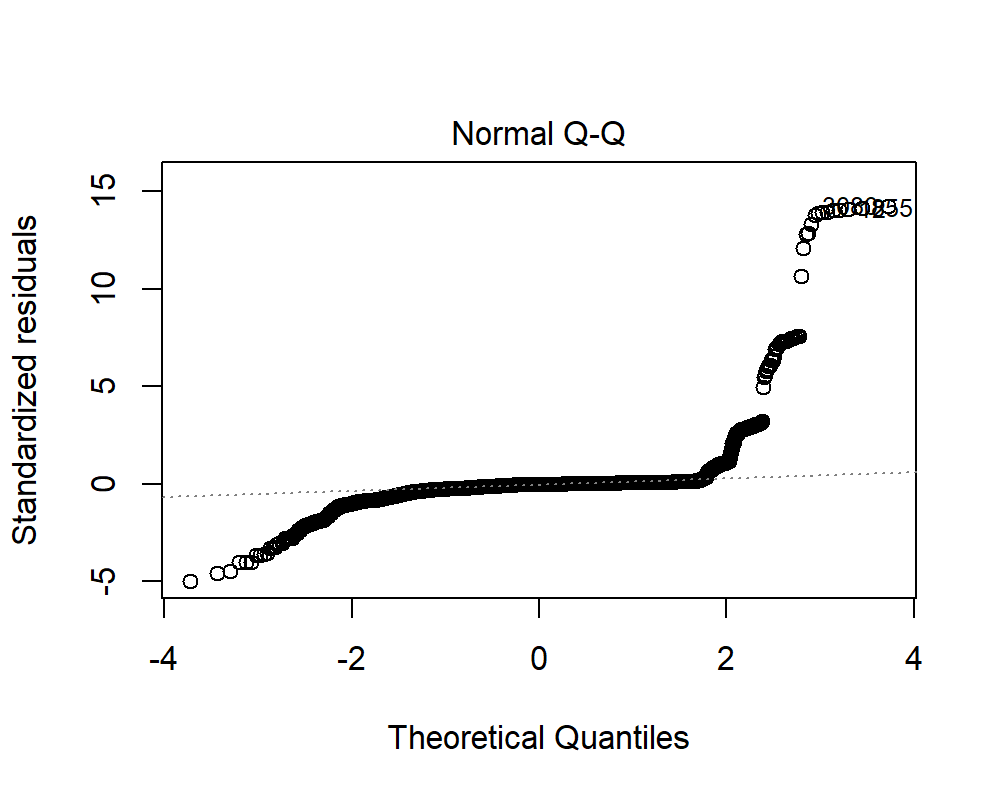
plot(x = vis\_step,  
 which = 1,  
 sub.caption = NA)



## (c)

Check the normality assumption.

plot(x = vis\_step,  
 which = 2,  
 sub.caption = NA)



The residuals are largely deviated from the normal curve.

**Hypothesis**:

The residuals are normally distributed.

The residuals are not normally distributed.

vis\_step\_resid <- sample(x = residuals(vis\_step),  
 size = 1500,  
 replace = FALSE)  
  
shapiro.test(x = vis\_step\_resid)

Shapiro-Wilk normality test  
  
data: vis\_step\_resid  
W = 0.32985, p-value < 2.2e-16

Since, p-value < 0.05, we reject and conclude that the residuals are not normally distributed.

## (d)

Are the errors correlated?

**Hypothesis**:

There is no correlation among the residuals.

The residuals are auto-correlated.

durbinWatsonTest(vis\_step)

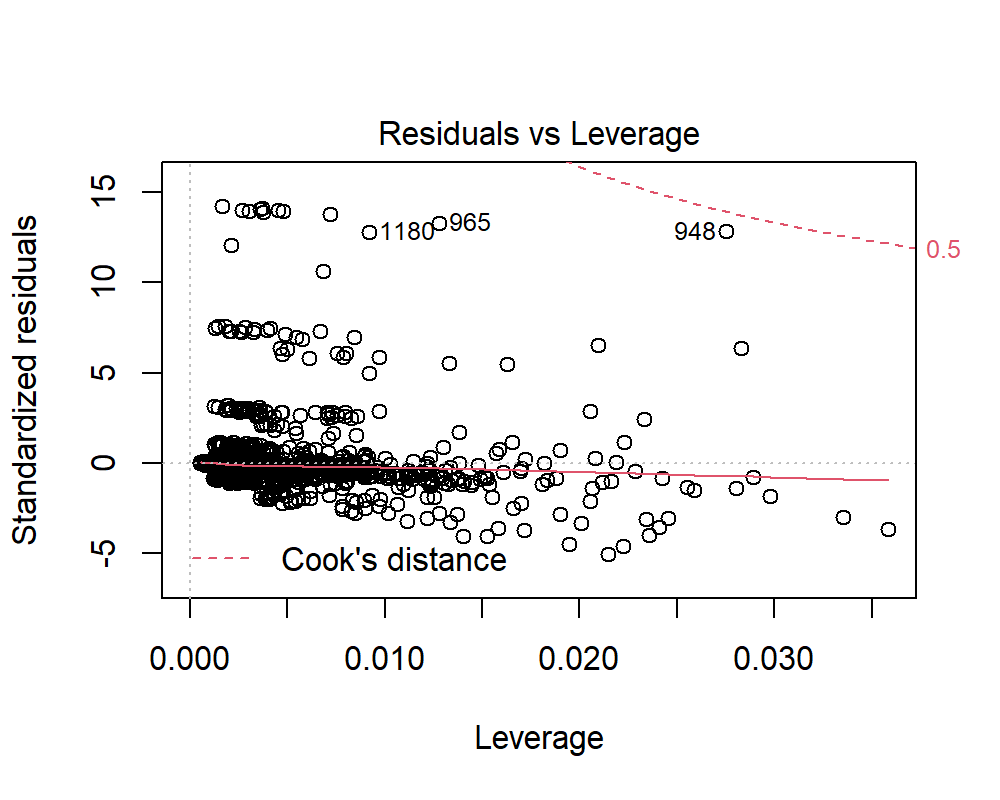
lag Autocorrelation D-W Statistic p-value  
 1 0.01671781 1.96655 0.226  
 Alternative hypothesis: rho != 0

Since the p-value > 0.05, we do not reject and conclude that there is no correlation among the residuals.

## (e)

Check for leverage points, outliers, influential points.

plot(x = vis\_step,  
 which = 5,  
 sub.caption = NA)



The errors are not equally distributed about zero. There are many observations having high positive residuals. Those are the observations that highly divert from the rest of the data.

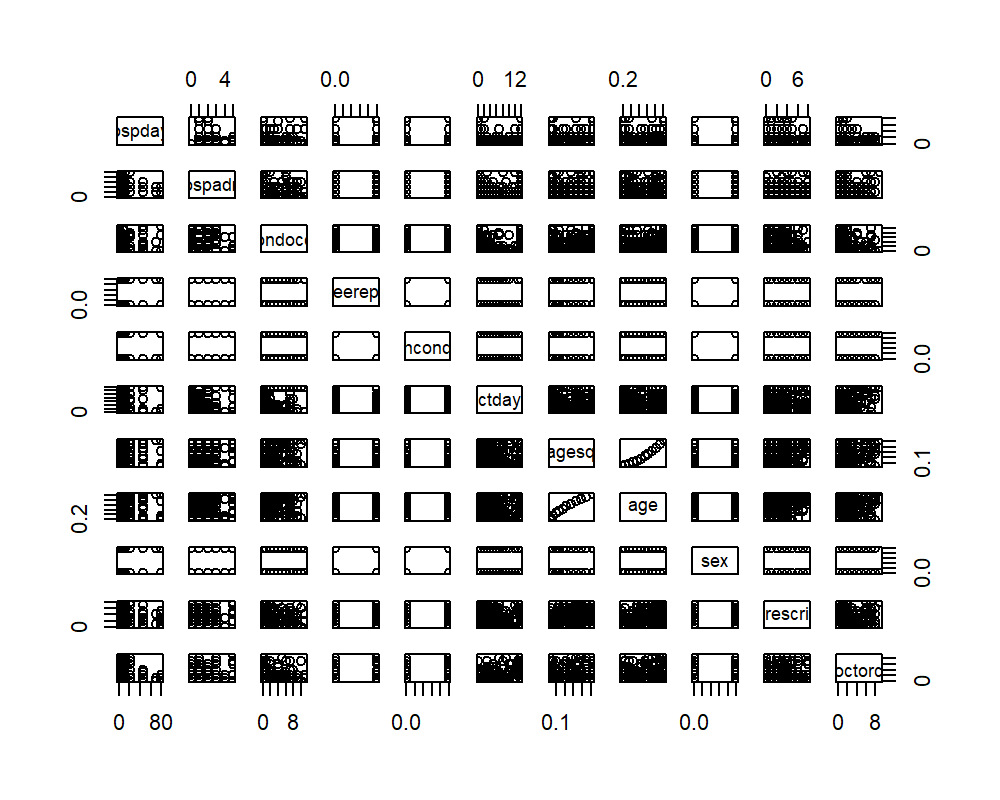
## (f)

Check the structure of the relationship between the predictors and the response.

vis\_step$call$formula[[3]]

hospadmi + nondocco + freerepa + chcond2 + actdays + agesq +   
 age + sex + prescrib + doctorco

plot(dvisits[, c("hospdays", "hospadmi", "nondocco",  
 "freerepa", "chcond2", "actdays", "agesq",  
 "age", "sex", "prescrib", "doctorco")])



There is no significant relationship among the variables in order to support the model for a good fit.

vif(vis\_step)

hospadmi nondocco freerepa chcond2 actdays agesq age sex   
 1.141787 1.089105 1.678198 1.151981 1.320788 64.890428 62.692399 1.132647   
 prescrib doctorco   
 1.486281 1.302793

The variables, **age** and **ageq** exhibit very high multi-collinearity which is very obvious. So, one of the variables have to be removed from the model.

Our PhD students (and Masters students who are interested in doing academic research) can also check out the following research article about this dataset to get more information about economics of healthvare and potential research topics in this direction. However, this material is completely optional, not required for this class (but provided for students interested in research).

# Question 6

The following data provides the COVID-19 cases per state since January:

<https://covidtracking.com/api/v1/states/daily.csv>

The purpose is to predict “the total number of cases in US per day” with linear regression. Please use the data till the end of September for training and the rest for testing. Perform diagnostics on your model and show that your model is a good model. This is a harder question (such as the optional homework 3), so a “perfect model” may not exist; the purpose is to do “our best”.

## Importing cases data

cases <- read.csv(file = "https://covidtracking.com/api/v1/states/daily.csv")  
  
cases <- cases %>%  
 mutate(date = as.Date(x = as.character(date),  
 format = "%Y%m%d"))

## Splitting the data into train and test

cases\_train <- cases %>%   
 filter(month(date) < 10)  
  
cases\_test <- cases %>%   
 filter(month(date) >= 10)

covid\_model <- lm(formula = "total ~ positive",  
 data = cases\_train)  
  
summary(covid\_model)

Call:  
lm(formula = "total ~ positive", data = cases\_train)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-2902935 -50371 -18730 60492 5352816   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 2.038e+04 5.718e+03 3.565 0.000365 \*\*\*  
positive 1.172e+01 5.001e-02 234.309 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 553000 on 11727 degrees of freedom  
 (152 observations deleted due to missingness)  
Multiple R-squared: 0.824, Adjusted R-squared: 0.824   
F-statistic: 5.49e+04 on 1 and 11727 DF, p-value: < 2.2e-16

## Predictions using the model

cases\_test <- cases\_test %>%   
 mutate(total\_pred = predict(object = covid\_model,  
 newdata = cases\_test))

## Model Evaluation

# Model Accuracy on train data  
train\_rmse <- sqrt((sum(covid\_model$residuals ^ 2)) / nrow(cases\_train))  
train\_rmse

[1] 549413.2

# Model Accuracy on test data  
test\_resid <- cases\_test$total - cases\_test$total\_pred  
  
test\_rmse <- sqrt((sum(test\_resid ^ 2)) / nrow(cases\_test))  
test\_rmse

[1] 1971424

Since, RMSE of the model on the train data is less than the RMSE of the model on the test data, we can infer that the model underfits the data.