COVID-19 DATASET

FINAL PROJECT SUBMISSION

AUTHORS: Parth Shah, Mandar Dhande, Abhishek Vakharia
DATE: May 16, 2021.

Prompt 1: THE DATASET

We will be using the COVID-19 dataset to perform our analysis. This dataset is largely derived from studies run in hospitals and nations affected with SARS-COV-2, that generally only admit seriously affected patients to hospitals. However, it should be possible to derive reasonably accurate estimates of these quantities by (a) accounting for the prevalence of asymptomatic patients, and (b) only including sufficiently representative studies.

The dataset is a combination of multiple studies carried across different countries. Each row of the dataset represents a cohort of patients. While some studies examine a single cohort, many of them examine multiple cohorts of patients. Given a cohort, each column represents attributes about this cohort of patients, roughly divided in the following categories:

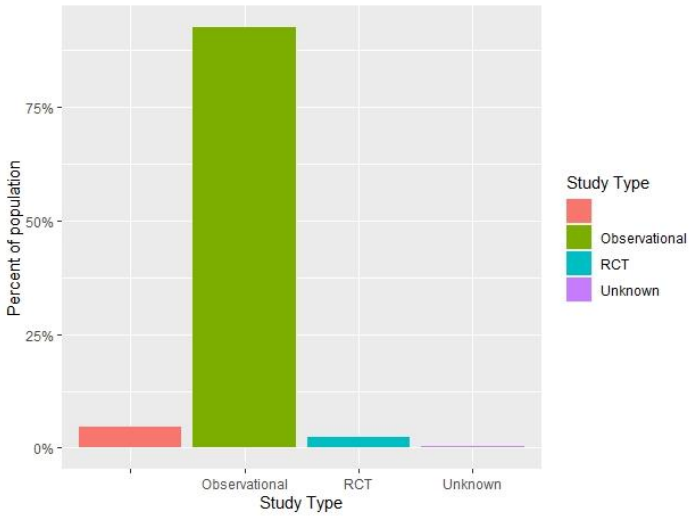
- demographic information (e.g., number of patients in the cohort, aggregated age and gender statistics)
- comorbidity information (e.g., prevalence of diabetes, hypertension, etc.)
- symptoms (including fever, cough, sore throat, etc.)
- treatments (including antibiotics, intubation, etc.)
- standard labs (including lymphocyte count, platelets, etc.)
- outcomes (including discharge, hospital length of stay, death, etc.)

We will be using the different visualization methods (using R) to start analyzing the data and further run to certain conclusions.

SOME SAMPLE SUMMARY STATISTICS TO BETTER EXPLAIN OUR DATASET

1. **STUDY TYPE**
   We have created a gg plot to show the distribution of the type of study that was conducted.

```
library(ggplot2)
ggplot(data,
       aes(x = Study.Type..Observational.or.Randomized.Clinical.Trial..RCT..,
           y = ..count.. / sum(..count..), fill=Study.Type..Observational.or.Randomized.Clinical.Trial..RCT..)) +
  geom_bar()+
  labs(x = "Study Type",
       y = "Percent of population",
       title  = "") +
  scale_y_continuous(labels = scales::percent) +
  guides(fill=guide_legend(title="Study Type"))
```

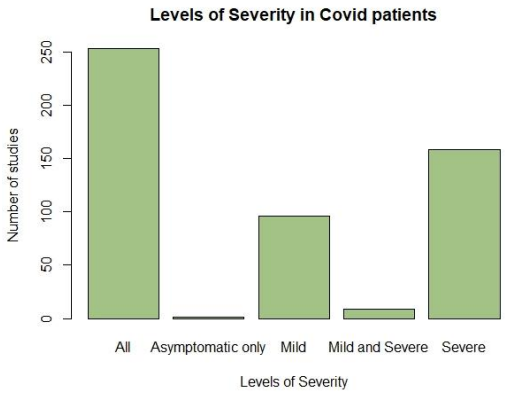


We can clearly observe that most of the studies that were conducted were Observational while only a few were Randomized Clinical trial (where the study is an experiment, and the environment is controlled by the experimenter). We also come across a few unknown values where the type of study remains unknown.

2. SEVERITY

We decided to plot a bar plot to put forth the distribution of the severity of COVID-19 found in the patients across the studies.

```
#Getting rid of the NA values
Severity = data$Severity[!(is.na(data$Severity) | data$Severity=="")]

#Creating vectors
S_mild = c('Mild', 'Mild only', 'Mild Only')
S_severe = c('Severe', 'Severe/Critical Only', 'Severe/critical only')

#Gathering values with the same level of severity
Severity[Severity %in% S_mild] = "Mild"
Severity[Severity %in% S_severe] = "Severe"
Severity[Severity %in% 'Both'] = "Mild and Severe"


#The barplot for severity
barplot(table(Severity)
        , main = 'Levels of Severity in Covid patients'
        , ylab = 'Number of studies', col = rgb(0.4,0.6,0.2,0.6)
        , xlab = 'Levels of Severity')
```

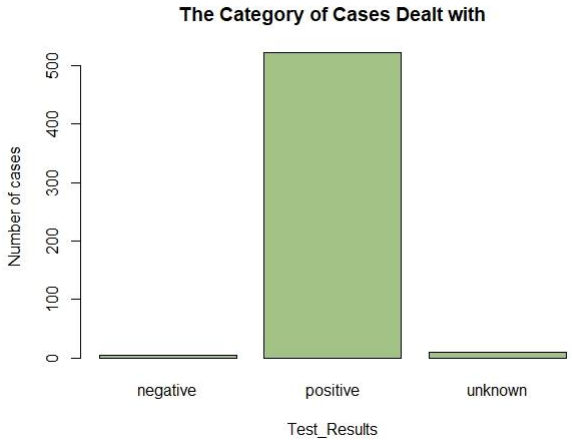


The severity levels in the dataset were not consistent hence performed classification of the mentioned levels before plotting the bar plot. One of them being, 'Mild only', 'Mild Only', 'Mild' where they all belong to the same category, thus classifying it as Mild Only.
In this variable 'All' represents the studies which examine the patients with all levels of severity them being: Mild, Severe and Asymptomatic. Similarly, 'Mild and Severe' represents the studies examining patients with both levels of severity.

3. POSITIVE/NEGATIVE CASES

We decided to plot a bar plot to check the distribution of the type of COVID-19 cases found in the patients across the studies.

```
#Getting rid of the NA values
Cases = data$Positive.negative.cases[!(is.na(data$Positive.negative.cases) | data$Positive.negative.cases=="")]

#Creating vectors
positive = c('Positive only', '', 'Positive','','Positive Only')
negative = c('Negative only')
unknown = c('Positive and Negative/Unconfirmed')

#Gathering values with the same case type
Cases[Cases %in% positive] = "positive"
Cases[Cases %in% negative] = "negative"
Cases[Cases %in% unknown] = "unknown"

#The barplot for Cases
barplot(table(Cases)
        , main = 'The Category of Cases Dealt with'
        , ylab = 'Number of cases', col = rgb(0.4,0.6,0.2,0.6)
        , xlab = 'Test_Results')
```

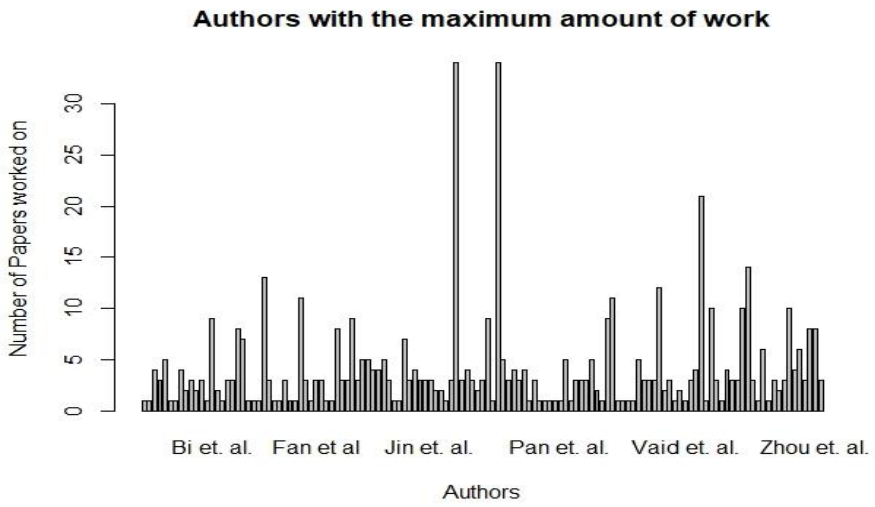


We can clearly see that the number of positive cases is way more than the negative cases. In fact, the number of negative cases is less than the unknown cases as well. The type of cases also gives more credibility, if the study was conducted on patients who were affected rather than the ones did not display any symptoms or were not affected.

4. **AUTHORS**

We decided to plot a bar plot to see how extensively any given author has studied the disease and how has it affected different categories of patients.

```
#The barplot for Author
barplot(table(data$Author)
        , main = 'Authors with the maximum amount of work'
        , ylab = 'Number of Papers worked on'
        , xlab = 'Authors' )
```



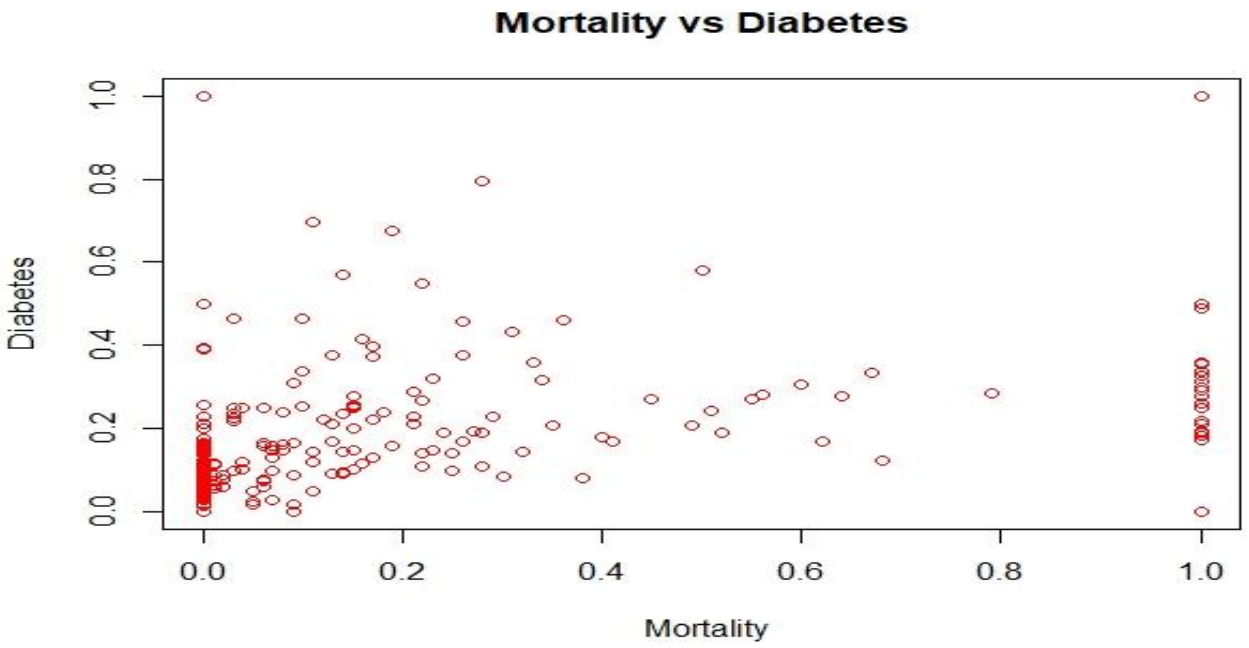
Here, we observe that two authors have worked and have carried out way more research than any other author, hence making their work way more credible than the rest. We can derive such a conclusion since more the studies you are a part of, more is the spectrum of knowledge about that topic. Hence the work by these authors will be way more in depth and more factful than the others.

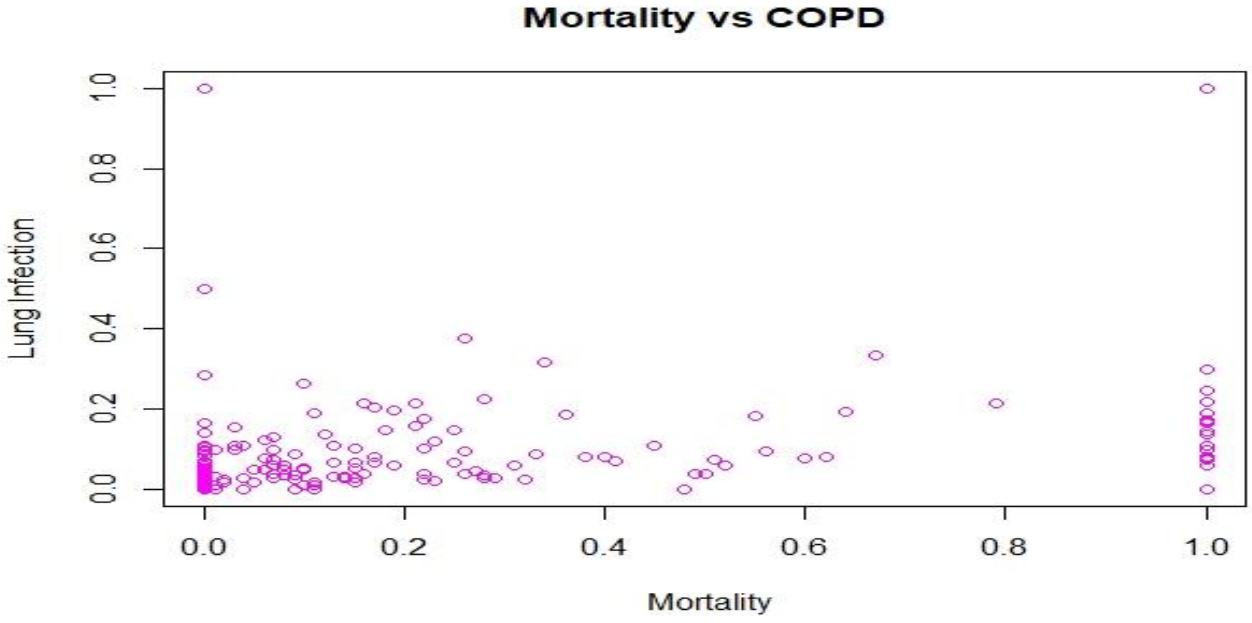
Prompt 2: Statement of the question problem our team is investigating

"Which medical precondition affected the mortality rate and influenced the medication/treatment administered for COVID-19?"

Here, we can use a regression model, probably MLR, using all the medical preconditions as explanatory variables, and observe their individual or combined impact on the mortality rate.

This question intrigued us immensely, due to the fact that, we could determine the approximate extent of each medical precondition among COVID patients and how it ultimately affected the mortality rate. The future scope of this question is brilliant and could really prove to be beneficial in saving many lives, if we can successfully establish what preconditions are the most critical and need urgent care. Consequently, better and timely treatment can be administered that would help combat this deadly virus.


Mortality vs Diabetes
Diabetes
Mortality
0.0
0.2
0.4
0.6
0.8
1.0


Mortality vs COPD
Lung Infection
Mortality
0.0
0.2
0.4
0.6
0.8
1.0

Prompt 3:

Multiple linear regression (MLR), also known as multiple regression, is a statistical technique that predicts the outcome of a response variable by combining many explanatory variables. MLR aims to model the linear relationship between the explanatory (independent) variables and the response (dependent) variable.

Since it includes more than one explanatory variable, multiple regression is essentially an extension of ordinary least-squares (OLS) regression.

Multiple linear regression (MLR) is a good model to use for our dataset as it helps us gauge the change in effect of dependent variables when an independent variable change.

The advantages of using this modelling technique are plenty. First is the ability to determine the relative influence of one or more explanatory variables to the response variable. For instance, we could determine the effect each medical precondition has on a COVID-19 patient's health. Additionally, we can establish whether certain symptoms were/are definitely present in COVID patients.

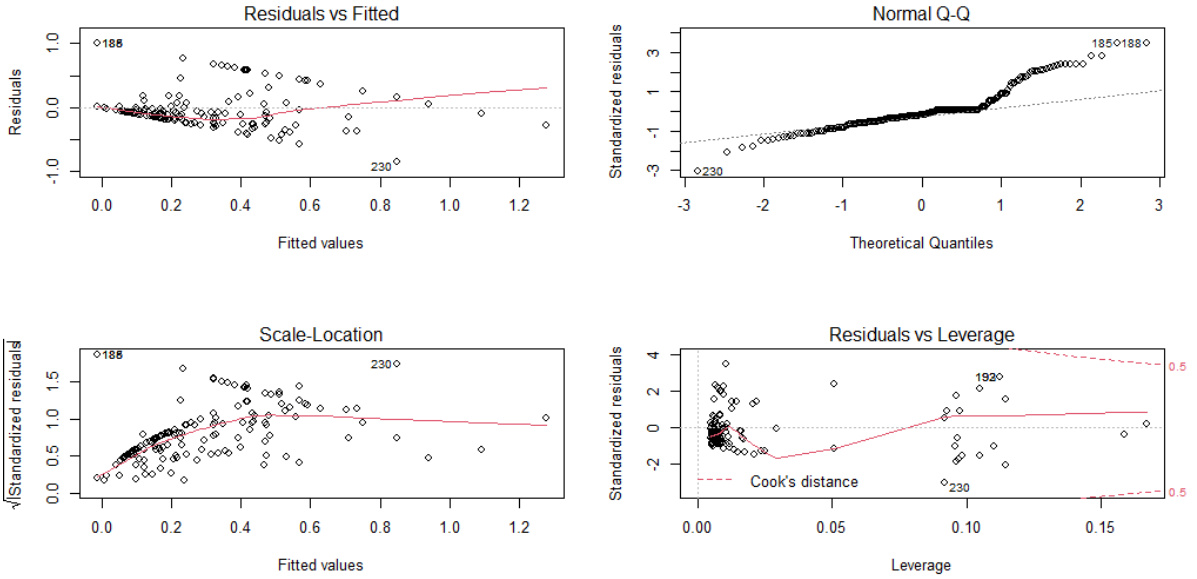
Beta regression is a technique that has been proposed for modelling of data for which the observations are limited to the open interval (0, 1)

Beta regression can be conducted with the *betareg* function in the *betareg* package. With this function, the dependent variable varies between 0 and 1, but no observation can equal exactly zero or exactly one. The model assumes that the data follow a beta distribution.

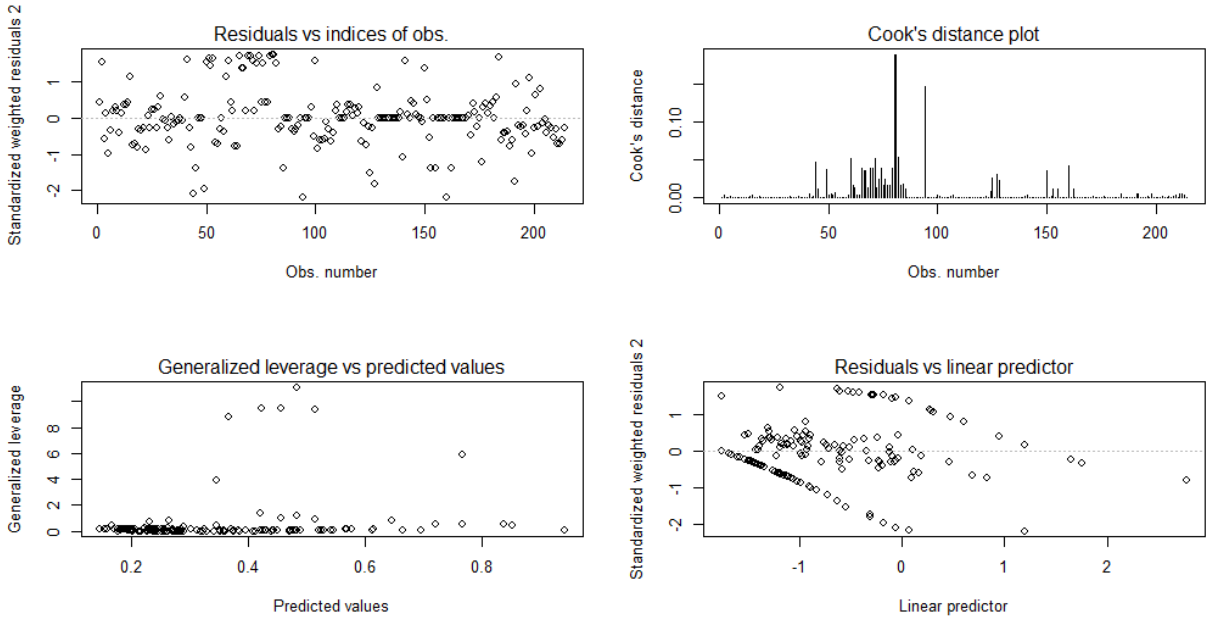
The advantage of using this modelling technique is that all continuous values between 0 and 1 are taken into consideration. This not only saves us from spending time performing data transformation but also makes the process of [data handling simpler. Additionally, beta regression, when compared to logistic regression is better because logistic regression allows only 0/1 values while beta regression permits some variance in our data.

Prompt 4:

1) As we observe the Residuals v/s Fitted plot and Residuals v/s Leverage, we notice that there a few points where the difference between the predicted value and the measured value is significant. These points are nothing but the outliers in our regression analysis. For instance, the points 230, 188, and 192 are the outliers as they lie far away from the best fit line. We see that the Cook's distance exceeds by a value of 0.15 to the right. This further proves the presence of outliers.



For beta regression, the plot shows that the cook's distance exceeds the value by 0.10, indicating there are outliers present.



2) For the first fitting model i.e. Multiple Linear Regression, we started off by creating a new data frame that consisted of 5 columns (4 explanatory and 1 response). The explanatory variables are nothing but the four main medical precondtions that we are considering for our analysis, 'Hypertension' , 'Diabetes' , 'Cancer' , 'Cardiovascular disease' . 'Mortality' is our response or dependent variable. Moving on, we went ahead with renaming the columns for simplicity purposes and ease of use.To improve accuracy and eliminate data redundancy, we removed all the 'NA' values from our dataset and stored them in a new data frame 'data_beta2'.

```
#Making a new frame
data_beta = data.frame(cbind(data$Mortality, data$Hypertension ,data$Diabetes ,
                             data$Cardiovascular.Disease..incl..CAD., data$Cancer..Any.))

#Renaming the columns.
colnames(data_beta)<-c("Mortality","Hypertension", "Diabetes" , "CardioVascularDisease" , "Cancer")

#Omitting rows with NA values.
data_beta2 = na.omit(data_beta)
```

For fitting using beta regression, we are transforming the mortality rate to a range of continuous values between 0 and 1. It's important to note that, all the values with an exact 0 are transformed to 0.001 and all the values with an exact 1 are transformed to 0.99. This makes it possible to include all the values in our regression analysis and returns accurate results.

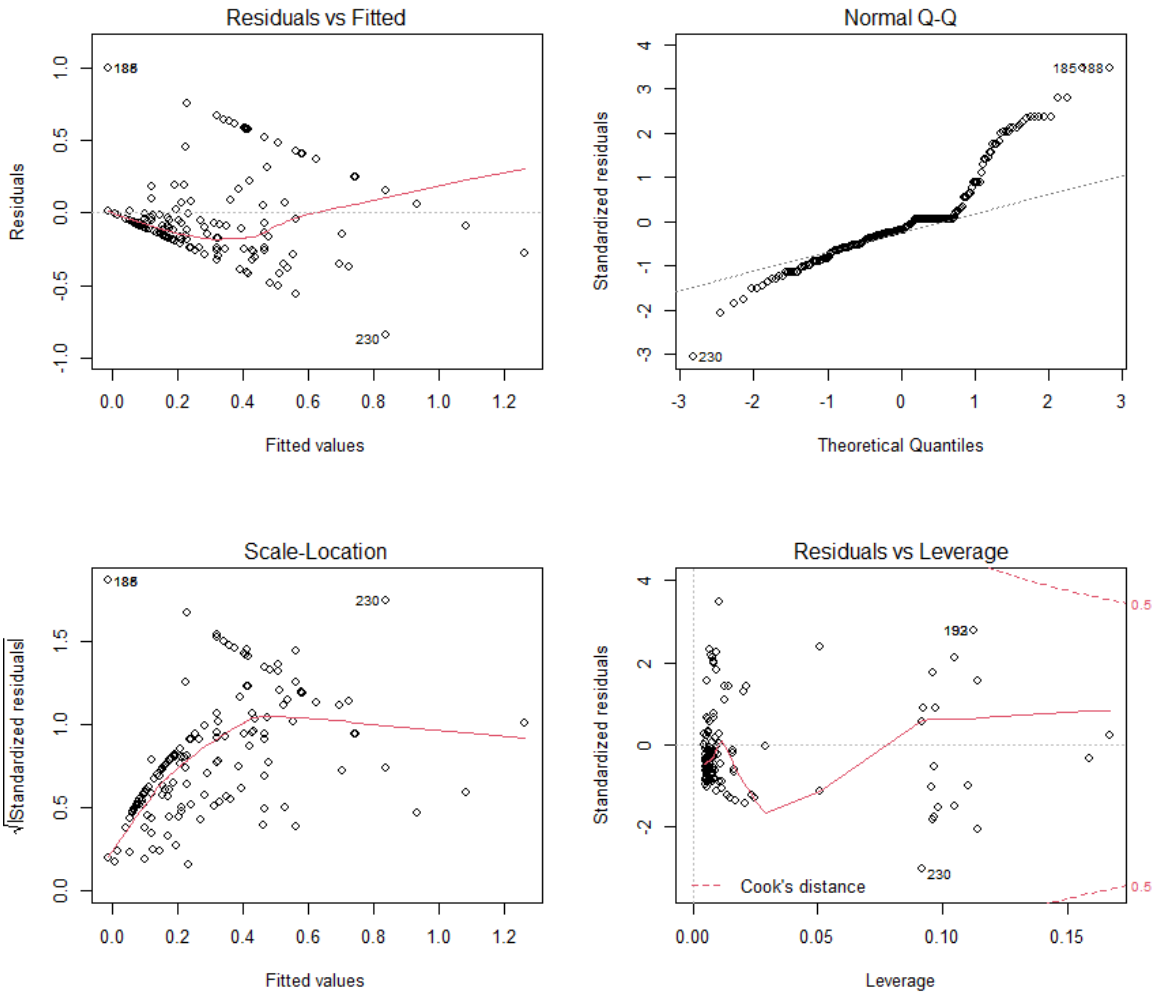
```
# Here we transform the data since only values between 0 and 1 are considered hence the values
# that are exactly 0 are transformed to 0.001
# while the ones that are an exact 1 are transformed to 0.99
data_beta2$Mortality [data_beta2$Mortality == 0] = 0.001
data_beta2$Mortality [data_beta2$Mortality == 1] = 0.99
```

3) We have chosen our 'best' parameters by making use of the concept of Forward Selection. Forward selection is a form of stepwise regression that starts with a blank model and gradually adds variables. Every time you take a step forward, you add the one variable that improves your model the most.

It is one of two widely used stepwise regression methods; the other, almost opposite, is backward elimination. Starting with a model that contains every possible variable, you delete the unnecessary variables one by one.

In forward selection, we consider the small p value which indicates a significant association with the response variable.

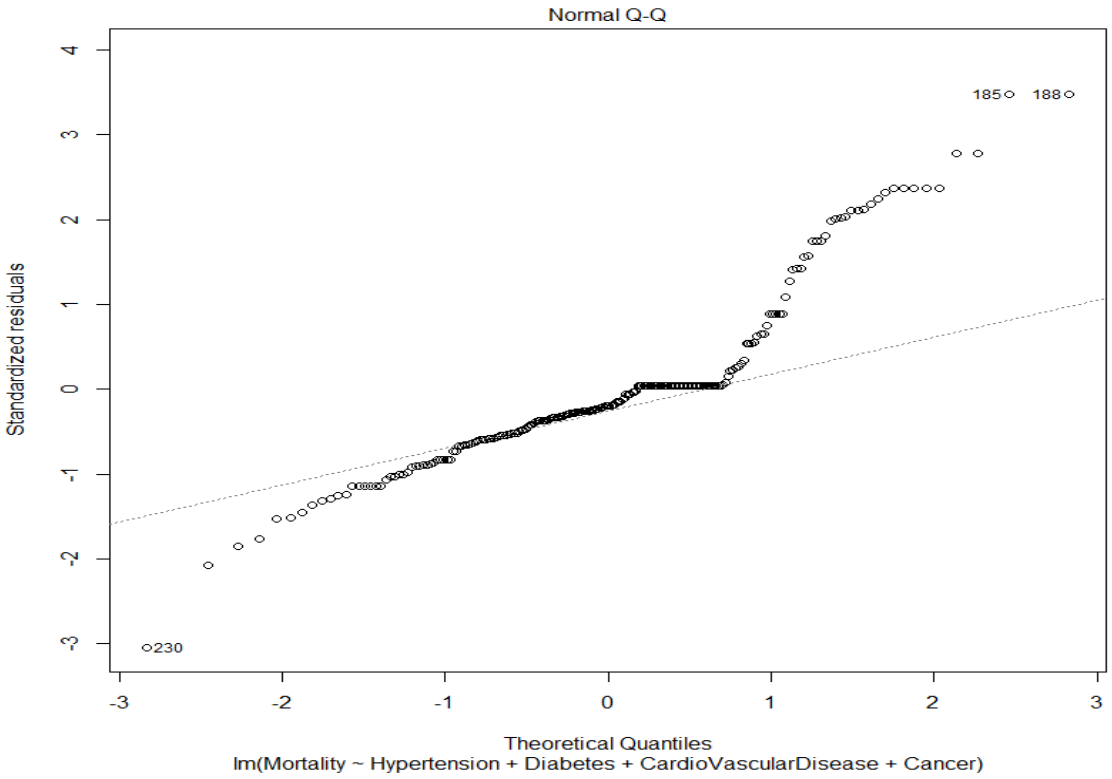
4)



For Beta Regression, we haven't actually gone ahead and made a plot but obtained a summary that returns a phi coefficient for each of the explanatory variable. As we see the below image, the phi coefficient for this fitting model is 0.9094 which signifies that the model is a very good fit for our dataset.

```
> print (model)

Call:
betareg(formula = Mortality ~ Hypertension + Diabetes + CardioVascularDisease + Cancer, data = data_beta2)

Coefficients (mean model with logit link):
          (Intercept)          Hypertension              Diabetes  CardioVascularDisease                Cancer
              -1.7596                1.1280                1.5831                1.8239                0.5642

Phi coefficients (precision model with identity link):
 (phi)
0.9094

> # We get a phi coefficient value of 0.9094 , which tells us that it a great model for this kinf of a data set
>
```

5) In the below QQ plot for Multiple Linear Regression, we observe that most of the data lies well along the best fitted line. Additionally, there is some skewness present on the right side of the graph owing to the presence of several outliers. Thus, it would be safe to conclude that if not the ideal or perfect fit, this is most definitely a very good fit for our data.



Prompt 5:

For Multiple Linear Regression, using forward selection, we have selected parameters that have a very low p-value, hence they have a significant association with the response variable and we also see that the normal Q-Q plot has variables along the regression line indicating a very good fit if not perfect for dataset.

For Beta-regression we get a p-value of 0.9094, which indicates the model is a very good fit for the dataset.

```
> #Here we carry out intrapolation of the data by defining values in the range of the explanatory variables
> predict (mlr2, newdata = data.frame(Hypertension = 0.75 ,  Diabetes = 0.45 , Cardiovascular.Disease..incl..CAD. = 0.20 , Cance
r..Any. = 0.10))
        1
0.5621539
> # With these values the model predicts that there is a 56.2139% chance of Mortality
> #Now we try to extrapolate the data by giving values beyond the scope for it
> predict (mlr2, newdata = data.frame(Hypertension = 1.1 ,  Diabetes = 1.1 , Cardiovascular.Disease..incl..CAD. = 1.1, Cancer..An
y. = 1.10))
       1
1.67469
> # When we extrapolate the data with the given values it throws a prediction beyond 100% - 167.469%
>
```

We have carried out interpolation of the data by defining values in the range of our explanatory variables i.e., medical preconditions. This model states that there is a 56.2139% chance of a person passing away having one of these preconditions.

```
0.5578115
> #We see that there is a slight change in the prediction , here it is 55.78% chance of mortality , hence the output can be consid
ered more
> #accurate since it is only that data which out model is defined to run on
>
```

To improve accuracy, we eliminated all the NA's from our dataset and then applied interpolation. This returned us with a 55.78% chance of a patient passing away due to COVID-19 having one or more medical preconditions that we are considering.

Prompt 6: BIASES IN COLLECTION OF DATA

1. We observe that the studies have been carried out in the initial months of COVID 19: most of them in April and May and we have come a long way from there, we even have a few vaccines that have been approved now, so the observations would have changed drastically since then. Hence the bias of the studies being conducted in the initial months also exists.

2. There are no consistent values in any of the pre-conditions specified in the dataset. As a result, it is impossible to know how many people with pre-existing conditions, such as diabetes, chronic diseases, cardiovascular conditions and such others may experience symptoms. Similarly, since the data is inconsistent and includes missing values when it comes to the use of drugs like Remdesivir, hydroxychloroquine and/or chloroquine and other such drugs. It is impossible to make an accurate estimate of the drug's efficacy. Hence another bias.

3. Since most of the studies are observational rather than experimental, we depend on the observations of a particular author who may or may not possess observational biases.

- **Measurement error** – Different authors concentrate their studies in different regions with specific areas of interest. For instance, a particular author carrying a study on mortality rate in the Wuhan province of China, is likely to record the data only for the deaths caused due to COVID-19, ignoring other insignificant data for that study. This indicates that the author could overlook important parameters such as symptoms, medical preconditions or treatment administered, which are all very significant from a holistic viewpoint. Consequently, this leads to an introduction of anomalies or inconsistencies.

- **Sampling error** – Most of the studies have been carried out in and around China. Additionally, Wuhan province has been an epicenter of this pandemic. Thus, the trends or findings that we obtain after carrying out the study cannot be applied or generalized to the entire population, comprising of patients from different provinces and countries. For example, our dataset has sample studies that involve patients from countries such as France, Germany, Italy etc. Hence, we cannot establish any kind of association with this finding to people all over the world.

- **Modelling error** – When we use Multiple Linear Regression to model our data, the only downside is that our analysis will be biased due to presence of several missing and null values. The trends or observations that we notice will only consider the values that are present. As a result, our aim to successfully predict mortality rate among COVID-19 patients will be inaccurate. In order to mitigate these issues, we must ensure robust collection of data over several regions to avoid biases and have consistency. Consequently, this would help us make insightful inferences from our study. Furthermore, since mortality rate as a column in our dataset has a fixed interval range (0,1), thus beta regression as a modelling technique makes much more sense. If not this method, then we would need to perform a log-based transformation of this column before we can go ahead with fitting.

Other probable caveats in our model:

- The data is highly inconsistent which is evident. One of the major biases that is visible is that it mainly has studies carried out in and around China. Hence, we can maybe only generalize the results for the people in China (that too only a few provinces). Thus, we cannot draw any generalizations for the world.

- Since the data is mainly focused on the population in China and its few provinces, the cultural habits (mainly the eating habits, the way of living and other such factors) of people on those provinces are only considered and not the rest of the world. Hence there is that bias.