Regression Analysis on COVID, Health, and Population Metrics

Antonio R, Aditya M, Aadam L, Christopher L

For this analysis we are investigating a COVID-19 data set that was provided to us from class. We are essentially using linear regression models to find patterns that give us an insight into some relationships of the variables in the data set. These linear regression models will, on a high level, fit lines and curves of best fit to the data. After we gather the results from these models we interpret their importance, as well as their validity. Linear models come with some assumptions, so to test their validity we need to show that it is not incorrect to make these assumptions given this data set.

We can answer multiple questions through a regression analysis of the COVID-19 data. - One of the questions we want to analyze is how the percentage of the population that died from COVID-19 in a country interacts with that country's human development index (HDI). - We also want to find what is the relationship between new deaths and percentage of vaccinated population, and compare it to new deaths with respect to the percentage of fully vaccinated population.

The dataset we are using includes data on confirmed COVID cases, deaths, hospitalizations, testing, vaccinations, and other metrics across several countries since January 4, 2021. Our goal is to answer the previous questions by leveraging this data set.

An individual observation in the data (a row) corresponds to the number of new covid cases, total number of covid cases, total number of death, and other covid, health, and population related metrics for a country in a single day since January 4, 2021. Our dataset has 1107 observations.

Our responses (dependent variables) are the variables we want to find more about, in our case, these will be the percentage of population that died from COVID-19 in a country, and new COVID-19 deaths.

Our covariates (independent variables) are the variables we will analyze with our responses to get further insight into the relationships we want to analyze: To analyze the percentage of population that died from COVID-19 in a country we will use HDI as a covariate. To analyze the number of new deaths, we will use the percentage of vaccinated population and the percentage of fully vaccinated population as covariates.

Some of our covariates are not in the original dataset, we will manipulate the dataset by aggregating, multiplying, and dividing data to get our responses and covariates:

```
covid = read.csv(file='covid.csv')
location <- covid$location

date <- covid$date

approx_percentage_new_positive_tests <-
↪  ((covid$new_tests*covid$positive_rate)/covid$population)*100

approx_percentage_total_positve_tests <-
↪  ((covid$total_tests*covid$positive_rate)/covid$population)*100

tests_per_case <- covid$tests_per_case

percentage_vaccinated <- (covid$people_vaccinated/covid$population)*100

percentage_fully_vaccinated <-
↪  (covid$people_fully_vaccinated/covid$population)*100

percentage_new_vaccinated <- (covid$new_vaccinations/covid$population)*100

population_density <- covid$population_density

percentage_population_over_65 <- (covid$aged_65_older/covid$population)*100

gdp_per_capita <- covid$gdp_per_capita

percentage_cardiovascular_death <- covid$cardiovasc_death_rate/1000

diabetes_prevalence <- covid$diabetes_prevalence

hospital_beds_per_100 <- covid$hospital_beds_per_thousand/10

life_expectancy <- covid$life_expectancy

human_development_index <- covid$human_development_index

stringency_index <- covid$stringency_index

percentage_total_deaths <- (covid$total_deaths/covid$population)*100

perecentage_new_deaths <- (covid$new_deaths/covid$population)*100

temp <- cbind.data.frame(covid$location, covid$population)
```

```r
agg <- aggregate(covid$new_deaths, by=list(covid$location), mean)

for(con in unique(covid$location)) {
  temp[,2][temp[,1] == con] = (agg[,2][agg[,1] == con]/temp[,2][temp[,1] ==
 ↪  con][1])*100
}

percentage_avg_new_deaths_by_country = temp[,2]

# Here we are creating the new csv and dataframe covid_new using the
 ↪  previous variables.
covid_new = cbind.data.frame(date, location,
 ↪  approx_percentage_new_positive_tests,
 ↪  approx_percentage_total_positve_tests, tests_per_case,
 ↪  percentage_vaccinated, percentage_fully_vaccinated,
 ↪  percentage_new_vaccinated, population_density,
 ↪  percentage_population_over_65, gdp_per_capita,
 ↪  percentage_cardiovascular_death, diabetes_prevalence,
 ↪  hospital_beds_per_100, life_expectancy, human_development_index,
 ↪  stringency_index, percentage_avg_new_deaths_by_country,
 ↪  percentage_total_deaths, perecentage_new_deaths)
write.csv(covid_new, "./covid_new.csv",row.names= FALSE)
```

Model 1: Percentage of Population that Died from COVID With Respect to the Human Development Index

Response: Percentage of Average Population that died from COVID in a country, Covariate: Country's HDI

Here we aggregate the data, remember, the data set includes data for each day after the mentioned date, so we need to aggregate these over each country.

```r
covid_new = read.csv(file='covid_new.csv')
agg_pop_death_percentage = aggregate(
    x = covid_new$percentage_total_deaths,
    by = list(covid_new$location),
    FUN = mean
  )
agg_hdi = aggregate(
    x = covid_new$human_development_index,
    by = list(covid_new$location),
    FUN = mean
  )
```

Here we create the model, and fit our regression line (line of best fit).

```
pop_death_per_country <- agg_pop_death_percentage$x
hdi <- agg_hdi$x
linmod <- lm(pop_death_per_country~hdi)
b0 <- linmod$coef[1]
b1 <- linmod$coef[2]
df.y <- data.frame(
  "Mean percentage of dead population" = mean(pop_death_per_country),
  "Standard Deviation of percentage of dead population" =
  ↪  sd(pop_death_per_country),
  "Minimum percentage of dead population" = min(pop_death_per_country),
  "Maximum percentage of dead population" = max(pop_death_per_country)
)
show(df.y)
```

```
##   Mean.percentage.of.dead.population
## 1                         0.08397941
##   Standard.Deviation.of.percentage.of.dead.population
## 1                                          0.05077081
##   Minimum.percentage.of.dead.population Maximum.percentage.of.dead.population
## 1                           0.008498168                               0.1851294
```

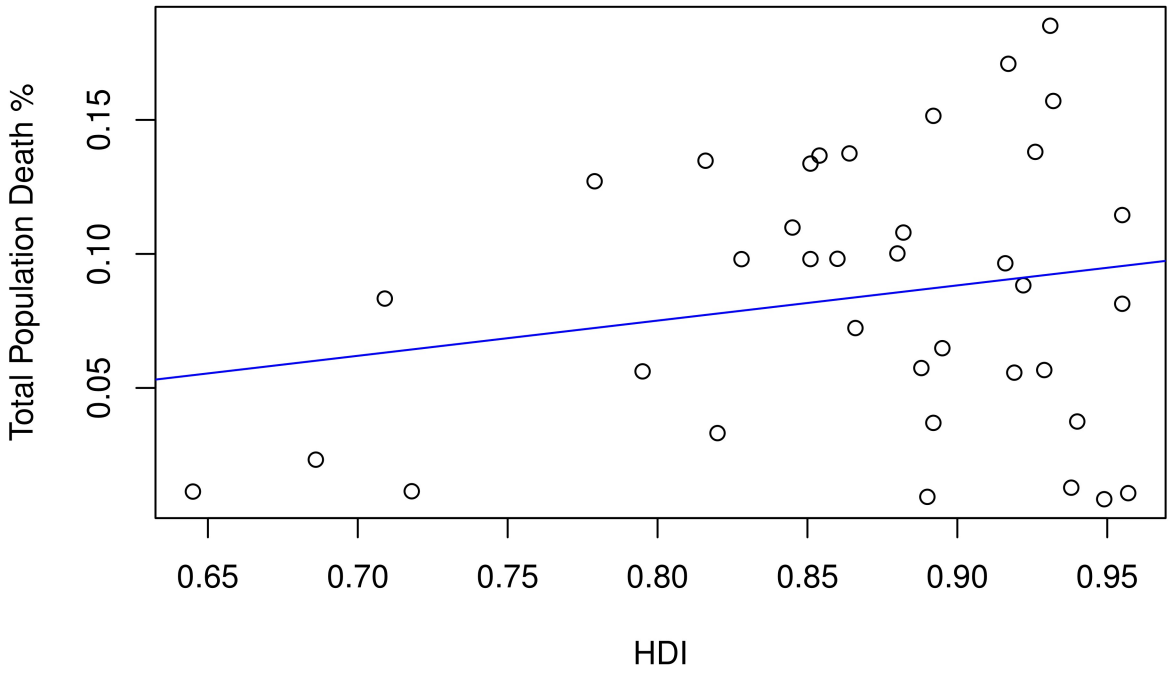
```
df.x <- data.frame(
  "b0" = b0,
  "b1" = b1,
  "Mean HDI" = mean(hdi),
  "Standard Deviation of HDI" = sd(hdi),
  "Minimum HDI" = min(hdi),
  "Maximum HDI" = max(hdi)
)
show(df.x)
```

```
##                      b0       b1  Mean.HDI Standard.Deviation.of.HDI
## (Intercept) -0.02997058 0.131377 0.8673514                0.07818149
##             Minimum.HDI Maximum.HDI
## (Intercept)       0.645       0.957
```

Here we graph our response with our covariate and the line of best fit from the model.

```
plot(hdi, pop_death_per_country, main="Figure 1: Line of best fit from Model
↪  1", xlab="HDI", ylab="Total Population Death %")
abline(b0, b1, col="blue")
```

Figure 1: Line of best fit from Model 1



Here is some more data on our model.

```
standard.error <- summary(linmod)$coef[2,2]
p.value <- summary(linmod)$coef[2,4]
df.p.and.error <- data.frame(
  "Standard Error" = standard.error,
  "P Value" = p.value
)
show(df.p.and.error)
```

```
##   Standard.Error   P.Value
## 1      0.1074983 0.2298255
```

We will also make a boxplot of the reponse and covariate to identify outliers.

```
par(mfrow=c(1,2))
boxplot(hdi, main="Figure 2: HDI", horizontal=TRUE)
boxplot(pop_death_per_country, main="Fig 3: Total Population Death %",
↪  horizontal=TRUE)
```

Figure 2: HDI

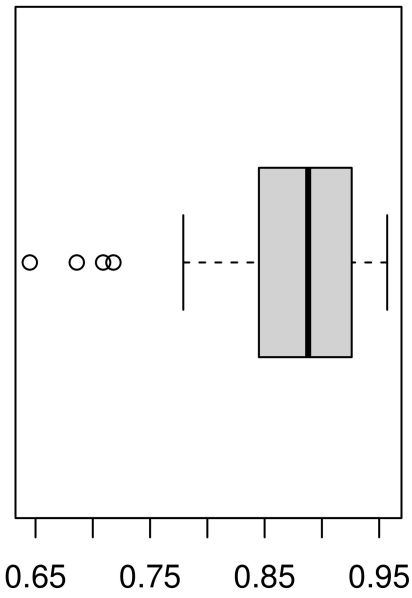
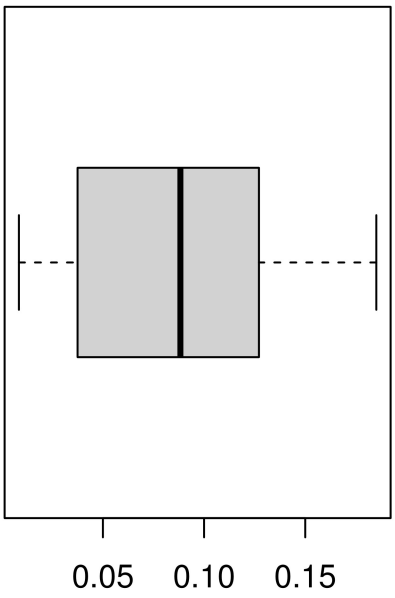


Fig 3: Total Population Death %



From these boxplots we can identify no outliers for the country total population death percentage, and only 4 outliers for the HDI, we still need to consider these in our analyisis since they give us valuable information.

Models 2 and 3: New Deaths with Respect to the Percentage of Vaccinated and Fully Vaccinated Population

Response: Average New Deaths Percentage, Covariate: Percentage of population that is vaccinated. Response: Average New Deaths Percentage, Covariate: Percentage of population that is fully vaccinated.

Similar to Model 1, we can aggregate the data by country, but in this case we can analyze both the aggregated and non-aggregated "raw" variables.

```
agg_new_deaths <- aggregate(covid_new$perecentage_new_deaths,
↪  by=list(covid_new$location), mean)$x
agg_fully_vaccinated_population <-
↪  aggregate(covid_new$percentage_fully_vaccinated,
↪  by=list(covid_new$location), mean)$x
agg_vaccinated_populations <- aggregate(covid_new$percentage_vaccinated,
↪  by=list(covid_new$location), mean)$x
```

Observation 1

Correlation between the percentage new deaths which is the number of new deaths as a percentage of the total population and percentage of population vaccinated (one case is fully vaccinated and another in vaccinated with at least one dose).

Observation 2

Correlation between the percentage new deaths which is the number of new deaths as a percentage of the total population and percentage of population vaccinated (one case is fully vaccinated and another in vaccinated with at least one dose) All aggregated by country/ location.

We can graph the boxplot of all of the variables we will use for these models to identify outliers.

Fig 4: Percentage new deaths

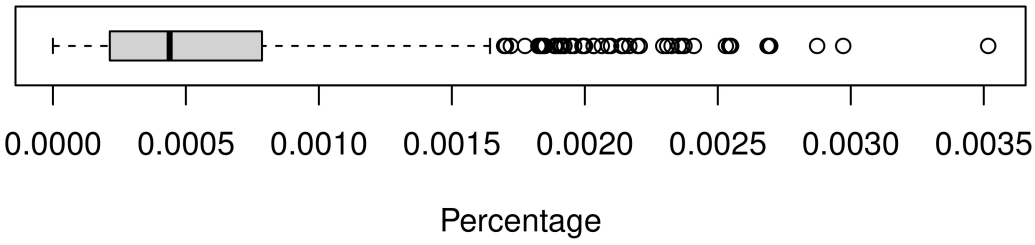


Fig 5: Percentage vaccinated population

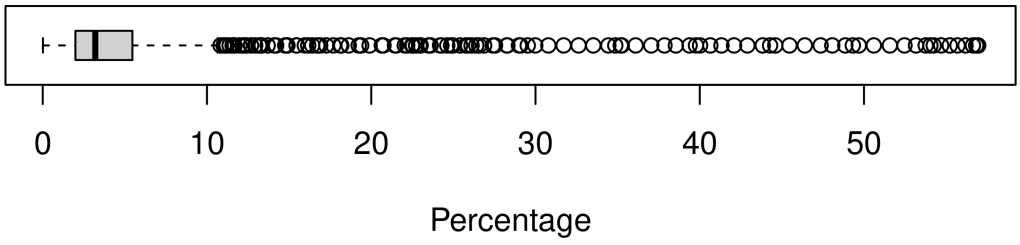


Fig 6: Percentage fully vaccinated population

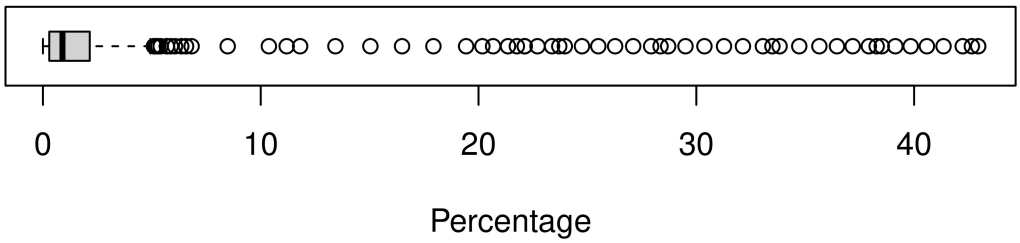


Fig 7: Aggregated Percentage new deaths due to Covid

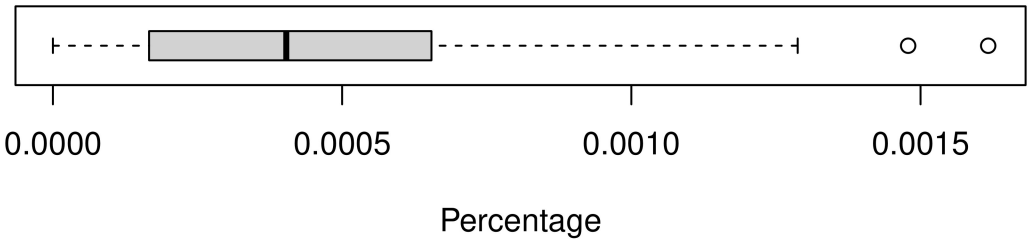


Fig 8: Aggregated percentage vaccinated population

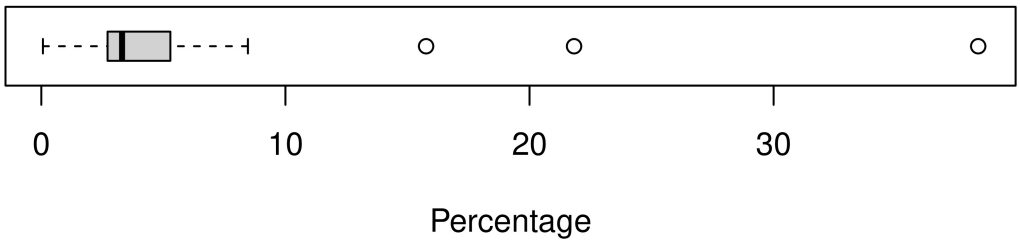
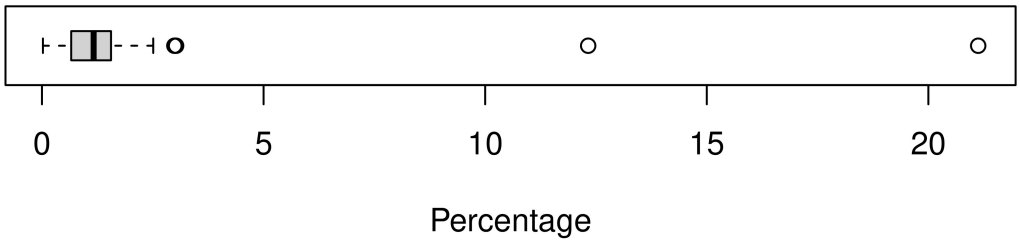


Fig 9: Aggregated percentage fully vaccinated population



We can see why aggregating the data is good approach. We won't discard the raw data, but overall aggregating per country allowed us to get cleaner results with less outliers.

There is also some further information about our variables and models we can summarize.

Summary of response and covariates mean,

```
print("Summary for observation 1")
```

```
## [1] "Summary for observation 1"
```

```
print("")
```

```
## [1] ""
```

```
paste0("Mean of New Death Percentage : ",
 ↪  mean(covid_new$perecentage_new_deaths))
```

```
## [1] "Mean of New Death Percentage : 0.000574512770199863"
```

```
paste0("Mean of Percentage of Vaccinated : ",
 ↪  mean(covid_new$percentage_vaccinated))
```

```
## [1] "Mean of Percentage of Vaccinated : 5.98819648381781"
```

```
paste0("Mean of Percentage of Fully Vaccinated : ",
↪  mean(covid_new$percentage_fully_vaccinated))
```

```
## [1] "Mean of Percentage of Fully Vaccinated : 2.33318530763579"
```

```
print("")
```

```
## [1] ""
```

```
print("Now observation 2, for aggregated data by country or location")
```

```
## [1] "Now observation 2, for aggregated data by country or location"
```

```
print("")
```

```
## [1] ""
```

```
paste0("Mean of New Death Percentage(Aggrigated):",mean(agg_new_deaths))
```

```
## [1] "Mean of New Death Percentage(Aggrigated):0.000479282033399978"
```

```
paste0("Mean of Percentage of
↪  Vaccinated(Aggrigated):",mean(agg_vaccinated_populations))
```

```
## [1] "Mean of Percentage of Vaccinated(Aggrigated):5.17469428202813"
```

```
paste0("Mean of Percentage of Fully
↪  Vaccinated(Aggrigated)",mean(agg_fully_vaccinated_population))
```

```
## [1] "Mean of Percentage of Fully Vaccinated(Aggrigated)1.9906907822009"
```

Summary of response and covariates standard deviation,

```
print("Summary for observation 1")
```

```
## [1] "Summary for observation 1"
```

```
sd(covid_new$perecentage_new_deaths)
```

```
## [1] 0.0005252965
```

```
sd(covid_new$percentage_vaccinated)
```

```
## [1] 9.301557
```

```
sd(covid_new$percentage_fully_vaccinated)
```

```
## [1] 5.817093
```

```
print("Now observation 2, for aggregated data by country or location")
```

```
## [1] "Now observation 2, for aggregated data by country or location"
```

```
sd(agg_new_deaths)
```

```
## [1] 0.0004012848
```

```
sd(agg_vaccinated_populations)
```

```
## [1] 6.981281
```

```
sd(agg_fully_vaccinated_population)
```

```
## [1] 3.802026
```

Summary of response and covariates range,

```
print("Summary for observation 1")
```

```
## [1] "Summary for observation 1"
```

```
range(covid_new$perecentage_new_deaths)
```

```
## [1] 0.000000000 0.003516713
```

```
range(covid_new$percentage_vaccinated)
```

```
## [1]  0.01100007 56.95181849
```

```
range(covid_new$percentage_fully_vaccinated)
```

```
## [1] 1.918823e-05 4.294238e+01
```

```
print("Now observation 2, for aggregated data by country or location")
```

```
## [1] "Now observation 2, for aggregated data by country or location"
```

```
range(agg_new_deaths)
```

```
## [1] 0.000000000 0.001616939
```

```
range(agg_vaccinated_populations)
```

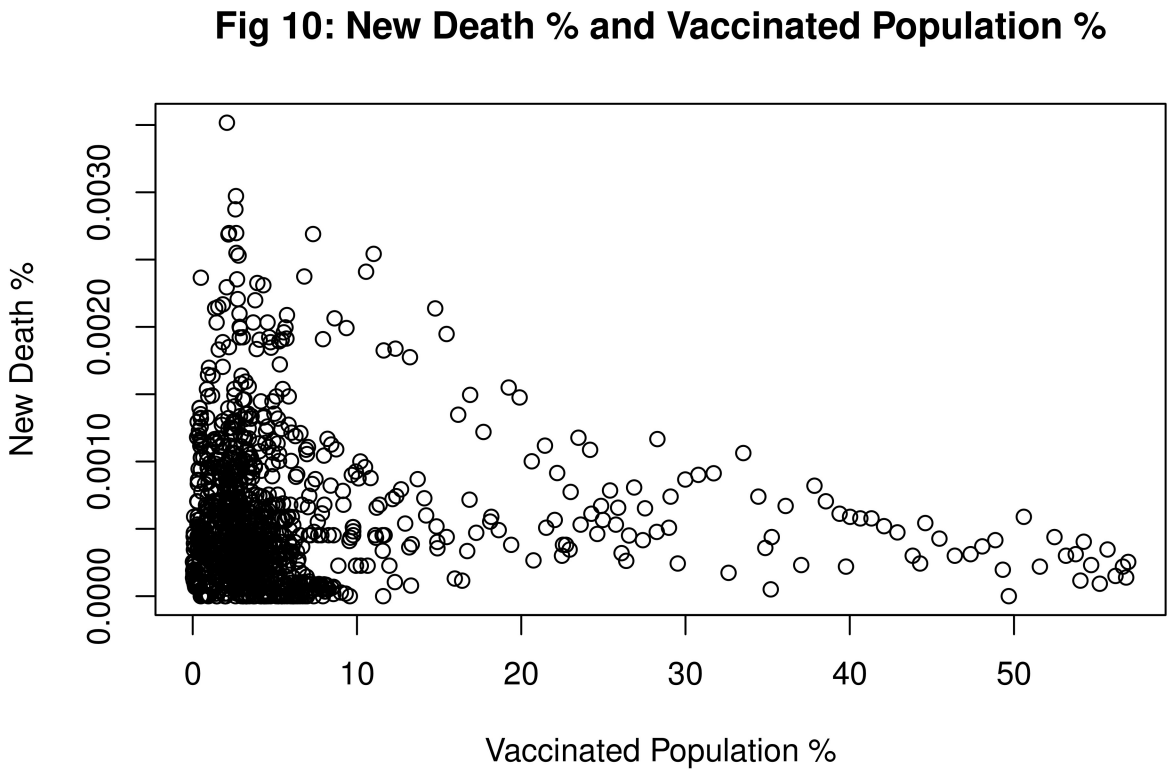
```
## [1]  0.06468428 38.38096537
```

```
range(agg_fully_vaccinated_population)
```

```
## [1]  0.0212315 21.1243446
```

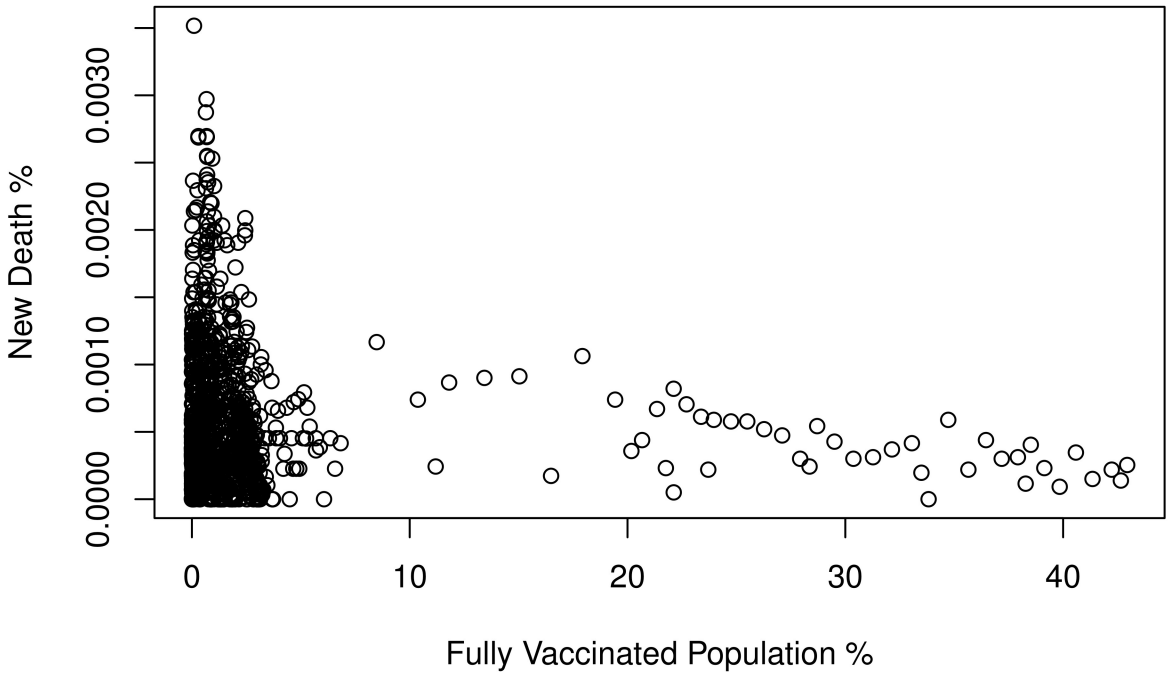
Scatter plots against each covariate,

```
plot(covid_new$percentage_vaccinated, covid_new$perecentage_new_deaths,
↪  main="Fig 10: New Death % and Vaccinated Population %", xlab="Vaccinated
↪  Population %", ylab="New Death %")
```



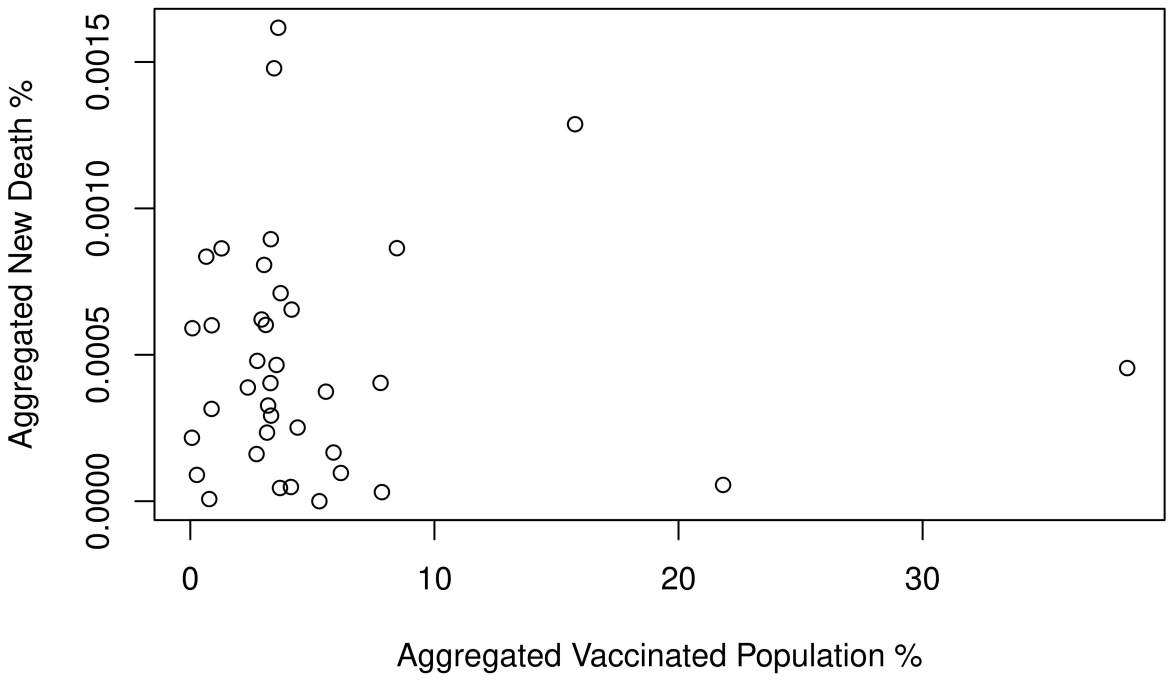

```
plot(covid_new$percentage_fully_vaccinated,
↪  covid_new$perecentage_new_deaths, main="Fig 11: New Death % and Fully
↪  Vaccinated Population %", xlab="Fully Vaccinated Population %",
↪  ylab="New Death %")
```

Fig 11: New Death % and Fully Vaccinated Population %



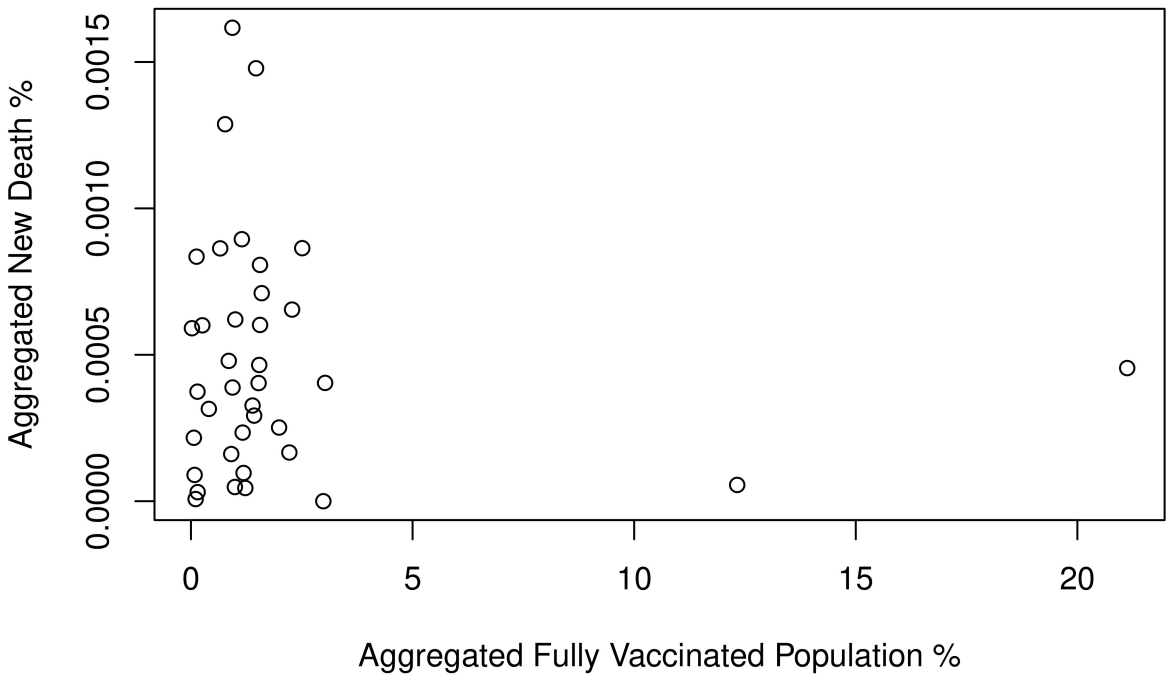

```
plot(agg_vaccinated_populations, agg_new_deaths, main="Fig 12: Aggregated
↪  New Death % and Aggregated Vaccinated Pop. %", xlab="Aggregated
↪  Vaccinated Population %", ylab="Aggregated New Death %")
```

Fig 12: Aggregated New Death % and Aggregated Vaccinated Pop. %




```
plot(agg_fully_vaccinated_population, agg_new_deaths, main="Fig 13: Agg. New
↪  Death % and Agg. Fully Vaccinated Pop. %", xlab="Aggregated Fully
↪  Vaccinated Population %", ylab="Aggregated New Death %")
```

Fig 13: Agg. New Death % and Agg. Fully Vaccinated Pop. %



Summary of estimated regression coefficients

```
print("Summary for observation 1")
```

```
## [1] "Summary for observation 1"
```

```
linmod <- lm(covid_new$perecentage_new_deaths ~
 ↪  covid_new$percentage_vaccinated + covid_new$percentage_fully_vaccinated)
print(linmod$coefficients)
```

```
##                             (Intercept)       covid_new$percentage_vaccinated
##                            5.661810e-04                            1.112396e-05
## covid_new$percentage_fully_vaccinated
##                           -2.497904e-05
```

```
print("")
```

```
## [1] ""
```

```
print("Now observation 2, for aggregated data by country or location")
```

```
## [1] "Now observation 2, for aggregated data by country or location"
```

```
linmod.2 <- lm(agg_new_deaths ~ agg_vaccinated_populations +
 ↪  agg_fully_vaccinated_population)
print(linmod.2$coefficients)
```

```
##                       (Intercept)      agg_vaccinated_populations
##                      4.448064e-04                     2.970524e-05
## agg_fully_vaccinated_population
##                     -5.989874e-05
```

In an attempt to analyze the points of interest outlined in the introduction, we created a few models relating different variables to total number of deaths per region and number of new deaths per region.

We use a total of three different linear models to analyze the relationship of the variables of interest to total and marginal number of deaths.

For each of the linear models we are assuming four things hold true for the errors from the model: Linearity of Errors, Independence of Errors, Normality of Errors, and Equality of Variance of Errors, or LINE for short.

When we say "error", we mean the difference between the observed value of an observation and the predicted value of an observation given by the model.

I will now explain these assumptions in more detail.

Linearity: we require the errors from the model to be roughly linear in shape when graphed about zero.They do not have to precisely form a line, but they should all be contained within a tight band about zero when plotted.

Independence: the errors should each be independent of other errors, meaning that having any knowledge about one observation's error should not give you any information about another observation's error.

Normality: we require that the errors in our model be normally distributed with an average value of 0.

Equal Variance: we require that the variance of the errors be consistent and evenly spread throughout the model.

The first variable which we decided to look into was the Human Development Index.

As stated before, the Human Development Index is a composite statistic measure of the life expectancy, human education, and per capita income of a country, which is used to place countries into different tiers of "development".

For reference, higher life expectancies, higher levels of education, and higher per capita incomes correlate to higher Human Development Indices.

We figured that this would be a good variable for determining and predicting the number of deaths in a region for countries with higher life expectancies tend to have better access to medicine and health care for treating diseases, countries with higher education are more

likely to raise awareness for "cleaner" practices to avoid the spread of diseases, and countries with higher per capita income are more likely to availability to better health practices.

Our initial model for these two variables plots Total Deaths (as a percentage with respect to total population) against our predictor variable, Human Development Index, aggregated by country.

The aggregation is just to remove duplicate observations for the same country.

We expect that having a higher HDI would imply that the total number of deaths is lower for a given country.

This plot will help us get an initial idea of the relationship between HDI and Total Deaths.

```
covid_new = read.csv(file='covid_new.csv')

agg_pop_death_percentage = aggregate(

    x = covid_new$percentage_total_deaths,

    by = list(covid_new$location),

    FUN = mean

  )

agg_hdi = aggregate(

    x = covid_new$human_development_index,

    by = list(covid_new$location),

    FUN = mean

  )
```

```
pop_death_per_country <- agg_pop_death_percentage$x

hdi <- agg_hdi$x

linmod <- lm(pop_death_per_country~hdi)

b0 <- linmod$coef[1]

b1 <- linmod$coef[2]

df.y <- data.frame(
```

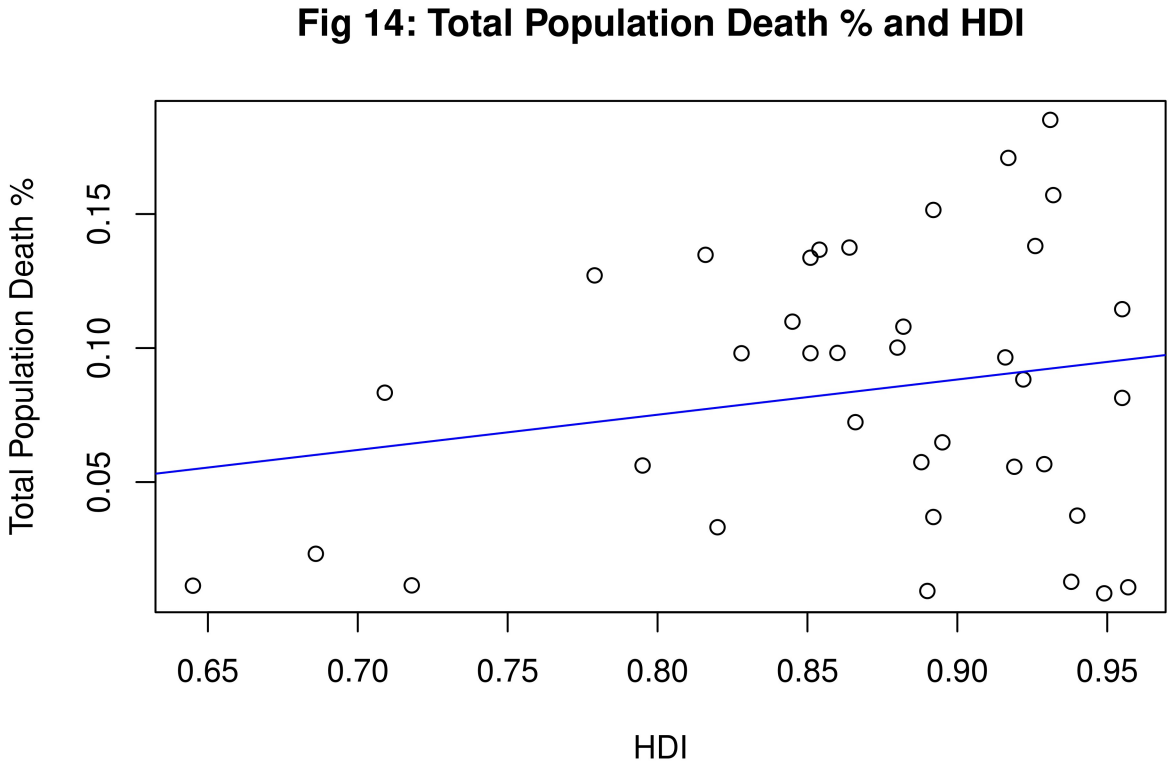
```
  "Mean percentage of dead population" = mean(pop_death_per_country),

  "Standard Deviation of percentage of dead population" =
  ↪  sd(pop_death_per_country),

  "Minimum percentage of dead population" = min(pop_death_per_country),

  "Maximum percentage of dead population" = max(pop_death_per_country)

)

show(df.y)
```

```
##   Mean.percentage.of.dead.population
## 1                         0.08397941
##   Standard.Deviation.of.percentage.of.dead.population
## 1                                          0.05077081
##   Minimum.percentage.of.dead.population Maximum.percentage.of.dead.population
## 1                           0.008498168                               0.1851294
```

```
df.x <- data.frame(

  "b0" = b0,

  "b1" = b1,

  "Mean HDI" = mean(hdi),

  "Standard Deviation of HDI" = sd(hdi),

  "Minimum HDI" = min(hdi),

  "Maximum HDI" = max(hdi)

)

show(df.x)
```

```
##                      b0       b1  Mean.HDI Standard.Deviation.of.HDI
## (Intercept) -0.02997058 0.131377 0.8673514                0.07818149
##             Minimum.HDI Maximum.HDI
## (Intercept)       0.645       0.957
```

```
plot(hdi, pop_death_per_country, main="Fig 14: Total Population Death % and
↪  HDI", xlab="HDI", ylab="Total Population Death %")

abline(b0, b1, col="blue")
```



We can see from the plot there appears to be a slight correlation between HDI and Total Deaths, except the correlation is in the opposite direction from what we predicted.

This may just a be an unfortunate consequence of the aggregation, but in either case it would be best to perform a test to determine the likelihood that the relationship shown in the graph is not an instance of chance. For this, we can perform a hypothesis test.

If there does not exist a relationship between HDI and Total Deaths, then we would expect the slope of the line to be equal to 0, whereas if there is a relationship between the two variables, then we would expect the slope to be anything but 0.

Therefore we pose the following hypothesis test on the slope of the model: $H_0: \beta_1 = 0$ and $H_a: \beta1 \neq 0$.

```
standard.error <- summary(linmod)$coef[2,2]

p.value <- summary(linmod)$coef[2,4]
```

```
df.p.and.error <- data.frame(

  "Standard Error" = standard.error,

  "P Value" = p.value

)

show(p.value)
```

```
## [1] 0.2298255
```

With the p-value being 0.230, we can say that the smallest α for which we can conclude the alternative hypothesis is 0.230.

This simply means that we can say with a confidence of 77% that the percent of the total population that died from COVID has a linear relationship with the Human Development index, despite it not being linear in the direction we hoped.

We made an attempt to adjust this linear model to help it better fit the our error assumptions for a linear model. Instead of plotting total deaths against HDI, we found that we get a significantly better model if we plot total deaths against $\text{HDI}^2+\text{ HDI}$.

When we perform this transformation on the model, it helps stretch out the data on the x-axis by a factor of their magnitude.

Therefore when the values on the x-axis are large, they get spread out more.

We decided that such a transformation might be better because the data points in the right side of the graph are more farther spread from the model than the points on the left hand side of the graph.

This means that in the transformed data, the model can better account for the trend of the data.

The second variable we chose to look at was percentage of population vaccinated/fully vaccinated, the difference being whether or not a person has had a complete vaccination vs only had a partial vaccination.

The logic behind this decision is pretty self explanatory: if a population has a high vaccination rate, then they are more likely to have a lower death rate.

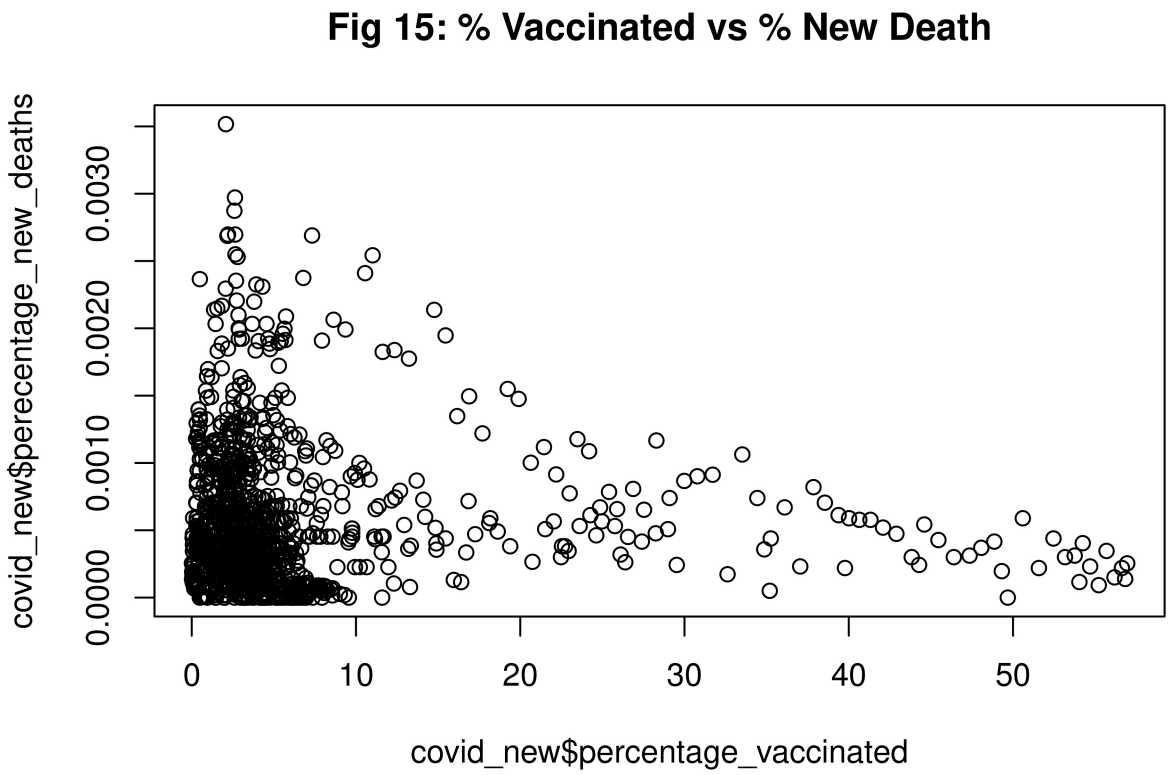
```
agg_new_deaths <- aggregate(covid_new$perecentage_new_deaths,
↪  by=list(covid_new$location), mean)$x

agg_fully_vaccinated_population <-
↪  aggregate(covid_new$percentage_fully_vaccinated,
↪  by=list(covid_new$location), mean)$x
```

```
agg_vaccinated_populations <- aggregate(covid_new$percentage_vaccinated,
↪  by=list(covid_new$location), mean)$x
```

Our initial models for these plots came out very ugly.

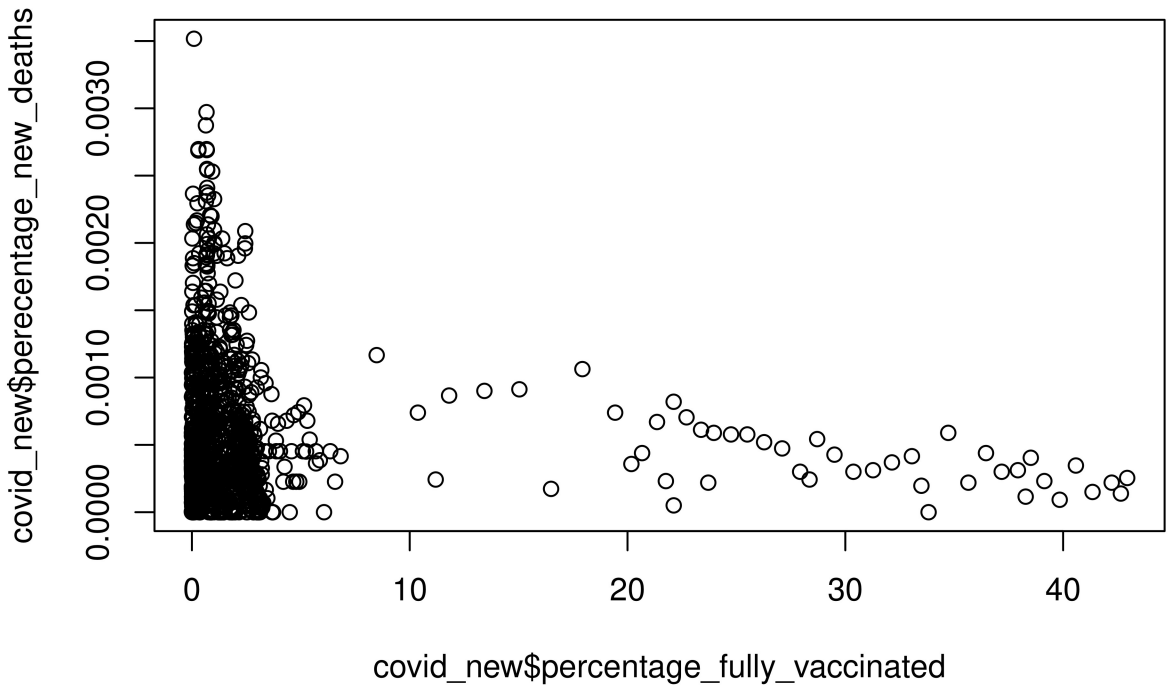
We will provide a copy of the plots below.

```
plot(covid_new$percentage_vaccinated, covid_new$perecentage_new_deaths,
↪  main="Fig 15: % Vaccinated vs % New Death")
```




```
plot(covid_new$percentage_fully_vaccinated,
↪  covid_new$perecentage_new_deaths, main="Fig 16: % Fully Vaccinated vs %
↪  New Death")
```

Fig 16: % Fully Vaccinated vs % New Death



As you can see in these plots, there are an extremely high number of data points concentrated in the range 0 to 10 on the x-axis.

This is due to the fact that we are not aggregating the data, so every single entry in our entire data set is included in this plot, including many early observations when the number of new covid deaths was very low.

Aside from this though, we do see a slightly negative linear trend of data points extending out from the mass on the left.

We can interpret this as a good thing, for this likely means that for all the observations in which the number of new deaths was significant, having a higher vaccinated population helps decrease the number of new deaths in the population.

Overall though, we cannot perform much useful analysis on the models since the concentrated data on the left would dilute the accuracy of the results.

One attempt we made to fix this was to combine the two models while transforming the predictor variables.

Instead of plotting new deaths against percentage of vaccinated/fully vaccinated population individually, we plotted against percentage of population vaccinated squared combined with the square root of percentage of population fully vaccinated.

When looking at the two models we see that the plot of new deaths against percentage

vaccinated has a sharper cone of data from left to right than that of the plot against fully vaccinated population.

By taking the square of the vaccinated data we can stretch out these data points thus giving us a plot where the cone of data from left to right is less steep.

Similarly when we take the square root of the fully vaccinated population, this makes the less steep cone of data points stretching from left to right more steep.

Now that we have performed data transformations on each of our two predictor variables, if we know splice their effects together into one model, the resulting model will take into account both of the transformations which make the two individual graphs look more similar into a combined graph which can have a better model overall.

The last variable which we chose to look at is actually just a revision of the second.

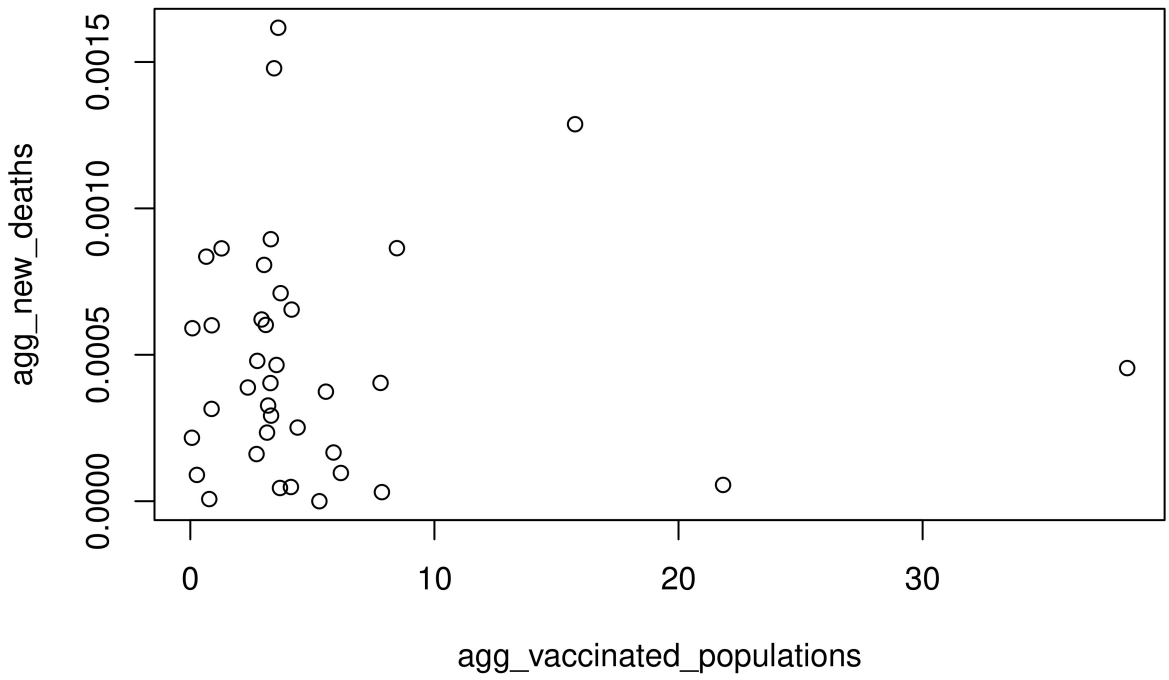
Since the plots in our second model are heavily concentrated on the left, we thought that if we were to aggregate the data by country, then we could help eliminate the congestion of data points which made it difficult to perform analysis.

The models for these plots unfortunately aren't a whole lot better, but they do allow us to perform some useful analysis.

Here are the plots containing the models for percentage vaccinated/fully vaccinated.

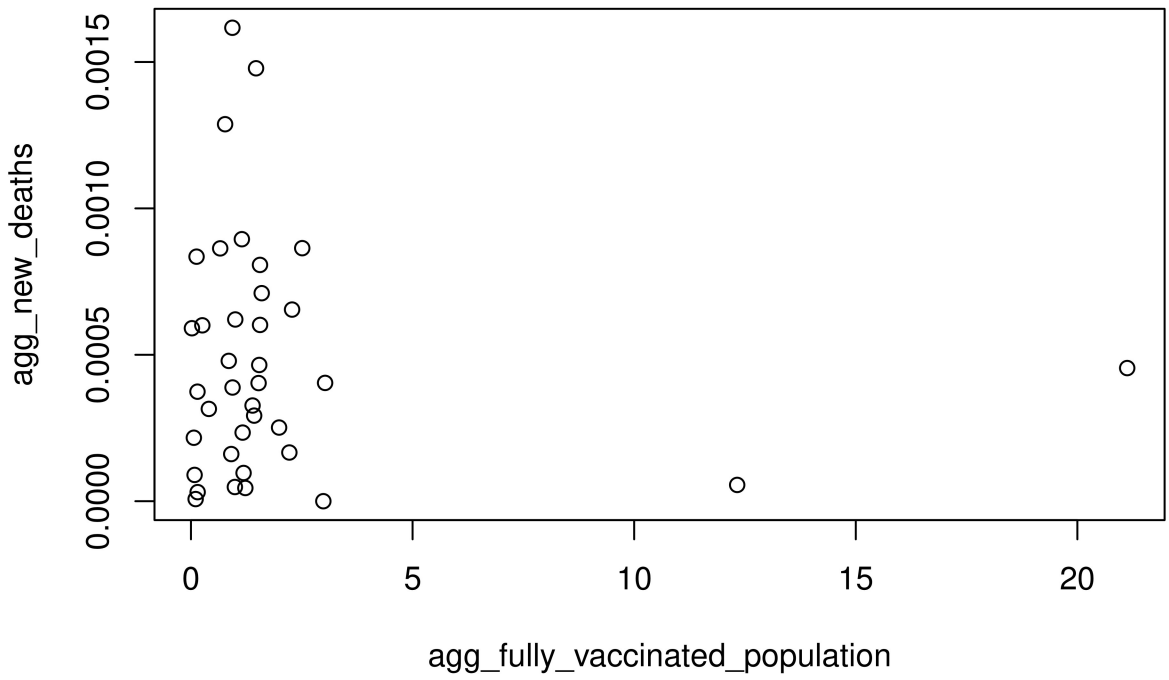
```
plot(agg_vaccinated_populations, agg_new_deaths, main="Fig 17: % Vaccinated
↪  vs % New Death (Aggrigated on Location)")
```

Fig 17: % Vaccinated vs % New Death (Aggrigated on Location)




```
plot(agg_fully_vaccinated_population, agg_new_deaths, main="Fig 18: % Fully
↪  Vaccinated vs % New Death (Aggrigated on Location)")
```

Fig 18: % Fully Vaccinated vs % New Death (Aggrigated on Locatior



We can see from the plots that the congestion on the left has significantly decreased and that among the few data points not on the left side of the plots, they correlate to low numbers of new deaths, indicating that among the countries which have higher percentages of vaccinated populations, their respective aggregated new death rate is low.

Although the sample size is low for this conjecture, it is still a promising result.

To adjust this model, we followed our decision made in the previous model combined the two plots together, but instead of transforming our predictor variables, we instead transformed our response variable.

The transformed data now plots aggregated new deaths squared against percentage of population vaccinated and percentage of population fully vaccinated.

Recall how the reasoning behind modifying the input data in the previous models was to stretch out the data on the x-axis.

Well we can see clearly in these plots that it would be much better to modify the data along the y-axis instead since most of the data points are found along the left side of the graph.

By squaring the number of aggregated new deaths, we stretch out these values in our previous plots along the y-axis.

In doing so, what appears is a slight but strong negative correlation between aggregated new deaths and aggregated (fully) vaccinated population.

This is of course the result which we had hoped would find to hold true in the data.

The Final Model for question 1

Final model estimated regression coefficients

Final model estimated regression coefficients are -1.729823, 4.326984, and -2.557500.

Final model standard errors

Final model standard errors are plotted below, with an average error of 5.500678e-19, standard error of 1.808786, and p-value of 2.242029e-02 along with 1.723184e-01 R squared error.

```
pop_death_per_country <- agg_pop_death_percentage$x
hdi <- agg_hdi$x

linmod <- lm(pop_death_per_country~poly(hdi, 2, raw=TRUE))
b0 <- linmod$coef[1]
b1 <- linmod$coef[2]
b2 <- linmod$coef[3]

print(c(b0, b1, b2))
```

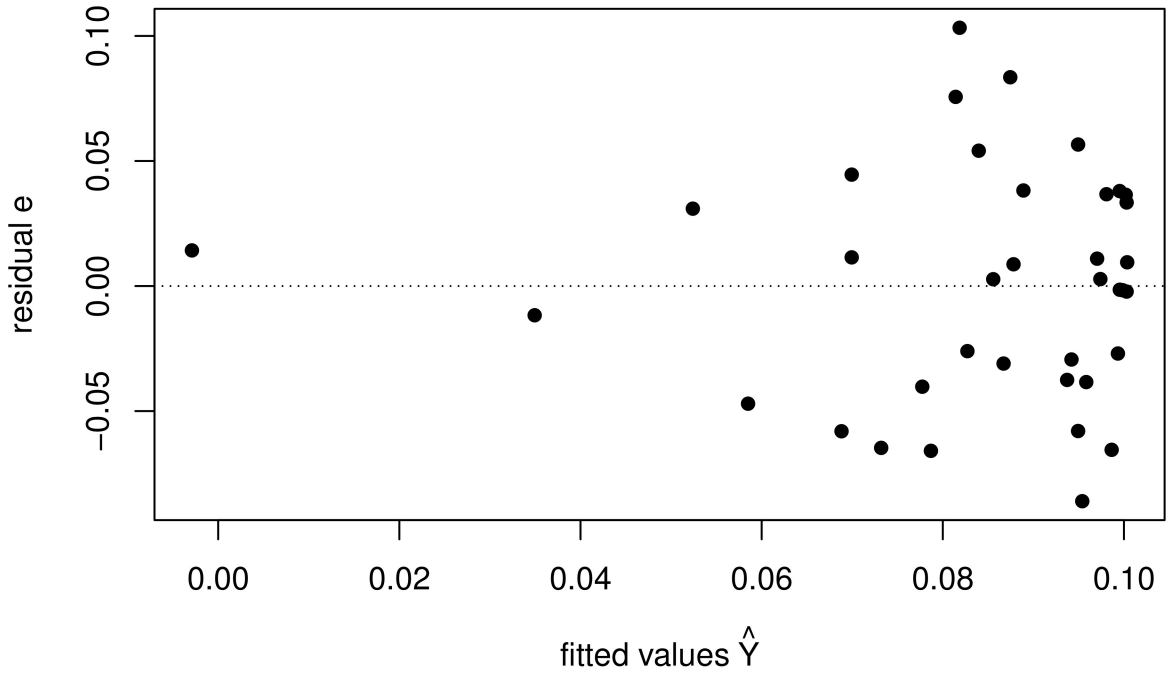
```
##               (Intercept) poly(hdi, 2, raw = TRUE)1 poly(hdi, 2, raw = TRUE)2
##                 -1.729823                 4.326984                -2.557500
```

```
Yhat <- linmod$fitted.values
e <- resid(linmod)

standard.error <- summary(linmod)$coef[2,2]
p.value <- summary(linmod)$coef[2,4]
print(c(mean(e), standard.error, p.value, summary(linmod)$r.squared))
```

```
## [1] 5.500678e-19 1.808786e+00 2.242029e-02 1.723184e-01
```

```
plot(Yhat,e, xlab = expression('fitted values' ~ hat(Y)),ylab="residual e",
↪  pch = 16, main="Fig 19: Fitted vs Residual")
abline(h = 0, lty = 3)
```

Fig 19: Fitted vs Residual



Final model test results

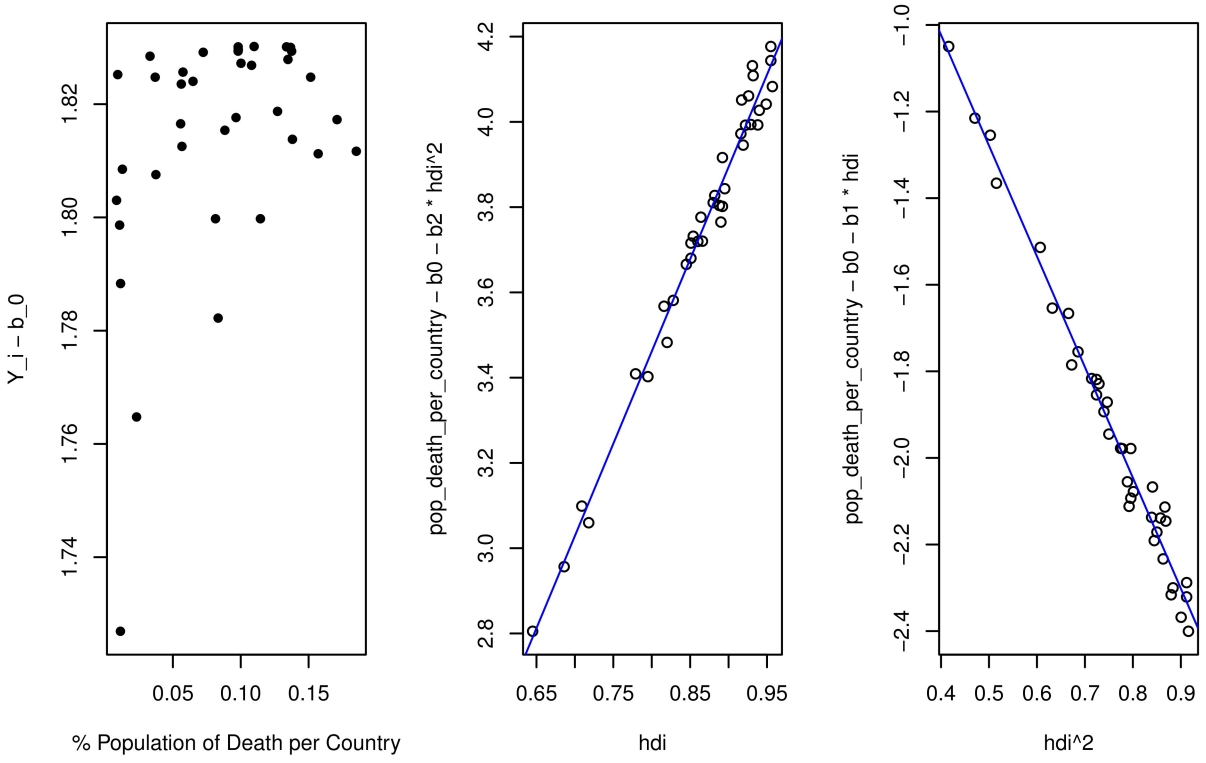
```
Y.hat <- linmod$fitted.values # Obtain fitted values
# Compute sums of squares
Y.bar <- mean(pop_death_per_country)
SSR <- sum((Y.hat - Y.bar)^2)
SSE <- sum((pop_death_per_country - Y.hat)^2)

n <- length(hdi)
p <- 3

# Compute mean squares
MSR <- SSR/(p-1)
MSE <- SSE/(n - p)

f.stat <- MSR/MSE
alpha <- 0.05
fquantile <- qf(1 - alpha, p-1, n - p)
print(c(f.stat, p-1, n-p, fquantile))
```

```
## [1]  3.539300  2.000000 34.000000  3.275898
```

Clearly the Fstat is 3.539300 which is more than the 0.95 quantile, 3.275898 we conclude H_a: $\beta_k \neq 0$ for some k > 0 and k $<$ p, i.e. there is a linear association (regression relation). It also implies that One or more of β_1, β_2 is non zero.

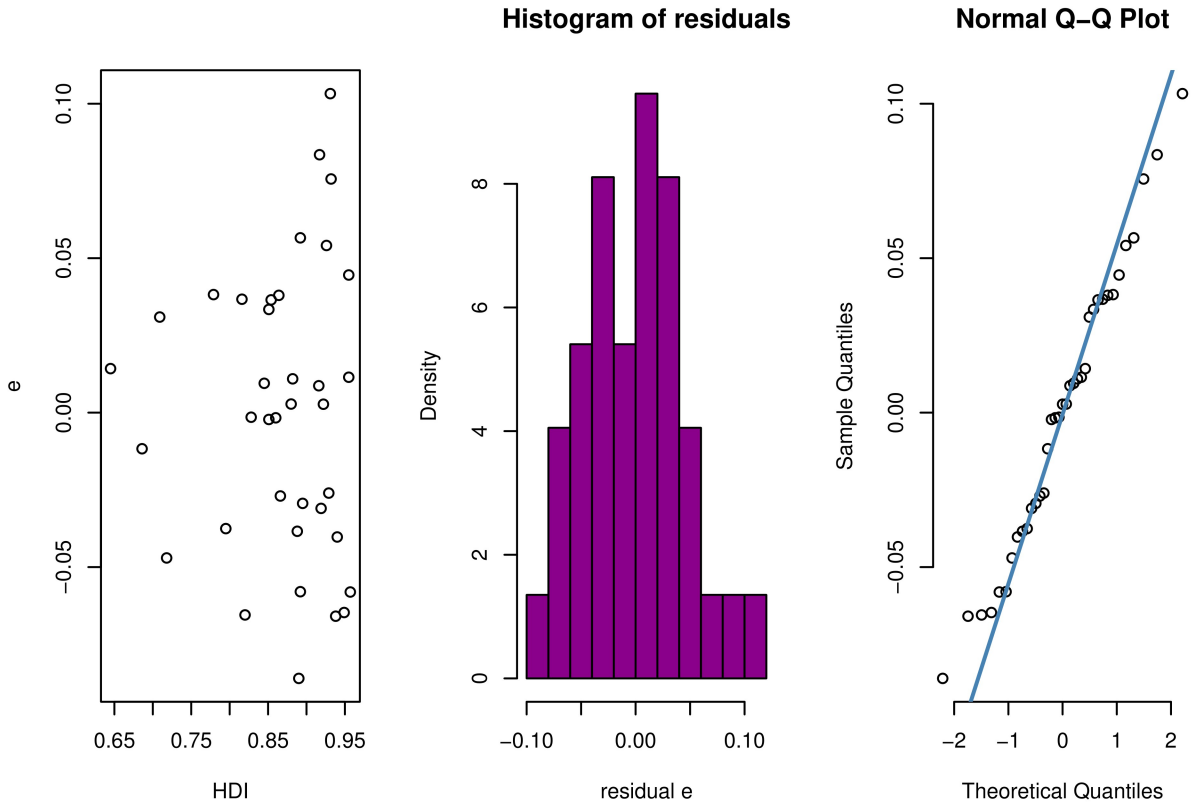
Detailed exploration of final model residuals

```
Y1 <- linmod$fitted.values - b0
e <- resid(linmod)
par(mfrow = c(1, 3))
plot(pop_death_per_country, Y1, xlab ="% Population of Death per Country"
 ↪  ,ylab=expression('Y_i - b_0'), pch = 16)
abline(h = 0, lty = 3)
plot(hdi, pop_death_per_country - b0 - b2*hdi^2)
abline(a = 0, b = b1, col = "blue")
plot(hdi^2, pop_death_per_country - b0 - b1*hdi)
abline(a = 0, b = b2, col = "blue")
```




```
e <- resid(linmod)
par(mfrow=c(1,3))
plot(hdi, e, xlab="HDI")
hist(e, main="Histogram of residuals", xlab="residual e", col="darkmagenta",
 ↪  freq=FALSE)
```

```
qqnorm(e, pch = 1, frame = FALSE)
qqline(e, col = "steelblue", lwd = 2)
```



Independence

For this regression model, the HDIs are clearly independent from each other since we have a single data point for each country in the dataset. Therefore we are not violating the independence assumption of regression models.

Equal variance

We are also assuming equal variance in our regression model. From the residuals against the fitted values plot, we can see that they make almost a uniform spread of residual at the later values of Y.hat. Clearly, there is a consistent vertical spread almost throughout the graph. There is a minor/ very small violation of equal variance due to this inconsistent spread in the beginning of Y.hat (One outlier).

Linearlity

There are no specific regions in the graph with a majority of positive or negative residuals, therefore the model doesn't exclusively underpredict or overpredict in a region, indicating

that the data tends to be linear. The same can be said from the $Y_i - b_0$ against % Population of Death per Country and $Y_i - b_0 - b_1X$ and $Y_i - b_0 - b_2X^2$ against X^2 and X

Normality

The QQ-Plot allows us to visualize the normality of the errors from the data. There is a close to none violation of normality in the QQ-Plot since the data is relatively evenly spread out accross the line, with no strafing tails.

Need for interactions assessed

None needed cause uni variable.

The Final Model for question 2

Final model estimated regression coefficients

Final model estimated regression coefficients are -6.785320e-04, 1.274800e-07, and -1.058172e-04.

Final model standard errors

Final model standard errors are plotted below, with an average error of 3.778659e-19, standard error of 6.477939e-08, and p-value of 4.932836e-02 along with 1.762273e-02 R squared error.

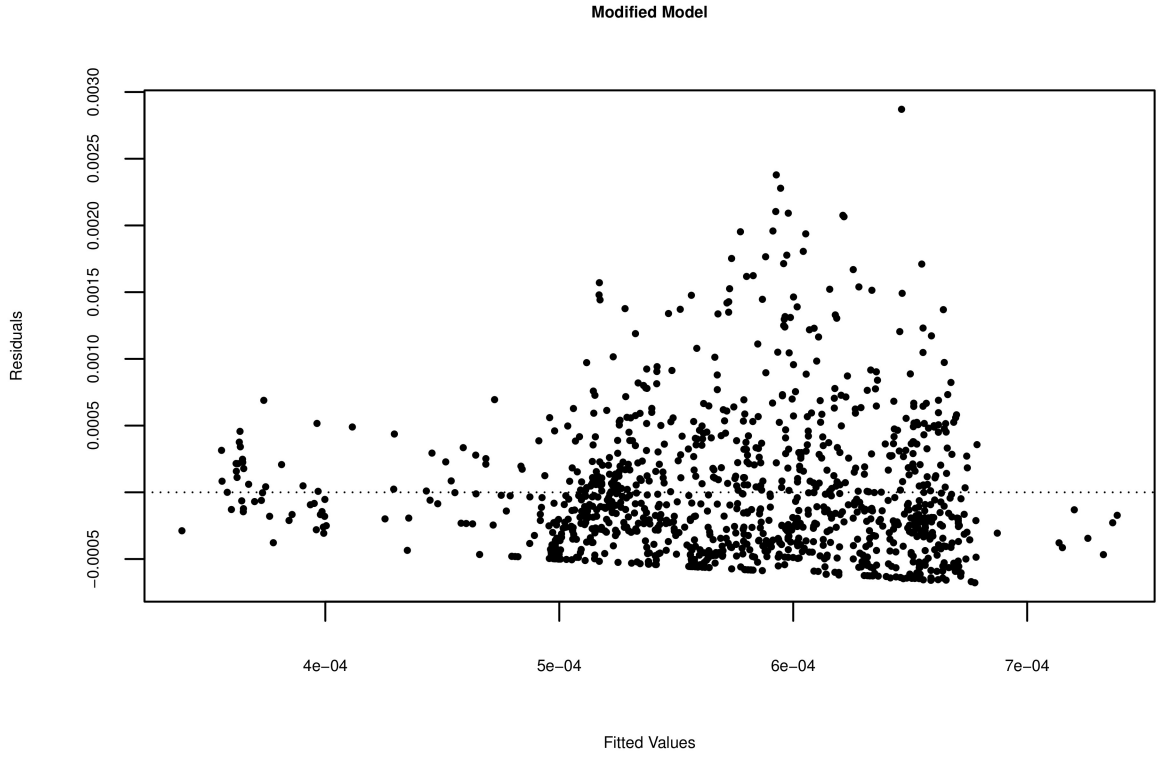
```
linmod2.modify <- lm(covid_new$perecentage_new_deaths ~
↪  I(covid_new$percentage_vaccinated ^ 2) +
↪  I(covid_new$percentage_fully_vaccinated ^ 0.5))
b0.modify <- linmod2.modify$coef[1]
b1.modify <- linmod2.modify$coef[2]
b2.modify <- linmod2.modify$coef[3]

print(c(b0.modify, b1.modify, b2.modify))
```

```
##                                   (Intercept)
##                                  6.785320e-04
##         I(covid_new$percentage_vaccinated^2)
##                                  1.274800e-07
## I(covid_new$percentage_fully_vaccinated^0.5)
##                                 -1.058172e-04
```

```
e.square <- linmod2.modify$residuals
Y.hat.square <- linmod2.modify$fitted.values

standard.error <- summary(linmod2.modify)$coef[2,2]
p.value <- summary(linmod2.modify)$coef[2,4]
print(c(mean(e.square), standard.error, p.value,
↪  summary(linmod2.modify)$r.squared))
```

```
## [1] 3.778659e-19 6.477939e-08 4.932836e-02 1.762273e-02
```

```
plot(Y.hat.square, e.square, pch = 16, xlab = "Fitted Values", ylab =
↪  "Residuals",
main = "Modified Model", cex=0.5, cex.lab=0.5, cex.axis=0.5, cex.main=0.5)
abline(h = 0, lty = 3)
```



Final model test results

```
Y.hat <- linmod2.modify$fitted.values # Obtain fitted values
# Compute sums of squares
Y.bar <- mean(covid_new$perecentage_new_deaths)
SSR <- sum((Y.hat - Y.bar)^2)
SSE <- sum((covid_new$perecentage_new_deaths - Y.hat)^2)

n <- length(covid_new$perecentage_new_deaths)
p <- 3

# Compute mean squares
MSR <- SSR/(p-1)
```

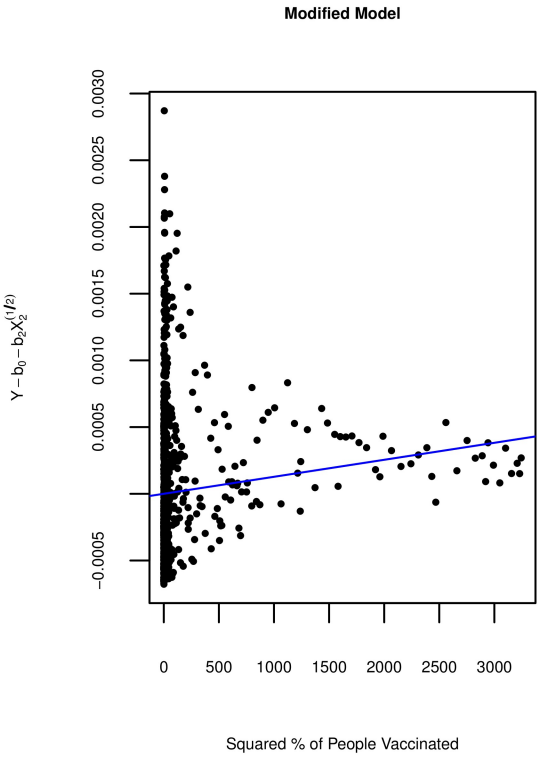
```r
MSE <- SSE/(n - p)

f.stat <- MSR/MSE
alpha <- 0.05
fquantile <- qf(1 - alpha, p-1, n - p)
print(c(f.stat, p-1, n-p, fquantile))
```

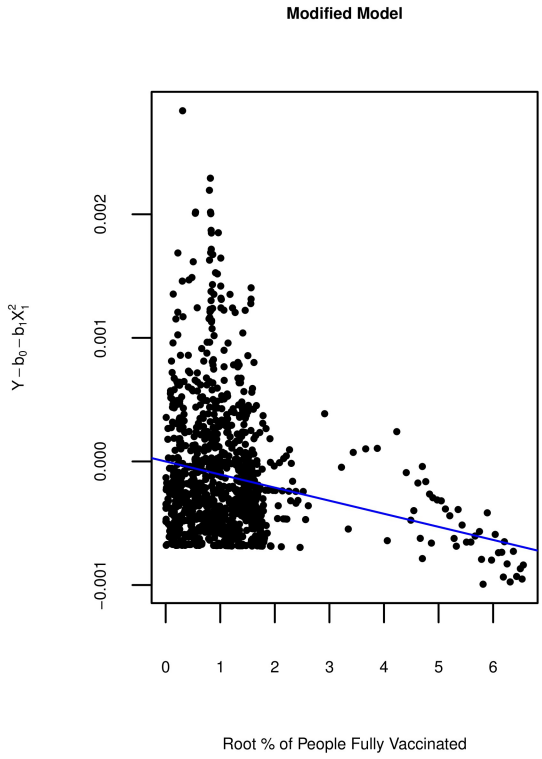
```
## [1]    9.902252    2.000000 1104.000000    3.003876
```

Clearly the Fstat is 9.902252 which is more than the 0.95 quantile, 3.003876 we conclude H_a: $\beta_k \neq 0$ for some k $>$ 0 and k $<$ p, i.e. there is a linear association (regression relation). It also implies that One or more of β_1, β_2 is non zero.

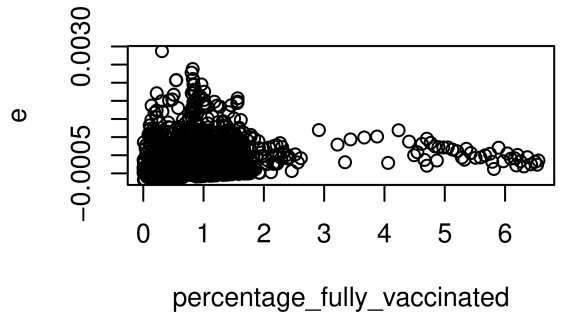
Detailed exploration of final model residuals

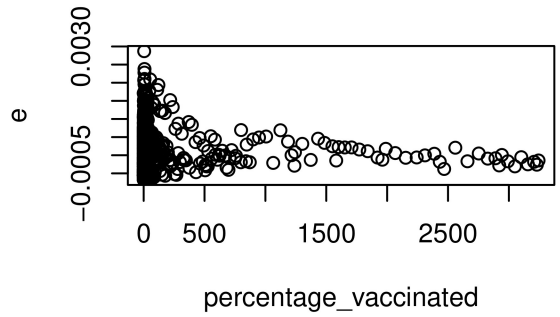
```r
Y <- covid_new$perecentage_new_deaths
par(mfrow = c(1, 2))
plot(covid_new$percentage_vaccinated^2, Y - b0.modify -
 ↪  b2.modify*(covid_new$percentage_fully_vaccinated^(1/2)), pch = 16, xlab
 ↪  = "Squared % of People Vaccinated",
ylab = expression(Y-b[0]-b[2]*X[2]^(1/2)), main = "Modified Model",
cex=0.5, cex.lab=0.5, cex.axis=0.5, cex.main=0.5)
abline(a = 0, b = b1.modify, col = "blue")
plot(covid_new$percentage_fully_vaccinated^(1/2), Y - b0.modify -
 ↪  b1.modify*covid_new$percentage_vaccinated^2, pch = 16, xlab = "Root % of
 ↪  People Fully Vaccinated",
ylab = expression(Y-b[0]-b[1]*X[1]^2), main = "Modified Model",
cex=0.5, cex.lab=0.5, cex.axis=0.5, cex.main=0.5)
abline(a = 0, b = b2.modify, col = "blue")
```

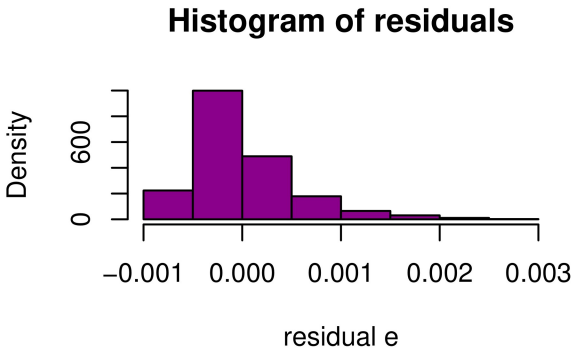


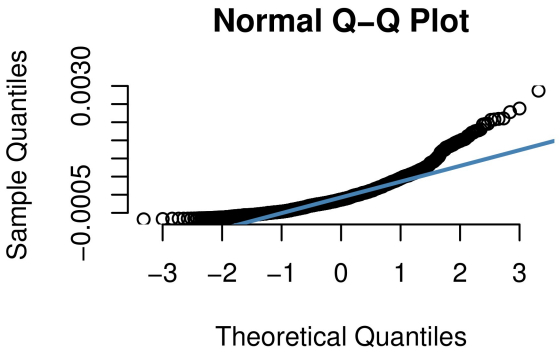



```
e <- resid(linmod2.modify)
par(mfrow=c(2,2))
plot(covid_new$percentage_fully_vaccinated^(1/2), e,
 ↪  xlab="percentage_fully_vaccinated")
plot(covid_new$percentage_vaccinated^2, e, xlab="percentage_vaccinated")
hist(e, main="Histogram of residuals", xlab="residual e", col="darkmagenta",
 ↪  freq=FALSE)
qqnorm(e, pch = 1, frame = FALSE)
qqline(e, col = "steelblue", lwd = 2)
```









Independence

For this regression model, the % of People fully Vaccinated and % of People Vaccinated is independent to % of new deaths. This is because the % of vaccinated is not grouped/ aggregated by country thus there are multiple points for a single country at different day's which are making the dependence to each other very plausible. Thus the relation between the residual error is can be dependent. Therefore we might be violating the independence assumption of regression models.

Equal variance

We are also assuming equal variance in our regression model. From the residuals against the fitted values plot, we can see that they make a cone-shape, clearly, there isn't a consistent vertical spread throughout the graph. There is a huge violation of equal variance due to this inconsistent spread.

Linearlity

There are specific regions in the graph with a majority of positive or negative residuals, therefore the model does under-predict or over-predict in a region, indicating that the data tends to be non-linear. The same can be said from the $Y_i - b_0$ against % Population of Death

In the case of $Y_i - b_0 - b_1$ * % of People Vaccinated there is significant over predication when the % of People fully Vaccinated increases and in the case of $Y_i - b_0 - b_2$ * % of People Fully Vaccinated there is significant under predication when the % of People fully Vaccinated increase

Normality

The QQ-Plot allows us to visualize the normality of the errors from the data. There is a significant tail violation of normality in the QQ-Plot since the data is relatively evenly spread out across the line, when more points should be centered in the middle of the line. The tails also strafe from the line.

Need for interactions assessed

Even on adding the interaction term, the p-value didn't suggest for keeping that variable.

```
n <- length(covid_new$perecentage_new_deaths)
linmod2_expanded <- lm(covid_new$perecentage_new_deaths ~
↪  I(covid_new$percentage_vaccinated ^ 2) +
↪  I(covid_new$percentage_fully_vaccinated ^ 0.5) +
↪  (covid_new$percentage_vaccinated *
↪  covid_new$percentage_fully_vaccinated))
Y.hat.full <- linmod2.modify$fitted.values
Y.hat.expanded <- linmod2_expanded$fitted.values
SSE.full <- sum((covid_new$perecentage_new_deaths - Y.hat.full)^2)
SSE.expanded <- sum((covid_new$perecentage_new_deaths - Y.hat.expanded)^2)
df.full <- n - 3
df.expanded <- n - 4
f.stat <- (SSE.full - SSE.expanded)/(df.full -
↪  df.expanded)/(SSE.expanded/df.expanded)
alpha <- 0.05
fquantile <- qf(1 - alpha, df.full - df.expanded, df.expanded)
pvalue <- 1 - pf(f.stat, df.full - df.expanded, df.expanded)
```

Clearly the Fstat is 1.335048e+00 which is less than the 0.95 quantile, 3.849903e+00. we conclude H_0: $\beta_4 = 0$ i.e. $X_1 * X_2$ can be dropped from the regression model

The Final Model for question 3

Final model estimated regression coefficients

Final model estimated regression coefficients are 3.128825e-07, 5.763682e-08, and -1.128989e-07.

Final model standard errors

Final model standard errors are plotted below, with an average error of -2.080667e-23, standard error of 3.673124e-08, and p-value of 1.258740e-01 along with 7.622458e-02 R squared error.

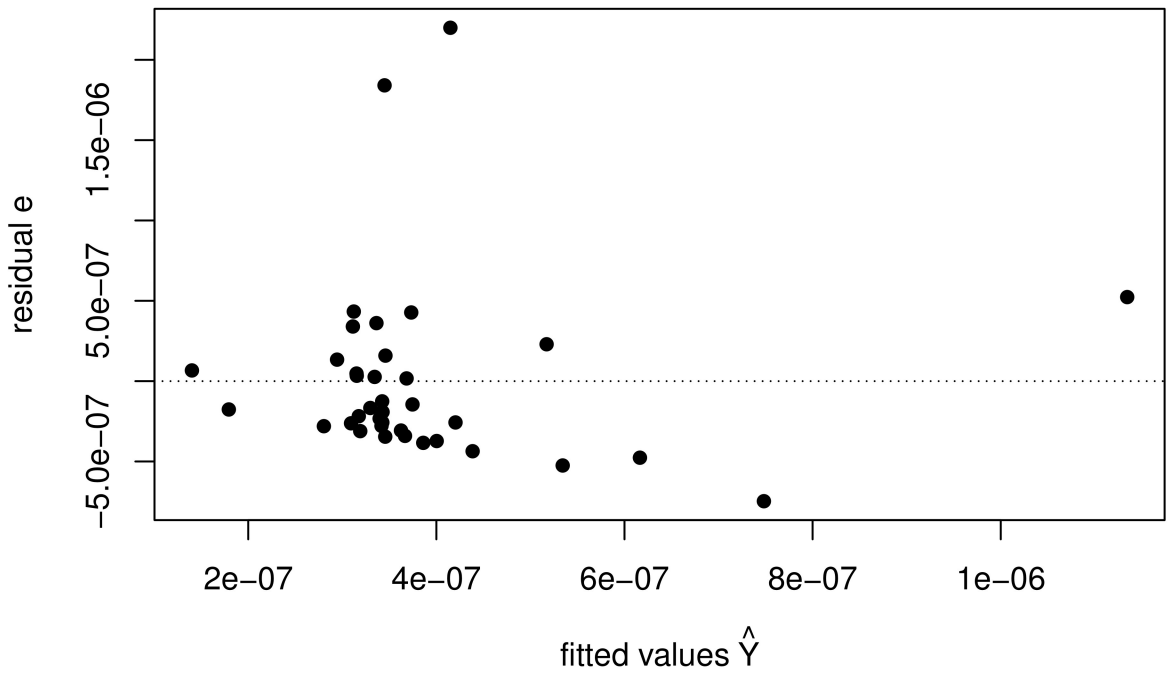
```
agg_new_deaths <- aggregate(covid_new$perecentage_new_deaths,
↪  by=list(covid_new$location), mean)$x
agg_fully_vaccinated_population <-
↪  aggregate(covid_new$percentage_fully_vaccinated,
↪  by=list(covid_new$location), mean)$x
agg_vaccinated_populations <- aggregate(covid_new$percentage_vaccinated,
↪  by=list(covid_new$location), mean)$x

linmod.2 <- lm(agg_new_deaths^2 ~ agg_vaccinated_populations +
↪  agg_fully_vaccinated_population)
b0 <- linmod.2$coef[1]
b1 <- linmod.2$coef[2]
b2 <- linmod.2$coef[3]

print(c(b0, b1, b2))
```

```
##                      (Intercept)      agg_vaccinated_populations
##                     3.128825e-07                     5.763682e-08
## agg_fully_vaccinated_population
##                    -1.128989e-07
```

```
Yhat <- linmod.2$fitted.values
e <- resid(linmod.2)

standard.error <- summary(linmod.2)$coef[2,2]
p.value <- summary(linmod.2)$coef[2,4]
print(c(mean(e), standard.error, p.value, summary(linmod.2)$r.squared))
```

```
## [1] -2.080667e-23  3.673124e-08  1.258740e-01  7.622458e-02
```

```
plot(Yhat,e, xlab = expression('fitted values' ~ hat(Y)),ylab="residual e",
↪  pch = 16)
abline(h = 0, lty = 3)
```



Final model test results

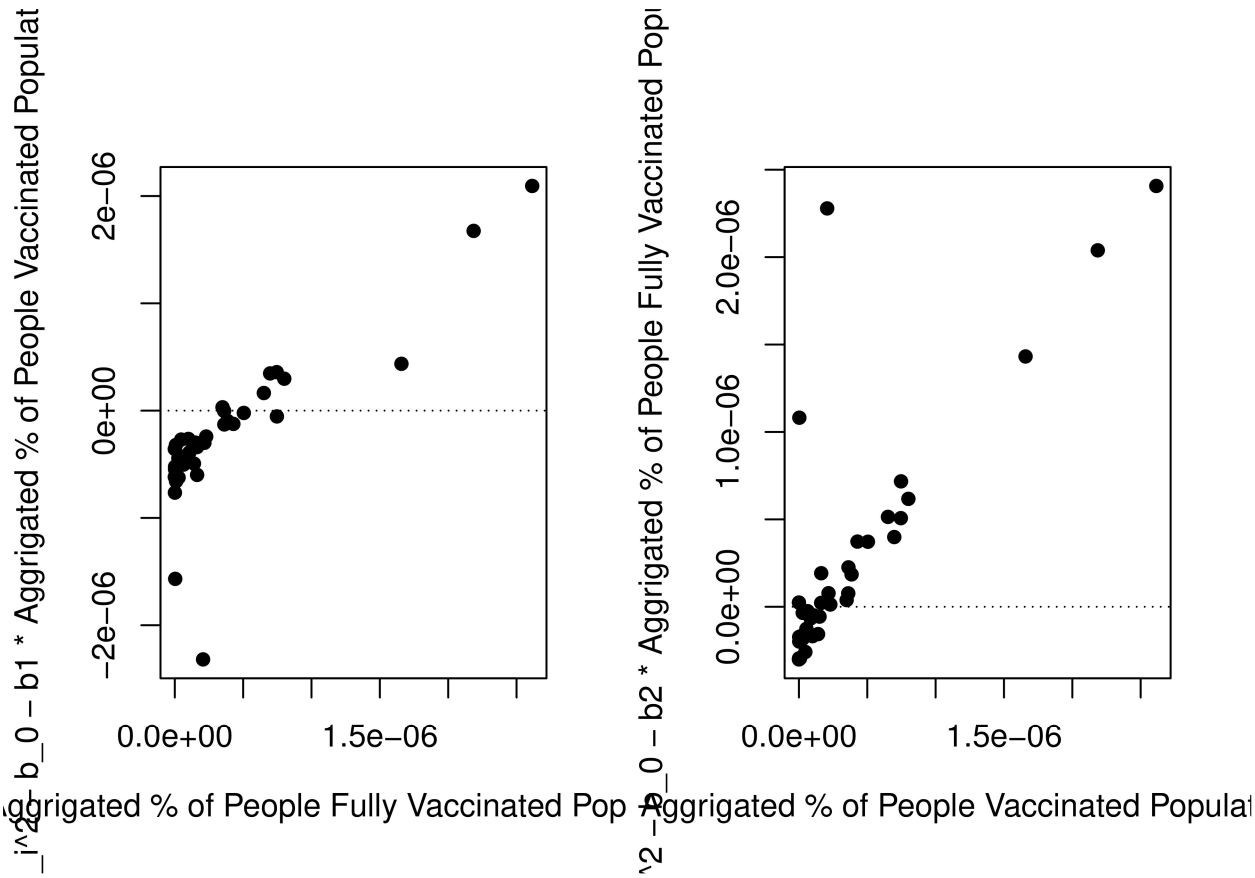
```
Y.hat <- linmod.2$fitted.values # Obtain fitted values
# Compute sums of squares
Y.bar <- mean((agg_new_deaths)^2)
SSR <- sum((Y.hat - Y.bar)^2)
SSE <- sum(((agg_new_deaths)^2 - Y.hat)^2)

n <- length(agg_new_deaths)
p <- 3

# Compute mean squares
MSR <- SSR/(p-1)
MSE <- SSE/(n - p)

f.stat <- MSR/MSE
alpha <- 0.1
fquantile <- qf(1 - alpha, p-1, n - p)
print(c(f.stat, p-1, n-p, fquantile))
```

```
## [1]  1.402741  2.000000 34.000000  2.465809
```

Clearly the Fstat is 1.402741 which is close but less than the 0.95 quantile, 2.465809 we conclude H_o: $\beta_k = 0$ for some k > 0 and k < p, i.e. It implies that One or more of β_1, β_2 is zero.

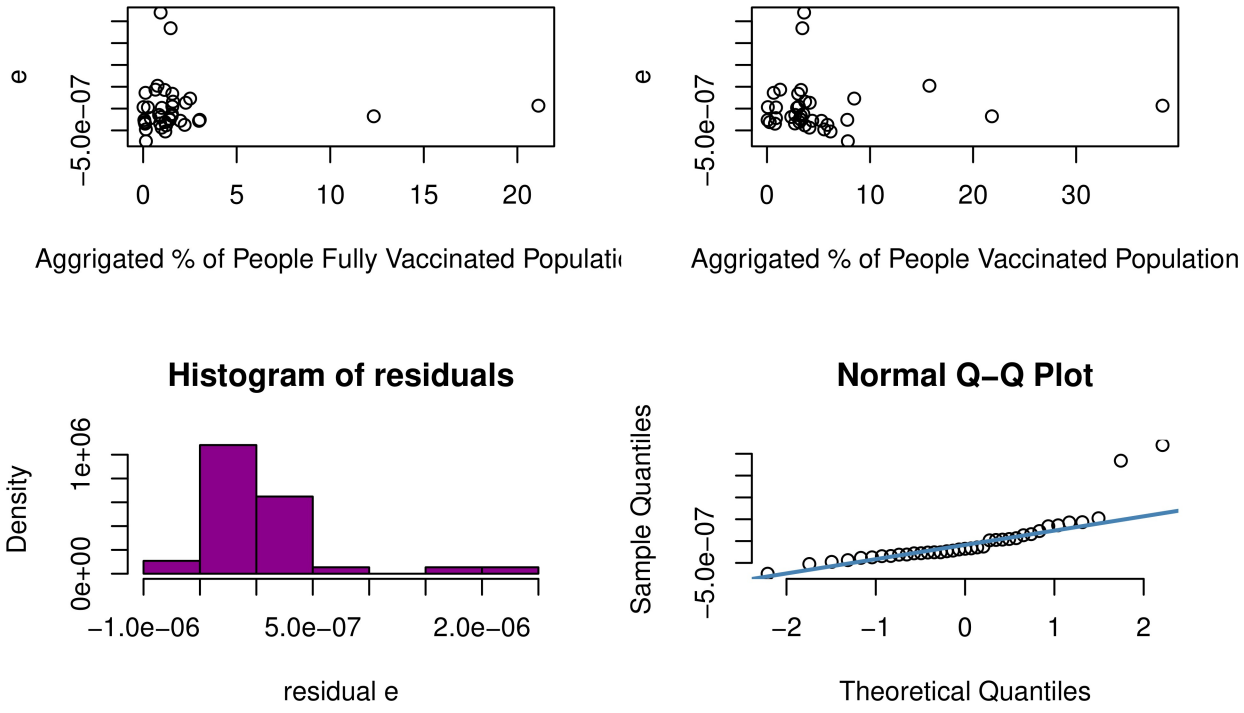
Detailed exploration of final model residuals

```
Y1 <- (agg_new_deaths)^2 - b0 - b1*agg_vaccinated_populations
par(mfrow=c(1,2))
e <- resid(linmod.2)
plot((agg_new_deaths)^2,Y1, xlab = expression('Aggrigated % of People Fully
↪  Vaccinated Population'),ylab='Y_i^2 - b_0 - b1 * Aggrigated % of People
↪  Vaccinated Populations', pch = 16)
abline(h = 0, lty = 3)
Y2 <- (agg_new_deaths)^2 - b0 - b2*agg_fully_vaccinated_population
plot((agg_new_deaths)^2, Y2, xlab = expression('Aggrigated % of People
↪  Vaccinated Populations'),ylab='Y_i^2 - b_0 - b2 * Aggrigated % of People
↪  Fully Vaccinated Population', pch = 16)
abline(h = 0, lty = 3)
```




```
e <- resid(linmod.2)
par(mfrow=c(2,2))
```

```
plot(agg_fully_vaccinated_population, e, xlab="Aggrigated % of People Fully
 ↪  Vaccinated Population")
plot(agg_vaccinated_populations, e, xlab="Aggrigated % of People Vaccinated
 ↪  Population")
hist(e, main="Histogram of residuals", xlab="residual e", col="darkmagenta",
 ↪  freq=FALSE)
qqnorm(e, pch = 1, frame = FALSE)
qqline(e, col = "steelblue", lwd = 2)
```



Independence

For this regression model, the % of People fully Vaccinated and % of People Vaccinated might be dependent to % of new deaths. This is because the % of vaccinated is grouped/ aggregated by country thus there is a point for a single country. And via the graph, the % of People fully Vaccinated and % of People Vaccinated have no dependence to % of new deaths thus affecting the relation between the residual errors (Showing no dependence). Therefore we aren't violating the independence assumption of regression models.

Equal variance

We are also assuming equal variance in our regression model. From the residuals against the fitted values plot, we can see that they make a thin uniform band with 3 outliers, clearly, there is a consistent vertical spread throughout the graph. There is no violation of equal variance due to this consistent spread.

Linearlity

There are no specific regions in the graph with a majority of positive or negative residuals, therefore the model doesn't exclusively under-predict or over-predict in a region, indicating that the data tends to be linear. The same can be said from the $Y_i^2 - b_0 - b_1 * \%$ of People Vaccinated against % Population of Death and $Y_i^2 - b_0 - b_2 * \%$ of People Fully Vaccinated against % Population of Death

Normality

The QQ-Plot allows us to visualize the normality of the errors from the data. There is a minor tail violation of normality in the QQ-Plot (cause of an outlier) since the data is relatively evenly spread out across the line (barring one or two). The tails don't strafe much from the line (Little on one end again not alot). And the distribution is not "compacted" on the ends. The quantiles from residuals distribution almost match quantiles from a normal distribution. Thus assumption of normality is not violated.

Need for interactions assessed

Even on adding the interaction term, the p-value didn't suggest for keeping that variable.

```
n <- length(agg_vaccinated_populations)
linmod2_expanded <- lm(agg_new_deaths^2 ~ agg_vaccinated_populations +
↪  agg_fully_vaccinated_population + agg_vaccinated_populations *
↪  agg_fully_vaccinated_population)
Y.hat.full <- linmod.2$fitted.values
Y.hat.expanded <- linmod2_expanded$fitted.values
SSE.full <- sum((agg_new_deaths^2 - Y.hat.full)^2)
SSE.expanded <- sum((agg_new_deaths^2 - Y.hat.expanded)^2)
df.full <- n - 3
df.expanded <- n - 4
f.stat <- (SSE.full - SSE.expanded)/(df.full -
↪  df.expanded)/(SSE.expanded/df.expanded)
alpha <- 0.05
fquantile <- qf(1 - alpha, df.full - df.expanded, df.expanded)
pvalue <- 1 - pf(f.stat, df.full - df.expanded, df.expanded)
print(c(f.stat, df.full - df.expanded, df.expanded, fquantile, pvalue))
```

```
## [1]  0.003546532  1.000000000 33.000000000  4.139252496  0.952871028
```

Clearly the Fstat is 1.2643724 which is less than the 0.95 quantile, 4.1392525. we conclude H_0: $\beta_4 = 0$ i.e. agg_vaccinated_populations * agg_fully_vaccinated_population can be dropped from the regression model

Conclusion

We started with raising a question of does a country's Human Development Index(HDI) show any relationship with number of people who died of covid in that country? Then to get answer to this question we first modified the data so we can actually compare data of different country with each other, as countries varry is population size and number of death won't be a justifiable with HDI which is already normalized for the population size. Therefor we opted for percentage of people who died of covid, and made a linear model to see if there exist a linear relationship between percentage of people who died of covid and the HDI of that country. We found out from doing the experiment decribed earlier that HDI and percentage of people who died of covid shows a good linear relation, and as we got that result there was no need of tranforming the data for model 1. And doing futher experiments on the model as described earlier we found out that there was no volation of independence, normality, or equal varience. Hence making our model 1 a good representation of relationship between HDI and death due to covid in that country. Now this is really suprising as model gives a positive relationship between HDI and deaths, and that means as the HDI increases the number of people who died due to covid increase. So does this means better developed countries had more deaths due to covid? Maybe, but most likely not. As developed countries cared more for their people and invested money in keeping track of actual number of deaths, we see this trend were reported number of deaths in developed nations are higher. This finding can just go to show more evidence on countries which are not developed enough didn't invest money in keeping track of actual number of deaths and we would never know the actual number of deaths that took place due to covid in those countries.

Next question we raised was does vaccines help in reducing the number of deaths? So to answer this question we again took a location wise approach, in which we normalized the data according to population by making them percentage of death, percentage of people vaccinated, and percentage of people fully vaccinated. And we made linear regression model on percentage of people vaccinate and percentage of people fully vaccinated to predict percentage of deaths, and we called it Model 2. Doing futher experimentations as described in the report we found out that there was clear violation of linearity in this model, and we needed a transformation so we transfomed one of our covariance to square of percentage of people vaccinated. This showed a great improvement in the model bu still showed violation of linearity when number of vaccinated people started getting bigger, and same volations were noticed for normality and equal variance too. Therefor to answer this same question we opted for Model 3 which is a model that shows relationship of square of percentage of population who died due to covid with percentage of people vaccinated and percentage of people fully vaccinated. And doing experiments on this model we found out it was a way better fit of data, showing no violation of linearity, independence, normality, or equal variance. So we decided to answer our question using this model. This model shows near zero but a positive relationship between percentage of people vaccinated and square of percentage of people who died due to covid.

Which means number to deaths in genral tends to go up as more people are vaccinated, we wouldn't come to a conclution of this being true as we have no ways of getting the actual data that can tell us the true story as conculded before we know we don't have correct data from lower developed contries. But on a positive note we see a positive relationship between percentage of people fully vaccinated and square of percentage of people who died due to covid. This show how effective being fully vaccinated is against covid, as even after having the false number from lower indexed contries which didn't show high number of vaccinations, with false data on deaths we still get a positive relationship. So we conclude our report with findings of effectiveness of vaccinations.