Regression Analysis on COVID, Health, and Population Metrics

Antonio R, Aditya M, Aadam L, Christopher L

We can answer multiple questions using regression analysis of the covid data. We will focus on new deaths and new positive cases, and we will use total deaths for some exceptions. We can analyze the statistical relation between these and various parameters like new vaccinations, age, population, etc using regression analysis of the covid dataset. For example, We could find the correlation between total death due to Covid in a country and number of vaccinated people.

This dataset includes data on confirmed COVID cases, deaths, hospitalizations, testing, and vaccinations across several countries since January 4, 2021. The goal of this project is to explore how a the number of new vaccinations varies with respect to other measures over time and across different counties.

An individual observation 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.

We could select multiple combinations of responses and observations to do a regression. The two main responses we will focus on are new deaths and new cases because with this we will be able to predict intervals for deaths and cases on a new day given the covariates. We will also use the total deaths for one specific case.

Our covariates are: days after January 4th, 2021; approximation of new postive tests; total positive tests; tests per case; percentage of the average population that died from COVID; percentage of people vaccinated; percentage of people fully vaccinated; new vaccinations; population density; percentage of population aged over 65/70; GPD per capita; cardiovascular death rate; diabetes prevalence; hospital beds per thousand; life expectancy; human development index; stringency index.

The number of observations is 1107.

Fig1: Histogram of New death due to COVID

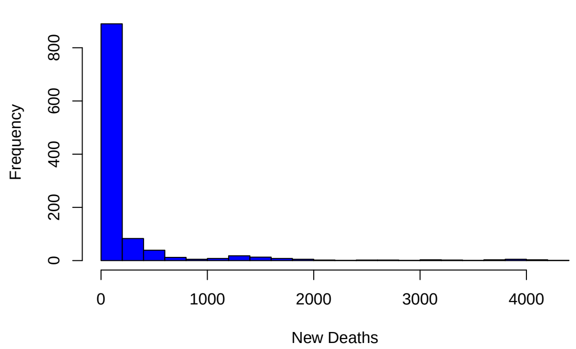
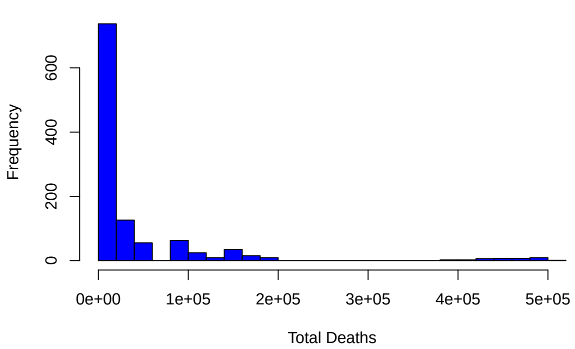


Fig2: Histogram of Total death due to COVID



One of the questions we want to analyze is how the percentage of the population that died from covid correlates to the human development index of a country.

We also want to find the correlation between new deaths and percentage of vaccinated population, and compare it to new deaths with respect to the percentage of fully vaccinated population.

Finally, we want to observe the correlation between new deaths and all of our covariates.

To do this, we need to create the new covariates that are not in the data by multiplying and dividing respective columns from the original dataframe.

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

Response: Percentage of Average Population that died from COVID, Covariate: HDI

```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
  )
```

```r
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]
```

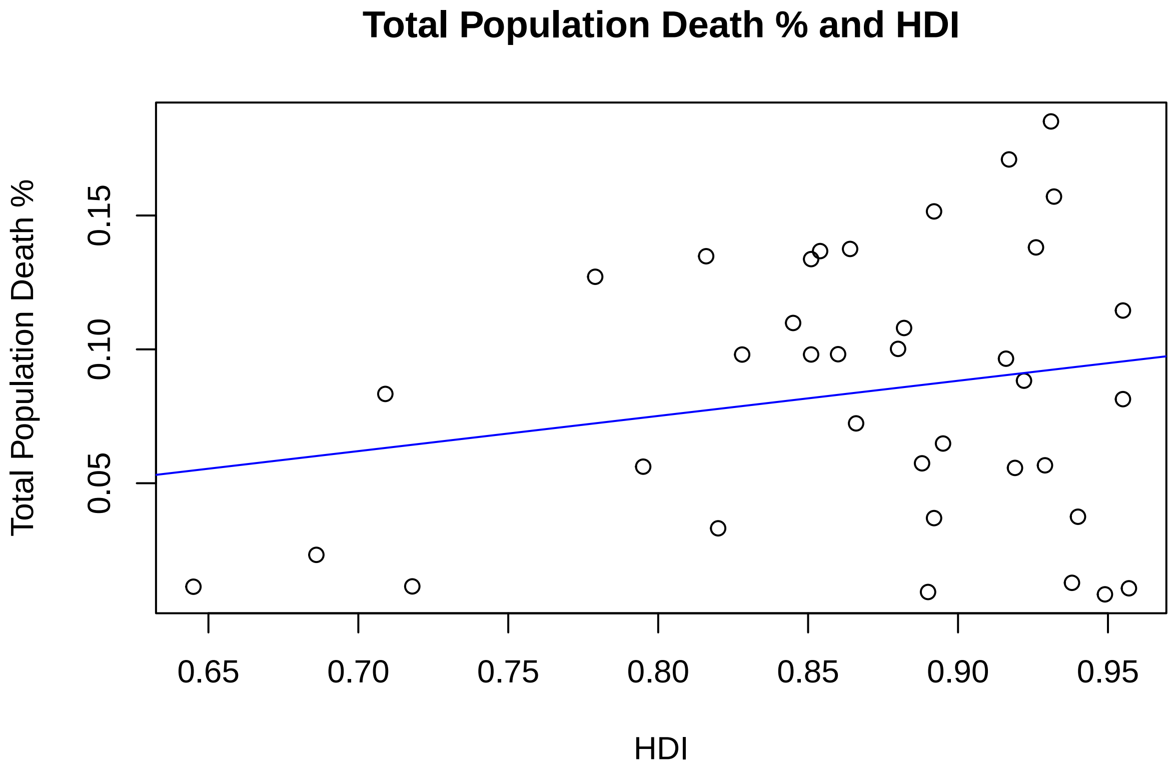
```
df.y <- data.frame(
  "Mean percentage of dead population" = mean(pop_death_per_country),
  "Standard Deviation of percentage of dead population" = sd(pop_death_per_country),
  "Minimum percentage of dead population" = min(pop_death_per_country),
  "Maximum percentage of dead population" = max(pop_death_per_country)
)
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="Total Population Death % and HDI", xlab="HDI", yl
abline(b0, b1, col="blue")
```



```
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 can make the following hypothesis test: $H_0 : \beta_1 = 0$ and $H_a : \beta 1 \neq 0$. With the p-value being 0.230, we can conclude that the smallest α for which we can conclude the alternative hypothesis is 0.230. Which simply means that we can say with a confidence of 0.77 that the percent of the total population that died from COVID has a linear relationship with the Human Development index.

Analysis of Results

To summarize, we started by looking at the regression model between the percentage of the population killed by COVID-19 grouped by country, and the HDI of that country. The resulting model produced a b_1 of approximately 0.131 which is not very large. Furthermore, when we plot the regression model against the data in a scatter plot, the regression model does not appear to fit the data very well, due to the large spread in each variable, seemingly irregardless of the other variable's value. However this is not enough to conclude that the HDI does not have a significant effect on percentage of deaths per country. To formally determine the significance of the explanatory variable on the response variable, we conducted a hypothesis test and calculated the p-value of the test, which came out to 0.230. As stated previously, this means that we can say with a confidence of 0.77 that the percent of total population killed by COVID-19 has a linear relationship with the HDI. A confidence of 0.77 will not often be large enough to conclude that HDI alone has a significant impact upon the percentage of deaths, so therefore we are more likely to conclude the opposite. However, as we explore the data further, it may be worth regressing percentage of deaths per country against HDI in conjunction with other covariates, as this might be more likely to yield a statistically significant result.

Evaluating The Model

The validity of the previous analysis depends on the assumptions we have made by using a regression model, so we need to verify which assumptions we violate.

<>

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 a fan-shape, clearly, there isn't a consistent vertical spread throughout the graph. There is a moderate violation of equal variance due to this inconsistent spread.

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

Normality

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

New Deaths with Respect to the Percentage of Vaccinated and Fully Vaccinated Population

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

```
agg_new_deaths <- aggregate(covid_new$perecentage_new_deaths, by=list(covid_new$location
agg_fully_vaccinated_population <- aggregate(covid_new$percentage_fully_vaccinated, by=l
agg_vaccinated_populations <- aggregate(covid_new$percentage_vaccinated, by=list(covid_n
```

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 atleast 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 atleast one dose) All aggregated by country/ location.

Summary of response and covariates mean,

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

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

```
mean(covid_new$perecentage_new_deaths)
```

```
## [1] 0.0005745128
```

```
mean(covid_new$percentage_vaccinated)
```

```
## [1] 5.988196
```

```
mean(covid_new$percentage_fully_vaccinated)
```

```
## [1] 2.333185
```

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

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

```
mean(agg_new_deaths)
```

```
## [1] 0.000479282
```

```
mean(agg_vaccinated_populations)
```

```
## [1] 5.174694
```

```r
mean(agg_fully_vaccinated_population)
```

```
## [1] 1.990691
```

Summary of response and covariates standard deviation,

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

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

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

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

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

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

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

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

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

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

```r
sd(agg_new_deaths)
```

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

```r
sd(agg_vaccinated_populations)
```

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

```r
sd(agg_fully_vaccinated_population)
```

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

Summary of response and covariates range,

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

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

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

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

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

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

```r
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)
```

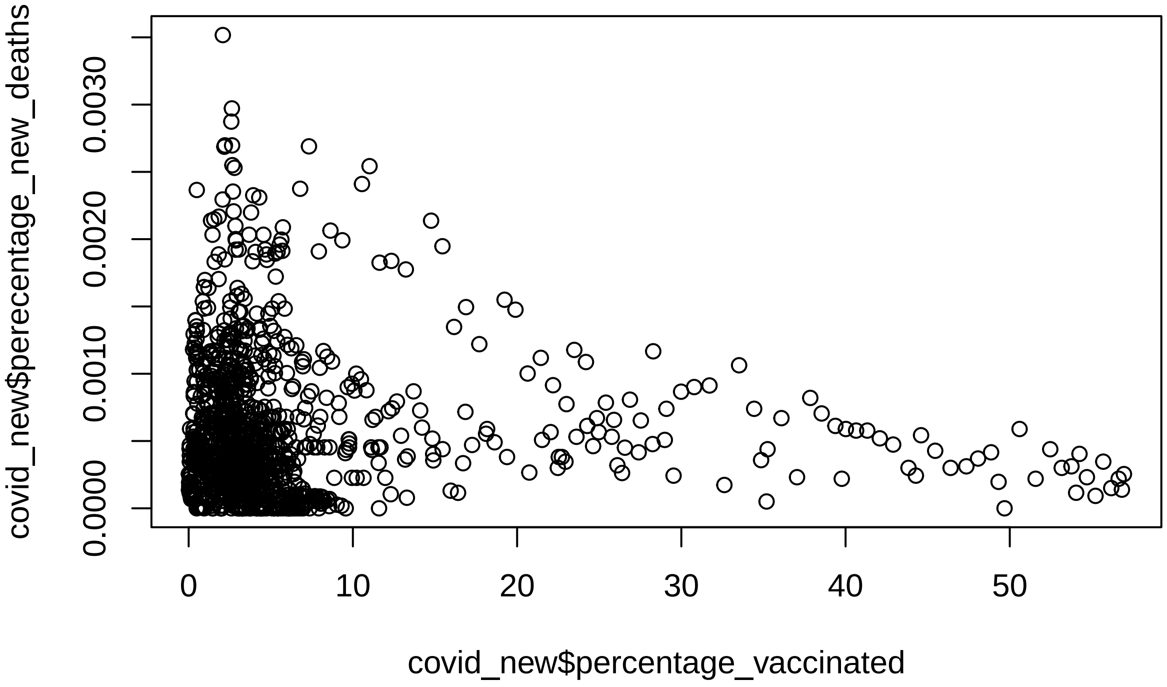
```
## [1]  0.0212315 21.1243446
```

Scatter plots against each covariate,

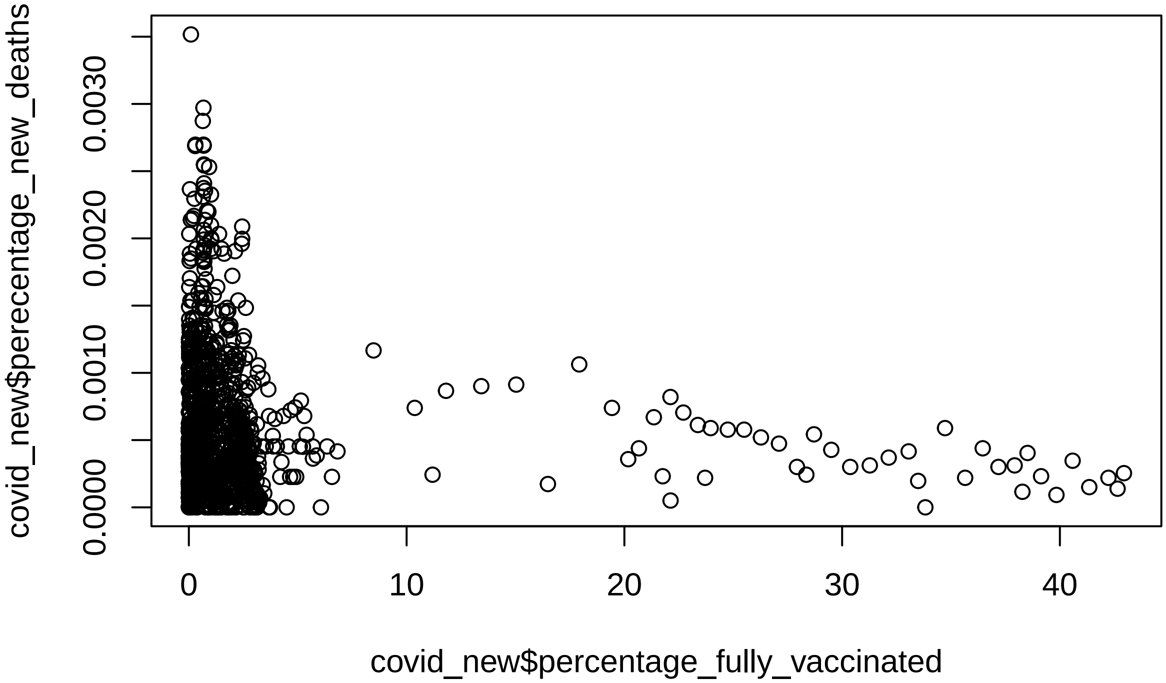
```
print("Summary for observation 1")
```

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

```
plot(covid_new$percentage_vaccinated, covid_new$perecentage_new_deaths)
```



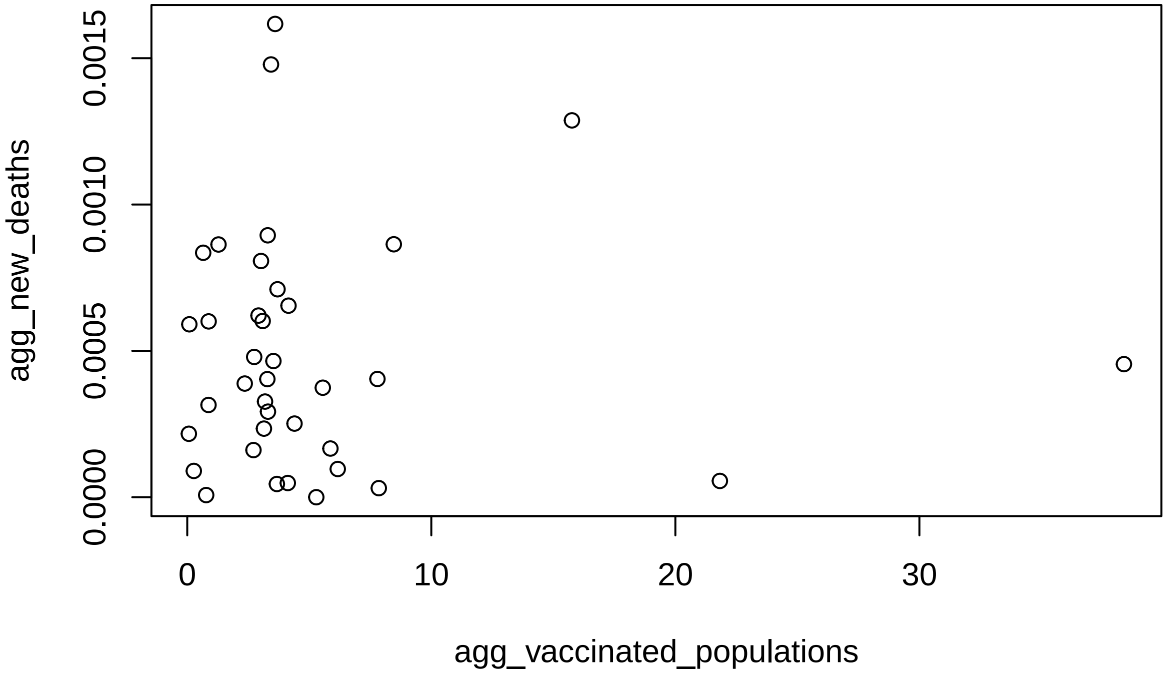
```
plot(covid_new$percentage_fully_vaccinated, covid_new$perecentage_new_deaths)
```



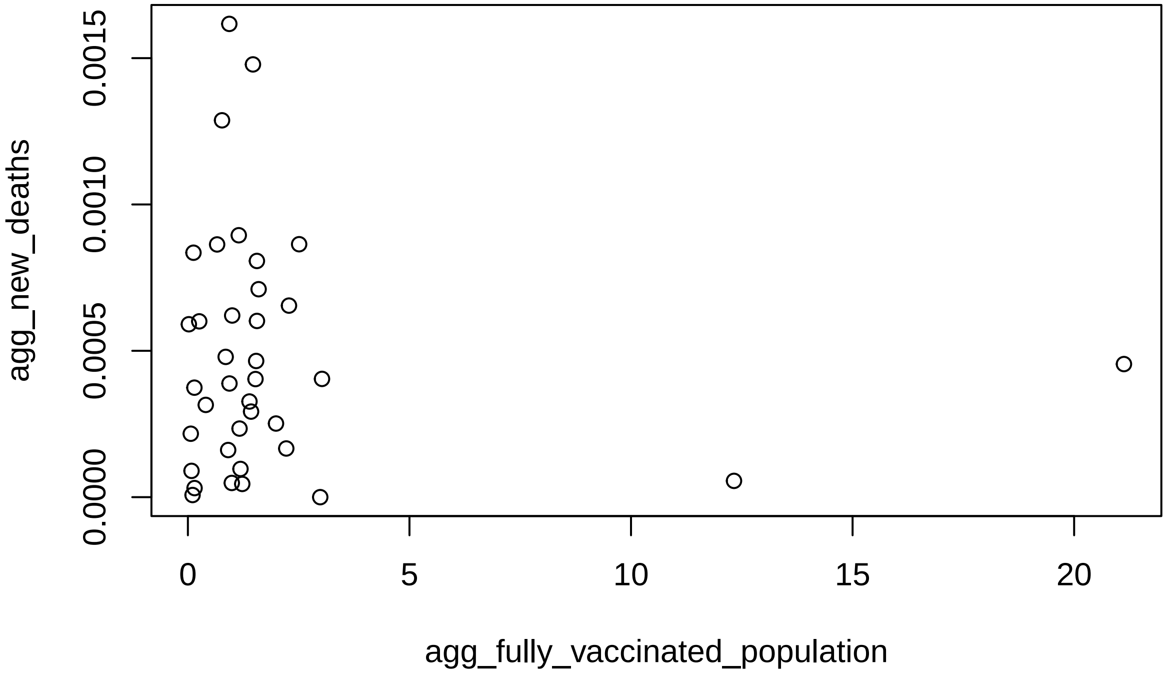
```
print("Now observation 2, for aggregated data by country or location")
```

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

```
plot(agg_vaccinated_populations, agg_new_deaths)
```



```
plot(agg_fully_vaccinated_population, agg_new_deaths)
```



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_
print(linmod$coefficients)
```

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

```
print("")
```

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

```
# summary(linmod)
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_popula
print(linmod.2$coefficients)
```

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

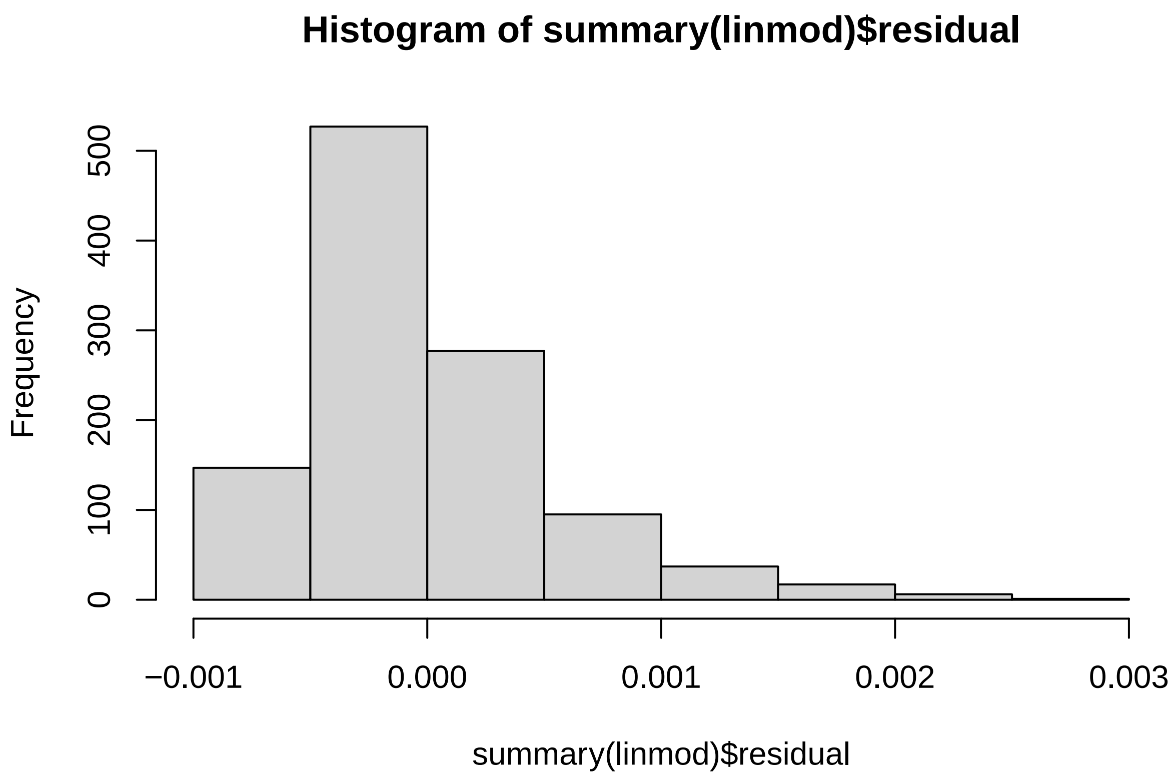
```
# summary(linmod.2)
```

Summary that includes estimated standard errors

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

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

```
hist(summary(linmod)$residual)
```



```
print("R squared value")
```

```
## [1] "R squared value"
```

```
print(summary(linmod)$r.squared)
```

```
## [1] 0.01884497
```

```
print("MSE")
```

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

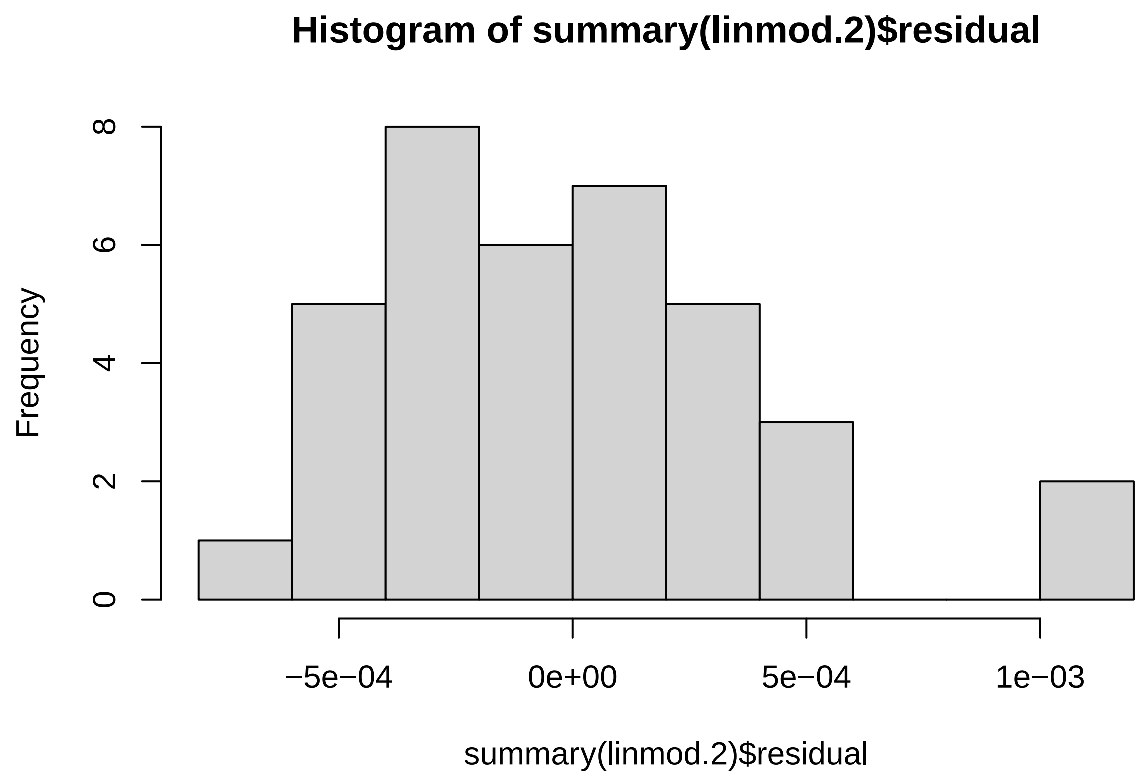
```
print(sigma(linmod)^2)
```

```
## [1] 2.712268e-07
```

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

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

```
hist(summary(linmod.2)$residual)
```



```
print("R squared value")
```

```
## [1] "R squared value"
```

```
print(summary(linmod.2)$r.squared)
```

```
## [1] 0.04719345
```

```
print("MSE")
```

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

```
print(sigma(linmod.2)^2)
```

```
## [1] 1.624552e-07
```

Summary that includes test results, here we found the p value and t statistic with a null hypothesis of $\beta = 0$

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

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

```
n <- nrow(covid_new)
b1 <- linmod$coef[2]
s.b1 <- summary(linmod)$coef[2, 2]
b2 <- linmod$coef[3]
```

```
s.b2 <- summary(linmod)$coef[3, 2]
t_stat_b1 <- b1/s.b1 # Compute the test statistic of B1
t_stat_b2 <- b2/s.b2 # Compute the test statistic of B2
p <- 3
df <- n - p # Degrees of freedom
pvalue <- 2*pt(-abs(t_stat_b1), df)
print(pvalue)
```

```
## covid_new$percentage_vaccinated
##                      0.002171847
```

```
pvalue <- 2*pt(-abs(t_stat_b2), df)
print(pvalue)
```

```
## covid_new$percentage_fully_vaccinated
##                           1.736434e-05
```

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

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

```
n <- length(agg_new_deaths)
b1 <- linmod.2$coef[2]
b2 <- linmod.2$coef[3]
s.b1 <- summary(linmod.2)$coef[2, 2]
s.b2 <- summary(linmod.2)$coef[3, 2]
t_stat_b1 <- b1/s.b1 # Compute the test statistic of B1
t_stat_b2 <- b2/s.b2 # Compute the test statistic of B2
p <- 3
df <- n - p # Degrees of freedom
pvalue <- 2*pt(-abs(t_stat_b1), df)
print(pvalue)
```

```
## agg_vaccinated_populations
##                  0.2457979
```

```
pvalue <- 2*pt(-abs(t_stat_b2), df)
print(pvalue)
```

```
## agg_fully_vaccinated_population
##                       0.2033894
```

###Analysis of Results

Since observation 2 is the aggregate of the data by country, we would expect the variance of the data to decrease in testing the second observation. After looking at the data analysis, we can see that this expectation was correct, as the the variance in the data decreases when we aggregate the data by country rather than looking at each individual observation all together. However, the values of the mean also decrease noticeably when we aggregate the data. The reasoning for this change is likely due to the fact that at most one or two of our countries have significantly larger values for percent deaths, and people vaccinated than other countries, thus the values of the means get skewed towards larger values. However, in the aggregate, there are less larger outlierish observations contributing to the mean so the value of the mean decreases down to a more likely range. When we go to look at the scatterplots of the data, we see that this explanation is likely correct, as you can see the number of data points which are farther from the concentrated center of the graph decreases from non-aggreagted data to aggregated data. However, attempting to fit these scatterplots with a regression model does not appear to yield any useful results because the data is heavily concentrated around the bottom left corner of the graph. This makes it difficult for a line of best fit to beneficial, despite the fact that the plots appear to be triangular in shape with regards to the upper bound on the relationship between vaccinated population and new deaths. A better way to explore this observation could be to instead of aggregate the data by country, analyze each country's data individually with respect to time to see if as more people get vaccinated, less new deaths appear.