Exploring COVID-19 Deathrates and Deathtolls

Jean-Sebastien Paul

Abstract

The outbreak of the COVID-19 coronavirus, caused by severe acute respiratory syndrome (SARS) coronavirus 2 (SARS-CoV-2), has so far killed over 276K people and infected a confirmed 3.97M. This open-ended project was intended to explore deathtolls and deathrates, hopefully providing insight as to what affects these and what can be done to reduce this.

Intoduction

Data Sources

Recovery Data

COVID-19 Recovery data was taken from John Hopkins University Center for Systems Science and Engineering[1]. This was in cumulative recovery format for each country on a date basis.

Predictor Data

COVID-19 Predictors were taken from Kaggle[2]. Predictors extracted from this included: Density[3], Urban Population, 2020 Population, Hospital Beds/1000 citizens[4], Sex Ratio (overall and based on age)[5], Lung diseases death rate[6] (overall and for both sexes), Median age, 2018 GDP[7], Crime Index[8], Smoking Rate (2016)[9], for as many countries as possible.

Testing Data

COVID-19 Testing data was downloaded from ourworldindata[10]. This was cumulative number of tests for each country available on a date basis.

Case and Death Data

COVID-19 Case and Death Data was extracted from the European Centre for Disease Prevention and Control[11]. This contained new cases and deaths each day for a large number of countries. Cumulative cases and deaths was added, as was deathrate.

[1] https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv

[2] https://www.kaggle.com/nightranger77/covid19-demographic-predictors

[3] https://www.worldometers.info/ - Data on Density, Population, Median Age, Urban Population

[4] https://data.worldbank.org/indicator/SH.MED.BEDS.ZS

[5] https://en.wikipedia.org/wiki/List_of_countries_by_sex_ratio, https://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS

[6] https://www.worldlifeexpectancy.com/cause-of-death/lung-disease/by-country/

[7] https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

[8] https://worldpopulationreview.com/countries/crime-rate-by-country/

[9] https://ourworldindata.org/smoking#prevalence-of-smoking-across-the-world

[10] https://ourworldindata.org/grapher/full-list-total-tests-for-covid-19

[11] https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

Logical Variables

Rich and Crime discrete variables were created based on the predictor data. Rich being 1 is defined as being above mean GDP/capita. Crime being 1 is defined as being above mean crime index (indicates more crime).

Combination

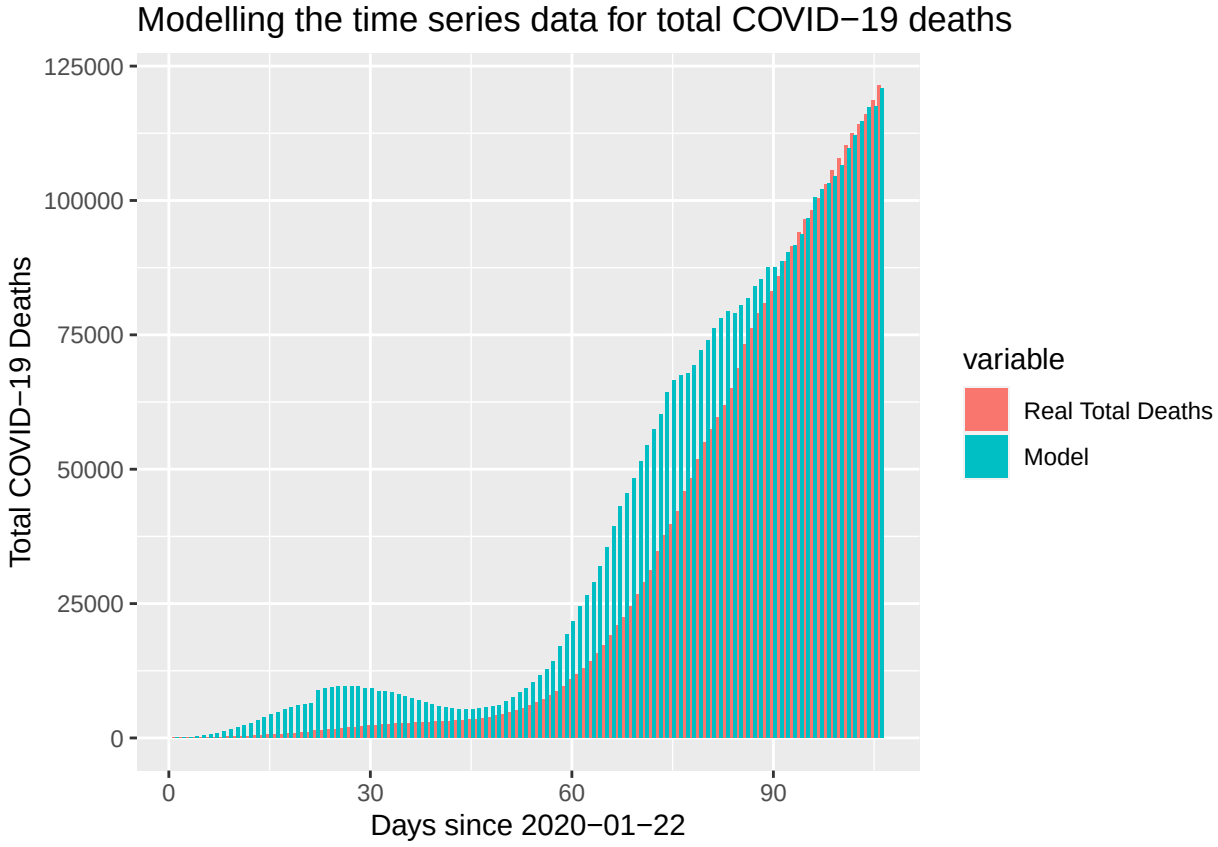
All these datapoints were joined by country and by date. 2 main dataframes were created. One contained data solely on cases, recoveries, testing and deaths. The other was a combination of this and the predictor data- this was smaller however on account of not all countries having predictor data on them available.

Topic 1: Exploration of the Distribution and Time Series change of Variables related to Death Rate from COVID-19 using probability distributions.

Part 1: Can we model the time series data of new deaths world wide as binomial or poisson distribution

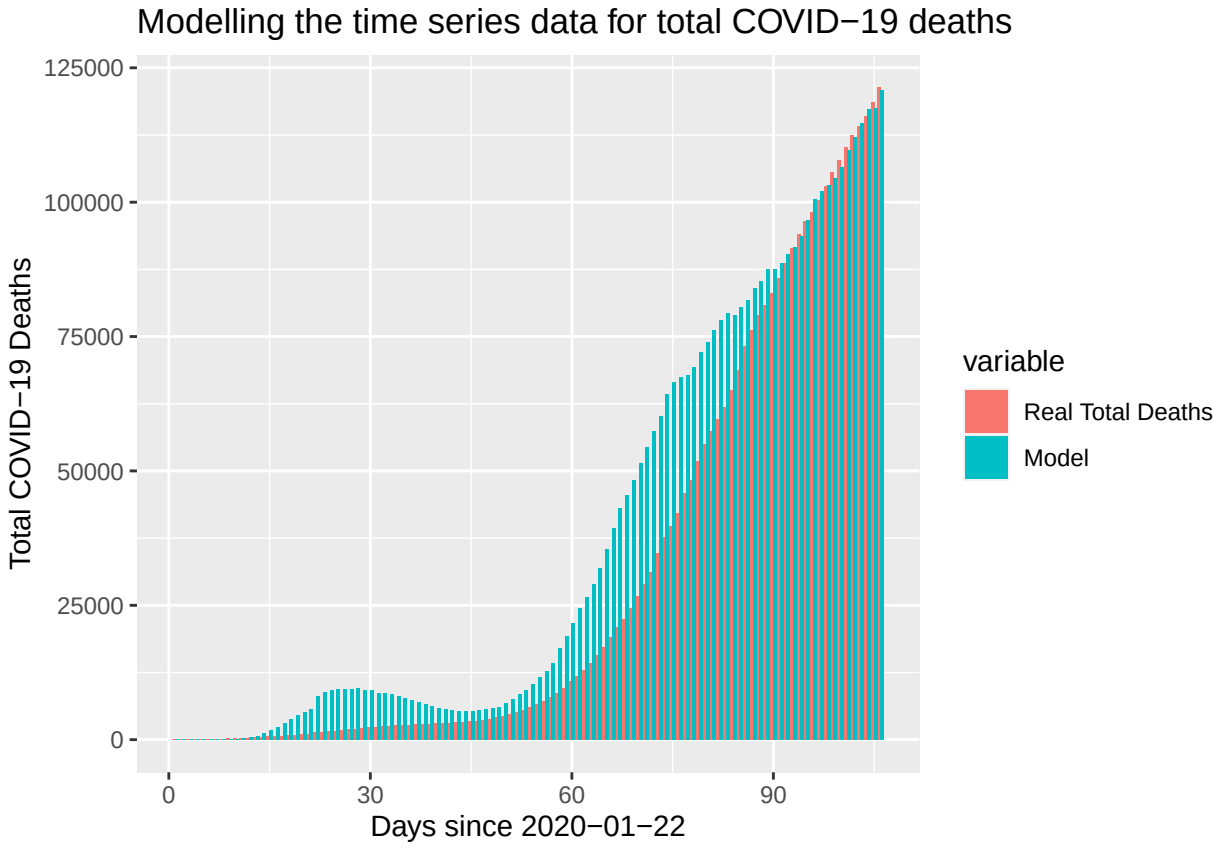
It was theorized that the time series data of new COVID-19 deaths worldwide could be modelled as fitting a poisson or binomial distribution. This would make sense as each case could be considered a Bernoulli random variable with some probability of surviving or dying. This probability was though to be relatively constant enough that the model should see some success.

However both models were weak. The best binomial model found was defined as: $TotalDeaths = ActiveCases * 0.17 * P(X \leq Day)$ where $X \sim Binom(k, DeathRate)$, k is defined as $ActiveCases/1000$ rounded to the nearest whole number, Day is days since 2020-01-22, and $DeathRate$ is considered to be the mean overall total death rate. The plot of this model is shown below.



This is clearly an ill-fitting model, overpredicting for the first 90 days. It erroneously also a predicted a first wave of COVID-19 deaths of sorts. Whilst it might have started to fit the real data well towards the end, it was overall poor.

The Poisson model did not fare better, with near identical results. The best model found was defined as $TotalDeaths = ActiveCases * 0.17 * P(X \leq Day)$ where $X \sim Binom(k, DeathRate)$ and Day is days since 2020-01-22. The plot of this model is shown below.

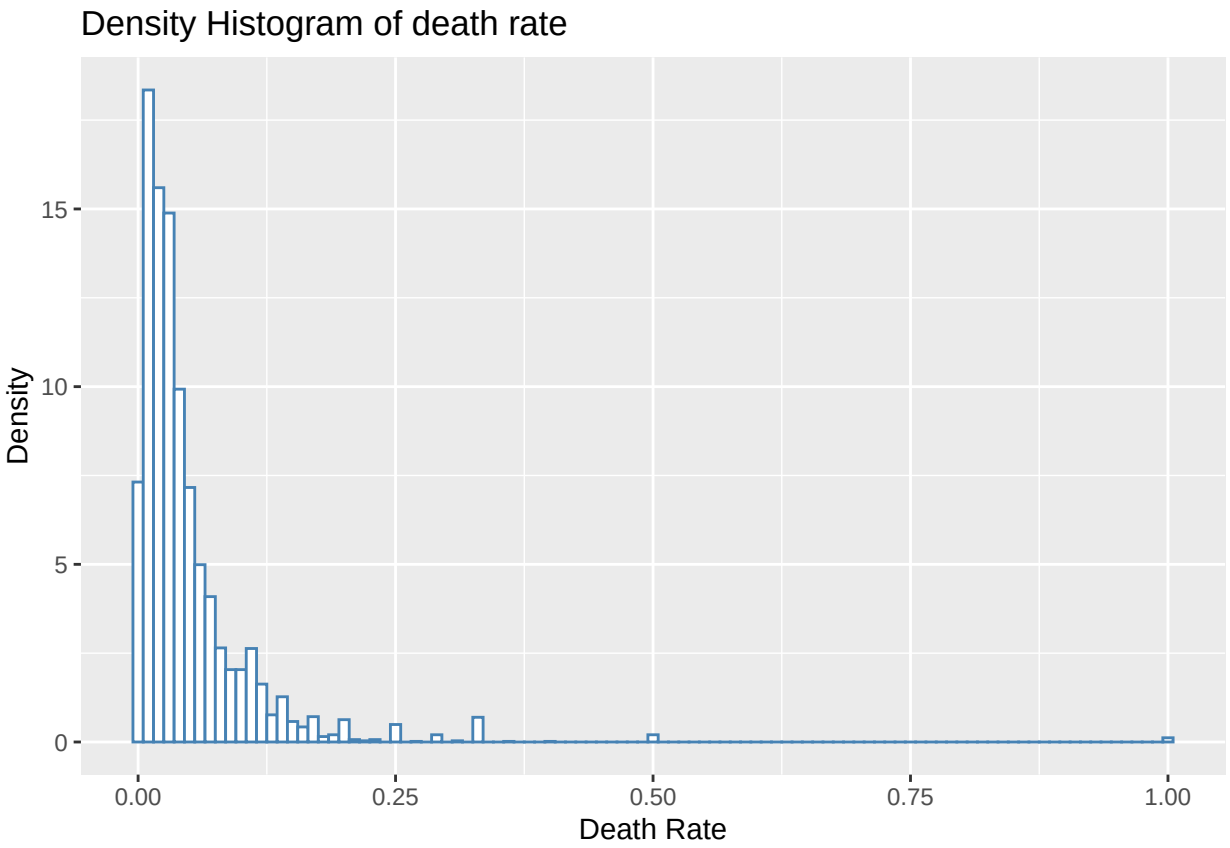


The same conclusions as the binomial model can be drawn for the poisson model. It is overall poor.

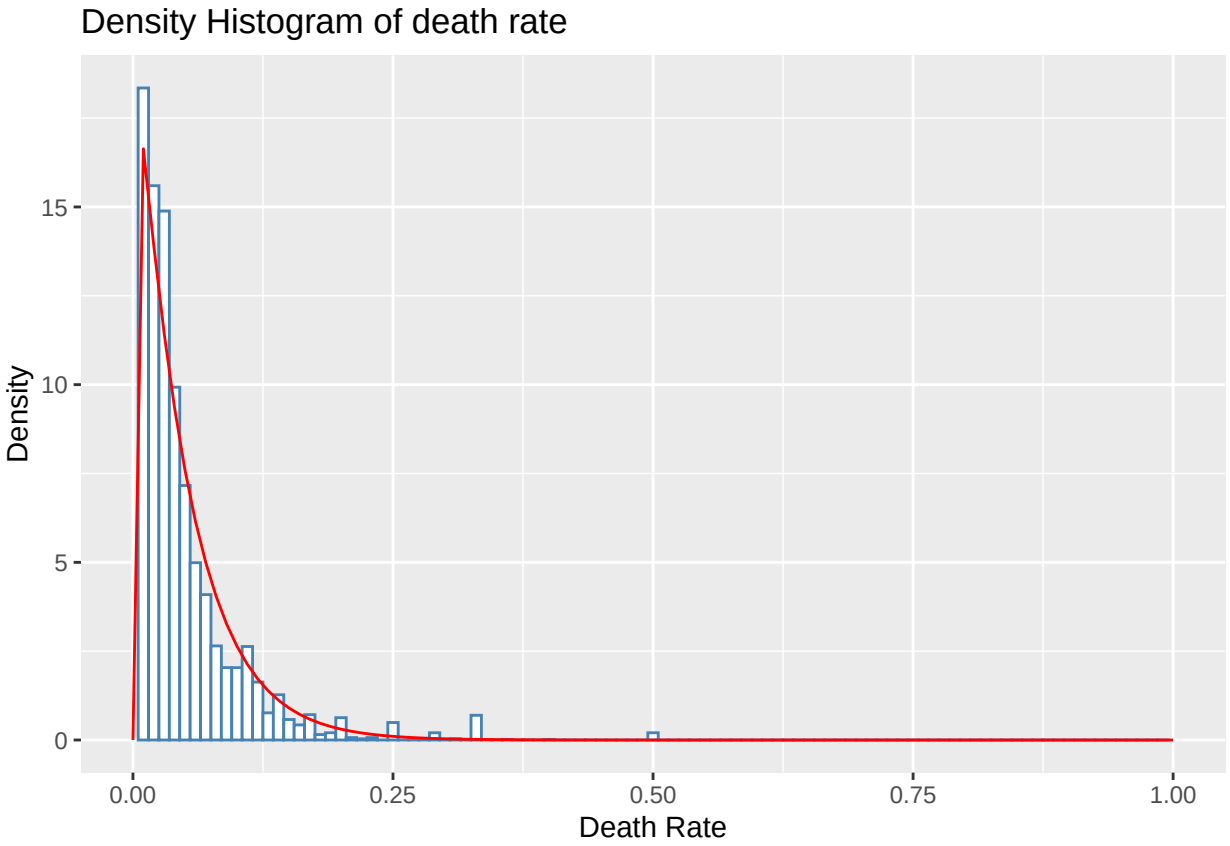
It was concluded therefore that modelling cumulative deaths via a probability distribution is unlikely to succeeed. Potentially multivariable regression could yield better models and reveal more about what affects death tolls.

Part 2: How is the frequency of death rates distributed: does this converge at a certain value or does it vary wildly? Can we model this via a probability distribution function?

Looking at a plot of death rates revealed seemingly low variance, with a high concentration of deathrate values towards about 2%, with positive skew.



This suggested that a gamma distribution would be a good fit. A gamma probability density function model was developed $f_X(x) = dgamma(shape = 1.06011822, rate = 21.99330112)$. This was overlayed onto the histogram.



This looked like a decent fit. However this was shown to be wrong by a chi squared test that returned a p-value of 1.651946e-37. It is thus extremely improbable that we observe a test statistic this extreme from the relevant chi square distribution, so under the current evidence we strongly rejected the null hypothesis that deathrates follows a gamma distribution and failed to model the distribution of death rates.

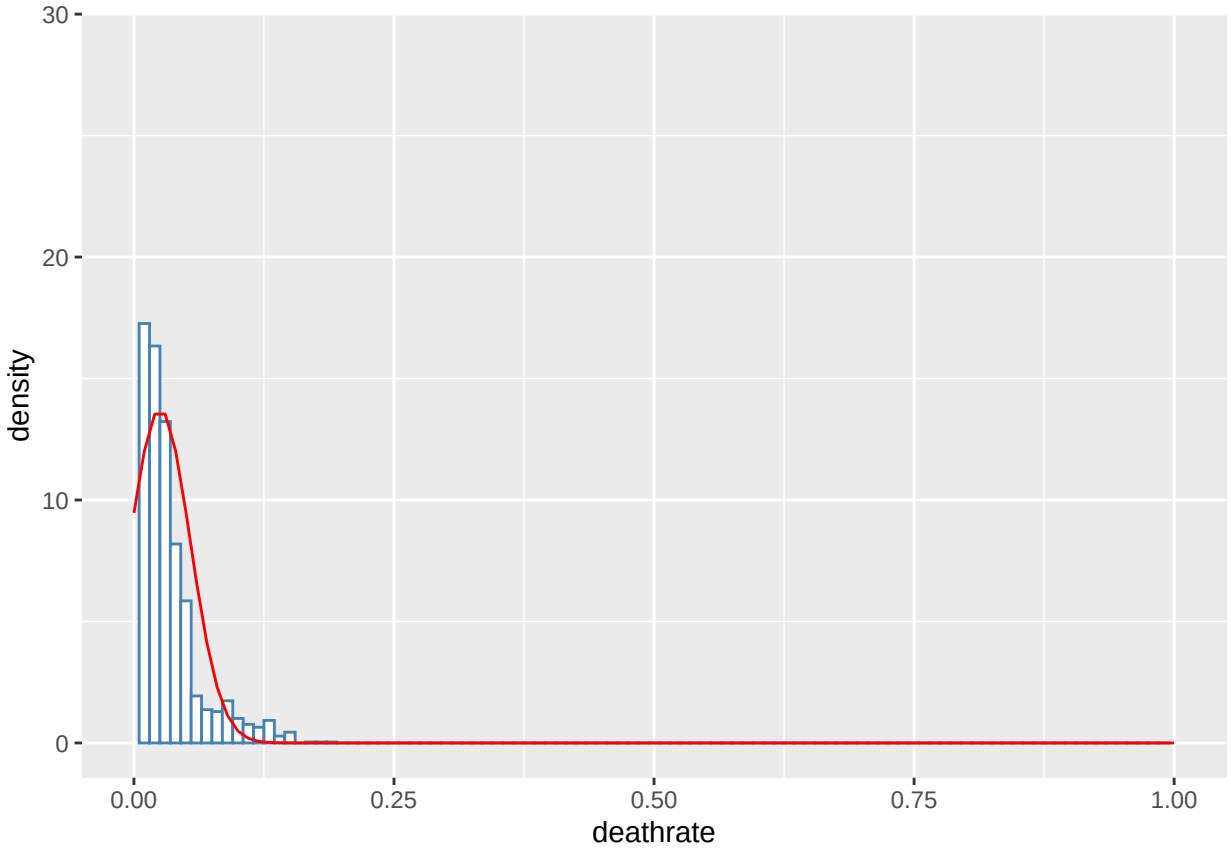
Topic 2: Exploring Correlations and Independences within COVID-19 Data

Part 1: Is there a significant difference in mean death rates between rich countries and poor countries?

Rich countries were defined as being above mean GDP/Capita, and poor as below. The mean death rates in rich and poor countries was investigated to see if the difference of 0.002490895 (rich being larger suprisingly) was significant. A t-test was first used to look at this.

```
##
##  Welch Two Sample t-test
##
## data:  data2$deathrate[rich] and data2$deathrate[poor]
## t = 1.7699, df = 1200.9, p-value = 0.07699
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0002701965  0.0052519868
## sample estimates:
##  mean of x  mean of y
## 0.02671302 0.02422213
```

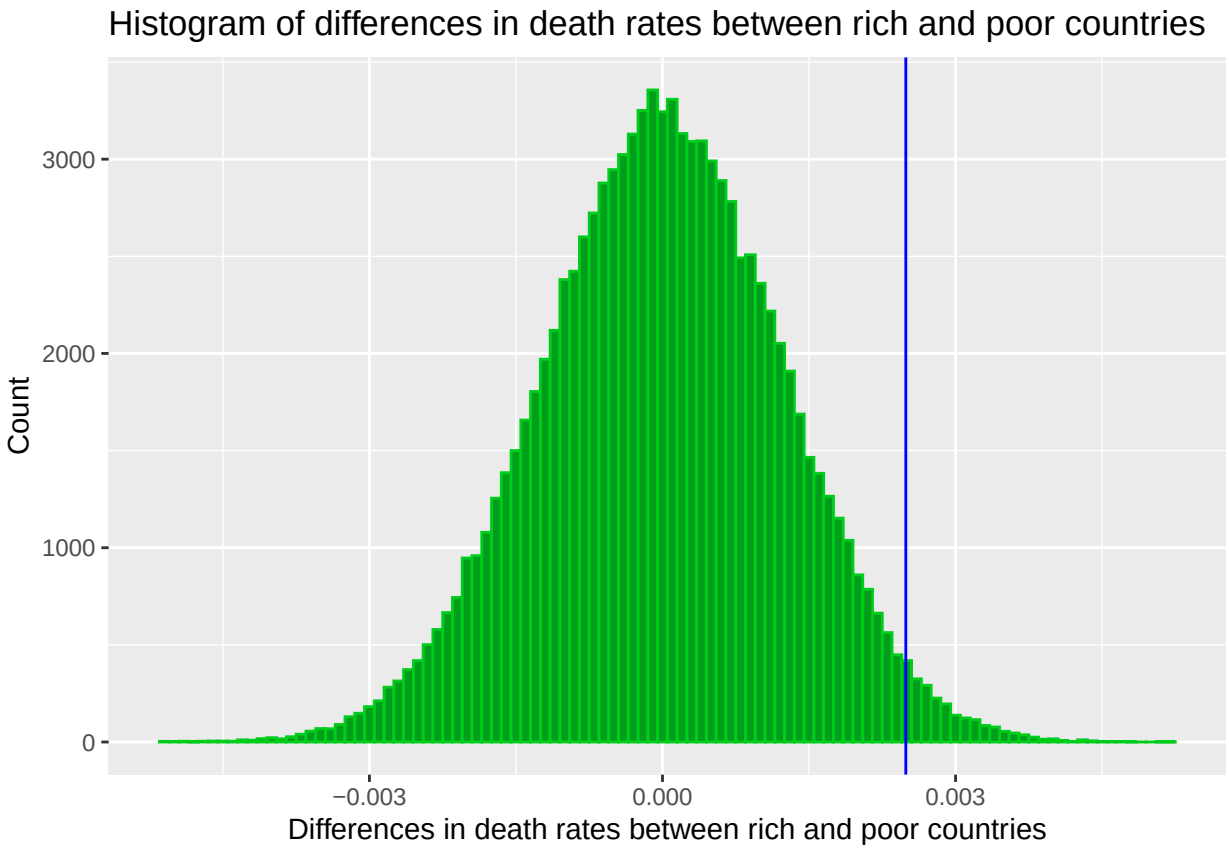
The T test returned a p-value 0.07699, seemingly leading to the suprising acceptance of the null hypothesis of no mean difference of death rates between rich and poor countries. However the t-test was discovered to be fundementally flawed. T-tests assume normality. COVID-19 Death-rates do not follow a normal distribution.




```
##
##  Shapiro-Wilk normality test
##
## data:  data2$deathrate
## W = 0.78878, p-value < 2.2e-16
```

This is clearly shown by the plot and the Shapiro-Wilk test for normality, with p-value <2.2e-16 meaning that we reject the null hypothesis of no significant difference with the normal distribution.

As such a Permutation test was to be used to test for significant mean difference between rich and poor country deathrates, with no prior assumptions needed.



There is a 4.199958% chance of discrepency by chance (by the p-value extracted from the test). We reject the null hypothesis of no mean difference at the 0.05 level of significance. This leads to suprising conclusion that richer countries have higher COVID-19 Deathrates. This is probably because poorer countries have less testing and are therefore less able to determine who has died from COVID-19, to such an extent that this would occur. The significant difference could be promising for modelling in topic 3 later on.

Part 2: Are crime and high death rates are independent?

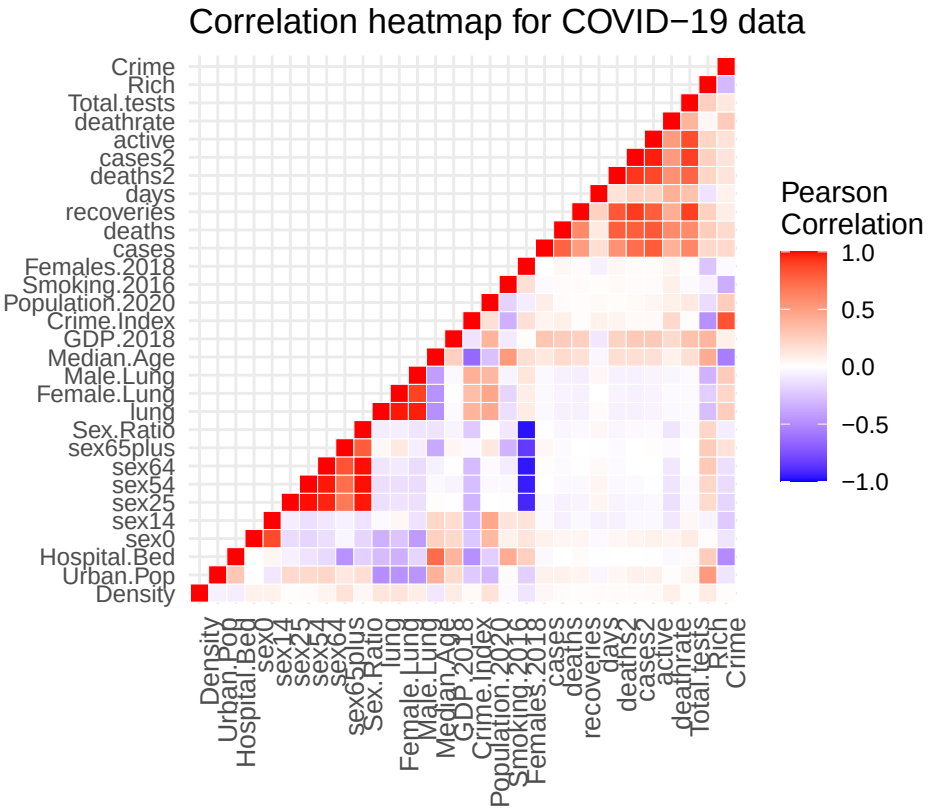
A new logical variable *hdr* was defined: being 1 for countries in the upper half of death rates, and 0 for countries in the lower half. A chi-squared test on this and *Crime* was used to investigate independence. Below is the contingency table for these variables.

	Crime	
	0	**1**
hdr		
0	731	509
1	502	737
#Total cases	1233	1246

The chi-squared test returned a p-value so small that the computer used rounded it to 0. This meant that the null hypothosis of independence was strongly rejected. Crime and high deathrates are unsuprisingly related, crime itself being correlated with many other critical variables in a society such as wealth and how obeyant a population is. This suggests crime will be useful in modelling in topic 3.

Part 3: Exploring Correlation of all our data variables amongst themselves

We can carry this out using a correlation heatmap. This could provide insights to what is likely useful for modelling in topic 3.



Death rate is actually not strongly correlated much at all with the other variables. Interestingly, we dont see high degrees of similarity. Perhaps this indicates that deathrate is relatively constant and hard to model. This indicates challenges for topic 3.

Topic 3: Modelling deathrate and deathtoll

Modelling deathrate via logistic regression

8 models were trained on 80% of the data, and tested on the other 20%. The best model was determined as that which had the lowest squared estimate of errors .

To reduce spread of outliers: *GDP*.2018, *Population*.2020, *Total.tests* were logged. *recoveries*, *cases*2 and *cases* were squarerooted however since they had 0 values in the data.

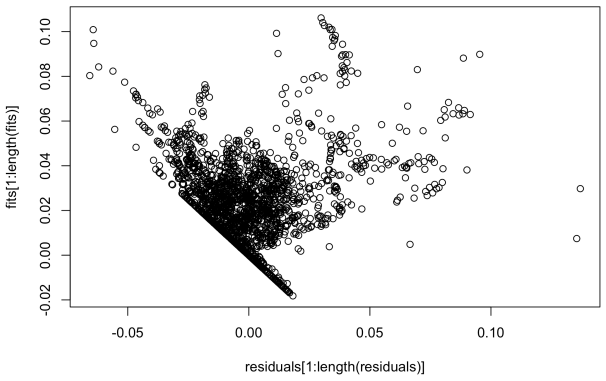
In order to deal with multicollinearity, the following were dropped one by one from the numeric data in accordance with their variance inflation factors: *lung*, *Sex.Ratio*, *sex*54, *GDP*.2018, *cases*2, *sex*14, *Male.lung*. This led to only these variables being used in modelling: *Density*, *Urban.Pop*, *Hospital.Bed*, *sex*0, *sex*25, *sex*64, *sex*65*plus*, *Female.Lung*, *Median.Age*, *Crime.Index*, *Population*.2020, *Smoking*.2016, *Smoking*.2016, *Females*.2018, *cases*, *recoveries*, *days*, *Total.tests*, *Rich*, *Crime*.

The first model was a logistic regression of deathrate onto all the available variables. The second model was a logistic regression model of deathrate onto the interaction terms of all the available variables to the second power. The third model was a backward/forward stepwise regression model on the first model with AIC criterion. The fourth model was a backward/forward stepwise regression model on the first model with BIC criterion. The fifth model was a lasso model of deathrate onto all the available variables. The sixth model

was a lasso model of deathrate onto all the available variables. The seventh model was a neural net with threshold 0.1 of deathrate onto all the available variables. Finally, the eighth model was simply a constant: the mean deathrate.

Comparison on the testing data revealed the sixth model (ridge) to be best, with the lowest SSE. However this model had a relatively poor R^2 value of 0.4729924. It's adjusted R^2 was very close though at 0.4519564, indicating that the model did not overfit and strongly explained some of the variance. It should be noted that diagnostic plots were not promising with the fitted values versus residual plot not looking random on the bottom left hand side. Moreover we received p-value <2.2e-16 in the shapiro wilk test for normality on residuals meaning that we rejected the null hypothesis of no significant difference with the normal distribution, and conclude that under the current evidence, the residuals do not follow a normal distribution; directly against the assumptions of linear regression. This may compromise the model.

```
                         s0
(Intercept)     -6.206662e-02
Density          2.019813e-06
Urban.Pop        1.417604e-05
Hospital.Bed    -1.090896e-03
sex0            -3.004354e-02
sex25           -1.442704e-02
sex64           -5.988526e-03
sex65plus        1.382621e-02
Female.Lung     -6.589495e-05
Median.Age       1.294715e-03
Crime.Index      3.520410e-04
Population.2020  3.988713e-03
Smoking.2016     3.525355e-04
Females.2018    -4.294787e-04
cases            5.020080e-04
recoveries       1.569107e-04
days             7.275809e-04
Total.tests     -4.694847e-03
Rich             5.173558e-03
Crime            5.292701e-03
```




```
        Shapiro-Wilk normality test

data:  residuals[1:length(residuals)]
W = 0.90598, p-value < 2.2e-16
```

Modelling total d via logistic regression

9 models were trained on 80% of the data, and tested on the other 20%. The best model was determined as that which had the lowest squared estimate of errors .

To reduce spread of outliers: *GDP*.2018, *Population*.2020, *Total.tests* were logged. *recoveries*, *cases*2 and *cases* were squarerooted however since they had 0 values in the data.

In order to deal with multicollinearity, the following were dropped one by one from the numeric data in accordance with their variance inflation factors: *lung*, *sex*54, *Sex.Ratio*, *sex*64, *Females*.2018, *Population*.2020, *days*, *recoveries*, *Male.Lung*, *sex*0. This led to only these variables being used in modelling: *Density*, *Urban.Pop*, *Hospital.Bed*, *sex*14, *sex*25, *sex*65*plus*, *Female.Lung*, *Median.Age*, *GDP*.2018, *Crime.Index*, *Smoking*.2016, *cases*, $cases2, *Total.tests*, *Rich*, *Crime*.

The first model was a linear regression of deathrate onto all the available variables. The second model was a forward/backward stepwise regression model on the first model with AIC criterion. The third model was a forward/backward stepwise regression model on the first model with BIC criterion The fourth model was a logistic regression model of deathrate onto the interaction terms of all the available variables to the second power. The fifth model was a forward/backward stepwise regression model on the fourth model with AIC criterion. The sixth model was a forward/backward stepwise regression model on the fourth model with BIC criterion The seventh model was a lasso model of deathrate onto all the available variables. The eighth model was a lasso model of deathrate onto all the available variables. The ninth model was a neural net with threshold 0.1 of deathrate onto all the available variables.

Comparison on the testing data revealed the fifth and sixth models to be best, with the lowest SSE. Looking at the fifth model we had a strong R^2 of 0.8900757 and similarly strong adjusted R^2 of 0.8864039, indicating that the model did not overfit and strongly explained much of the variance. It should be noted however that

diagnostic plots were not promising with the fitted values versus residuals plot not looking random on the left hand side and the standardized residuals not following a normal distribution apparently. This could compromise the model.

(Intercept)	cases2	Total.tests	cases	Rich
228.189069107	3.568295415	-56.545429397	-6.487649206	347.499609576
GDP.2018	Crime.Index	Crime	Median.Age	sex65plus
-31.319679579	13.299603688	341.719756698	1.292411876	175.796813259
Female.Lung	Hospital.Bed	Density	Urban.Pop	cases2:cases
-1.058921839	77.499103030	-0.204705383	-4.488417938	0.029668445
cases2:Total.tests	cases2:Rich	Total.tests:Rich	cases:Rich	Total.tests:GDP.2018
-0.436804874	0.407671504	-13.138084067	1.856821168	3.743321303
Rich:Crime.Index	Total.tests:Crime.Index	cases2:Crime.Index	Rich:Crime	cases:Crime
-2.395103785	-0.367275030	0.029253533	90.293177544	2.427982293
Total.tests:Crime	GDP.2018:Crime	cases:Median.Age	Total.tests:Median.Age	cases2:Median.Age
-8.610558233	-5.751998526	0.240818618	-0.846227016	0.014575200
Crime:Female.Lung	cases:Female.Lung	cases:Hospital.Bed	Crime:Hospital.Bed	Crime:sex65plus
-1.111661479	0.021529530	-0.331025462	24.091374697	138.332795190
Hospital.Bed:Density	Crime.Index:Hospital.Bed	Crime.Index:sex65plus	Total.tests:Urban.Pop	cases2:Hospital.Bed
-0.099962565	-1.074692716	-4.698680642	0.157080757	0.066798801
cases:Urban.Pop	Density:Urban.Pop	Crime.Index:Density	cases:GDP.2018	Median.Age:Density
-0.024931477	-0.003881585	-0.007664087	-0.175892530	0.035208354
Crime.Index:Crime	Rich:Urban.Pop	Rich:Hospital.Bed	Hospital.Bed:Urban.Pop	Female.Lung:Density
-5.563696569	1.360489623	10.174366765	-0.569077758	-0.002497032
Rich:Median.Age	Crime.Index:Female.Lung	Crime:Density	Median.Age:Urban.Pop	Crime:Urban.Pop
-8.683176828	0.062881024	0.154824422	0.088215908	-1.827968052
Crime.Index:Urban.Pop	Crime.Index:Median.Age	Female.Lung:Urban.Pop	Rich:Density	sex65plus:Urban.Pop
0.062456901	-0.153182321	-0.018124455	-0.180590772	-1.812260460
GDP.2018:Urban.Pop	cases2:sex65plus			
0.130764654	-0.229488353			

Figure 1: Model coefficients