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 the 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

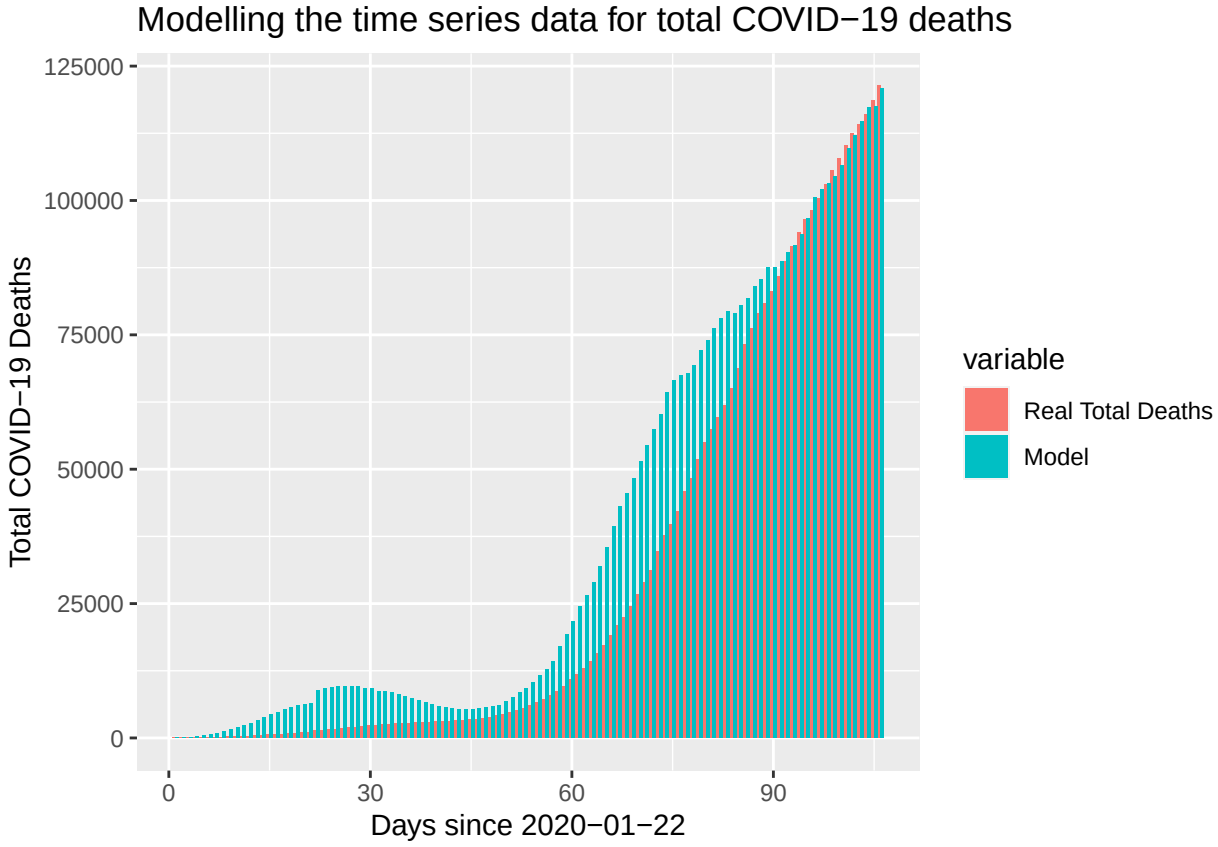
All these datapoints were joined by country and by date. 2 main dataframes were created. One contained data solely on cases, recoveries, testing and deaths. The other was a combination of this and the predictor data- this was smaller however on account of not all countries having predictor data on them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

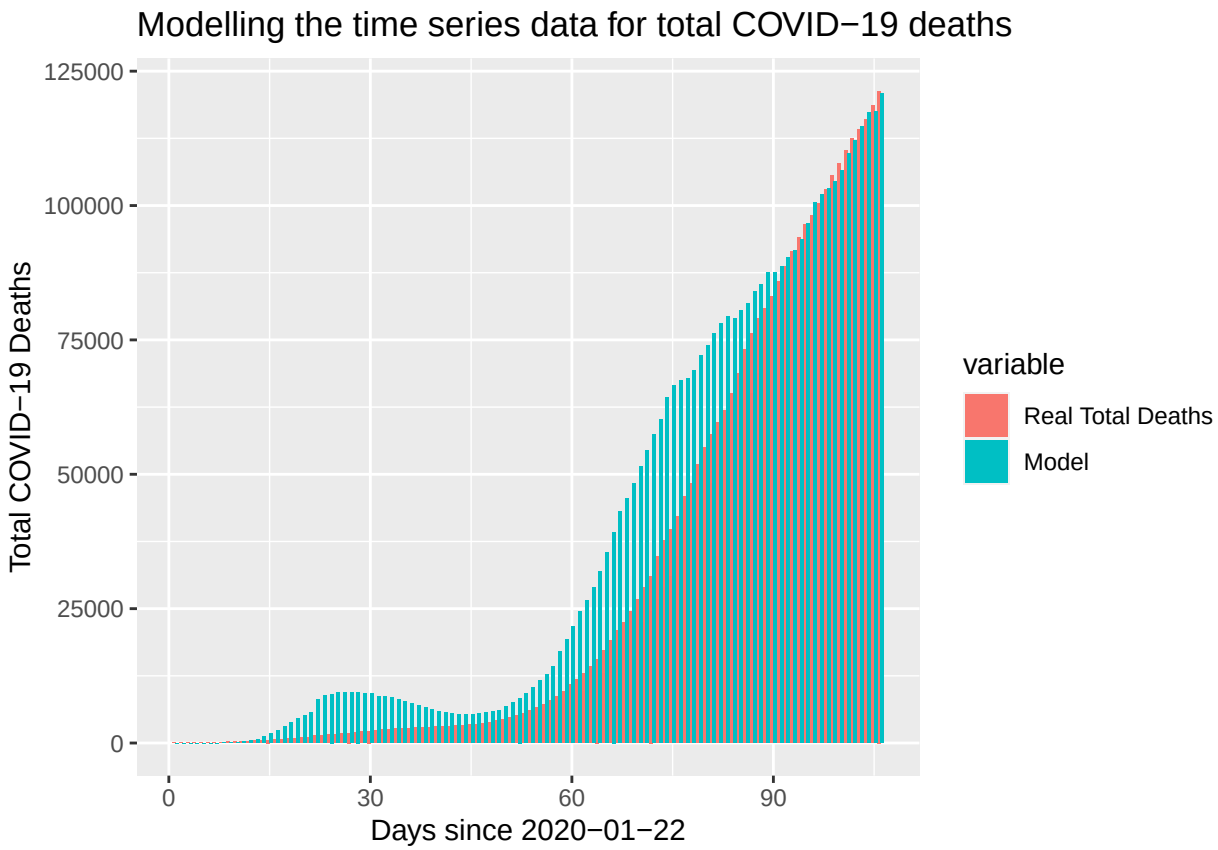
It was theorized that the time series data of new COVID-19 deaths worldwide could be modelled as fitting a poisson or binomial distribution. This would make sense as each case could be considered a Bernoulli random variable with some probability of surviving or dying. This probability was though to be 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. This is unsuprising as the binomial tends to the poisson in the limit of large n given many samples and provided that they are approximately independent The best model found was defined as $TotalDeaths = ActiveCases * 0.17 * P(X \leq Day)$ where $X \sim Pois(16.68705)$ and Day is days since 2020-01-22. The plot of this model is shown below.

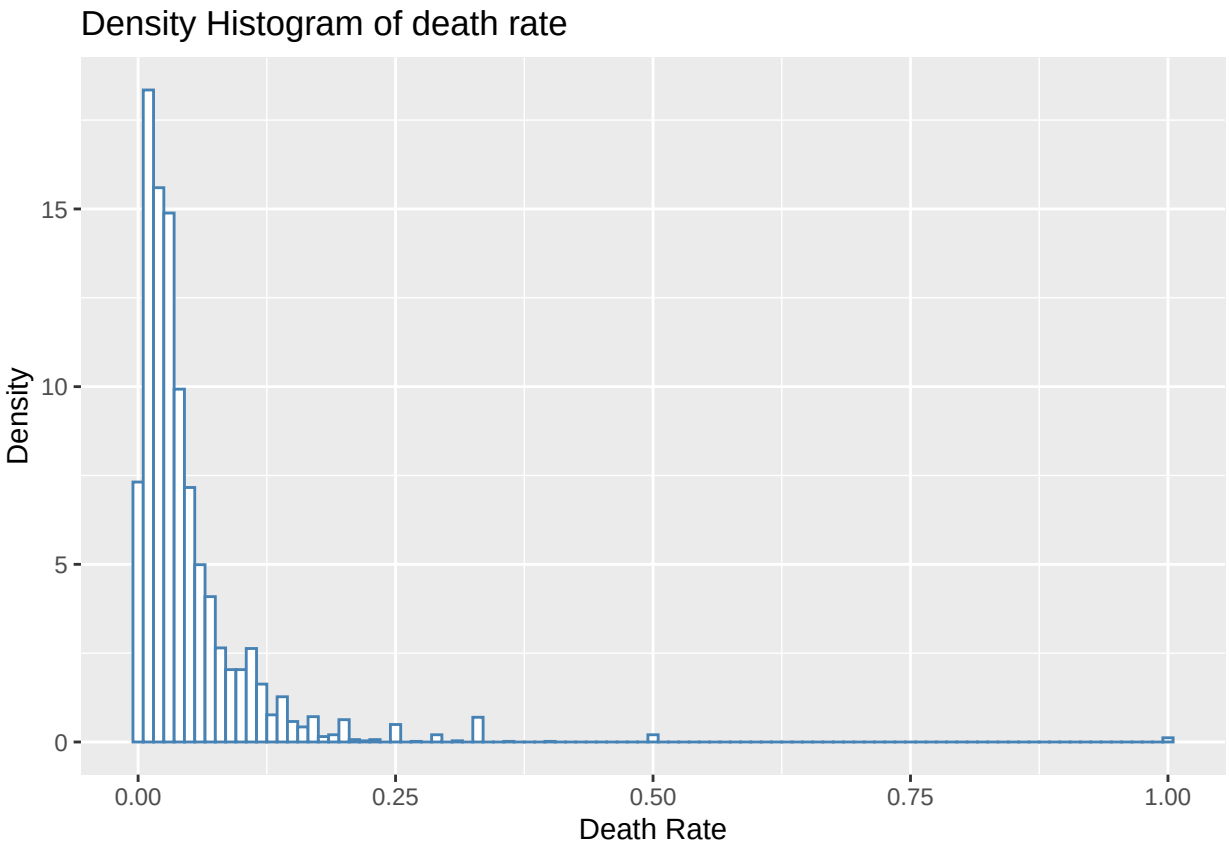


The same conclusions as the binomial model can be drawn for the poisson model. It is overall poor.

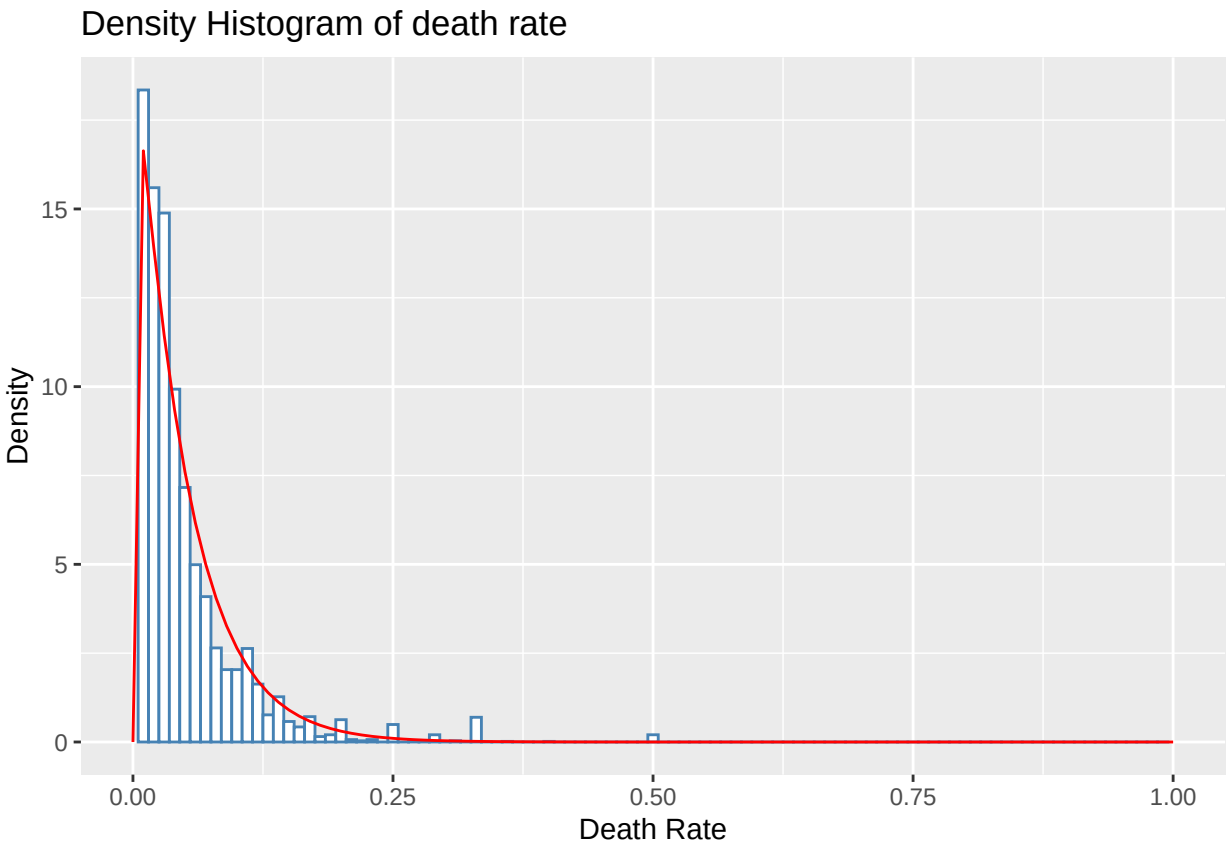
It was concluded therefore that modelling cumulative deaths via a probability distribution is unlikely to succeeed. Potentially multivariable regression could yield better models and reveal more about what affects death tolls.

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.

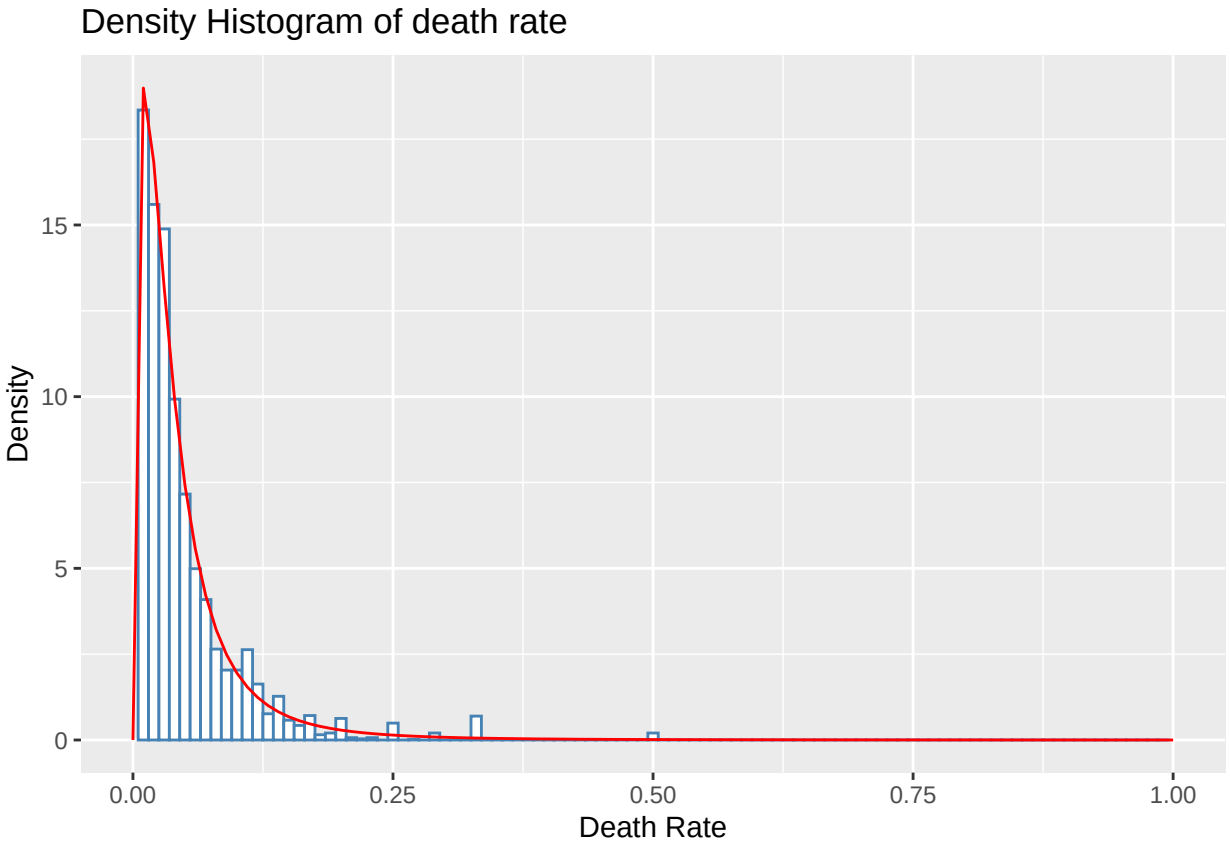


This suggested that a gamma distribution would be a good fit. A gamma probability density function model was developed $f_X(x) = dgamma(shape = 1.06011822, rate = 21.99330112)$. This was overlayed onto the histogram.



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. This is unsuprising as the blips seen (especially between 0.25 and 0.5) imply a standard distribution like the gamma distribution will not be valid. Potentially a heavy tail distribution could be used to alleviate this.

As such, a Burr Heavy tailed distribution was fitted with $f_X(x) = dburr(shape1 = 2.010724, shape2 = 1.354063, rate = 16.967252)$.



Running the Two-sample Kolmogorov-Smirnov test resulted in a p-value of 0.07323082. This meant we failed to reject the null hypothesis at the 0.05 level of significancethat the sample came from the relevant distribution. Therefore it was concluded that deathrate counts can indeed be approximated by the Burr distribution.

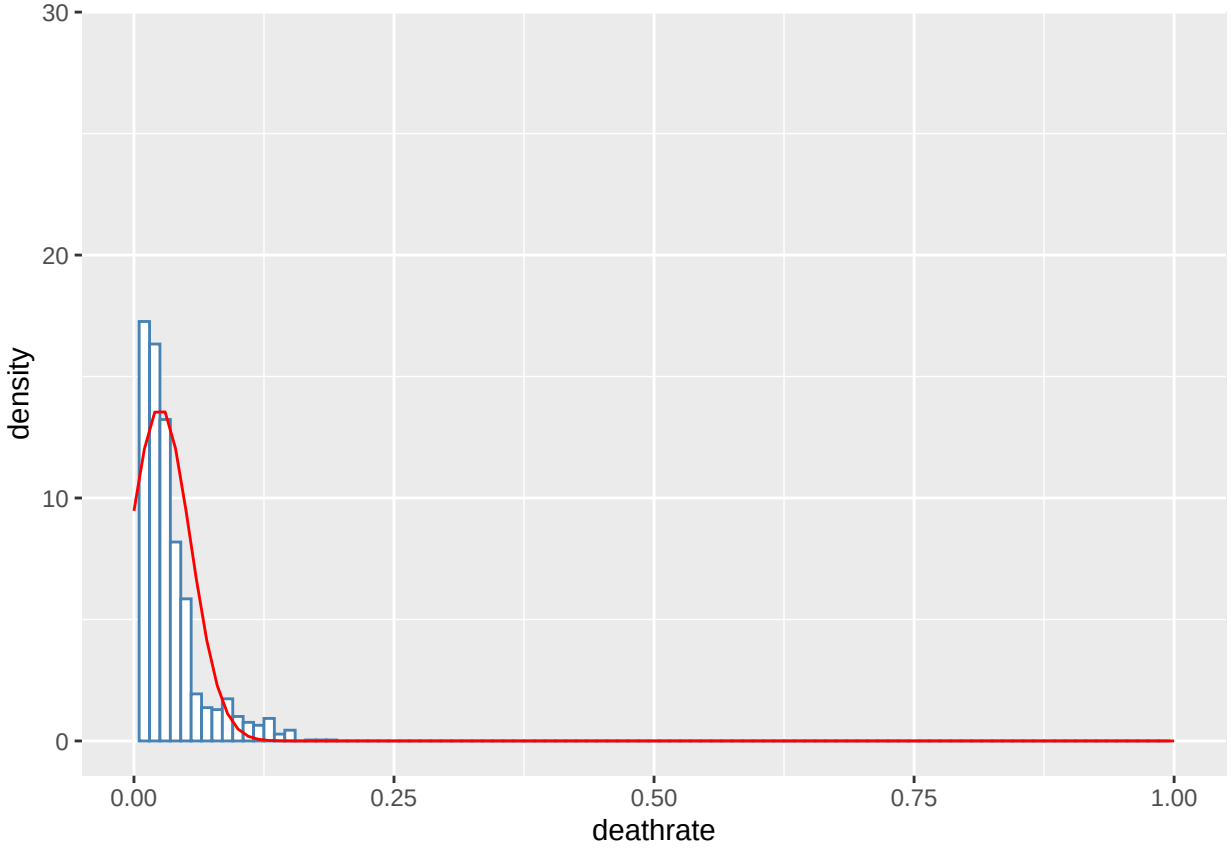
Topic 2: Exploring Correlations and Independences within COVID-19 Data

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
```

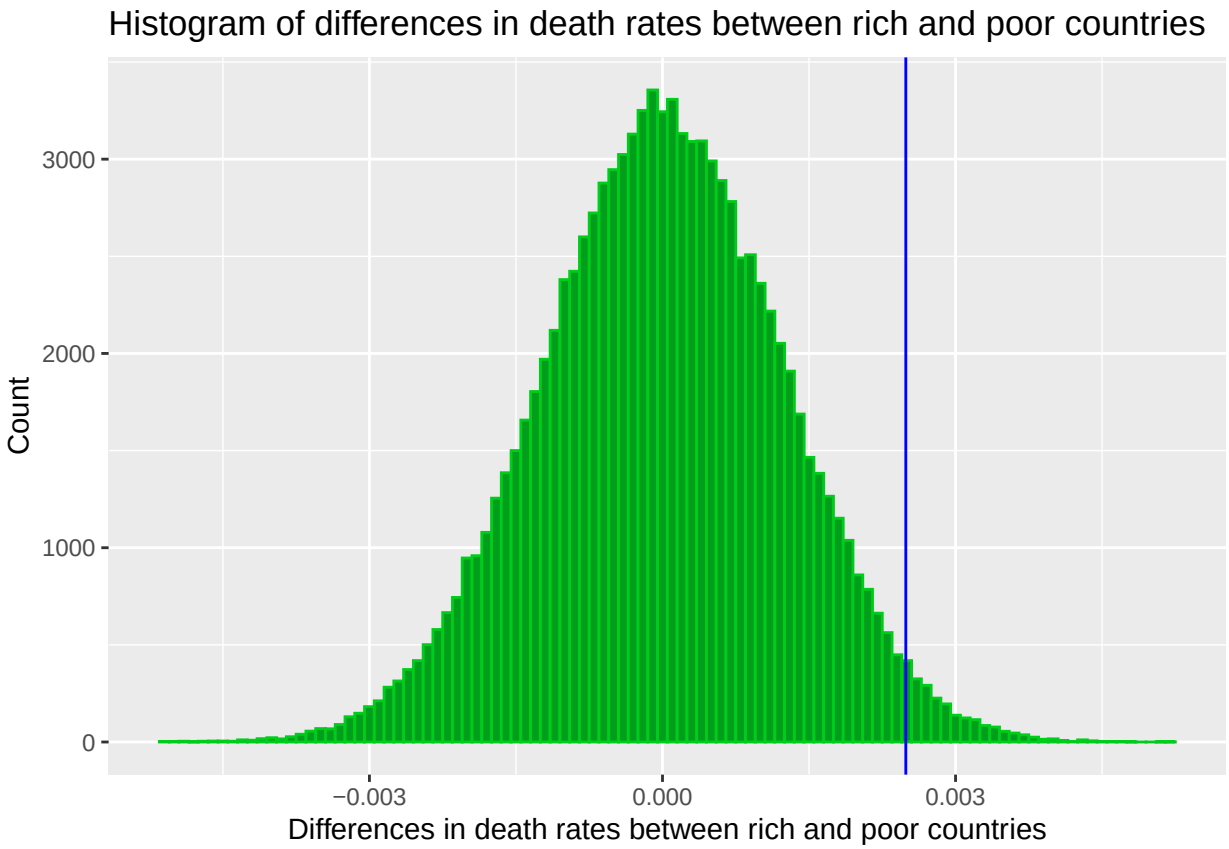
The T test returned a p-value 0.07699, seemingly leading to the acceptance of the null hypothesis of no mean difference of death rates between rich and poor countries. However the t-test was discovered to be fundementally flawed. T-tests assume normality. COVID-19 death-rates do not follow a normal distribution.




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

This is clearly shown by the plot of the overlayed normal distribution 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 a great extent. The significant difference could be promising for modelling in topic 3 later on.

Part 2: Are crime and high death rates independent?

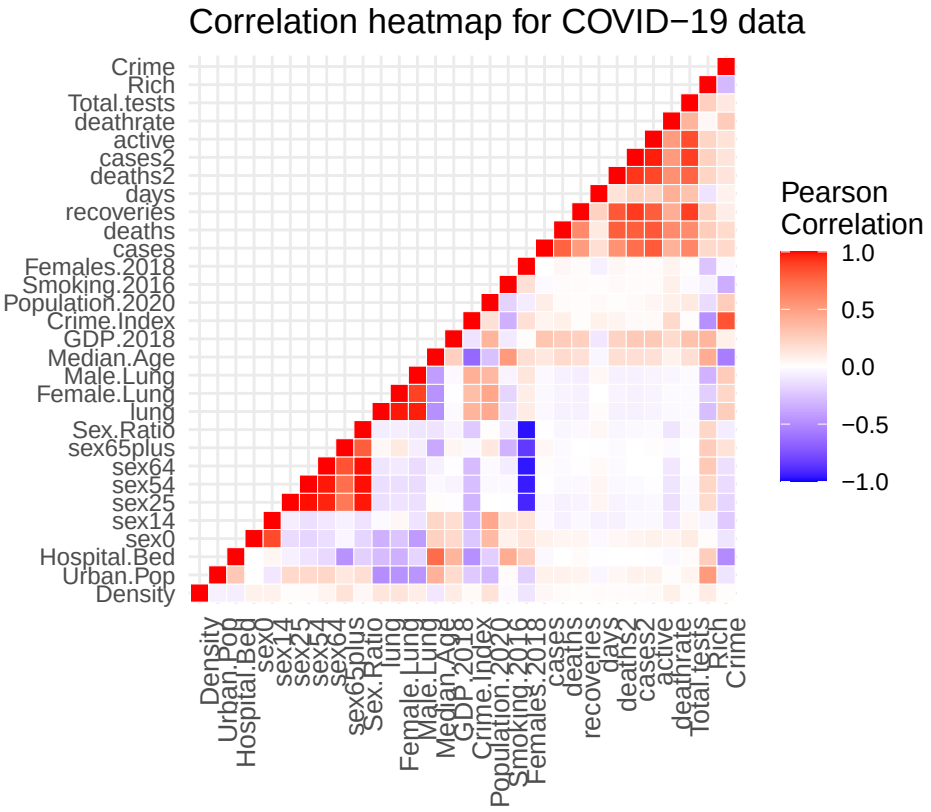
A new logical variable *hdr* was defined: being 1 for countries in the upper half of death rates, and 0 for countries in the lower half. A chi-squared test on this and *Crime* was used to investigate independence. Below is the contingency table for these variables.

	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, implying omitted variable bias is at play. 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.



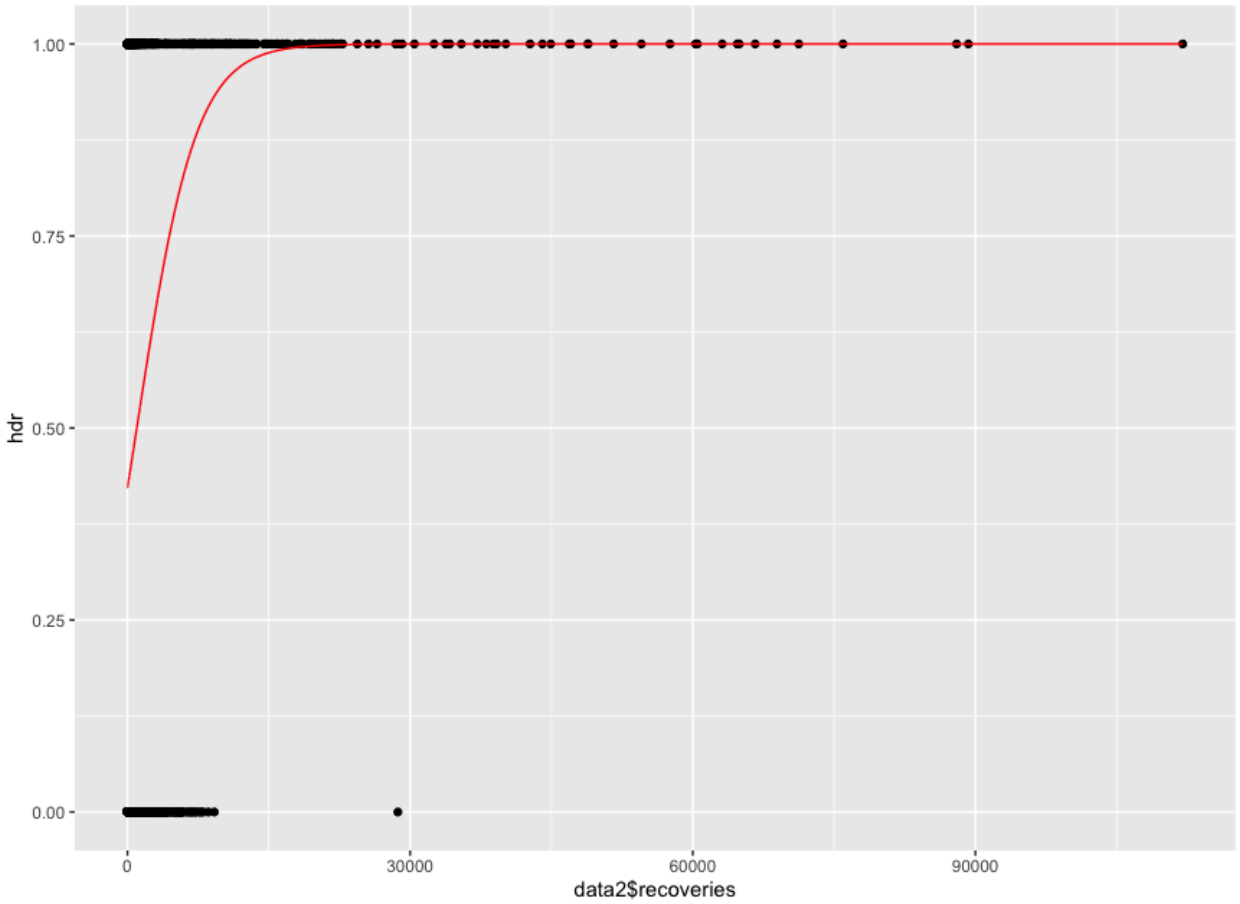
Death rate is actually not strongly correlated much at all with the other predictive variables. Perhaps this indicates that deathrate is relatively constant and hard to model. This indicates challenges for topic 3. Interestingly, we dont see high degrees of similarity overall.

Topic 3: Modelling deathrate and deathtoll

Modelling deathrate via logistic regression

Part 1: univariate logistic regression

The first attempt at doing this was with a univariate logistic model using recovery data on *hdr*, to try to model high death rates.



As can be clearly seen, this is not a great model, as reflected in the low McFadden Pseudo R^2 value of 0.07780584. A different approach will likely be needed to explain more of the variance in deathrate, hence multiple logistic regression.

Part 2: multiple logistic regression

7 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 on testing data deathrate values.

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.

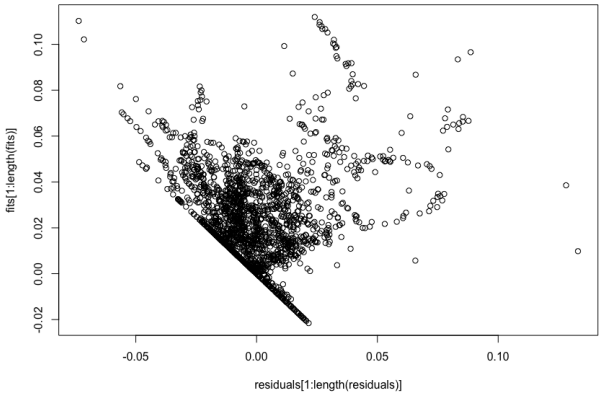
In order to deal with multicollinearity, the following were dropped one by one from the numeric data in accordance with their variance inflation factors: *lung*, *Sex.Ratio*, *cases*2, *sex*54, *GDP*.2018, *active*, *sex*14. This led to only these variables being used in modelling: *Density*, *Urban.Pop*, *Hospital.Bed*, *sex*0, *sex*25, *sex*64, *sex*65*plus*, *Female.Lung*, *Median.Age*, *Crime.Index*, *Population*.2020, *Smoking*.2016, *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 ridge model of deathrate onto all the available variables. The seventh model was a neural net model with threshold 0.1 of deathrate onto all the available variables.

Comparison on the testing data revealed the fifth model (Linear Lasso) to be best, with the lowest SSE

(5.116951e-06). Overall this model had a relatively poor R^2 value of 0.4917774. Its adjusted R^2 was very close however at 0.4703785, indicating that the model did not overfit and strongly explained some of the variance. It should be noted that in diagnostics that the residuals vs fitted values plot did not look random for low fitted values, and indicated heteroskedacity. Moreover, we received p-value <2.2e-16 in the shapiro wilk test for normality meaning that we rejected the null hypothesis of no significant difference with the normal distribution, and concluded that under the current evidence, the residuals do not follow a normal distribution. These violations of the standard linear regression assumptions could have compromised the model.

```
(Intercept)      2.369798e-01
Density          5.926432e-07
Urban.Pop        2.692032e-05
Hospital.Bed    -9.953676e-04
sex0            -1.681061e-01
sex25           -2.441041e-02
sex64           -2.091697e-02
sex65plus       -1.005761e-02
Female.Lung      4.859137e-04
Male.Lung       -4.320670e-04
Median.Age       1.721363e-03
Crime.Index      3.679629e-04
Population.2020  5.837790e-03
Smoking.2016     4.194620e-04
Females.2018    -3.342664e-03
cases            5.055285e-04
recoveries       1.823849e-04
days             9.362554e-04
Total.tests     -7.092612e-03
Rich             3.724101e-03
Crime            5.392028e-03
```




```
[1] "Shapiro Wilk for Normality of residuals"

        Shapiro-Wilk normality test

data:  residuals[1:length(residuals)]
W = 0.92833, p-value < 2.2e-16
```

Modelling new deaths via linear 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 .

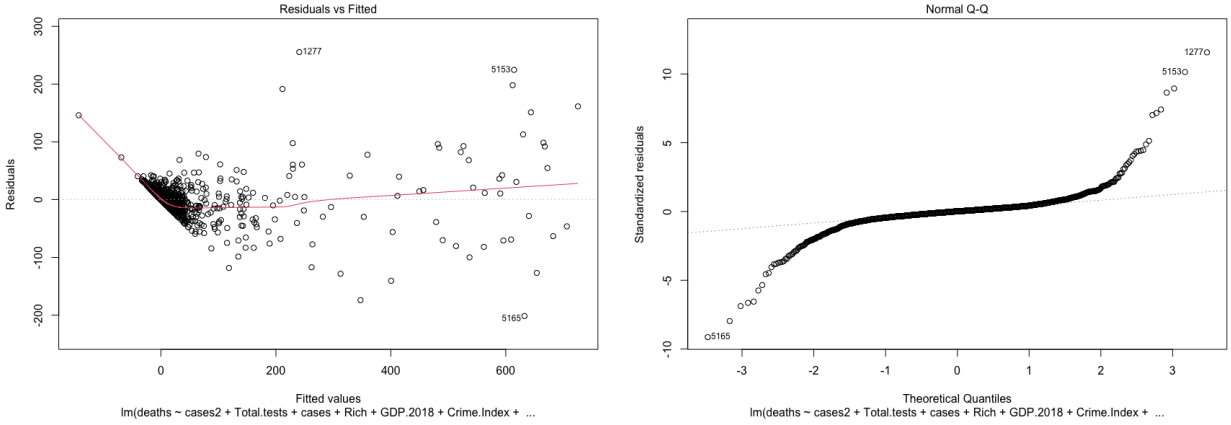
To reduce spread of outliers: *GDP*.2018, *Population*.2020, *Total.tests* were logged. *recoveries*, *cases*2 and *cases* were squarerooted however since they had 0 values in the data.

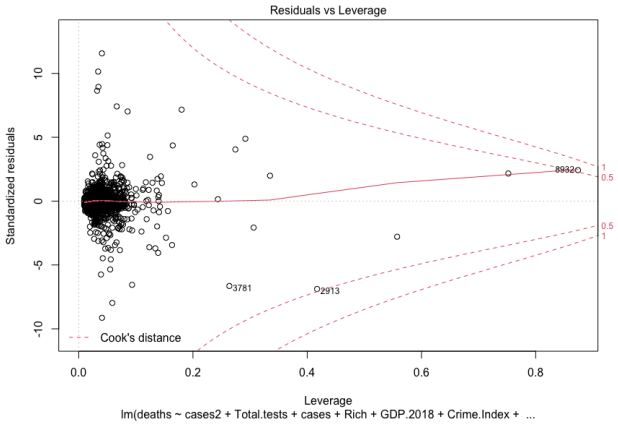
In order to deal with multicollinearity, the following were dropped one by one from the numeric data in accordance with their variance inflation factors: *lung*, *sex*54, *Sex.Ratio*, *sex*64, *Females*.2018, *Population*.2020, *days*, *recoveries*, *Male.Lung*, *sex*0. This led to only these variables 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 new deaths 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 linear regression model of new deaths 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 new deaths onto all the available variables. The eighth model was a ridge model of new deaths 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 SSEs (9.527257e+00). Looking at the fifth model we had a strong R^2 of 0.9140345 and similarly strong adjusted R^2 of 0.8996393, 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 (although it was well spaced as fitted values increased) and the standardized residuals not following a normal distribution. Moreover, an non constant variance test indicated heteroscedasticity, with p value <2.22e-16 and point 8932 was poor on the cook's distance plot. These violations of linear regression assumptions could compromise the model.

(Intercept)	cases2	Total.tests	cases	Rich	GDP.2018
-2.207879e+02	5.508652e-01	-2.352626e+00	-6.068852e+00	3.193559e+01	-7.308113e+00
Crime.Index	Crime	Median.Age	sex65plus	Female.Lung	Hospital.Bed
1.689037e+01	-4.842785e+02	4.965834e-02	-8.227278e+00	-1.176383e+00	6.276907e+01
Density	Urban.Pop	Smoking.2016	cases2:cases	cases2:Total.tests	cases:Rich
1.006348e+00	1.313938e+00	-2.426918e+00	2.479013e-02	-4.377941e-01	1.494914e+00
cases2:Rich	Total.tests:Rich	Total.tests:GDP.2018	cases:Crime.Index	Rich:Crime.Index	Rich:GDP.2018
3.972033e-01	-1.507378e+01	3.078009e+00	5.650597e-02	-1.412105e+00	1.834870e+01
Total.tests:Crime.Index	cases2:Crime.Index	cases:Crime	Total.tests:Crime	cases:Median.Age	Total.tests:Median.Age
-6.966193e-01	3.361440e-02	1.141026e+00	-4.427179e+00	2.596331e-01	-1.434116e+00
Crime:Female.Lung	cases:Hospital.Bed	Total.tests:Hospital.Bed	Crime:Hospital.Bed	Crime.Index:Hospital.Bed	sex65plus:Female.Lung
-9.631641e-01	-3.735456e-01	1.516257e+00	3.484454e+01	-1.519029e+00	5.201948e+00
Crime:sex65plus	Crime.Index:sex65plus	Rich:Median.Age	Median.Age:Female.Lung	Total.tests:Female.Lung	Female.Lung:Hospital.Bed
4.049440e+02	-1.132741e+01	-7.451036e+00	-1.198985e-01	-4.946955e-02	6.163335e-01
Hospital.Bed:Density	Total.tests:Urban.Pop	Crime:Density	Median.Age:Density	Median.Age:Urban.Pop	Hospital.Bed:Urban.Pop
-4.762897e-02	1.323364e-01	2.381504e-01	2.118370e-02	1.294173e-01	-4.864906e-01
sex65plus:Smoking.2016	cases2:Smoking.2016	Total.tests:Smoking.2016	cases:Urban.Pop	cases2:Median.Age	cases2:Female.Lung
5.759131e+00	2.316145e-02	-2.683479e-01	-2.442709e-02	3.623217e-02	1.203554e-02
Crime.Index:Density	cases2:Urban.Pop	GDP.2018:Density	Rich:Smoking.2016	GDP.2018:Urban.Pop	GDP.2018:Crime
-7.595600e-03	5.909522e-03	-4.719398e-02	-1.930650e+00	-2.284256e-01	1.036386e+01
Rich:Female.Lung	Crime.Index:Urban.Pop	cases2:Density	cases:Density	Rich:Density	Crime.Index:Crime
-1.133392e+00	5.423755e-02	-7.261213e-04	2.427362e-03	-1.633117e-01	-2.883045e+00
Crime:Urban.Pop	cases2:Crime	cases:GDP.2018	cases2:GDP.2018	Urban.Pop:Smoking.2016	Total.tests:sex65plus
-1.143966e+00	-1.555251e-01	-2.338681e-01	4.940298e-02	-1.748477e-02	-5.288368e+00






```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 10318.71, Df = 1, p = < 2.22e-16
```

Evaluation and Conclusion

Overall comment

This project hopefully shared some insight as to the distribution of COVID-19 deathrates and deathtolls, and led to some ideas of the root factors at play with the constructed models. Although deathrates frequency was succesfully modelled by a Burr distribution, probability distributions are likely not enough to successfully model and predict deathtolls. As such regression models are likely the way to go for better modelling. This project concluded with a not so successful attempt at this with a logistic regression for deathrates, but with also a good new deaths model being developed. Indeed this new deaths model was the best overall model developed.

Future work

The death models were plagued with diagnostic errors, implying that assumptions of standard linear regressions were being violated. It is possible to adjust the coefficents of one's model using the *sandwich* library, such that the assumptions of standard linear regression are met. This would clearly be something wise to do for greater application of any developed model. It would also be good to do some n-fold validation to better assess predictive performance overall.

Moreover, it could be interesting to look at and evaluate each coefficient in the model, and explain this from a biological/ epidemiological perspective. This could potentially help policy-makers better deal with the current outbreak and future ones.

Furthermore, the current model could use more variables! For instance a new dummy variable could be used to indicate whether quarantine restrictions are in place. This example in particular could help determine whether restrictions are effective. Indeed this and more variables would likely let the model explain more of the variance.

Finally, building a more in depth interlinked model system could be advantageous. One could perceive modelling susceptible individuals, and from that infectious individuals and then from that recovered individuals. This 3-way system would cover the main principles of how a virus spreads, and could potentially lead to dynamic simulations and more accurate models.