STEM Fellowship Big Data Challenge 2020

Correlation between days since first case of nCOVID-19 and case fatality rate in Southern European Countries

Stanley Z. Hua[1]

[1]University of Toronto

June 1, 2020

Abstract

In the midst of a pandemic, there is great demand for resources and information. Yet, as focus is placed on "flattening the curve", the question needs to be asked: 'How well are we taking care of those infected?' A possible measure for this is the ***case fatality rate***, which is defined as the proportion of infected who have died due to disease.[1] Through this paper, we hope to shed light on the question raised by uncovering potential relationships in up-to-date nCOVID-19 data. To address this, exploratory data analysis and statistical analysis on data retrieved from the R package COVID19[2] and The World Data Bank[3] were done using R.

Results imply a moderately strong correlation between the number of days since the first case and case fatality rate for respective countries in Southern Europe (with 1000 cases or above). We believe that this finding has potential in supplementing future studies that hope to analyze actions the governments have taken and will take to improve public health systems and policies in treating those infected. As more information is available every day, it is recommended that future studies, hoping to study case (or infection) fatality rates between countries, find and implement reliable estimates for the true number of nCOVID-19 cases in each country.

Keywords
covid-19, case fatality, southern europe

1 Introduction

As of late May 2020, news depict higher orders of government and even lower differing in their approaches to addressing the nCOVID-19 pandemic. Now, we are seeing a greater need for information and analysis to equip our leaders with the knowledge necessary to make informed decisions to improve continued response to the pandemic. Widespread media coverage has been successful in elucidating the harmful effects of the novel coronavirus 2019, and promoting measures for protecting oneself and lowering transmission. At the same time, hospitals and public health systems are being overwhelmed. Lack of testing facilities and housing facilities for the infected are major and current problems for many countries.

Case fatality rate is defined as the proportion of infected who have died due to disease.[1]

Knowing this, how do we predict if a country is doing relatively better or worse in handling its infected? This is a multi-faceted question that we hoped to contribute to by analyzing data currently available using R. In the beginning, preliminary exploratory data analysis was conducted. After filtering, data analysis was done on chosen variables including *days since first case, case fatality rate, time, proportion who recovered* and *population density*. Methods of data analysis used were simple linear regression via least squares method, and Kendall's tau correlation. These led to the discovery of a positive linear relationship between *case fatality rate* and *days since the first case* for countries in Southern Europe.

This is an interesting trend to observe given that two countries in this category (i.e. Italy and Spain) are currently known to have near the highest number of deaths worldwide. This implies that *case fatality rate* may increase linearly with time for these countries in Southern Europe. However, it is with hope that preferably

improvements to public health systems and policies will prove to disrupt this linear trend.

2 Materials & Methods

The following section outlines the actions and methods of this research. The methodology was done using R. Packages used include **COVID19, tidyverse, ggthemes** and **gridExtra**.(*dplyr and ggplot are included in the tidyverse package*)

Preparation

First, nCOVID-19 data was gathered from the COVID19 package[2], and 2018 population density data was gathered from The World Data Bank[3]. It is important to note that population density is from 2018, as this is the latest uniform data available. These were made available in RStudio. nCOVID-19 and 2018 population density data were filtered for countries containing valid data, and for the country ids in the "countries of interest" (a user-defined variable). A new data frame was created merging the two previous data frames (data frames must be parallel according to country id). Columns included country id, region, population, *proportion of the population infected*, *population density*, latest total cases, *latest case fatality rate*, *latest proportion who recovered*, and *days since first case.Case fatality rate* and *proportion who recovered* were calculated given by

$$\frac{death\ (or\ recovered)\ count}{total\ confirmed} \tag{1}$$

for any one time. Additionally, *proportion of the population infected* and days since first case were computed separately, where *days since the first case* was created (as a parallel list) by subtracting the latest day by the first day when confirmed were greater than 0 for each country, and a similar process for *proportion of the population infected*.

In preparation for testing, lists of regions worldwide were created containing their respective countries (in 3 letter ISO country code format), including North America, Europe, Africa, Asia and their sub-regions.

Exploratory Data Analysis

Preliminary exploratory data analysis was conducted by creating time series plots of dependent variables *case fatality rate and proportion who recovered* versus *time* for countries of interest.

To lower variability, case fatality rate and proportion of those who recovered were computed with estimation of the true number of confirmed cases (supposedly including asymptomatic cases). This was done by sampling the possible proportion of asymptomatic cases using mean and standard deviation from the Diamond Princess cruise ship (mean: 17.9 and sd: 1.2).[4] Vertical dotted lines were placed when the first case began and when there was the nth case, which was made 1000. Each individually produced plot was arranged in a vertical grid plot, and if numbering more than five, plots were saved in a pdf. Plots were compared visually to see if the vertical line at nth case provided a decent minimum number of cases if the region between the vertical lines captures high variability (or amount of fluctuations).

Linear Regression and Statistical Tests

We define the possible independent variable as *population density* and *proportion of the population infected*, and the dependent variables as *case fatality* and *proportion who recovered*, while *days since first case* could be either.

Each possible combination x-y of independent and dependent variable is plotted for region (or sub-region or countries) of interest. Regions (or sub-region or countries) of interest with seemingly linear relationships are further analyzed using the built-in linear regression function and the built-in correlation function with method parameter assigned to Kendall's tau. The summary of the linear regression function is printed, displaying results from a t-test and the p-value. If significant ($p < 0.05$), relationship between variables for specific region (or sub-region or countries) and p-value is noted.

These are plotted individually with linear regression line. The p-values are compared, and the relationship (for specific region, etc.) with the greatest significance is kept.

Data analysis was done on 05/31/2020, where data analysis includes this date as the current maximum date from the COVID19 package.

3 Results

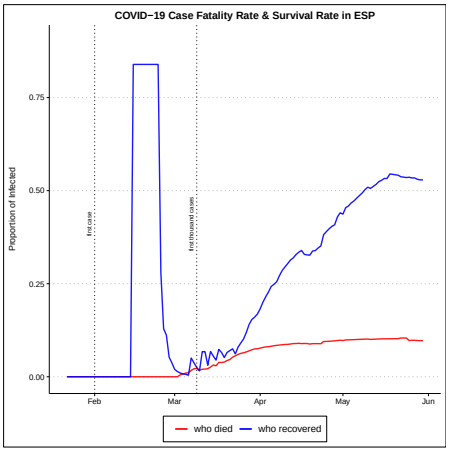

Figure 1: Exploratory Data Analysis of Spain

A time series plot for *case fatality rate* (proportion who died) and *proportion who recovered* is shown above for one of the Southern European countries. A great spike in *proportion who recovered* is observed due to low numbers of confirmed.

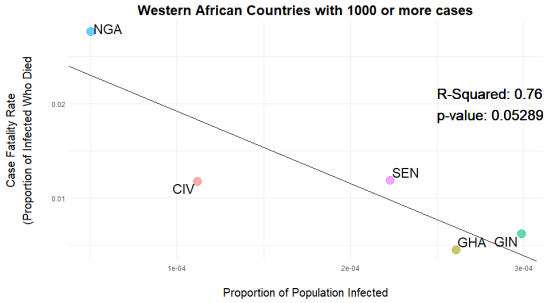

Figure 2: Linear Regression of *Case Fatality Rate* vs *Proportion of Population Infected* in Western African Countries

Analysis of the relationship between *case fatality rate* and *proportion infected* in Western African countries is not statistically significant (p-value > 0.05). However, kendell's tau correlation coefficient (-0.6) shows moderately strong negative correlation .

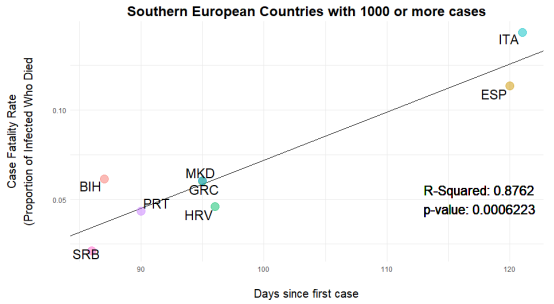

Figure 3: Linear Regression of *Case Fatality Rate* vs *Days since first case* in Southern European Countries

The linear regression graph above exhibits the main finding of the study, displaying a very statistically significant linear trend (p-value < 0.001) between *case fatality rate* and *days since the first confirmed nCOVID-19 case*. Kendall's tau correlation coefficient (0.55) shows moderately strong positive correlation.

4 Discussion

Time series plots of case fatality rate and proportion who recovered were created with the goal of observing the variability and randomness in the dependent variables over time. This was done given the differing progression and number of nCOVID-19 cases per country.

Method for estimating true number of cases was done by randomly sampling a normal distribution of mean: 17.6 and sd: 1.2, which is based on the results from analysis of the Diamond Princess cruise ship by Mizumoto et al. 2020. It is clear that there is a great possibility for error in this model by assuming every country is rightfully sampled by the nationalities of the Diamond Princess passengers. However, we believe the accuracy of the measure is as important in the preliminary exploration of the data. It served its purpose, which was to lower the height of spikes (great change in the variables over a short time) when cases are generally low. Nevertheless, preliminary exploratory data analysis showed that changes in case fatality rate over time is more stable once cases are greater than 1000 (Refer to Figure 1.). This value, though easily changeable, served as the minimum cases acceptable for countries to be included in later stages of the methodology.

Following this, tests of linear regression, and Kendall's tau were done on the aforementioned variable combinations to find relationships between variables for different sub-regions of the world. By default, simple linear regression is done via least squares method in R, and this

is used. Though linear regression is powerful enough to be understood by a trend line, its p-value is needed to test the probability that the result is observed, or in other words, its statistical significance. Additionally, the R-Squared value is included to express how well the data points are explained by the linear regression model. On the other hand, Kendall's tau correlation coefficient was chosen rather than spearman's rho and pearson's r correlation since 1) variables used are not normally distributed (Refer to Figure 2 and 3.), and 2) Kendall's tau is less sensitive to errors and outliers.

Figure 2 is an example of one of the tests done. The p-value observed was not statistically significant (less than 0.05), albeit the Kendall's tau correlation coefficient implies a strong positive correlation. Since its p-value is near to 0.05, it was kept as an example of other tests done and held in regard as a minor finding. For the Western African countries with more than the minimum number of cases, a trend is observed where *case fatality rate* decreases for countries with greater *proportions of the population who are infected*. What this implies for the field of epidemiology and other related fields is not part of the scope of this paper.

After the methodology was run for all variables, the most statistically significant correlation was found between *case fatality rate* and *days since the first case* for countries in Southern Europe with 1000 cases or more. (Refer to Figure 3.) The implications of this main finding is that the linear model predicts that if *days since the first case* increases (as it will over time), then *case fatality rate* will increase linearly according to the equation (where y is case fatality rate and x is days since first case):

$$y = 0.002685x - 0.196542 \tag{2}$$

This is assuming conditions that influence the variables in this model stagnate, and that countries with vastly different systems for handling the pandemic can be compared as such. The researcher acknowledges these assumptions serve as potential sources of error, especially for future studies.

In the end, it is with hope that this trend proves to show the need to improve current public health systems and measures to support those infected in countries not only limited to Southern Europe.

5 Conclusions

After analysis of open-source COVID19 data, a strong positive correlation between case fatality rate and days since the first case was discovered for countries in Southern Europe. This implies that case fatality rate may continue to increase linearly over time for the countries in Southern Europe.

This is a very urgent and concerning matter considering the highly populated countries in Southern Europe. These include Italy and Spain, both of which have high death tolls as of May 2020. This is an undesirable projection, but with improvements to care for those infected such as vaccines, this trend could change.

We highly recommend that future researchers study this trend over time with particular interest in the effect of introducing vaccines and new public health policies. Additionally, it is recommended that all variables mentioned in this research are also studied in the context of these widespread changes.

Acknowledgements

Special thanks to Johann Espino, Zachary Hizon, Nafis Mohammad, Angitha Mriduraj, and Angela Niu for reviewing.

References

[1] Jason Oke and Carl Heneghan. Global covid-19 case fatality rates, May 2020.

[2] Emanuele Guidotti and David Ardia. Covid-19 data hub. Working Paper, 2020.

[3] World Development Indicators, Statistical Capacity Indicators, Education Statistics All Indicators, Gender Statistics, Health Nutrition, and Population Statistics. Population density (people per sq. km of land area), 2018.

[4] Kenji Mizumoto, Katsushi Kagaya, Alexander Zarebski, and Gerardo Chowell. Estimating the asymptomatic proportion of coronavirus disease 2019 (covid-19) cases on board the diamond princess cruise ship, yokohama, japan, 2020. *Eurosurveillance*, 25(10), 2020.