COVID-19 early days

A Multivariate Linear Regression and Decision Tree Analysis Approach to understanding the early progression of the novel corona virus based on country demographics

Lindo Khoza
willkhoza@gmail.com

April 12, 2020

Abstract

This analysis is based on the Novel Coronavirus (COVID-19) Cases from a South African perspective repository [5]. The intention is to understand the influence of demographic variables on the early progression of the COVID-19 within different countries. As such, models used are chosen for their simplicity and transparency.

This report can only serve as a snapshot in time as the data process is live and unfolding. The diagrams and fitted models will change, this is captured in the repository. Alternatively, the annexure that accompanies the report is dynamically generated, and the plots and information reflected there is updated accordingly.

The repository sources data from the 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE repository [4].

The analysis shows that older populations have a tendency travel more, resulting in then having more exposure to COVID-19. This coupled with other pecks of being old, appears to have been a statistcally significant driver of the COVID-19 pandemic affecting older people with the extreme intensity which it has.

Background

COVID-19 is an infectious disease caused by a variant of the coronavirus family. The disease is associated with respiratory illnesses.

The primary action taken by governments to curb the spread of the disease is to encourage people to wash hands frequently, to social distance, and to restrict overall physical interaction with other people.

0.1 Variable Description

0.1.1 Response Variables

The multivariate linear regression uses demographic predictor variables to model the following response variables:

1. t_1: The number of days it takes for a country to report a confirmed infection following the earliest report date in china.
2. t_2: The number of days it takes for a country to report it's first known death as a result of the virus following it's first confirmed infection.

The earliest confirmed case of COVID-19 is unofficially reported to be 17 November 2019 [1].

0.1.2 Demographic Predictors

The following demographic variables are used to fit the model:

1. Population: country population
2. Annual Change: annual change of population in proportions
3. Annual Change Absolute: annual change of population in absolute terms
4. Density: P/km^2
5. Land Area: km^2
6. Migrants: Net migration in the country
7. Fertility Rate
8. Median Age
9. Urban Population: Proportion of population living in urban areas
10. World Share
11. Migration index = $sgn(\text{Migrants}) \times ln(abs(\text{Migrants}))$
12. lnPopulation = $ln(\text{Population})$

The demographic data is with respect to the year 2020, as reported by worldometer [3].

0.1.3 Additional Predictors

An additional predictor was retrieved from wikipedia [2] due to the association of the virus with travelling:

1. Passengers: Annual number of airline passengers per country
2. lnPassengers = ln(Passengers)

The data has been processed and stored in the repository.

0.2 Summary Statistics

Overview

Data from approximately 169 countries was used to fit the linear model. Atleast 54 of these countries have not recorded a death as a result of COVID-19. Also, atleast 11 countries have missing demographic data, especially countries with very small population sizes such the Holy See.

Data Source Design

In the JHU repository [4], data from 4 cruise ships was also observed from the original JHU repository, as such these data points will systematically not have demographic information.

Furthermore, the data capturing period is from 22 January 2019 onwards, as such, countries which experience left truncation (China, Taiwan, South Korea, US, Japan, Thailand) are excluded from the model fitting dataset.

Central Statistics

Table 1: Response Variable Summary Statistics

	Min	Q_1	Q_2	Q_3	Max	Mean
t_1	67	102	110	119	135	108.4
t_2	0	8	15	22	61	17.31

Data Snippet

Table 2: Snippet of data after preprocessing [5]

Country	t_1	t_2	inception	confirmed	death
Zimbabwe	124	3	17/11/2019	20/03/2020	23/03/2020
South Africa	109	22	17/11/2019	05/03/2020	27/03/2020

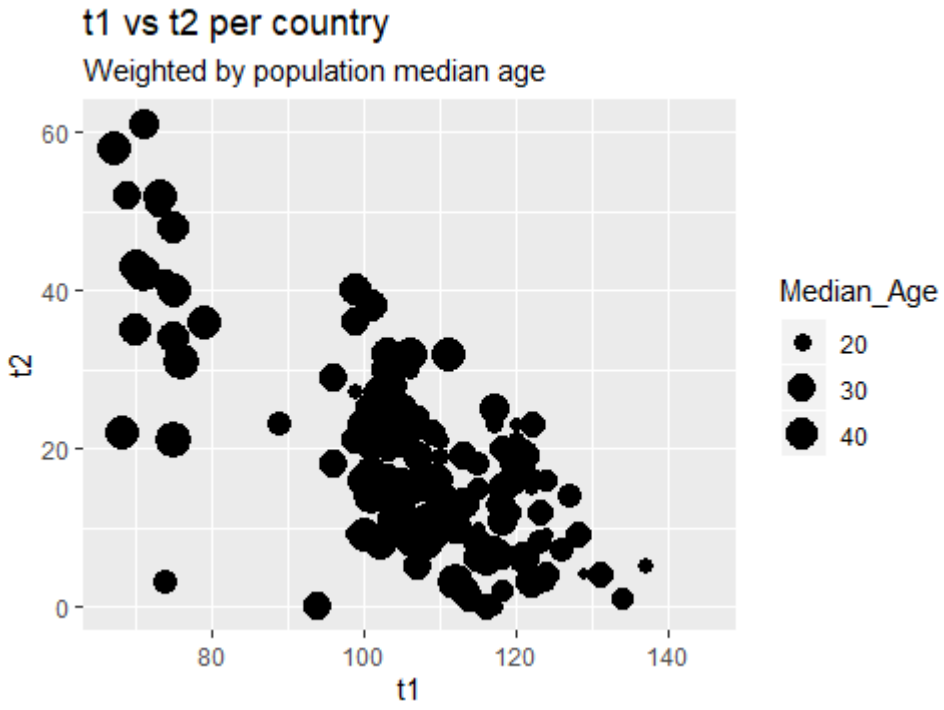
0.2.1 Dependence and Concordance

1. $t_1 \sim t_2$

   The pearson correlation between the 2 response variables is -0.78. This shows that t_1 is inversely related to t_2, such that taking a long to report a first confirmed infection is associated with taking a short amount of time to report a death soon afterwards.

   For instance, the case of Zimbabwe and South Africa as shown in 2. Zimbabwe 124 days to report a first case, only to report a death 3 days later. Where else South Africa took 109 days to report a first case, and reported a death after 22 days.

   Assuming the COVID-19 force of mortality to not be dependent on the time the first case is confirmed, it may appear that countries who take long to report instances of the virus are experiencing an oversight issue, such that by the time they commence their mortality investigation by reporting a first case, some COVID-19 lives are actually reaching the end on their lifetimes.

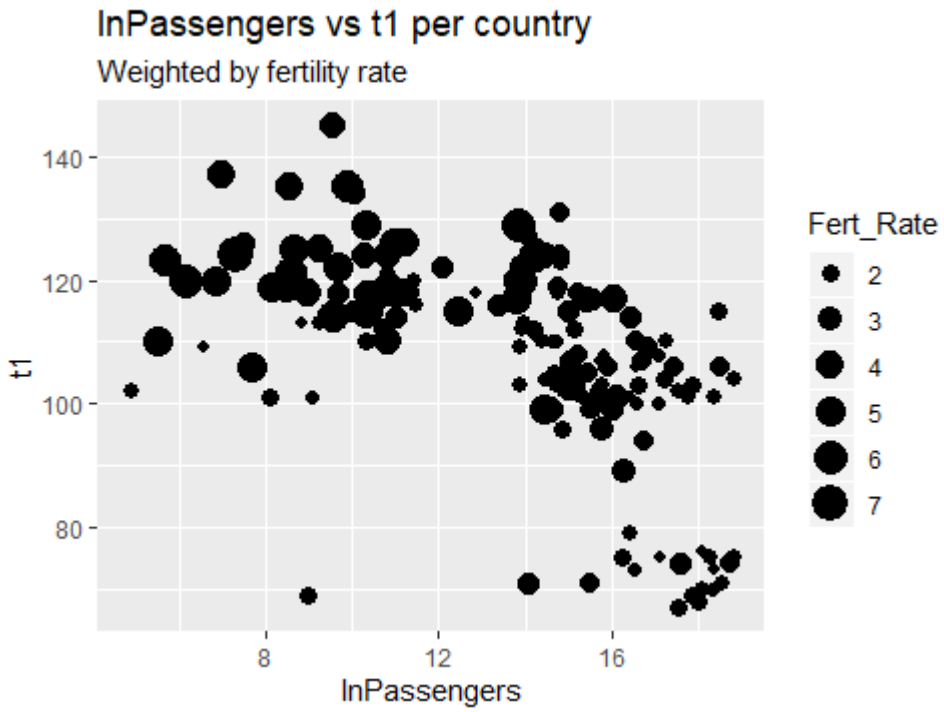


2. Passengers $\sim t_1$

   The pearson correlation between Population and t_1 is -0.565. This shows that t_1 is also inversely related to the number of airline passengers, such that a large number number of airline passengers is associated with reporting a first confirmed case sooner.

   Furthermore, we observe that ln(Passengers) axis separates the t_1 into axis into 2 categories, i.e. The segment below 12 and the segment above 12. These identify countries with low and high numbers of air passengers respectively, with a strange realization that countries with low numbers of air passengers tend to have higher fertility rates. This narative is further reinforced by the next relationship, which shows that countries with a low number of airpassengers, tend to have younger populations.

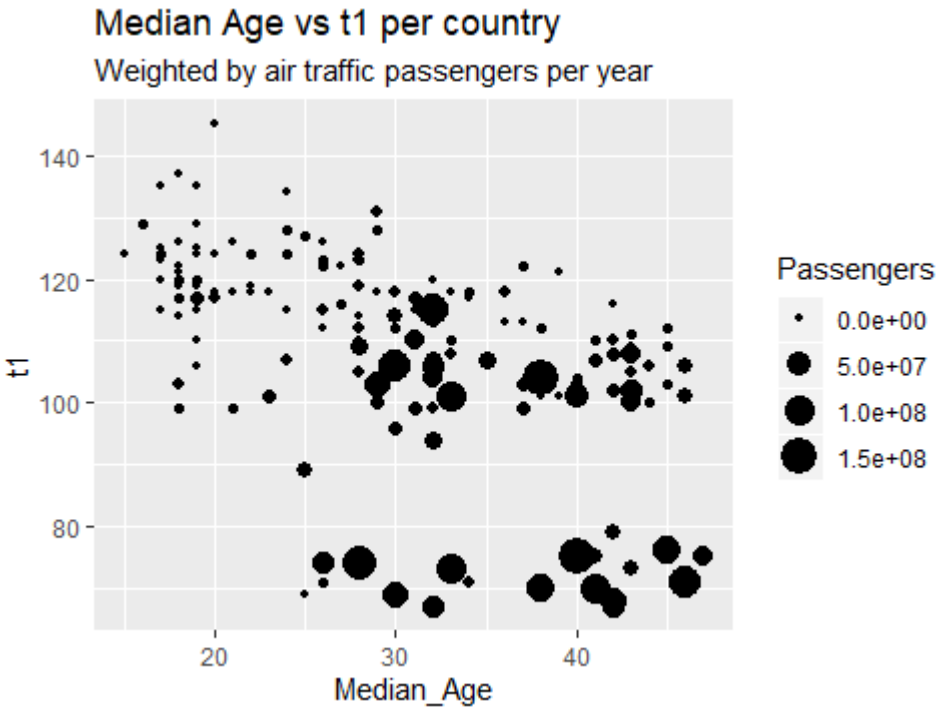
   i.e: Older populations have a tendency travel more (resulting in more exposure to COVID-19) compared to younger populations, they also have more interaction with other networks of older people. This, coupled with having weaker immune systems, may have been a major driver of the COVID-19 pandemic affecting older people with the extreme intensity which it has.



3. Median Age $\sim t_1$

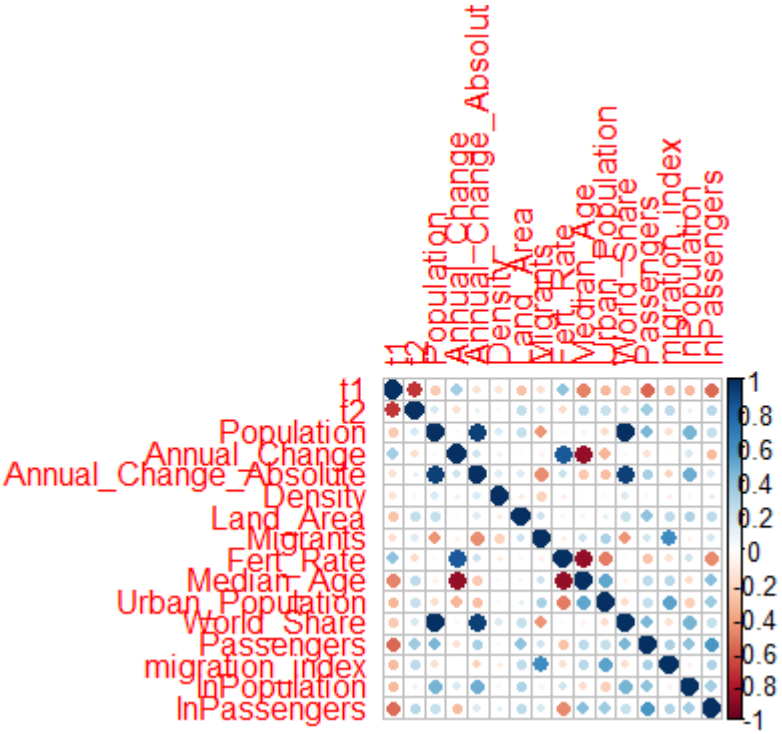
   The pearson correlation is -0.45.

   To reiterate, older people appear to travel more, and even those who may not travel are more likely to interact with older people who do as opposed to their young counter parts. This, coupled with them having weaker immune systems, appears to be one of the drivers of networks of old people being exposed and harshly affected by the virus.



4. . ∼ . (correlogram)

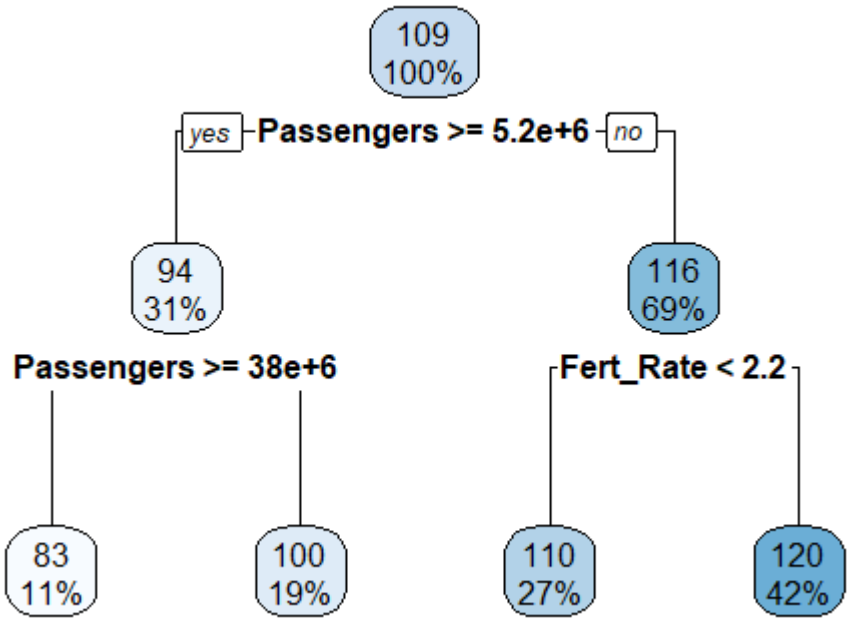
   For completeness, see below correlogram for summary of all other relationships with respect to the pearson correlation.



0.3 t_1: Decision Tree

The following decision tree model was fitted for variable t_1

Figure 1: Decision Tree model for t_1 with complexity parameter of 0.025



The decision tree re-iterates the theory that older people tend to travel, and as a result have more exposure to COVID-19 as opposed to their younger counter parts.

The model begins by identifying that countries with a large number of air travellers are the purest segment to identify the countries with the lowest time to report a first case. It further segments countries who have less fertility rates and annual growth rates as those who take the least time to report a first case. Both these cases point out to older populations, who are not as fertile, but have the means to use air travel frequently.

0.4 t_2: Multivariate Linear Regression

The following stepwise multivariate linear regression model was fitted for variable t_2

$$t_2 = 118.61624 - 0.72402t_1 - 0.21348X_{MedianAge} - 1.14529X_{lnPopulation}$$

It produces an R^2 of 0.626, which means the linear model accounts for approximately 62.6% of the variation in t_2. This will suffice for the purposes of this report.

The linear model uses t_1, Median Age, and ln(Population) to explain t_2. Such that an increase in any one of these predictors is expected to reduce the amount of time it takes to observe the first death.

All other predictors have been found to not be statistically significant in accordance to the stepwise procedure based on the AIC criterion.

Bibliography

[1] Ma Josephina *China's first confirmed Covid-19 case traced back to November 17*. South China Morning Post, 13 March 2020.
https://www.scmp.com/news/china/society/article/3074991/coronavirus-chinas-first-confirme

[2] List of countries by airline passengers
https://en.wikipedia.org/wiki/List_of_countries_by_airline_passengers

[3] Countries in the world by population (2020)
https://www.worldometers.info/world-population/population-by-country/

[4] 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE
https://github.com/CSSEGISandData/COVID-19

[5] Lindo Khoza *Novel Coronavirus (COVID-19) Cases from a South African perspective*
https://github.com/willkhoza/COVID_SA