Linear Regression Model to Study the Factors Affecting COVID-19 Mortality in 21 European Countries

Lampros Katsiamitros, Evangelia N. Petraki

Department of Economics, National and Kapodistrian University of Athens, Greece

evpetra@econ.uoa.gr

Abstract

The current research concerns data analysis with machine learning techniques using python programming language. More specifically, factors such as GDP per capita, median age, percentage of smokers, prevalence of diabetes and cardiovascular disease, are being studied in 21 European countries for their effect on mortality from COVID-19 with the help of a linear regression model. The dataset used in this research comes from ourworldindata.org and it covers the period from the end of January 2020 to September 15, 2022. In the first part of this study, the correlation of stringency index with deaths from COVID-19 is analyzed and the results show that 12% of total deaths per million is correlated with the index, while new deaths per million is correlated by 15%. In the second part, a linear regression model is constructed for the analysis of the abovementioned factors. In conclusion, the model explains 14.9% of total deaths per million from COVID-19.

Keywords. Data Analysis, COVID-19 Mortality, Linear Regression, Machine Learning, Stringency Index.

Stringency Index

The COVID-19 Stringency Index is a measure of the severity of government policies implemented to control the spread of the disease. It is based on several indicators, such as school closures, travel bans and workplace closures. [3]

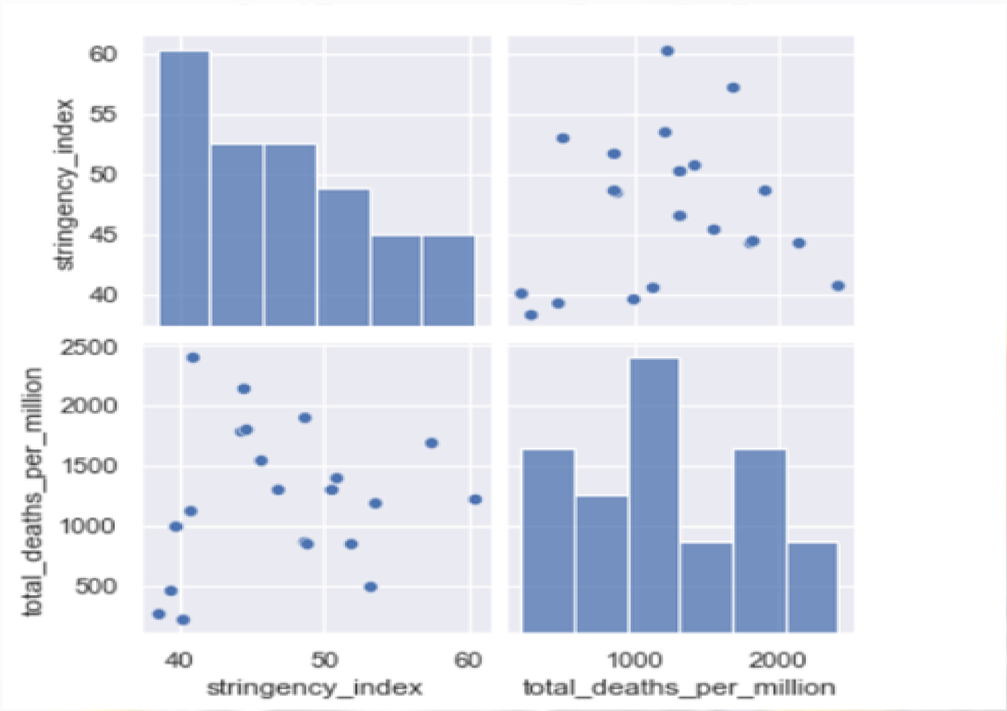
Data

They have come from various sources such as John Hopkins University in the USA, WHO and ECDC. They consist of 216,816 records and 67 variables and concern 243 countries and regions. Each entry is for one day and one country/region.

Correlation of stringency index and deaths from COVID-19

The methodology followed is as follows:

- 21 European countries are examined, more specifically Finland, Denmark, Switzerland, Norway, Sweden, Hungary, Belgium, Czech Republic, Slovakia, Poland, France, Ireland, United Kingdom, Netherlands, Portugal, Spain, Germany, Cyprus, Austria, Italy and Greece.
- The stringency index average is calculated for each country
- For each country, the average of total deaths per million is calculated (total deaths per million)
- For each country, the average number of new deaths per million (new deaths per million) is calculated
- Finally, a comparison is made between the mean of the stringency index and the means of the two other variables respectively.
- The results are as follows: The correlation of new deaths per million with stringency index is 0.15. Similarly, the correlation of total deaths per million with stringency index is 0.12.



Conclusions

There is correlation between the index and deaths from COVID-19, but small, which contradicts the literature, which reports a strong correlation. The different results of the present study are explained by considering a much longer period, with a large spread of the virus and large variations in the stringency index, thus its small impact on deaths from COVID-19.

Multiple Linear Regression Model

Data from 21 European countries are analyzed and the factors under consideration are: gdp per capita, median age, diabetes prevalence, cardiovascular death rate and smokers. The model is built using multiple linear regression of the Statsmodels python library.

First, the data are cleaned, grouped by month and the average of each variable is found. The model that was built concerns all 21 countries and not each country separately. Finally, the data are normalized.

From the results, it follows that the model is statistically significant, as the probability of the null hypothesis (H0), given by Prob (F-statistic) = 4.85×10^(-17), is very small. The R2 index, which shows how correlated the dependent variable total deaths per millions is with the linear combination of the independent variables, is 0.149.

With the help of the P-value, at a confidence level of 5%, only the variable smokers have a P-value = 0.823 > 0.05 and is removed from the model. It is run again with the remaining variables and remains statistically significant as Prob (F-statistic) = 9.58×10^(-18). The R2 index remains at 0.149. In this model, all variables are statistically significant at the 5% confidence level.

The conclusions drawn are the following:

The variable gdp per capita has a negative effect on the dependent variable (total deaths per million) with a coefficient of -918.157.

The media age variable has a positive correlation with the dependent variable, with a coefficient of 397.002.

The diabetes prevalence variable has a negative effect on total deaths per millions, with a coefficient of -794,899.

The cardiovasc death rate variable has a positive correlation with the dependent variable, with a coefficient of 876.124.

Conclusions

GDP per capita was found to have a negative effect on total deaths per million. Richer countries therefore have the ability through advanced infrastructure and systems to limit mortality from the COVID-19 disease. The variable median age has a positive effect on the dependent variable. Therefore, countries with an increased median age of the population are more vulnerable to deaths from COVID-19. In addition, the prevalence of diabetes in the population has a negative impact on total deaths per million, which is unexpected as the sources report diabetes as a major cause of disease burden. The variable cardiovasc death rate has a positive effect on the dependent variable, which also results from the literature.

In conclusion, the model explains 14.9% of mortality from COVID-19. This finding is very important, because it demonstrates that non-health factors (GDP per capita, median age) also affect deaths.

Selected References

1. Bhaskaran K., Bacon S., Evans S. J., Bates C. J., Rentsch C. T., MacKenna B., Tomlinson L., Walker A. J., Schultze A., Morton C. E., Grint D., Mehrkar A., Eggo R. M., Inglesby P., Douglas I. J., McDonald H. I., Cockburn J., Williamson E. J., Evans D., Curtis H. J., Hulme J.W.,Parry J.,Hester F.,Harper S.,Spiegelhalter D.,Smeeth L., Goldacre B. (2021). Factors associated with deaths due to COVID-19 versus other causes: population-based cohort analysis of UK primary care data and linked national death registrations within the OpenSAFELY platform. The Lancet regional health. Europe, July 2021, Vol. 6, 100109. https://doi.org/10.1016/j.lanepe.2021.100109
2. Ciotti M., Ciccozzi, M., Terrinoni, A., Jiang, W. C., Wang, C. B., & Bernardini, S. The COVID-19 pandemic. Critical reviews in clinical laboratory sciences, 2020, 57(6), 365–388. https://doi.org/10.1080/10408363.2020.1783198
3. COVID-19 Government Response Tracker Retrieved Apr. 5, 2023, from www.bsg.ox.ac.uk/research/covid-19-government-response-tracker
4. Data on Covid-19 (coronavirus) by Our World in Data Retrieved Sep. 15, 2022, from https://github.com/owid/covid-19-data/tree/master/public/data