THE INFLUENCE OF COVID-19 ON THE OECD ECONOMY

BENEDEK PÓSFAY

INTRODUCTION

Predicting how the economy of a country changes over time is a difficult task. However, the aim of this project is to analyse data from 29 OECD member countries, and try to come up with a model, that can accurately guess the wellbeing of a country's economy.

The model will use monthly aggregated coronavirus data along OECD's Composite Leading Indicators (CLI) Index. What monthly COVID-19 variables' change can have noticeable effect on the economic status of OECD member countries? By answering the question, we will have a clearer understanding of the relationships between the input variables, and we will be able to foresee economic growth or decrease in advance and prepare for it accordingly. The model could be used as a tool for anyone who has interest in predicting a country's economic status.

DATA CHARACTERISTICS

There are 9 monthly observations collected from 29 countries resulting in a total of 261 observations of 22 input variables. There are 14 coronavirus related variables and 8 demographic indicators.

The data also includes our output variable, the CLI Index, which describes how a given country's economy performs based on leading economic indicators and predictions, resulting in an index which at 100 points means that the economy is not moving in either direction. If the index falls below 100, it indicates a decrease in performance, if above 100, we can expect growth. However, the change in the index is relatively small from month to month, most of the times below 2 points.

The selected Covid-19 input variables are the following:

- **Total cases**: Total confirmed cases of COVID-19
- **New cases**: New confirmed cases of COVID-19
- **New cases smoothed**: New confirmed cases of COVID-19 (7-day smoothed, summed up)
- **Total deaths**: Total deaths attributed to COVID-19
- **New deaths**: New deaths attributed to COVID-19
- **New deaths smoothed**: New deaths attributed to COVID-19 (7-day smoothed, summed up)
- **Total cases per million**: Total confirmed cases of COVID-19 per 1,000,000 people
- **New cases per million**: New confirmed cases of COVID-19 per 1,000,000 people
- **New cases smoothed per million**: New confirmed cases of COVID-19 (7-day smoothed, summed up) per 1,000,000 people
- **Total deaths per million**: Total deaths attributed to COVID-19 per 1,000,000 people
- **New deaths per million**: New deaths attributed to COVID-19 per 1,000,000 people
- **New deaths smoothed per million**: New deaths attributed to COVID-19 (7-day smoothed, summed up) per 1,000,000 people
- **Reproduction rate**: Real-time estimate of the effective reproduction rate (R) of COVID-19.
- **Stringency index**: Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)

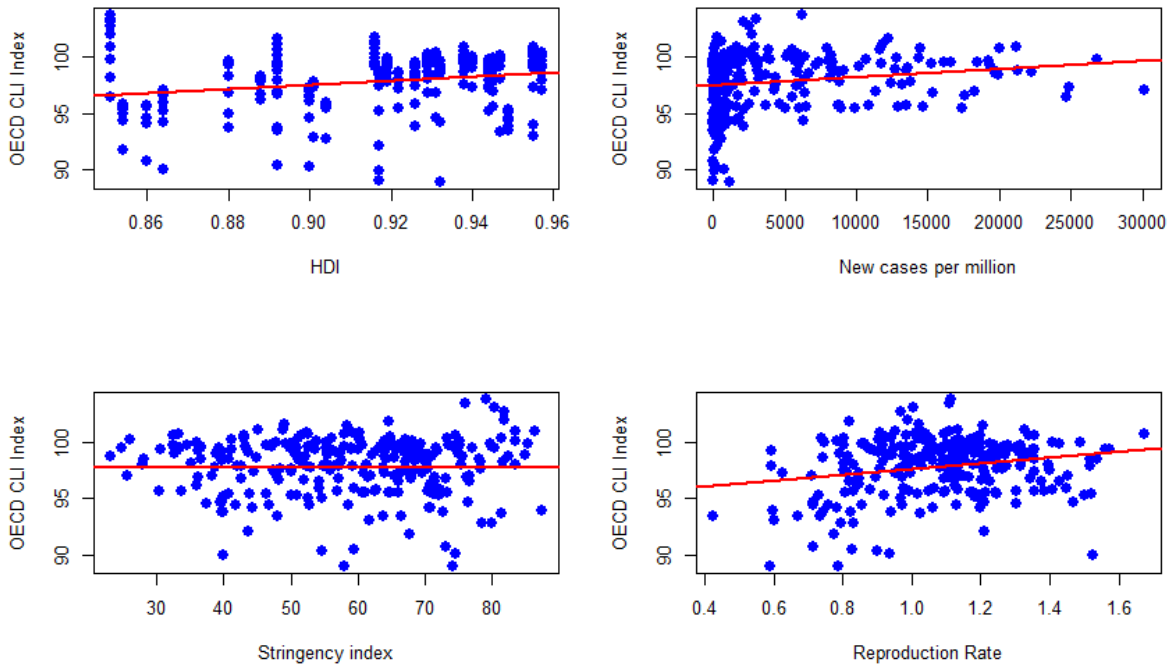
These are the included demographic indicators:

- **Population**: Population in 2020
- **Median age**: Median age of the population, UN projection for 2020
- **GDP per capita**: Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available
- **Female smokers**: Share of women who smoke, most recent year available
- **Male smokers**: Share of men who smoke, most recent year available
- **Hospital beds per thousand**: Hospital beds per 1,000 people, most recent year available since 2010
- **Life expectancy**: Life expectancy at birth in 2019
- **Human Development Index (HDI)**: A composite index measuring average achievement in three basic dimensions of human development — a long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506

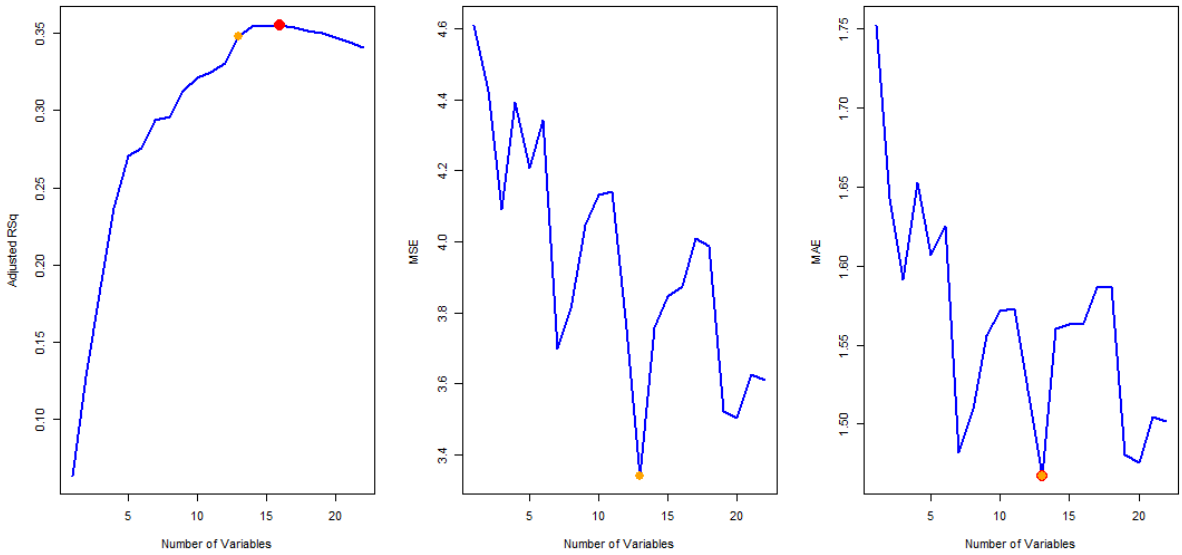
We must keep in mind that the demographic predictors are static per country, meaning that they are not changing from one month to another, rather they are included to give the model a better understanding of the current demographic status of a given country.

Model Selection

Doing a few simple linear regressions and scatter plots, we can discover some interesting characteristics in the relationships between some predictors and the *CLI Index*. Quite surprisingly, there was a relatively high correlation between the *Reproduction rate* and the *index*. Another similarly surprising observation is that the *Stringency index* (which measures the strictness of a government) has no influence on the *CLI Index* at all, while in theory, these measures have a huge impact on the actual economies. The *HDI* or the *New cases per million* predictors, however, somewhat correlate with the *CLI Index*, indicating that our model should contain these variables.



Running best subset selection on the data to eliminate insignificant input variables, we arrive at the following conclusion: when selecting the following 13 out of the 22 predictors, the model reached the smallest MSE (3.341) and MAE (1.467) and a close-to-maximum adjusted R-squared of 0.348.



The following input variables were selected in the model, with their respective estimated coefficients:

Predictor	Coefficient	Coefficient * Mean(input variable)
Intercept	67.504	-
Total cases	2.324e-6	1.509
New cases	-4.984e-6	-0.923
Total deaths	-1.058e-4	-1.907
New deaths	2.091e-3	6.736
New deaths smoothed	-2.033e-3	-6.488
Total deaths per million	3.421e-3	1.296
Reproduction Rate	3.077	3.335
Population	3.073e-8	1.146
Median age	-0.198	-8.42
Female smokers	-0.135	-2.959
Male smokers	0.166	4.912
Hospital beds per thousand	-0.297	-1.437
HDI	36.595	33.506

At first, the variance of the coefficients may look large, but it is mainly due to the variance of the different input variables. Therefore, applying some normalization to the coefficients (i.e., multiplying them by the respective mean values from the dataset) can show their real relevance compared to each other. Immediately, *HDI* appears to be by far the most influential variable, while *New cases*, *Population*, *Total deaths per million*, *Hospital beds per thousand* and *Total cases* are the least relevant ones. Another key observation is that the coefficients of *New deaths* and *New deaths smoothed* are cancelling each other out, simply because the values of *New deaths* and *New deaths smoothed* are very close to each other in a single observation. As a result, excluding these two predictors from the model should not make a significant change in its performance.

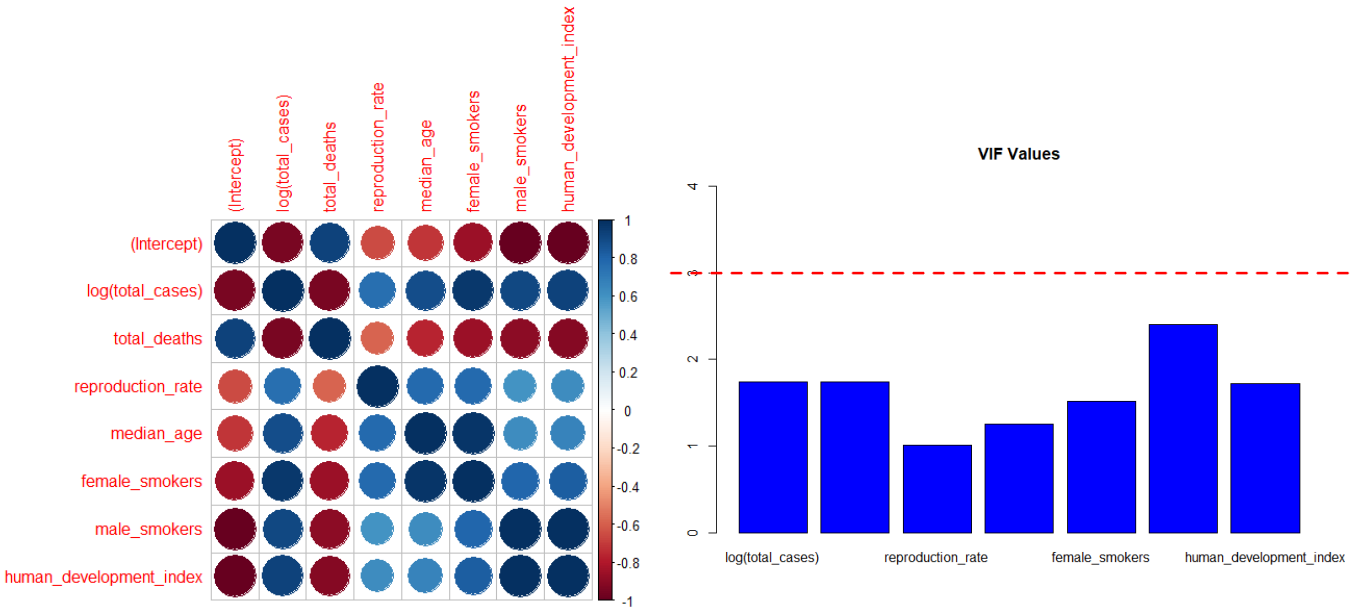
By doing some more testing and investigations into multicollinearity, the removal of *New cases*, *New deaths*, *New deaths smoothed*, *Total deaths per million*, *Population* and *Hospital beds per thousand* improved the model's fit, and the significance of the remaining variables. Carrying out multiple linear regressions with the remaining 7 predictors and further applying logarithmic transformation to the *Total cases* resulted in an overall better model. The estimated coefficients, their standard errors and their t-values and p-values are summarized below.

Coefficients	Estimate	Standard Error	t-value	p-value
(Intercept)	63.12	7	9.017	3.34e-16
log(Total cases)	0.835	0.1	8.345	2.07e-14
Total deaths	-1.903e-5	4.204e-6	-4.526	1.11e-5
Reproduction rate	3.04	0.706	4.304	2.78e-5
Median age	-0.269	0.0579	-4.641	6.78e-6
Female smokers	-0.126	0.0277	-4.542	1.03e-5
Male smoker	0.142	0.0277	5.123	7.9e-7
HDI	35.21	6.621	5.317	3.19e-7

As we can see from the p-values, all the predictors are significant and therefore cannot be excluded from the model. The adjusted R-squared of the new linear regression has increased to 0.392, while the F-statistic is 17.74 on 7 and 175 DF (with the p-value<2.2e-16), so we can reject the null hypothesis.

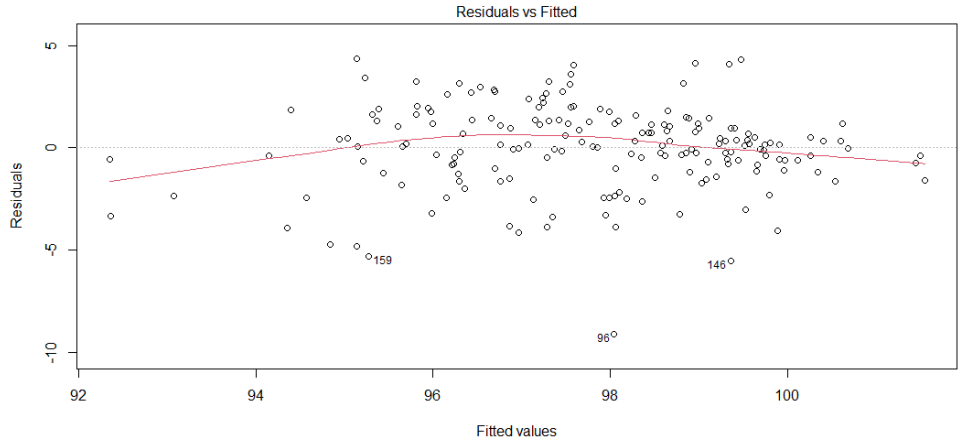
Model Evaluation

The fitted model explains almost 40% of the variance in the data. To make sure, that we cannot improve the model by removing predictors, we can look at their VIF values and their correlation matrix.



As expected, the predictors seem important and their VIF values are below 3, indicating no (or insignificant) multicollinearity.

Looking at the Residuals vs Fitted plot, we can observe that the mean of the error terms is around zero and the variance is about constant. There are three outliers (159, 146, 96), that may suggest further investigation to those observations.



After removing the three extreme cases, the adjusted R-squared of the model further increased to 0.438. The comparison of the coefficients of the originally fitted model and the new model is summarised in the table below.

Coefficients	Original model	New model	Change in %
(Intercept)	63.12	63.83	1.12%
log(Total cases)	0.835	0.822	1.56%
Total deaths	-1.903e-5	-1.752e-5	7.93%
Reproduction rate	3.04	2.944	3.16%
Median age	-0.269	-0.275	2.23%
Female smokers	-0.126	-0.126	<1%
Male smoker	0.142	0.137	3.52%
HDI	35.21	35.27	<1%

Clearly, the changes in the coefficients are small (with the maximum change being 7.93%), therefore, removing these observations did not improve the model significantly.

Summary and Conclusion

The aim of this paper was to find a relationship between Covid-19 related data and a country's economic status. OECD's *CLI Index* was set to be the response variable, as it is a composite index, that can accurately describe the monthly change in an economy. Our model received additional input variables, that were demographic indicators, to help it better understand a country's state. The predictors, then, were narrowed down by eliminating insignificant and multicollinear variables using best subset selection. Using linear regression, a model was fitted, that could predict around 40% of the variance in our data. This raises the question, whether such error is significant in predicting economic status, as even a small change in the *CLI Index* could mean a huge spike in unemployment or a large downfall in export rates. The data used, was also only from OECD member countries, so the model was spared from extreme economies. However, it should be able to give an idea about the direction of the economy. Further analysis can be conducted to produce a model which instead of predicting an actual value, tries to categorize the changes in the *CLI Index*, and classify the observations as such, thus, turning it into a classification problem.