A Nonparametric Approach to an Iterative Problem: Are political stability indicators more predictive of 2020 GDP changes than COVID statistics?

Henry Manley, Heather Ginsburg, Jordi Socher, Sophie Keller, Sam Meakem

May 14, 2021

Abstract

The COVID-19 pandemic has placed the global economy into disarray, creating uncertainty in the future state of labor and capital markets. With such uncertainty in this forecast, the current report attempts to investigate the effect of the pandemic on a country's GDP per capita. Such an estimate can help inform the likelihood and speed of recovery and is calculated with a slew of machine learning techniques. Data from 2019 and years preceding the pandemic, which consists of both COVID and economic indices, were evaluated. Best subset selection was utilized to determine the best predictors of GDP change, which were then passed into a horse race of three models. The data was analyzed after each regression, contributing to marginal model improvements. We find that the nonparametric loess model best predicts the percent change in GDP from 2019 to 2020 across the panel of all countries Furthermore, all economic indices explored in this study were negatively impacted by the COVID crisis. We conclude that there is a strong negative correlation between GDP change and the severity of COVID in a given country, yet political indicators contributed to an even stronger model than those COVID indicators alone.

0.1 Introduction

The COVID-19 pandemic that shook the world has had a devastating effect on not only the health of populations and countries across the globe, but on various other factors including the global economy. The economic crisis and depression of labor and capital markets brought on by the pandemic has affected each country uniquely, largely due to preexisting economic inequalities. This spread can be seen in Figure 1 which geographically maps the number of total cases per 100000 people in each country. From this baseline analysis, the United States, Brazil, and Spain had some of the highest case counts per capita, while Mongolia, China, and many African countries experienced much lower figures.

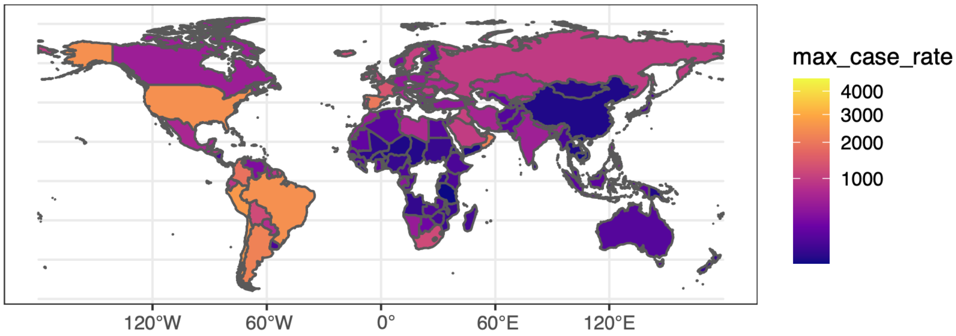


Figure (1) A World in Cases

As many economic entities were forced to halt all work when the pandemic began, the economy nosedived. Now, as countries have wide spread vaccine availability and distribution plans, the percentage of the population vaccinated, and new COVID strains and persistence, it begs the question: "How will *this* affect the economy?"

The pandemic is not over and it will likely take a long time to rebuild national economies and the interconnected web that is the global economy. This study aims to quantify and predict the pandemics' long-term effects on the global economy and how it relates to governmental restrictions, accountability, and stability throughout the pandemic using a horserace of machine learning techniques such as linear and local polynomial regression, with features yielded by subset selection.

0.2 Data Analysis

0.2.1 Implementation

Prior to any feature of model selection, the original data, which is from the study "The Impact of COVID-19 Pandemic on the Global Economy: Emphasis on Poverty Alleviation and Economic Growth," needed to be cleaned [**1**]. It was determined that the 2020 GDP per capita column from the original dataset had inconsistencies due to using the 2011 dollar as a reference point. To account for this, the International Monetary Fund data was used for 2020 GDP data and replaced the original values for 2020 GDP [**2**]. Additionally, 2019 GDP data was pulled from the UN's database to be used in the models as a comparison baseline [**4**]. From this baseline, our outcome variable of interest was created: the percentage change in GDP per capita from 2019 to 2020, by country. Figure 2 suggests a moderately negative correlation between each of the rankings and a country's change in GDP.

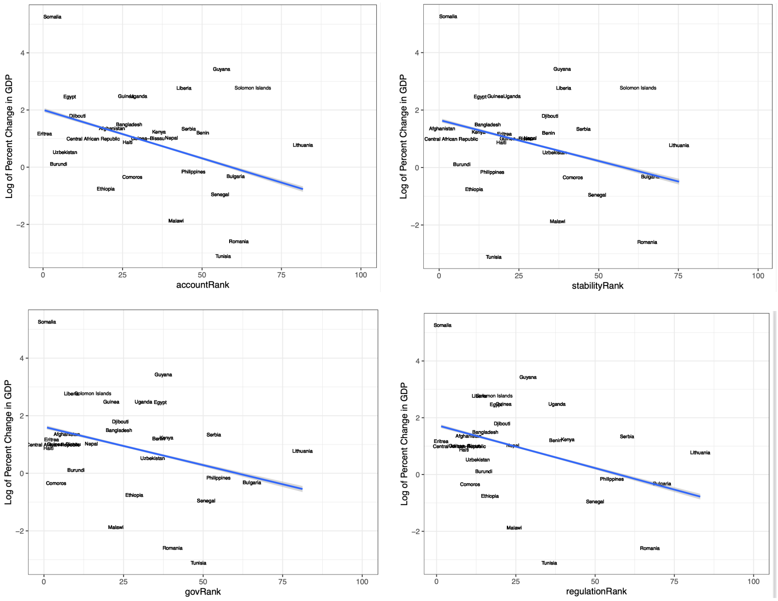
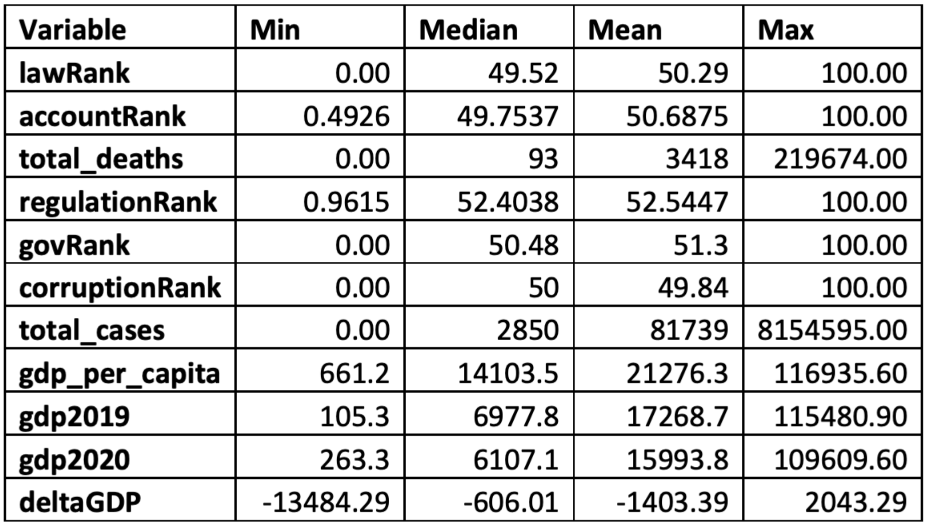


Figure (2) Institutional Stability vs. GDP

Next, data was merged from the World Bank measuring various political stability indicators such as rule of law (lawRank), voice and accountability (accountRank), political stability and absence of violence and terrorism (stabilityRank), the total number of deaths from COVID at the date the data was taken (total_deaths), the regulatory quality (regulationRank), government effectiveness(govRank), control of corruption (corruptionRank), the total number of cases of COVID at the date the data was taken (total_cases) [**3**]. More detailed explanations of what each of these variables represent can be found in Figure 8 of the Appendix. We chose to incorporate this data to determine if COVID statistics like case and death count proved to be stronger or weaker predictors of GDP changes than these political indicators.

This comprised the entirety of the data used and was sufficient to exploit the use of machine learning techniques to make robust predictions and gather meaningful results. See Figure 3 for data summary statistics, noting the median and mean for political stability measures that are not equal to 50 points. This is singlehandedly the effect of omitting missing values in our data cleaning process-marginal, but important to note.



Variable	Min	Median	Mean	Max
lawRank	0.00	49.52	50.29	100.00
accountRank	0.4926	49.7537	50.6875	100.00
total_deaths	0.00	93	3418	219674.00
regulationRank	0.9615	52.4038	52.5447	100.00
govRank	0.00	50.48	51.3	100.00
corruptionRank	0.00	50	49.84	100.00
total_cases	0.00	2850	81739	8154595.00
gdp_per_capita	661.2	14103.5	21276.3	116935.60
gdp2019	105.3	6977.8	17268.7	115480.90
gdp2020	263.3	6107.1	15993.8	109609.60
deltaGDP	-13484.29	-606.01	-1403.39	2043.29

Figure (3) Summary Statistics

0.2.2 Selection

We performed best subset model selection which compares all possible models using a set of specified predictors and outputs valuable information regarding what the dominant variables are. We chose to use the best subset method over other selection techniques since it assesses all possible variable combinations and thus provides more information. It is the optimal choice in terms of selecting the variables that actually contribute to the fit and disregarding the variables that do not contribute to the fit. This ensured that we had the full capability to get the most meaningful results from our models. The code for this appears in the file selection.R.

The results showed that the variables were selected in the following order: lawRank, accountRank, stabilityRank, total_deaths, regulationRank, govRank, corruptionRank, and lastly total_cases. The selection also indicates which number of variables would result in the highest R^2 value and averaging across all days; this was four variables. This means that the variables that would give us the most meaningful results to explore are: lawRank, accountRank, stabilityRank, and total_deaths. We took these variables and explored them further to determine how they were correlated. Specifically, and as evidenced by Figure

4, we would expect to see relatively low correlation between these four variables since the selection process would rule out one of two variables that covaried with the endogenous variable.

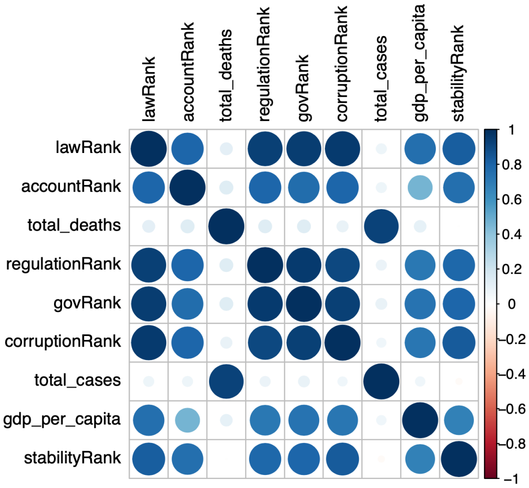


Figure (4) Feature Correlation Matrix

0.2.3 Model

Running and Plotting the Initial Models

Our analysis was done iteratively by day, horseracing three proposed models that incrementally made improvements on its predecessor, each day over the time series (January 1, 2020 - October 19, 2020). All models can be found in model.R. Model estimates were used to calculate R^2, feature coefficients, and Test MSE over time, where each day in our analysis included a set of countries and values of their respective features. This iterative methodology was used largely in part to avoid issues of constancy in our endogenous variable over the course of the year, and to capitalize on daily changes in COVID indices. Hence, we were able to isolate the days in which our model best (and worst) predicted changes in a country's GDP, and hence the global economy. In theory, data could have only been collected on one day for our model to predict the percentage change in GDP from 2019 to 2020 by $\sim 94\%$.

In part due to the data sparsity and noise as well as the causal nature of the research question being investigated, our model is mainly interested in measuring the change in the covariance of model features with the change in GDP over the course of the pandemic.

Model #1

To begin our analysis, we ran a naive linear regression model with robust standard errors clustered at the country level. This model, as specified below, included a handful of arbitrarily selected regressors, and yielded a relatively weak prediction. total_deaths + total_cases + stringency_index + accountRank:

$$\Delta GDP_{c,d} = \alpha_0 + \beta_1 Deaths_{c,d} + \beta_2 Cases_{c,d} + \beta_3 STI_c + \beta_4 Account_c + \epsilon_{c,d} \tag{1}$$

Model #2

To improve upon the relatively weak prediction power, we considered a best subset selector to yield the best features to include. This selector, using a combination of both Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge Regression, significantly improved the estimate of change in GDP. This elastic net selector subjects the naive linear model as follows:

$$\Delta GDP_{c,d} = \alpha_0 + +\epsilon_{c,d} \tag{2}$$

Model #3

At this point, it became clear that the relationship between our features and the endogenous variable may not be linear, perhaps being better estimated by a nonparametric function. Retaining the use of the subset selector, our third and final model included a local polynomial regression with a bandwidth calculated for each day. This model follows this specification:

$$\begin{aligned} &\Delta GDP_{c,d} = h^{-1} \textstyle\sum_{i=1}^{n} [\int_{s_{i-1}}^{s_i} K \dfrac{x-u}{h} du] y_i + \epsilon_{c,d} \\ &Where, s_i = \dfrac{x_{i-1} + x_i}{2} \end{aligned} \tag{3}$$

As seen in Figure 6, this model trends both above and below Model 2, with particular erratic behavior in months following May 2020. This is likely due to the increased sensitivity to changes in COVID indices, where a nonparametric function is more flexible but also likely overfit. To reiterate, due to the causal tones in the presented research question, we were willing to accept a perhaps otherwise unusually low level of variance in our model; we wanted the best non-linear estimate of the change in GDP in terms of the selected variables. Noise in all models in the early months of the pandemic is to be expected, with such vicious growth in cases and deaths that result in askew predictions and high Test MSE.

Results from the Initial Models

The results from the initial linear regression run on all predictor variables and on the subset with the chosen predictors from subset selection can be observed in Figure 5. To little surprise, the linear regression using the four predictors selected by best subset selection outperforms the naive linear model with an average R-squared value almost 7.5 times greater.

Not to mention, all features have statistically significant coefficients ($\alpha < 0.05$) with minimized standard errors; an indication of a proper selection process and predictive power in the select features. Interestingly, as shown in the graph of P-values below, P values decrease over time to demonstrate statistical significance for all variables in early March. The combination of the increased model performance and statistical significance in our regressors evidences a model that avoids the curse of dimensionality and may robustly predict changes in GDP. Despite initial excitement in model performance, what was less certain was the data's linearity- how do we know that the change in GDP is linearly proportional to the number of COVID cases/deaths and levels of political stability in a given day during the pandemic? In fact, it was observed that this was not the case, following preliminary data visualization. It was clear that the model could be improved.

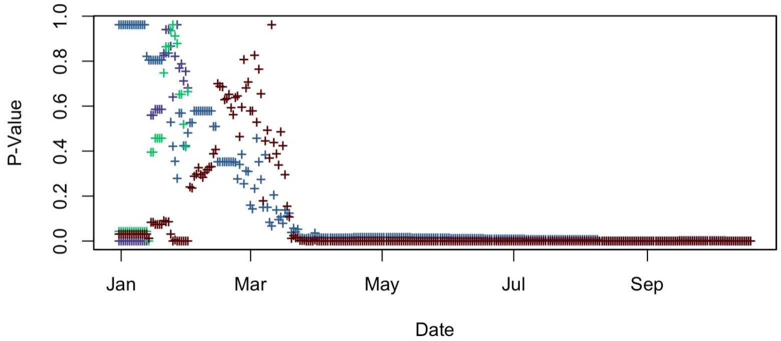
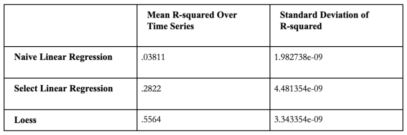


Figure (5) Results from Initial Models

How to Improve Models

After initial analysis, we determined that there was room to further iterate on our model by introducing a more flexible approach. Given gross uncertainty in the true shape of the hypothesized relationship between the change in GDP and our features, we opted for a nonparametric local polynomial regression model. This model, as mentioned previously, considers inherent nonlinear regularities in the data. Our final model used loess nonparametric regression on the best feature subset, being computed for every day over the time series and compared to previous models. The loess regression uses a uniformly distributed kernel of bandwidth w, calculated at runtime.



	Mean R-squared Over Time Series	**Standard Deviation of R-squared**
Naive Linear Regression	.03811	1.982738e-09
Select Linear Regression	.2822	4.481354e-09
Loess	.5564	3.343354e-09

Figure (6) Regression Results

0.2.4 Results from the Improved Model

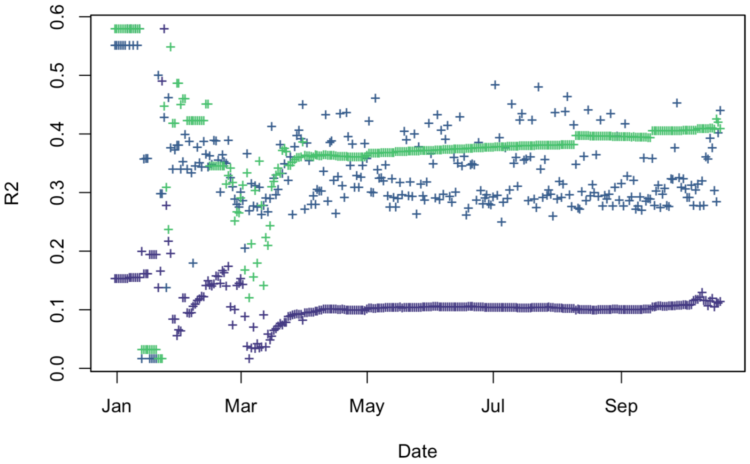


Figure (7) Model Horse Race
Loess *SubsetLinear* *NaiveLinear*

After running the local polynomial regression our findings show that GDP can be predicted from data using this model. The higher R-squared for the non-parametric model indicates that this model fits our data best. This Mean R-squared value of .5564 can be classified as a moderate indicator for the fit of the data, but is more significant compared to the other models that were run on the data. This value is almost double that of the linear regression with predictors selected via subset selection and over fourteen times that of the basic linear regression with all predictors, as observed in Figure 6 and Figure 7.

0.3 Conclusion

A main takeaway from our attempts to use linear models to predict GDP is that the relationship between GDP and other factors such as total COVID cases and corruption ranks is not as straightforward as we originally may have thought. One consideration when first producing our models was that we wanted to focus our study on determining the causal relationship between our input variables and GDP, rather than trying to predict GDP and measure accuracy. As a result, we did not split the data into training and testing. Not splitting our data allowed us to focus our models on discerning the effect of our input variables on GDP. Furthermore, performing subset selection showed that the model would be more accurate using a variety of input, showing that one would be ineffective in trying to predict GDP using only COVID data such as total cases. Running both a standard linear regression model and a model using the inputs produced by subset selection showed the improvement from using more specific inputs to predict GDP.

As shown by the above graphs, the non-parametric approach is able to supply a much more accurate model for predicting GDP. By using the inputs supplied by the subset selection for the non parametric model, we were able to create a model that could predict GDP more accurately. This is because the flexibility of the model was able to accommodate the highly variable nature of the inputs. In the model horse race graph, this is evidenced by the fact that the linear models converge very quickly to the same value and are not able to adjust to varying data. The non-parametric model, on the other hand, is fitted so that it takes into account the high variation in data. Furthermore, it can be observed that all of our models started out very erratically. This can be attributed to the extremely volatile growth of COVID-19 cases and related data at the beginning of the pandemic. As time went on, the data became more consistent and the models performed better since trends began to emerge in total deaths and cases. Each of our models has this property, but it is clear that the non-parametric model is the best suited to our data. Its statistically significant coefficients and the higher R-squared shows that the non-parametric model has a moderately good fit with our data, especially when compared to linear models. There is a clear causal effect of institutional stability data and COVID-19 case and death data on change in GDP; however, the relationship is not linear and needed a much more flexible model to demonstrate the correlation. From our model, we can determine that a country's GDP was likely to experience a negative change throughout the pandemic due to an increase in COVID-19 cases and deaths and depending on their institutional stability.

Other areas of interest that we had with this data were questions about how GDP combined with COVID data could predict ratings of institutional stability. We originally wanted to examine how the relationships between our variables influenced each other in all directions; in addition to how our variables effected change in GDP, we wanted to examine the reverse causality as well and find information about the relationship between COVID data and stability rankings. However, we decided to refine our study and focus on finding meaningful correlations involving the effect on change in GDP. Now having found data on the relationships that can predict GDP, it would be interesting to determine if using GDP to predict other variables would have the same flexible nature as was demonstrated by this study.

0.4 Github

The code for this study can be found at:
`https://github.com/henrymanley/covidSTSCI`

References

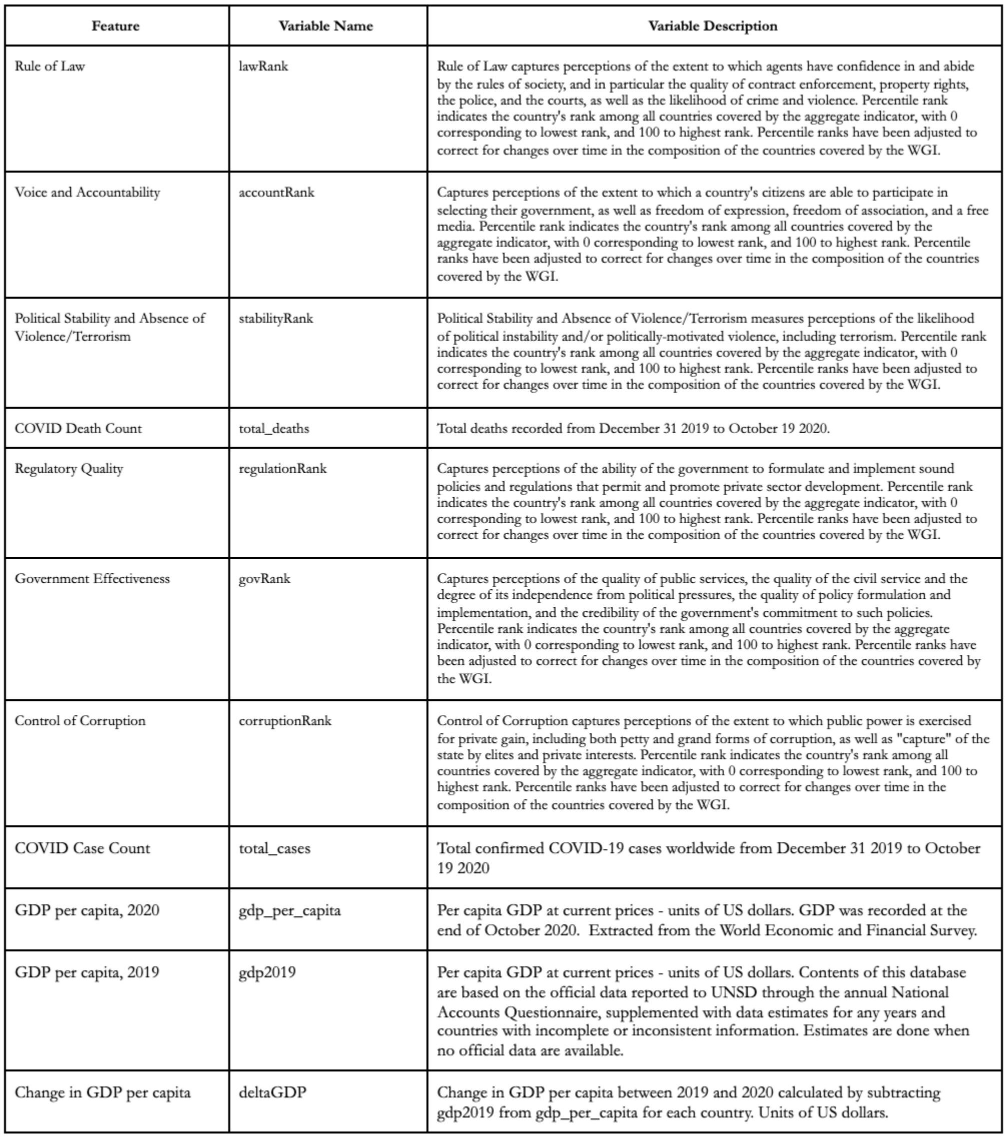
[1] Asare Vitenu-Sackey, P., Barfi, R. (2021). The Impact of Covid-19 Pandemic on the Global Economy: Emphasis on Poverty Alleviation and Economic Growth. The Economics and Finance Letters, 8(1), 32–43. https://doi.org/10.18488/journal.29.2021.81.32.43

[2] International Monetary Fund: World Economic Outlook Database. (2020, October). Report for selected countries and subjects. Retrieved from https://www.imf.org/en/Publications/WEO/weo-database/2020/October/weo-report

[3] The World Bank. (2019). Worldwide governance indicators. https://databank.worldbank.org/source/worldwide-governance-indicators

[4] United Nations Statistics Division (2019). UNdata — record View — per capita GDP at current prices - US dollars. //data.un.org/Data

.1 Appendix



Feature	Variable Name	Variable Description
Rule of Law	lawRank	Rule of Law captures perceptions of the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, property rights, the police, and the courts, as well as the likelihood of crime and violence. Percentile rank indicates the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank. Percentile ranks have been adjusted to correct for changes over time in the composition of the countries covered by the WGI.
Voice and Accountability	accountRank	Captures perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media. Percentile rank indicates the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank. Percentile ranks have been adjusted to correct for changes over time in the composition of the countries covered by the WGI.
Political Stability and Absence of Violence/Terrorism	stabilityRank	Political Stability and Absence of Violence/Terrorism measures perceptions of the likelihood of political instability and/or politically-motivated violence, including terrorism. Percentile rank indicates the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank. Percentile ranks have been adjusted to correct for changes over time in the composition of the countries covered by the WGI.
COVID Death Count	total_deaths	Total deaths recorded from December 31 2019 to October 19 2020.
Regulatory Quality	regulationRank	Captures perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development. Percentile rank indicates the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank. Percentile ranks have been adjusted to correct for changes over time in the composition of the countries covered by the WGI.
Government Effectiveness	govRank	Captures perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. Percentile rank indicates the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank. Percentile ranks have been adjusted to correct for changes over time in the composition of the countries covered by the WGI.
Control of Corruption	corruptionRank	Control of Corruption captures perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as "capture" of the state by elites and private interests. Percentile rank indicates the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank. Percentile ranks have been adjusted to correct for changes over time in the composition of the countries covered by the WGI.
COVID Case Count	total_cases	Total confirmed COVID-19 cases worldwide from December 31 2019 to October 19 2020
GDP per capita, 2020	gdp_per_capita	Per capita GDP at current prices - units of US dollars. GDP was recorded at the end of October 2020. Extracted from the World Economic and Financial Survey.
GDP per capita, 2019	gdp2019	Per capita GDP at current prices - units of US dollars. Contents of this database are based on the official data reported to UNSD through the annual National Accounts Questionnaire, supplemented with data estimates for any years and countries with incomplete or inconsistent information. Estimates are done when no official data are available.
Change in GDP per capita	deltaGDP	Change in GDP per capita between 2019 and 2020 calculated by subtracting gdp2019 from gdp_per_capita for each country. Units of US dollars.

Figure (8) Variable Analysis