The effects of HDI and GDP per capita on COVID deaths: A regression analysis

Rabia Gondur

December 15, 2022

1 Abstract

The COVID-19 pandemic had a drastic effect on many countries around the world. It changed the way that the world operates, from education to economy, and forced people to adjust to a new normal. The large amount of fatalities that COVID brought led researchers to investigate the possible causes behind the death rates due to pandemic. Although there are a number of studies that investigated a wide range of factors, this paper will focus on 2 categories and their effects on COVID fatalities: HDI and GDP per capita. This paper will investigate the effects of Human Development Index and GDP per capita on COVID deaths using multiple linear regression. Previous empirical work and cross sectional studies show that there is a negative correlation between the COVID cases/deaths and GDP per capita and a positive correlation between the COVID cases/deaths and HDI. Therefore, guided by these relationships, this paper will further perform regression analysis to see the variety of statistical and econometric findings. The hypothesis of this paper is that countries with higher GDP per capita and lower HDI are more likely to have lower COVID deaths. The results from the regression analysis display that Weighted Least Squares (WLS) is a better method for the regression compared to Ordinary Least Squares (OLS) for this dataset due to the issue with heteroscedasticity in the data. The overall analysis of the data shows that the relationships between the independent variables and the dependent variable align with the hypothesis of the paper, however, only HDI was statistically significant at the 5% level. Therefore, we fail to reject the null hypothesis for GDP per capita, but reject the null hypothesis for HDI.

2 Introduction

COVID-19 is a highly infectious disease that is caused by the SARS-CoV-2 virus. Due to its fast spread around the world, the COVID-19 pandemic was one of the deadliest outbreaks in recent human history. As of November 2022, it caused the death of more than 6 million people worldwide. Although it is hard to identify everything that contributed to the death rates, there are a variety of factors we can look at that could potentially shed a light on what contributed to COVID deaths. To name a few: economic status, education levels, access to healthcare, and age, are all potential factors that might have affected a person's outcome upon contracting COVID. Gross domestic product per capita, GDP per capita, is a measure of all the final goods and services in a country for a specific time period divided by the country's population. Since GDP per capita can be a good financial metric of a country, it can provide important insight into how the economic situation in a country can affect COVID fatalities. One example of the relationship between GDP per capita and COVID fatalities can be illustrated using the Preston curve [Figure 6]. The Preston curve is an important graph that demonstrates the relationship between income and life expectancy and is used in many areas of social science and econometrics [1]. The graph displays that countries with higher GDP per capita are more likely to have higher life expectancy compared to countries with lower GDP per capita. [2].

One study done by Pardhan et. al (2020) investigated the effects of GDP per capita on COVID cases for 38 European countries. Although they collected the data for 2 months, they were able to make some associations between a country's economic status and their likelihood of having COVID cases. They found that countries with higher GDP per capita had lower rates of new COVID cases, thus GDP per capita and COVID cases had an inverse relationship. Another study done by Coccia (2021), investigated the effects of COVID lockdown protocols and countries' healthcare spendings on COVID fatalities. One of their findings from the study was that countries that spend a higher percentage of their GDP on healthcare had less COVID fatalities as well as a better chance of shielding the negative effects of COVID on the overall economy [3].

Human development index, HDI, captures a collection of important metrics related to human development. It measures 3 key elements of human development, also called dimensions: economic status, life expectancy, and education. Life expectancy is the measured average number of years that a newborn is expected to live if they pass through the critical periods [4]. Economic status is measured through GNI, gross natural income, which reflects how much of the distribution of income among households, or individuals deviates from a perfectly equal distribution [4]. Finally, education is measured by the adult literacy rate. In the end all of these dimensions are equally weighted to give the final HDI. Although this metric does not capture every aspect of human progression and growth, it provides a convenient measure to summarize the general gross outlook of a country.

One study done by Mirahmadizadeh et. al (2022) investigated the relationship between human development index (HDI) and epidemiological indicators of COVID, ranging from rates of COVID cases to rates of COVID deaths for 157 countries [4]. The study showed that HDI had a positive correlation with the rate of COVID cases and rate of COVID deaths. Thus, the study displayed that as the HDI of a country increased, so did their rates of COVID cases and fatalities with a p-value < 0.001. [4]. Another study done by Valev (2020) looked into the relationship between the total COVID cases/deaths and 10 economic, demographic, and social indicators for 45 countries. One of these indicators was HDI. The study found a positive correlation between HDI and the total COVID cases/deaths. The correlation coefficient for COVID cases and deaths were 0.724 and 0.680 respectively, suggesting not only a positive correlation but a strong positive correlation between COVID cases/deaths and HDI [5].

After a brief literature review on the effects of HDI and GDP per capita on mortality, specifically COVID mortality, the following hypothesis can be made: Countries with higher GDP per capita and lower HDI are more likely to have lower COVID deaths.

3 Data Information

The data used in this paper was taken from *Kaggle* and it was titled: 'Impact of Covid-19 Pandemic on the Global Economy' [6]. This dataset had information from 2019-2020 on 170 countries of human development index(HDI), stringency index(STI), total deaths(TD), total cases(TD), population(POP) and GDP per capita(GDPCAP) [Table 1]. The new/transformed data had everything except the HDI in the natural log form [Table 2]. Thus lnPOP represents $ln(POP)$, lnGDPCAP represents $ln(GDPCAP)$, lnSTI represents $ln(STI)$, lnTD represents $ln(TD)$, and lnTC represents $ln(TC)$. There was no given explanation as to why some features were logged but one of the reasons could be due to the high multicollinearity of the original data. Another possible reason could be that many studies that investigated similar features to this paper also used logged form for the variables, suggesting that logged features perform better than the level features. The data frame was further modified for the regression analysis of this paper.

4 Data Analysis

The first step was to visualize the data in order to see the overall structure, *Table 1.* The data frame had Nan and 0 values in some columns, which were then cleaned from the data frame. Since the original data was a collection of almost two years of information for each country, the new data frame was organized in a way where each country would have only one year's worth of data. The reason why one of the years had to be eliminated was because only 51 countries out of 170 had information from 2019, thus in order to keep things equal for each country, the information from the year 2019 was omitted from the regression. In order for every country to have a single value for each features, the average of each category for each country was taken and reflected on the final data frame. The final cleaned and organized data had information from 151 countries from the year 2020. Therefore, the data was transformed into cross-sectional study. The proposed regression equation is:

$$ln(TD) = \beta_0 + \beta_1 HDI + \beta_2 ln(GDPCAP) + \epsilon \tag{1}$$

The null and alternative hypothesis is as follows:

$H_0 : \beta_1 = \beta_2 = 0$

$H_1 : \beta_1 \neq \beta_2 \neq 0$

4.1 Specifications

The first specification for this regression was the choice of independent and dependent variables. Section 2, Introduction, of this paper provided a brief review of the background and literature around COVID fatalities, HDI, and GDP per capita. Thus, the literature review and economic theories around the subject suggest that HDI and GDP per capita as independent variables and TD as dependent variable is a reasonable choice. Moreover, the empirical work also suggests that the expected sign of β_1 (coefficient of HDI) would be positive, and the expected sign of β_2 (coefficient of lnGDPCAP) would be negative. Null and alternative hypothesis for the expected sign of β_1 is as follows: $H_0 : \beta_1 \leq 0$ and $H_1 : \beta_1 > 0$. Null and alternative hypothesis for the expected sign of β_2 is: $H_0 : \beta_2 \geq 0$ and $H_1 : \beta_2 < 0$

The second specification for this regression was the functional form. When all of the data features are in the level form, the regression suggests a high multicollinearity. Therefore, most of the values were logged to help with multicollinearity and allow for better visualization. The theory also supports this decision.

The Preston curve shows a curved relationship between the income and life expectancy, thus GDPCAP and TD were both logged (and transformed to lnGDPCAP and lnTD respectively). The cleaned/transformed data already had TD and GDPCAP in the logged form, therefore they did not have to be changed. For the purposes of clarity, the column names were changed to better reflect the mathematical transformations. HDI was not logged. The reason for this is because when HDI gets logged, all of the HDI values become negative and it causes the regression to have a worse R^2 and p-values. Although both R^2 and p-values were low and close to each other for logged and not logged HDI cases, the not logged HDI was used since it had a relatively better performance.

The third specification is the properties of the error term. The OLS assumes several properties to be met for the error term. However, this data does not follow all of the assumptions. In this regression, there might be omitted variable bias, thus the independent variables might be correlated with the error term. This assumption comes from the fact that there are only two independent variables in the regression and the R^2 of the regression is quite low, suggesting that these independent variables might not be explaining enough of the variation in total deaths (more on this can be found in the *Results* section). The regression model is linear in its coefficients. The intercept is not suppressed hence it can be assumed that $E(\epsilon) = 0$. The multicollinearity was managed using the logged form of most of the variables thus it can be assumed that there is no perfect multicollinearity. The White test was performed on the data to check for heteroscedasticity and the results displayed that there was heteroscedasticity present on the data. The null and alternative for the White test is as follows: H_0 : Homoscedasticity is present, $p > 0.05$ and H_1 : Heteroscedasticity is present $p < 0.05$. Thus, to correct for the heteroscedasticity, Weighted Least Squares was used (more on this can be found in the *Results* section).

4.2 Results

The multiple linear regression model can be seen in *Equation 2* and the model summary can be seen in *Figure* 4. The equation displays that if GDPCAP increases by 1% , TD decreases by 0.285 percent, holding HDI constant. If HDI increases by 1 unit, TD increases by 564.8 percent, holding GDPCAP constant. The regression equation with coefficients after running linear regression using OLS is:

$$ln(TD) = 2.655 + 5.648(HDI) - 0.285(ln(GDPCAP)) \tag{2}$$

The α value was set to 5% since most of the previous empirical work on COVID cases/deaths vs HDI and GDP per capita used $\alpha = 0.05$ in their regressions. The p-value for GDP per capita and HDI was 0.458 and 0.058 respectively, therefore none of the independent variables were statistically significant at the 5% level. However, HDI was statistically significant at the 10% level. To check if HDI was significantly positive, the p-value of HDI was divided by 2 to get the one-tail. The one tail p-value for HDI was $0.058/2 = 0.029$. Thus, HDI was significantly positive at the 5% level. To check if lnGDPCAP was significantly negative, the p-value of lnGDPCAP was divided by 2 to get the one-tail. The one tail p-value for lnGDPCAP was $0.458/2 = 0.229$. Thus, lnGDPCAP was not significantly negative at the 5% level. The regression has an R^2 of 0.103 or 10.3%. This means that 0.103 or 10.3% of the variation in total death is explained in the regression. The Durbin-Watson coefficient (d) of the model is 2.251. Since d of 2 means no serial correlation, and the d coefficient of the model is only 0.251 more than this, we fail to reject the null hypothesis and conclude that there is no serial correlation.

The White test was performed to check for heteroscedasticity. The test statistic was 12.675 with a p-value $= 0.0266$. Since $0.0266 < 0.05$, we reject the null hypothesis, meaning that heteroscedasticity is present. One of the first methods to aid with heteroscedasticity is transforming the variables to their log form, however, since the variables are already transformed, that method does not provide much help. Another method that is widely used if the log transformation of the variables does not aid with heteroscedasticity is using Weighted Least Squares (WLS). The method used for this regression was Ordinary Least Squares (OLS) which makes certain assumptions, one of them being that there is constant variance in the errors. Since the White test displays that our data violates this assumption, a better and more accurate way of performing the regression would be by using WLS. Contrary to OLS, WLS does not assume that there is constant variance in the errors. Since WLS weights the observations proportional to the inverse of the error variance for that observation, it overcomes the issue of heteroscedasticity (non-constant variance). Upon performing WLS, a new model summary was formed (*Figure 5*), and the previous regression equation was changed to Equation 3:

$$ln(TD) = 0.051 + 11.336(HDI) - 0.430(ln(GDPCAP)) \tag{3}$$

This new multiple linear regression model with WLS shows that if GDPCAP increases by 1%, TD decreases by 0.430 percent, holding HDI constant. If HDI increases by 1 unit, TD increases by 1133.6

percent, holding GDPCAP constant. The p-value for GDP per capita and HDI was 0.354 and 0.005 respectively, therefore, only HDI was statistically at the 5% level. In fact, HDI was also statistically significant even at the 1% level. To check if HDI was significantly positive, the p-value of HDI was divided by 2 to get the one-tail. The one tail p-value for HDI was $0.005/2 = 0.0025$. Thus, HDI was significantly positive at the 5% and 1% level. To check if lnGDPCAP was significantly negative, the p-value of lnGDPCAP was divided by 2 to get the one-tail. The one tail p-value for lnGDPCAP was $0.354/2 = 0.177$. Thus, lnGDPCAP was not significantly negative at the 5% level. The R^2 for the new estimated regression with WLS was 0.255 or 25.5% which is better than the R^2 of the old estimated regression with OLS. The R^2 for the new estimated regression displays that 0.255 or 25.5% of the variation in the total death is explained by the regression. The Durbin-Watson coefficient (d) of the model is 2.593 which is still close to 2, thus the new estimated regression also does not allude to any serial correlation. The most important reasons behind why WLS was substituted for OLS was to fix the issue of heteroscedasticity, thus the White test was performed to check this. The White test showed that the test statistic was 6.328 with a p-value = 0.2756. Since $0.2756 > 0.05$, we fail to reject the null hypothesis. Thus homoscedasticity is present, and our data does not suffer from heteroscedasticity anymore.

5 Discussion/Conclusion

This paper investigated the effects of HDI and GDP per capita on COVID deaths. The literature review of various studies supported the hypothesis of this paper which was: Countries with higher GDP per capita and lower HDI are more likely to have lower COVID deaths. To test this hypothesis, first the data was cleaned, and then reorganized prior to running the regression. After that, OLS was used to get the regression model. However, upon further investigation, the White test showed that there was heteroscedasticity present in the data. Therefore, WLS was used to get a more accurate representation of the regression and to fix the heteroscedasticity. After performing WLS, the issue with heteroscedasticity was solved. The new regression equation, *Equation 3*, displayed that HDI had a linear relationship with lnTD and GDPCAP had an inverse relationship with lnTD. This agrees with the previous empirical work on HDI and GDP per capita, since higher HDI is associated with higher COVID deaths and higher GDP per capita is associated with lower COVID deaths. However, the only variable that was statistically significant at the 5% was HDI. Thus, although the final regression shows that increase in HDI is associated with increase in total deaths (p-value=0.005), there isn't enough evidence to show that increase in GDP per capita is associated with decrease in total deaths. Although the literature shows that higher HDI leads to more deaths, it is hard to justify the reason behind this since higher HDI would mean higher scores for education, life expectancy, and economy. However, one reason for why the literature and this analysis show that higher HDI is associated with higher deaths could be because of how HDI is calculated. Since HDI is a cumulative result of 3 dimensions, each of them might be effecting the result and causing it to skew in a certain direction, hence a higher HDI is associated with higher TD. One way to see a clearer effects of HDI on COVID deaths for the future studies could be examining the individual scores for each dimensions as well as the overall score. Another conclusion from the this study was that GDP per capita did not appear to have a statistically significant effect. One of the reasons why it didn't could be the lack of data, or the way the data was collected. A future study can investigate a wider range of datasets and get the GDP per capita information of more countries.

6 Appendix

Table 1: *First 5 rows of the raw data, not logged*

COUNTRY	DATE	TC	TD	STI	POP	GDPCAP	HDI
Afghanistan	2020	16374.541219	520.383513	49.316165	38928341.0	1803.987	0.498
Albania	2020	4104.860577	120.759615	68.107452	2877800.0	11803.431	0.785
Algeria	2020	21109.051402	904.654206	75.205374	43851043.0	13913.839	0.754
Angola	2020	1200.652850	48.637306	79.024041	32866268.0	5819.495	0.581
Argentina	2020	176960.654206	3876.794393	88.403318	45195777.0	18933.907	0.825
...	...	...	...	...	...	...	...
Venezuela	2020	22257.207921	187.816832	83.926188	28435943.0	16745.022	0.761
Vietnam	2020	1005.950000	28.975000	68.137000	97338583.0	6171.884	0.694
Yemen	2020	1299.987879	364.400000	44.279758	29825968.0	1479.147	0.452
Zambia	2020	4909.537634	116.741935	49.719624	18383956.0	3689.251	0.588
Zimbabwe	2020	2827.114286	76.114286	79.827762	14862927.0	1899.775	0.535

Table 2: *First 5 rows of the updated data, everything except HDI is logged*

COUNTRY	DATE	HDI	lnTC	lnTD	lnSTI	lnPOP	lnGDPCAP
Afghanistan	2020	0.498	6.828144	4.187536	3.060082	17.477233	7.497754
Albania	2020	0.785	7.565514	4.130528	3.927348	14.872537	9.376146
Algeria	2020	0.754	7.005156	4.799373	3.206073	17.596309	9.540639
Angola	2020	0.581	5.679145	2.783673	4.112616	17.307957	8.668969
Argentina	2020	0.825	9.268997	6.045323	4.054678	17.626514	9.848710
...	...	...	...	...	...	...	...
Venezuela	2020	0.761	8.268582	3.895227	4.343886	17.163165	9.725856
Vietnam	2020	0.694	4.950369	0.892618	3.723534	18.393706	8.727759
Yemen	2020	0.452	6.027499	4.808522	3.609601	17.210890	7.299221
Zambia	2020	0.588	7.165626	3.383976	3.617026	16.726989	8.213179
Zimbabwe	2020	0.535	6.196063	2.902239	4.363253	16.514381	7.549491

Figure 1: *Correlation Table*

	DATE	HDI	lnTC	lnTD	lnSTI	lnPOP	lnGDPCAP
DATE	NaN	NaN	NaN	NaN	NaN	NaN	NaN
HDI	NaN	1.000000	0.287971	0.315680	-0.402258	-0.126487	0.955126
lnTC	NaN	0.287971	1.000000	0.890903	0.047296	0.567412	0.264714
lnTD	NaN	0.315680	0.890903	1.000000	-0.141021	0.604921	0.284340
lnSTI	NaN	-0.402258	0.047296	-0.141021	1.000000	-0.103188	-0.392106
lnPOP	NaN	-0.126487	0.567412	0.604921	-0.103188	1.000000	-0.143980
lnGDPCAP	NaN	0.955126	0.264714	0.284340	-0.392106	-0.143980	1.000000

Figure 2: *A scatter plot of 151 countries' HDI vs total death*

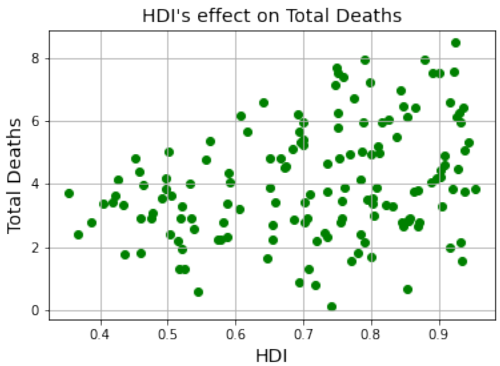


Figure 3: *A scatter plot of 151 countries' GDP per capita vs total death, where GDPCAP is in log form*

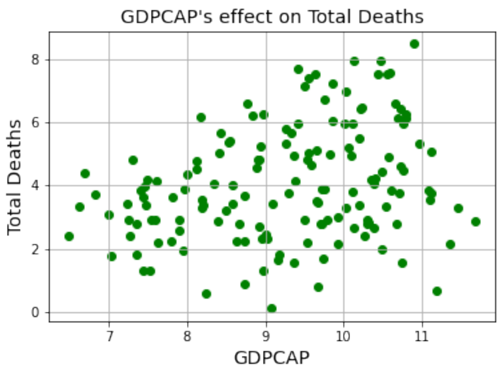


Figure 4: *Model Summary, OLS*

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     TD   R-squared:                       0.103
Model:                            OLS   Adj. R-squared:                  0.091
Method:                 Least Squares   F-statistic:                     8.499
Date:                Thu, 01 Dec 2022   Prob (F-statistic):           0.000321
Time:                        11:30:36   Log-Likelihood:                -292.50
No. Observations:                 151   AIC:                             591.0
Df Residuals:                     148   BIC:                             600.0
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.6551      1.660      1.599      0.112      -0.625       5.936
GDPCAP        -0.2851      0.383     -0.745      0.458      -1.041       0.471
HDI            5.6476      2.953      1.912      0.058      -0.188      11.483
==============================================================================
Omnibus:                        2.476   Durbin-Watson:                   2.251
Prob(Omnibus):                  0.290   Jarque-Bera (JB):                1.836
Skew:                           0.081   Prob(JB):                        0.399
Kurtosis:                       2.485   Cond. No.                         223.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 4.1: *White Test Results (OLS)*

```
Test Statistic': 12.67517178031914, 'Test Statistic p-value': 0.026620552484670722
```

Figure 5: *Model Summary, WLS*

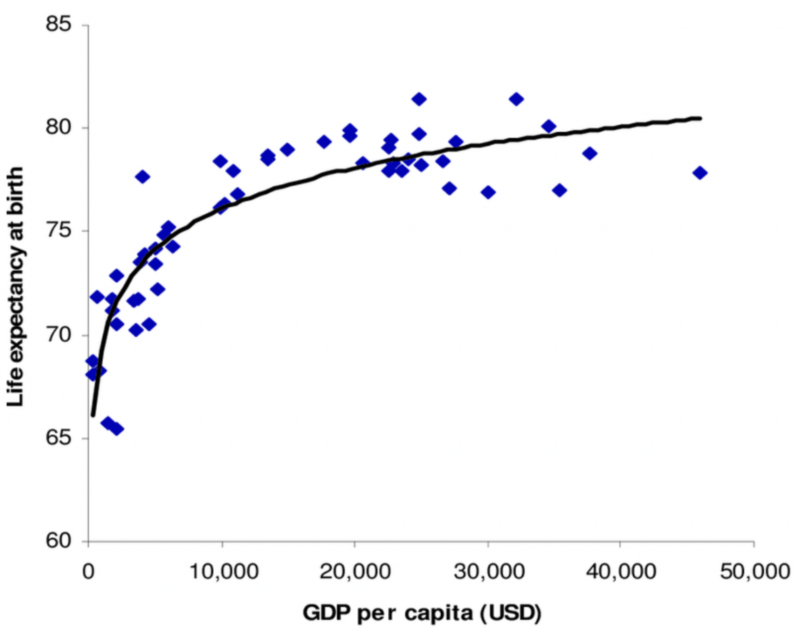
```
                            WLS Regression Results
==============================================================================
Dep. Variable:                   lnTD   R-squared:                       0.255
Model:                            WLS   Adj. R-squared:                  0.244
Method:                 Least Squares   F-statistic:                     25.26
Date:                Mon, 05 Dec 2022   Prob (F-statistic):           3.64e-10
Time:                        20:39:07   Log-Likelihood:                -441.98
No. Observations:                 151   AIC:                             890.0
Df Residuals:                     148   BIC:                             899.0
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0506      1.719      0.029      0.977      -3.346       3.447
lnGDPCAP      -0.4300      0.463     -0.929      0.354      -1.345       0.485
HDI           11.3362      3.997      2.836      0.005       3.437      19.235
==============================================================================
Omnibus:                       57.712   Durbin-Watson:                   2.593
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1179.890
Skew:                           0.718   Prob(JB):                    6.17e-257
Kurtosis:                      16.619   Cond. No.                         256.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 5.1: *White Test Results (WLS)*

```
Test Statistic': 6.327527382375161, 'Test Statistic p-value': 0.2756406942188171
```

Figure 6: *Illustration of the Preston Curve from Lewis et. al (2009)*



References

[1] W. Jack and M. Lewis, *Health investments and economic growth: Macroeconomic evidence and microeconomic foundations*. World Bank Washington, DC, 2009.

[2] J. Virk and D. Holmes, "Effect of healthcare systems on inequality in the uk and the us," *Open Journal of Social Sciences*, vol. 10, no. 11, pp. 1–8, 2022.

[3] M. Coccia, "The relation between length of lockdown, numbers of infected people and deaths of covid-19, and economic growth of countries: Lessons learned to cope with future pandemics similar to covid-19 and to constrain the deterioration of economic system," *Science of The Total Environment*, vol. 775, p. 145801, 2021.

[4] A. Mirahmadizadeh, M. Ghelichi-Ghojogh, M. Vali, K. Jokari, H. Ghaem, A. Hemmati, F. Jafari, S. S. Dehghani, A. H. Hassani, A. Jafari, *et al.*, "Correlation between human development index and its components with covid-19 indices: a global level ecologic study," *BMC Public Health*, vol. 22, no. 1, pp. 1–8, 2022.

[5] D. Valev, "Relationships of total covid-19 cases and deaths with ten demographic, economic and social indicators," *medRxiv*, 2020.

[6] P. A. Vitenu-Sackey and R. Barfi, "The impact of covid-19 pandemic on the global economy: emphasis on poverty alleviation and economic growth," *The Economics and Finance Letters*, vol. 8, no. 1, pp. 32–43, 2021.