

Factors Affecting COVID-19 and the Impact of Different Dimensions

Mingxuan Du ^{1,†}, Yujia Hao ^{2,*}, Ruohan Wu ^{3,†}

¹ School of gifted young University of Science and Technology of China, Hefei, China

² Department of Mathematics University of Sydney, Sydney, Australia

³ Department of Economics, University of Melbourne, Melbourne, Australia

* Corresponding Author Email: yhao9808@uni.sydney.edu.au

[†]These authors contributed equally.

Abstract. As COVID-19 unfolds, it strongly affects individuals, nations from all aspects worldwide, e.g., life quality, country's population, students' study, household income, GDP, tourist industry. This paper analyzes a detailed analysis from three directions: tourism industry, population, and student academic performance during COVID-19. As for the student performances, four different factors that affect children's academic performance are analyzed. According to the analysis, father's education level, family income, and student achievement are positively significant, whether it is COVID-19 and student achievement are negatively significant. Regarding to country's population, the correlation between COVID-19 new cases and the effect of population, median age, and GDP per capita in each country are estimated. Based on Linear regression model for total cases, it shows the population size and GDP per capita have positive effect on increase in total cases among countries while the median age does not have significant effect. For the last case of tourism industry, data from provinces and states of China as cases are used and linear models on cross-sectional data from 2020 to 2021 are constructed, the findings reveal that the tourism revenue is significantly affected by the number of locally confirmed cases of the COVID-19. Besides, the results also show that the amount of national 5A scenic sites, longitude, and latitude all have a major effect on city tourist revenue, while the COVID-19 pandemic has resulted in a decrease in tourism revenue. Overall, these results shed light on guiding further exploration of investigating the impacts of COVID-19.

Keywords: COVID-19; Student Performance; Population; GDP; tourism industry; Linear Regression Model.

1. Introduction

Corona virus disease (COVID-19) is a disease caused by severe acute respiratory syndrome known as corona virus [1]. As COVID-19 unfolds, it has negatively affected the daily life of human beings, education, the country's tourism, Economics, and the overall development of the world [2-6]. On 15 June 2020, this pandemic affected nearly 8 million people. On this date, as the World meter corona virus updates information showed that we have 7,984,067 total COVID-19 cases globally and 435,181 deaths [7].

During COVID-19 Period, students are required to learn remotely during the coronavirus, and teachers have to teach remotely through software. The education and teaching process realized through the network, teachers, and students are not required to be strictly asynchronous in time and space, which objectively results in poor teaching interaction. In addition, the quality of online teaching, students' learning attitudes, and psychological and external conditions have been affected. In addition, household income and parental education levels are both more or less affected by children's grades [8, 9]. Moreover, different countries are at different stages of the COVID-19 pandemic and CFR can change depending on how each country responds to the epidemic. For example, UK, as of 25th June, had over 306,000 confirmed cases with over 43,000 deaths (CFR: 14.0%). With the discovery of COVID-19, many scenic areas have to be closed, people cannot travel to the other city in holidays.

In this study, all datasets are obtained from Kaggle. Through a survey of college students in China, it is found that over 60% of students believe that online learning attitude is the primary influencing factor of learning effect, and more than 30% of students are dissatisfied with online teaching. Most students feel that online learning is interesting and less active, and other factors have seriously affected students' academic performance. In 205 countries and the number of populations, median ages and GDP per capita from starting date of each country until 31st May, 2020. And, there has been an unprecedented decline in tourism revenue because of plummeting tourist arrivals. Some studies have been done to find the impact of a global pandemic using a series of quantitative analysis (ARIMA, correlation, and regression) on longitudinal data [1]. Subsequently, this study aimed to achieve the following objectives:

To explore the relationship between children's performance and four factors, namely household income, whether the child had Covid, the father's education level, and the mother's education level, and how these four different factors affect children's academic performance.

To analyze the relation between total COVID-19 new cases based on population in each country, median age and GDP per capita using a linear regression model.

To figure out the relationship between tourism income of cities and the injected population based on linear model and explore the impact of four factors (local cumulative number of confirmed cases, the number of National 5A Scenery site of the province, latitude and longitude on the tourism revenue.

The rest part of the paper is organized as follows. The Sec. II. will Student Performance. The Sec. III will COVID-19 new cases based on population in each country. The Sec. IV will Tourism income of cities and the injected population. Eventually, a brief summary will be given in Sec. V.

2. Student Performance

2.1. Data

Data for studying student achievement can be searched from Kaggle, which records variables including student numbers, gender, family income, parents' education level, whether they have COVID-19, and students' grades during a semester during the epidemic. There are a total of 1401 students in this data. By calculating the average grades of the three total courses of the students as the dependent variable of this survey, the independent variable will select four dependent variables, namely family income, father's education level, mother's education level, and whether the student has COVID-19 Case. A multiple linear regression model will be used to analyze this data and established to analyze the mutual independence of the data by the R-square of the data, VIF, Durbin Watson, un-standardized regression coefficient, standardized regression coefficient, P-value, multiple multicollinearities. Furthermore, Conformity to normal distribution and significance of variables to verify the model's integrity. From Table. I, the R square is 0.470, and the adjusted R square is 0.468, which means that four variables explain 47% of the dependent variable-Average score. Besides, the MSE is 0.256743 in TABLE I. It represents a very high degree of matching between the actual value and the predicted value, which also means that this study is reasonable.

Table 1. Model Summary

Model Summary						
Mode	R	R Square	Adjust R Square	Significant F change	Durbin–Watson statistic	Mean squared error (MSE)
1	0.685479	0.469882	0.468362	1.8479E-190	1.972	0.256743

Table 2. Coefficients summary

Model	Unstandardized coefficient	Statistical Significance	Multicollinearity Statistic
	B		VIF
(Constant)	73.071	0	
Covidpos	-2.655	4.1559E-10	1.125336
Zscore: household income	5.853	1.8819E-148	1.128389
Father education	1.643	2.1147E-14	1.045232
Mother education	0.485	0.064665	1.044963
Independent Variable: Average Score			

2.2. Build Multiple Linear Regression

Multiple Linear Regression is built to Find Relationships Between Variables and Dependent Variables.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i \quad (1)$$

The expression of the multivariate linear model can be derived from the un-standardized regression coefficients in Table. II, which is

$$Y_i = 73.071 - 2.655X_{1i} + 5.853X_{2i} + 1.643X_{3i} + 0.486X_{4i} + \epsilon_i \quad (2)$$

2.3. Analyze the Significance of Variables

By observing the significance values of the four variables, their respective significance is obtained. The coefficients of the three variables covidpos, household income, and father education are all less than 0.05, which means that these three variables are all significant. The variable's (covidpos) unstandardized regression coefficient is -2.655, and variable (covidpos) is very significant again. By analysing these two values, it can be concluded that the variable (covidpos) is negatively correlated with the average grade of the dependent variable and is highly significant. This means that students with covid will have lower grade point averages than students without covid.

The unstandardized regression coefficients for the two variables household income and father education are 5.853 and 1.643, respectively, more significant than zero. These two positive values mean that the two variables a and b positively correlate with the dependent variable, student grades. The larger the two variables (a and b) are, the higher the student's average grade. On the contrary, the lower the students' average grades. The significance coefficient of the variable mother education is more significant than 0.05, which means that the variable has no apparent significance with the average grade of the dependent variable.

2.4. Evaluations of the model

Verifying that the established model is meaningful, the detection is consistent with the characteristics of linear regression and correctly corresponds to the content of previous reports. The Durbin-Watson statistic always has a value between 0 and 4. A value of 2.0 means that no autocorrelation was detected in the sample. Values from 0 to less than 2 indicate positive autocorrelation, and values from 2 to 4 indicate negative autocorrelation [7]. To satisfy the regression analysis, it is also necessary to ensure that each sample is independent of each other. Seen from Table. I, the Durbin Watson statistic value is 1.972, which is approximately 2. Then, it can be shown that these samples are independent of each other.

In order to test whether the model has multicollinearity, check the size of VIF. Usually, ten is the judgment boundary; when $VIF < 10$, there is no multicollinearity; when $10 \leq VIF < 100$, there is a robust multicollinearity Linear; when $VIF \geq 100$, the model has severe multicollinearity. The four

variables covidpos, household income, father education, mother education calculate the corresponding VIF value through the VIF formula shows below.

$$VIF_i = \frac{1}{1-R_i^2} \quad (3)$$

The calculated VIF values of the four variables are the values in the last column of Table II. All VIF values are less than 10, which means there is no multicollinearity between the variables, which also means this model fits a linear regression model. As shown in Fig.1, most of the residuals are in line with the normal distribution of residuals. Although a small part stands out, these residuals are randomly scattered and distributed, and they can also conform to the normal distribution of residuals.

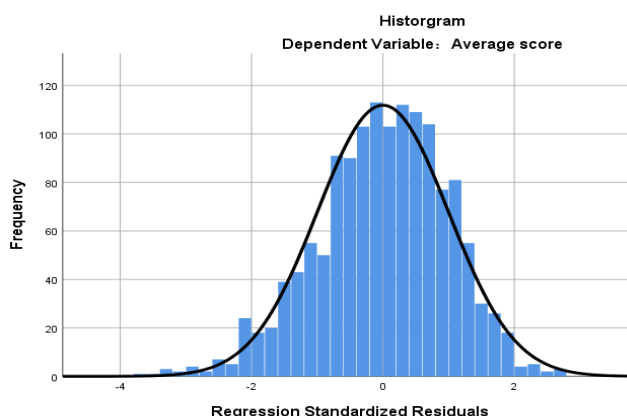


Figure 1. Residual histogram.

3. COVID-19 new cases based on population in each country

3.1. Data

The COVID-19 report data were collected from Kaggle. The time period of data was from starting date of each country to 31st May, 2020. The data included the total number of new cases, new cases per million, total number of new death and new deaths per million, date of recorded. Additionally, the datasets records variables such as population, median age, age 65 older, age 70 older GDP per capital and diabetes prevalence from 205 countries and others. In this study, linear regression model is adopted to analyze the relationship between total cases and population, median age and GDP per capita based on the available data from starting date of each country to 31st May, 2020.

3.2. Regression Model

In fact, regression analysis may be the most widely used statistical technique. It is a statistical modelling to deal with the relationship between variables for investors. The regression model has many variants such as linear regression, polynomial regression, and others used to model applications in numerous fields, including biological sciences, finance, insurance areas and others.

The Multiple linear regression model was used in this study to determine the relations between total cases and variables for COVID-19 from starting date of each country to 31st May, 2020 in 205 countries.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad (4)$$

where Y is COVID-19 new cases until 31st May, 202; X_1, X_2 and X_3 are independent variables with their corresponding coefficients β_1, β_2 and β_3 ; and β_0 is the intercept coefficient in this model.

3.3. Empirical analysis

Datasets can be searched from Kaggle. This data records variables such as population, median ages and GDP capita. There are a total of 205 countries in the datasets. And cross-sectional is used to

analyze in this study and the total cases is recorded from starting date of each country to 31st May, 2020. A multiple linear regression model will be used in this data and established to analyze the mutual independence of the data by the R-square of the data, VIF, Durbin Watson, standardized regression coefficient, significance and multicollinearities. Furthermore, Conformity to normal distribution and significance of variables to verify the model's integrity. As shown in Table. III, the value of R square is 0.114 and the adjusted R square is 0.099 which means that three variables explain 11.4 % of the dependent variable-Average score.

Table 3. Model Summary

Model Summary						
Mode	R	R Square	Adjust R Square	Significant F change	Durbin–Watson statistic	Mean squared error (MSE)
1	0.338	0.114	0.099	1.39837.287	2.031	0.74180997

Table 4. Coefficients Summary

Model	Unstandardized coefficient	Statistical Significance	Multicollinearity Statistic
	B		VIF
(Constant)	-20716.728	0.615	
population	0.000	0.000	1.010
median_age	528.328	0.735	1.723
Gdp_per_capita	1.445	0.045	1.725
Dependent Variable: total_cases			

3.4. Evaluations

The Durbin-Watson statistic in Table III is 2.031 indicating that there is no autocorrelation detected in the sample. Values from 0 to less than 2 points to positive autocorrelation and values from 2 to 4 means negative autocorrelation [7]. It is important to ensure each sample is independent with each other when building multi-linear regression. The Durbin-Watson statistic satisfies the basic assumption in this study. In order to test whether the model has multicollinearity, usually, we check the size of VIF. The three variables population, median age and GDP per capita correspond to VIF = 1.010, 1.723 and 1.725 respectively. It indicates that there is no multi-collinearity between 3 variables. The significance coefficient of the variable median age is 0.735, which is more than 0.05 indicating that the variable has no apparent significance with the total new cases. The rest two variables (population and GDP per capita) have p-value equal to 0.000 and 0.045, respectively. It indicates that median age is better to be removed when building up linear regression model in this study.

3.5. Build Multiple Linear Regression

Build Multiple Linear Regression to Find Relationships Between Variables and Dependent Variables. Based on Figure 2, the multivariate linear model can be derived from the standardized coefficients.

The variable X_1 (population) standardized regression coefficient is 0.272 and the variable X_2 (gdp_per_capita) is 0.213. By analyzing these two values, it can be concluded that the variables are positively correlated with the total cases in each country. The expression of the multivariate linear model can be derived:

$$Y = -7723.454 + 0.272X_1 + 0.213X_2 \quad (5)$$

where Y is COVID-19 new cases until 31st May, 2020; X_1, X_2 and X_3 are independent variables with their corresponding coefficients β_1, β_2 and β_3 ; β_0 is the intercept coefficient in this model.

Table 5. Coefficients Summary Adjusted

Model	Unstandardized coefficient	Statistical Significance	Multicollinearity Statistic
	B		VIF
(Constant)	-7723.454	0.611	
population	0.000	0.000	1.003
Gdp_per_capita	1.602	0.004	1.003
Dependent Variable: total_cases			

4. Tourism income of cities and the injected population

4.1. Data

There are totally 31 columns in the dataset. The tourism revenue data were collected from ceicdata.com, and the COVID-19 report data were collected from Kaggle. Then they were combined. The time period of data was from starting date of each country to the end of 2020. The data include the local cumulative number of confirmed cases, the number of National 5A Scenery site of the province, latitude and longitude. In this project, the tourism revenue is set as the response variable and use linear regression model to analyze the relationship between the response variable and four explanatory variables we have mentioned. Data was found from ceicdata.com and Kaggle.com and some combination was done.

4.2. Model

The Multiple linear regression model was built to determine the relations between the response variable and explanatory variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \quad (6)$$

where Y is the tourism Revenue of the province, X₁ is the local cumulative number of confirmed cases, X₂ is the number of National 5A Scenery site of the province, the X₃ is the Latitude, X₄ is the Longitude.

4.3. Empirical analysis

The expression of the multiple-linear model can be derived from the coefficients in above table, which is

$$Y = -456.759 - 0.086 X_1 + 32.809 X_2 - 14.742 X_3 + 9.477 X_4 \quad (7)$$

Viewing the above table, we accept all four regression coefficients are significant with test level $\alpha=0.1$. Moreover, we accept X₁ is significant in test level $\alpha=0.001$. From the regression analyses, $R^2=0.66$ and adjusted $R^2=0.6077$, which suggested 66% of variation of the tourism revenue could be explained by the four factors. The p value is 7.76e-06, which is small enough to suggest that the regression equation is significant. According to correlation analysis, the p-value of test of correlation between tourism revenue and confirmed cases is 0.8085 which means linear dependence relation is not strong, and the MSE is 22428.84. With this in mind, we still need larger dataset to optimize the model,

5. Conclusion

In summary, this paper investigates factors influencing COVID-19 cases as well as the impact of COVID-19 on student performance and the tourist industry. To be specific, this paper shows there was a significant positive correlation between COVID-19 cases and different related variables. In the multiple linear regression model, variables such as the number of populations from each country

significantly affected COVID-19 new cases, while the median age does not show much significance. According to the analysis, the researcher recommended that government administrative give more awareness and protection to society, as well as pay more attention to economic development. In general, the findings of this study may aid countries in developing pandemic preparedness plans. Based on the evaluation, COVID-19 cases caused a huge decline in tourism revenue over a prolonged period of time. In this regard, the tourism industry should take steps to recover from the crisis caused by the COVID-19 pandemic. By analyzing the data for studying students' achievement, it is concluded that the family economic level and education level impact children's academic performance and have a positive impact. Furthermore, COVID-19 also negatively impacts children's academic performance.

There are also limitations to studying the relationship between average student achievement and four different factors. There are too few factors to represent students across the world. Students don't get enough grades to represent all their classmates in the world. Neither family income nor parental education is representative of the world, just this school. Since there may be some students who are poor in their families and whose parents are not highly educated, but have excellent academic performance. Besides, the formula of this construction only considers four variables, and the degree of confidence is low. Nevertheless, this research has some defects and shortcomings. In detail, due the limitations of original datasets, there are only 2 factors to represent the possible increase in total cases which causes the relatively small R^2 . Additionally, across world comparisons between countries are only possible when simultaneously considering while the virus does not arrive in all countries simultaneously. Moreover, the absolute numbers are incomparable due to different population sizes. In the third model, the p value of the correlation test between the number of COVID-19 cases and tourism revenue is large, indicating that the linear link is weak. Furthermore, the dataset utilized to determine the linear relationship is insufficient. Nevertheless, expanding the dataset is difficult because there is no unified index that reflects the tourist resources of cities across countries, while the index is a necessary variable in the model.

In the future, more information on COVID-19 will become accessible, thereby we can fit better models by incorporating additional variables. Overall, our findings show that COVID-19 has a detrimental impact on student performance and tourism income, and they serve as a warning to authorities to take steps to eliminate those negative effects.

References

- [1] A. Argawu, "Regression Models for Predictions of COVID-19 New Cases and New Deaths in Ethiopia", *International Journal of Theoretical and Applied Mathematics*. Vol. 6, No. 5, 2020, pp. 53-63.
- [2] M. B. Fini, what dentists need to know about COVID-19 [J]. *Oral oncology*, vol. 105, 2020, 104741.
- [3] G. Goggin, and K. Ellis. "Disability, communication, and life itself in the COVID-19 pandemic." *Health Sociology Review* vol. 29.2, 2020, pp. 168-176.
- [4] M. W. Hussain, T. Mirza, and M. M. Hassan. "Impact of COVID-19 pandemic on the human behavior." *International Journal of Education and Management Engineering* vol. 10.8, 2020), pp. 35-61.
- [5] D. Banerjee, "Psychological preparedness for the COVID-19 pandemic, perspectives from India." *Psychiatry research* vol 288, 2020, 112999.
- [6] Karasmanaki, Evangelia, and Georgios Tsantopoulos. "Impacts of social distancing during COVID-19 pandemic on the daily life of forestry students." *Children and youth services review* vol. 120, 2021, 105781.
- [7] W. Kenton, "Durbin Watson Statistic Definition", Investopedia, 2021. [Online]. Available: <https://www.investopedia.com/terms/d/durbin-watson-statistic.asp>.
- [8] S. Gross, "Early Mathematics Performance and Achievement: Results of a Study Within a Large Suburban School System", *The Journal of Negro Education*, vol. 62, no. 3, p. 269, 1993.
- [9] Z. Li and Z. Qiu, "How does family background affect children's educational achievement? Evidence from Contemporary China", *The Journal of Chinese Sociology*, vol. 5, no. 1, 2018.

- [10] W. Lim and W. To, "The economic impact of a global pandemic on the tourism economy: the case of COVID-19 and Macao's destination- and gambling-dependent economy", *Current Issues in Tourism*, vol. 25, no. 8, pp. 1258-1269, 2021.