Using Multilinear Regression to Identify Novel Factors Associated with COVID-19

Xiaohan Chen

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation

This Thesis is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.
Using Multilinear Regression to Identify Novel Factors Associated with COVID-19 Death Rate

By

Xiaohan Chen

A Thesis
Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science

Major: Computer Engineering

The University of Memphis
December 2021
Copyright© Xiaohan Chen

All rights reserved.
ACKNOWLEDGEMENT

I would like to take this opportunity to deliver my thankful to those people who helped and supported me during the research and writing process. Firstly, Professor Dr. Banerjee, for his guidance and support throughout the research and writing of the thesis. He used to encourage me to move forward and inspired me into a higher level and complete my graduate education. Secondly, I would like to express my appreciation to my committee for their efforts and contributions to the work. Finally, I would also like to extend my gratitude to my family that fully support throughout my graduate life.
ABSTRACT

As of November 2021, more than five million people have died worldwide due to COVID-19. In this thesis, we consider a multilinear regression model to identify a small set of novel factors associated with COVID-19 death rate in 168 countries. From well-established sources, we collected data on eight factors encompassing death rate, physical and mental health, and economic and political status. Upon satisfying the assumptions, the multilinear regression model selected three out of the eight factors: obesity level, global freedom score, and per capita nominal GDP. While obesity has been identified by other studies as a risk factor for COVID-19 death, the other two selected factors are novel and associate the attitude and lifestyle of people of different countries with COVID-19 death rate. This association may help governments to devise policies to mitigate the spread of infection due to COVID-19 as well as other pandemics.
Contents
1 Introduction......................................................................................................................... 6
2 Data Collection .................................................................................................................. 7
   2.1 Death Rate (per 1M Pop.).......................................................................................... 7
   2.2 Estimating Excess Mortality...................................................................................... 7
   2.3 Happiness Level ....................................................................................................... 9
   2.4 Physical Activity Level............................................................................................. 10
   2.5 Food Energy Intake Level (kilocalories per capita) ................................................. 11
   2.6 Obesity Level ........................................................................................................... 11
   2.7 Life Expectancy ....................................................................................................... 11
   2.8 Per Capita nominal GDP ......................................................................................... 12
   2.9 Per Capita Income ................................................................................................... 12
   2.10 Human Freedom Index ......................................................................................... 13
   2.11 Economic Freedom Index ...................................................................................... 13
   2.12 Global Freedom Scores (Political Rights, Civil Liberties), Internet Freedom Scores,
       Democracy Scores ................................................................................................. 14
3 Multilinear Regression ...................................................................................................... 15
   3.1 The Model ............................................................................................................... 15
   3.2 Assumptions in Regression ...................................................................................... 15
   3.3 Checking and Fixing the Assumptions if Violated ..................................................... 16
       3.3.1 Linear and Additive ....................................................................................... 16
       3.3.2 Statistical Independence of Residuals ............................................................. 17
       3.3.3 Heteroskedasticity ....................................................................................... 18
       3.3.4 Multicollinearity ........................................................................................... 20
       3.3.5 Variance Inflation Factor .............................................................................. 21
       3.3.6 Normal Distribution of Error Terms ............................................................... 22
   3.4 Comparing Multilinear Regression With and Without Satisfying Assumptions ......... 26
4 Conclusions ...................................................................................................................... 27
1 Introduction

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. It has caused more than five million deaths worldwide. One of the ways to reduce death rate and slow down transmission is to use statistical methods to analyze the data and identify the factors highly correlated with death rate, and then respond appropriately to these factors. The contributions of this thesis are as follows:

1. We reviewed the literature on identifying factors related to COVID-19 death rate. We selected eight factors encompassing death rate, physical and mental health, and economic and political status, only some of which have been considered in other studies. We collected data for 168 countries corresponding to these eight factors from trusted sources.

2. We implemented a multilinear regression model for this data. Upon satisfying all the assumptions of multilinear regression, the model selected three out of the eight factors: obesity level, global freedom score, and per capita nominal GDP. While obesity has been identified by other studies as a risk factor for COVID-19 death, the other two selected factors are novel and associate the attitude and lifestyle of people of different countries with COVID-19 death rate.

3. We show that the prediction from the multilinear regression model that satisfies all the assumptions is more accurate than the same model but not satisfying the assumptions, even though the latter employs more variables than the former.
2 Data Collection

2.1 Death Rate (per 1M Pop.)

The total number of deaths due to COVID-19 in a country is the cumulative number of deaths among detected COVID-19 cases. The cumulative number used in this thesis was reported on September 24, 2021. The list of countries is based on the United Nations Geoscheme [31]. Death rate of a country is the total number of deaths divided by the population of that country multiplied by one million.

2.2 Estimating Excess Mortality

The approach of estimating the excess COVID-19 death rate is based on measurement of the excess death rate during the pandemic week by week, compared to the expectation based on past trends and seasonality. However, the total excess death rate is not equal to the excess COVID-19 death rate. Excess mortality is influenced by six drivers of all-cause mortality that are related to the pandemic and the social distancing mandates that come with the pandemic. These six drivers are [1]:

a) The excess COVID-19 death rate, that is, all deaths directly related to COVID-19 infection;

b) The increase in mortality due to needed health care being delayed or deferred during the pandemic;

c) The increase in mortality due to increases in mental health disorders including depression, increased alcohol use, and increased opioid use;

d) The reduction in mortality due to decreases in injuries because of general reductions in mobility associated with social distancing mandates;
e) The reductions in mortality due to reduced transmission of other viruses, most notably influenza, respiratory syncytial virus, and measles;
f) The reductions in mortality due to some chronic conditions, such as cardiovascular disease and chronic respiratory disease, occur when frail individuals who would have died from these conditions died earlier from COVID-19 instead.

Excess mortality has been suggested as the most appropriate indicator that can be used to measure the overall burden of the pandemic in terms of mortality.

In the IHME (Institute for Health Metris and Evaluation) estimation of COVID-19 infections, hospitalizations, and deaths to date, they had used officially reported COVID-19 deaths for nearly all locations. IHME has developed projections for total and daily deaths, daily infections and testing, hospital resource use, and social distancing due to COVID-19 for numbers of countries. In the data spreadsheet, I collected the total deaths as a factor for the death rate comparison.

The analysis of estimation of excess mortality follows four key steps [30]:

1. **Estimating excess mortality compared to expected mortality for locations where all-cause mortality data have been reported during the pandemic.** For all locations where weekly or monthly all-cause mortality has been reported since the start of the pandemic (March 2020), IHME estimate how much mortality increased compared to the expected death rate in all locations with sufficient data.

2. **Estimating the fraction of excess mortality that is direct COVID-19 deaths.** Based on a range of studies and consideration of other evidence, IHME estimate
the fraction of excess mortality that is from excess COVID-19 deaths as opposed to the five other drivers that influence excess mortality.

3. **Estimating the ratio of excess mortality to reported COVID-19 deaths.** IHME build a statistical model that predicts the ratio of excess COVID-19 deaths to reported COVID-19 deaths based on covariates and spatial effects.

4. **Generating predictions of excess COVID-19 mortality for all locations.** Using this statistical relationship to predict the ratio of excess to reported COVID-19 deaths in places without data on excess COVID-19 deaths and then multiply the reported COVID-19 deaths by this ratio to generate estimates of excess COVID-19 deaths for all locations.

### 2.3 Happiness Level

Since 2002, the World Happiness Report has used statistical analysis to determine the world's happiest countries. The report looks at countries with respect to their performance of six variables [16]:

- Gross domestic product per capita
- Social support
- Healthy life expectancy
- Freedom to make your own life choices
- Generosity of the general population
- Perceptions of internal and external corruption levels

To properly compare each country's data, the researchers created a fictional country—christened Dystopia—"the world's least-happy people." They then set Dystopia as the rock bottom value in each of the six categories and measured the scores of the real-world
countries against this value. All six variables were then blended to create a single combined score for each country [16].

The country with the highest score was Finland, which is a European country. Finland scored 7.809 out of a total possible score of 10. Every single country in the top five spots for happiest countries in the world are European countries. Finland is followed in order by Denmark, Norway, Iceland, and the Netherlands. All six variables are considered when defining a score for each country. The lowest-scoring country in the World Happiness Report of 2020 is Afghanistan. With a total ranking of 2.567, Afghanistan has a low life expectancy rate, paired with low gross domestic product rates per capita [16].

2.4 Physical Activity Level

Insufficient physical activity is defined as adults not meeting the WHO recommendations on physical activity for health, i.e. at least 150 min of moderate-intensity, or 75 min of vigorous-intensity physical activity per week, or any equivalent combination of the two [17].

The LANCET estimated the prevalence of insufficient physical activity in adults aged 18 years and older, in 168 countries. Physical activity data collected using wearable devices, such as accelerometers or pedometers, were not included because of the limited comparability with self-reported data. Where available, they used individual-level data to calculate the prevalence of insufficient physical activity, taking the sampling designs into account. Where raw data were not available, they used aggregated data as reported. They included all data that met the inclusion criteria and that were provided before the end of September 2017 [18].
Despite the importance of physical activity (PA) for health and well-being being recognized for more than a century, 23% of adults remain insufficiently active across the globe (World Health Organization, 2018). [4]

2.5 Food Energy Intake Level (kilocalories per capita)

Food consumption is the amount of food available for human consumption as estimated by the United Nations Food and Agriculture Organization (FAO) Food Balance Sheets. However, the actual food consumption may be lower than the quantity shown as food availability depends on the magnitude of wastage and losses of food in the household, for example during storage, in preparation and cooking, as plate-waste or quantities fed to domestic animals and pets, thrown or given away. [19]

2.6 Obesity Level

Overweight and obesity are defined as abnormal or excessive fat accumulation that may impair health. In this work, obesity level is measured as the percentage of a country's population considered to be obese [20]. According to the most recent data available from the World Health Organization (WHO), the island country of Marshall is the most obese in the world with obesity affecting 52.9% of the adult population. Vietnam is the least obese country with 2.1% of the population classified as obese. Among OECD (Organization for Economic Co-operation and Development) countries, the United States is the most obese (36.2%) [21].

2.7 Life Expectancy

Life expectancy is a statistical measure of the average time an individual is expected to live, based on the year of its birth, its current age, and other demographic factors
including sex. Also, it is the key metric for assessing population health. The most commonly used measure is life expectancy at birth (LEB). Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life. [22] Data is collected from Worldometers [32].

2.8 Per Capita nominal GDP

Gross domestic product (GDP) is the market value of all final goods and services from a nation in a given year. Countries are sorted by nominal GDP estimates from financial and statistical institutions, which are calculated at market or government official exchange rates.

The number does not consider differences in the cost of living in different countries, and the results vary greatly from one year to another based on fluctuations in the exchange rates of the country's currency. Such fluctuations change a country's ranking from one year to the next, even though they often make little or no difference to the standard of living of its population. [23] All data are collected on the Worldbank website. [24] The data is between the year of 2019 and 2020 depends on the country and in current United States dollars.

2.9 Per Capita Income

The median income is a number that falls in the middle of the nation’s income distribution. In other words, half of the nation’s adult residents have disposable income higher than this number, while the other half has disposable income that falls below this number.
2.10 Human Freedom Index

The Human Freedom Index presents the state of human freedom in the world based on a broad measure that encompasses personal, civil, and economic freedom. Human freedom is a social concept that recognizes the dignity of individuals and is defined here as negative liberty or the absence of coercive constraint. Because freedom is inherently valuable and plays a role in human progress, it is worth measuring carefully. The Human Freedom Index is a resource that can help to more objectively observe relationships between freedom and other social and economic phenomena, as well as the ways in which the various dimensions of freedom interact with one another. [10]

On a scale of 0 to 10, where 10 represents more freedom, the average human freedom rating for 162 countries in 2018 was 6.93. The data and report are co-published by the Cato Institute and the Fraser Institute. [10]

2.11 Economic Freedom Index

The foundations of economic freedom are personal choice, voluntary exchange, and open markets. As Adam Smith, Milton Friedman, and Friedrich Hayek have stressed, freedom of exchange and market coordination provide the fuel for economic progress. Without exchange and entrepreneurial activity coordinated through markets, modern living standards would be impossible. Potentially advantageous exchanges do not always occur. Their realization is dependent on the presence of sound money, rule of law, and security of property rights, among other factors. Economic Freedom of the World seeks to measure the consistency of the institutions and policies of various countries with voluntary exchange and the other dimensions of economic freedom. The data and report
are co-published by the Cato Institute, the Fraser Institute in Canada and more than 70 think tanks around the world. [28]

2.12 Global Freedom Scores (Political Rights, Civil Liberties), Internet Freedom

Scores, Democracy Scores

The Freedom in the World report is composed of numerical ratings and supporting descriptive texts for 195 countries and 15 territories. External analysts assess 210 countries and territories, using a combination of on-the-ground research, consultations with local contacts, and information from news articles, nongovernmental organizations, governments, and a variety of other sources. Expert advisers and regional specialists then vet the analysts’ conclusions. The final product represents the consensus of the analysts, advisers, and Freedom House staff. Freedom House has assessed the condition of political rights and civil liberties around the world. It is used on a regular basis by policymakers, journalists, academics, activists, and many others. For each country and territory, Freedom in the World analyzed the electoral process, political pluralism and participation, the functioning of the government, freedom of expression and of belief, associational and organizational rights, the rule of law, and personal autonomy and individual rights. [14]

Freedom House measures the level of democratic governance in 29 countries from Central Europe to Central Asia through its annual Nations in Transit report. The democracy score incorporates separate ratings on national and local governance, electoral process, independent media, civil society, judicial framework and independence, and corruption.
3 Multilinear Regression

Linear regression analysis is widely used of all statistical techniques. Multiple linear regression (MLR) can be simplified as multiple regression, is a statistical technique that uses multiple explanatory variables to predict the response outcome variable. The goal of multiple linear regression is to build a model showing the linear relationship between the explanatory (independent) variables and response (dependent) variables [11].

3.1 The Model

The multiple linear regression model is:

\[ y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon \]

where \( y \) is the predicted value of the dependent variable, \( \beta_0 \) the y-intercept (value of \( y \) when all other parameters are set to 0), \( \beta_i \) is the regression coefficient of the \( i^{th} \) independent variable \( X_i \) (i.e. the effect that increasing the value of the independent variable has on the predicted value of \( y \)), and \( \epsilon \) is the model’s error or residual. [25]

3.2 Assumptions in Regression

Regression is a parametric approach, and is restrictive in nature. It usually fails to deliver accurate prediction when a data set fails to fulfill a set of assumptions. Therefore, it is essential to validate these assumptions. There are five key assumptions [13]:

(i) Linearity and additivity of the relationship between dependent and independent variables,

(ii) Statistical independence of the errors,

(iii) Homoscedasticity (constant variance) of the errors

(a) versus the predictions,
(b) versus any independent variable,

(iv) Multicollinearity,

(v) Normality of the error distribution.

3.3 Checking and Fixing the Assumptions if Violated

The assumptions are checked using regression plots and statistical tests.

3.3.1 Linear and Additive

To fit a linear model, we need to transform a non-linear, non-additive dataset into a linear model, otherwise the regression algorithm would fail to capture the trend mathematically.

Given a data set \( \{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^n \) of \( n \) statistical units, where \( \{x_{i1}, \ldots, x_{ip}\}_{i=1}^n \) are the predictors and \( \{y_i\}_{i=1}^n \) are the responses, the additive model takes the form [8]:

\[
Y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon
\]

The \( f_j(X_j) \) are unknown smooth functions fitted to the data. An additive model can be fitted using the backfitting algorithm proposed by Buja, Hastie and Tibshirani [8].

Look for residual vs fitted value plots (explained below). Also, you can include

![Residual vs Fitted Value Plots](image.png)

Figure 1: Residual vs fitted values plot (adopted from [12]).
polynomial terms \((X, X^2, X^3)\) in your model to capture the non-linear effect.

### 3.3.2 Statistical Independence of Residuals

The presence of correlation in error terms drastically reduces model’s accuracy. This usually occurs in time series models where the next instant is dependent on previous instant. If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error. [12]

If this happens, it causes confidence intervals and prediction intervals to be narrower. Narrower confidence interval means that a 95% confidence interval would have lesser probability than 0.95 that it would contain the actual value of coefficients [12].

The autocorrelation (Box and Jenkins, 1976) function [9] can be used for the following two purposes:

1. To detect non-randomness in data.
2. To identify an appropriate time series model if the data are not random.

When we are given measurements, \(Y_1, Y_2, \ldots, Y_N\) at time \(X_1, X_2, \ldots, X_N\), the lag \(k\) autocorrelation function is defined as [9]:

\[
r_k = \frac{\sum_{i=1}^{N-k} (Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}
\]

**How to Check:**

The Durbin-Watson (DW) Test may be used to check whether the statistical independence of residuals assumption is satisfied. The DW statistic is a test for autocorrelation in the residuals from a statistical model or regression analysis. The DW
statistic will always have a value ranging between 0 and 4. A value of 2.0 indicates there is no autocorrelation detected in the sample. Values from 0 to less than 2 correspond to positive autocorrelation while values from 2 to 4 correspond to negative autocorrelation [6].

\[ \text{DW} = 2 \] would be the ideal case here (no autocorrelation)

\[ 0 < \text{DW} < 2 \] implies positive autocorrelation

\[ 2 < \text{DW} < 4 \] implies negative autocorrelation

<table>
<thead>
<tr>
<th>Omnibus:</th>
<th>93.175</th>
<th><strong>Durbin-Watson:</strong> 2.099</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob(Omnibus):</td>
<td>0.000</td>
<td><strong>Jarque-Bera (JB):</strong> 664.038</td>
</tr>
<tr>
<td>Skew:</td>
<td>1.922</td>
<td>Prob(JB): 6.18e-145</td>
</tr>
<tr>
<td>Kurtosis:</td>
<td>11.950</td>
<td>Cond. No. 3.30e+04</td>
</tr>
</tbody>
</table>

Figure 2: OLS summary result

The Ordinary least squares (OLS) summary gives us the Durbin Watson value among other useful insights.

**How to Fix:**

Add a column that lagged with respect to the Independent variable, center the variable (i.e. subtract all values in the column by its mean). From Figure 2, Durbin-Watson is \(~2\) which is very close to the ideal case. So, the assumption is satisfied and no modification of the dataset is required.

3.3.3 **Heteroskedasticity:**

In statistics, heteroskedasticity (or heteroscedasticity) happens when the standard deviations of a predicted variable, monitored over different values of an independent variable or as related to prior time periods, are non-constant. With heteroskedasticity, the
tell-tale sign upon visual inspection of the residual errors is that they will tend to fan out over time [3], as shown in Figure 3.

![Figure 3: Heteroskedasticity (adopted from [3]).](image)

Heteroskedasticity often arises in two forms: conditional and unconditional. Conditional heteroskedasticity identifies non-constant volatility related to prior period's (e.g., daily) volatility. Unconditional heteroskedasticity refers to general structural changes in volatility that are not related to prior period volatility. Unconditional heteroskedasticity is used when future periods of high and low volatility can be identified [3].

The error term in our regression model is homoscedastic if the variance of the conditional distribution of \( \mu_i \) given \( X_i \), \( \text{Var}(\mu_i|X_i = x) \), is constant for all observations in the sample [5]:

\[
\text{Var}(\mu_i|X_i = x) = \sigma^2 \ \forall \ i = 1, \ldots, n
\]

If instead there is dependence of the conditional variance of \( \mu_i \) on \( X_i \), the error term is said to be heteroskedastic. We then write [5]:

\[
\text{Var}(\mu_i|X_i = x) = \sigma_i^2 \ \forall \ i = 1, \ldots, n
\]
Hoskvedasticity is a special case of heteroscedasticity [5]. If heteroscedasticity is present, the residual vs fitted values plot can tell. If the plot shows a funnel shape pattern, then we say that heteroscedasticity is present. Here, we have plots of residuals vs fitted values on assumptions. We see a funnel like pattern in this section, so the heteroscedasticity is present. [12]

![Figure 4: Residual vs fitted values plot before transformed.](image1)

![Figure 5: Residual vs fitted values plot after transformed.](image2)

**How to Fix:**

We can do a non-linear transformation of the dependent variable such as log(Y) or square root of Y. Also, we can use weighted least square method to tackle heteroskedasticity. After transformed, we do not see a funnel like pattern in the after section, so no heteroscedasticity. [12]

3.3.4 Multicollinearity

This phenomenon exists when the independent variables are found to be moderately or highly correlated. In a model with correlated variables, it becomes a tough task to figure out the true relationship of a predictors with response variable. In other words, it becomes difficult to find out which variable is contributing to predict the response variable. [12]
Another point, with presence of correlated predictors, the standard errors tend to increase. And with large standard errors, the confidence interval becomes wider leading to less precise estimates of slope parameters. [12]

3.3.5 Variance Inflation Factor

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. This ratio is calculated for each independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model [26].

VIF can be calculated as [27]:

\[
VIF_i = \frac{1}{1 - R_i^2}
\]

where \( R_i^2 \) is the coefficient of determination of the regression equation in step one, with \( X_i \) on the left-hand side, and all other predictor variables (all the other X variables) on the right-hand side [27].

How to Check:

In case of very less variables, one could use heatmap, but that isn’t so feasible if there is large number of columns. Another common way to check would be by calculating VIF (Variance Inflation Factor) values. If

\[ VIF=1, \text{ Very Less Multicollinearity} \]
\[ VIF<5, \text{ Moderate Multicollinearity} \]
\[ VIF>5, \text{ Extreme Multicollinearity} \]

How to Fix:
We can easily remove the variables with high multicollinearity, or if we can find out which 2 or more variables have high correlation with each other, we could simply merge these variables into one. And make sure that all variables’ VIF < 5.

### 3.3.6 Normal Distribution of Error Terms

In probability theory, a normal (or Gaussian or Gauss or Laplace–Gauss) distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is [29]:

\[
    f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}
\]

The parameter \( \mu \) is the mean or expectation of the distribution (and its median and mode), while the parameter is its standard deviation. The variance of the distribution is \( \sigma^2 \). A random variable with a Gaussian distribution is said to be normally distributed and is also called a normal deviate [29].

If the error terms are non-normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares. Presence of non – normal

![VIF for selected independent variables](image1)

![VIF for all independent variables](image2)
distribution suggests that there are a few unusual data points which must be studied closely to make a better model [12].

**How to Check:**

Use distribution plot on the residuals and see if it is normally distributed. Or we can look at Q-Q plot (shown Figure 9). We can also perform statistical tests of normality.

**How to Fix:**

If the residuals are not normally distributed, we can try non–linear transformation of the dependent or independent variables.

![Figure 8: Comparison of residual distributions before and after eliminating variables.](image-url)
This Quantile-Quantile (Q-Q) is a scatter plot which helps to validate the assumption of normal distribution in a data set. We can infer if the data is generated from a normal distribution by using this plot. If it is, the plot will be a straight line. Absence of normality in the errors is shown as deviations from the straight line. [12]

From OLS results (Figure 11), we obtain the following coefficients:
Obesity Level: 0.1757
Global Freedom Scores: 0.0459
Per Capita nominal GDP: -0.00003
Intercept: 2.35485

Pearson correlation coefficient (all variables): 0.7241
Pearson correlation coefficient (three selected variables): 0.6938

From the multilinear regression model, three out of the eight factors are selected for 168
countries: obesity level, global freedom score, and per capita nominal GDP. Obesity level has highest coefficient factors of 0.1757 between these independent variable factors.

3.4 Comparing Multilinear Regression With and Without Satisfying Assumptions

Figure 12 shows the results from fitting a multilinear regression model to the same dataset without satisfying the assumptions. Four out of the eight factors are selected for 168 countries: obesity level, global freedom score, per capita nominal GDP, and life expectancy. When assumptions of regression are all satisfied, the first three factors are selected but the last one (life expectancy) is not.

The coefficient of determination, $R^2$, is the proportion of variation in the dependent variable that can be predicted by the independent variables. Adjusted-$R^2$ penalizes $R^2$ if useless explanatory variables are added to the model. The $R^2$ for multilinear regression without satisfying assumptions is 0.4 (Figure 12) and after satisfying assumptions is 0.929 (Figure 11). The adjusted-$R^2$ for the two cases are 0.385 and 0.928. Thus, based on $R^2$ and adjusted-$R^2$, satisfying assumptions improves the regression model.

We also found that the Pearson correlation coefficient between the actual and predicted death rates for the model with three selected variables (obesity level, global freedom score, per capita nominal GDP) after satisfying assumptions is 0.6938. For the model with four selected variables (obesity level, global freedom score, per capita nominal GDP, life expectancy) without satisfying any assumptions, the Pearson correlation coefficient is 0.6324. Thus, the prediction from the model that satisfies all assumptions is more accurate than the model that does not satisfy the assumptions even though the latter employs more variables than the former.
4 Conclusions

Within two years, COVID-19 has caused more than five million deaths worldwide. We collected data for 168 countries on eight factors, only some of which have been considered in other studies. We used multilinear regression to model the data. The model identified three factors to be highly correlated with death rate: obesity level, global freedom score, and per capita nominal GDP. While obesity has been identified by other studies as a risk factor for COVID-19 death, the other two selected factors are novel and associate the attitude and lifestyle of people of different countries with COVID-19 death rate. We also show that the prediction improves if the data is transformed to satisfy all the assumptions of multilinear regression. This work may help governments to devise policies to mitigate the spread of infection due to COVID-19 as well as other pandemics.
Figure 12: Results from fitting a multilinear regression model to the same dataset without satisfying the assumptions.
Reference


   https://en.wikipedia.org/wiki/Variance_inflation_factor

   https://www.cato.org/economic-freedom-world/2021

   https://en.wikipedia.org/wiki/Normal_distribution


   https://unstats.un.org/unsd/methodology/m49/