Impact of Various Factors on a Country's GDP using Regression Analysis

Econometrics Modelling and Forecasting:

Assignment Report

Submitted by,

Nehla Shajahan

MOR: Sem 2

South Campus

Roll no: 20/1613

INDEX

INTRODUCTION

Economists have used both theory and empirical research to explain the cause of economic growth. Since growth is a very dynamic process, studies that are based on cases hundreds of years ago might not be as relevant now. The technological changes in the last few decades have revolutionized the way countries improve their economy. Economic growth is measured by the increase in a country's total output or real Gross Domestic Product (GDP).

Economic growth is one of the most important indicators of a healthy economy. One of the biggest impacts of long-term growth of a country is that it has a positive impact on national income and the level of employment, which increases the standard of living. As the country's GDP is increasing, it is more productive which leads to more people being employed. This increases the wealth of the country and its population. Higher economic growth also leads to extra tax income for government spending, which the government can use to develop the economy. This expansion can also be used to reduce the budget deficit.

Additionally, as the population of a country grows, it requires growth to keep up its standard of living and wealth. Economic growth also helps improve the standards of living and reduce poverty, but these improvements cannot occur without economic development. Economic growth alone cannot eliminate poverty on its own.

Growth doesn't occur in isolation. Events in one country and region can have a significant effect on growth prospects in another. Here, we try to look at the impact of various factors such as infant mortality rate, literacy rate, access to phones, birthrate and deathrate of 195 countries on their respective GDPs.

VARIABLES USED FOR ANALYSIS

		Variable name
Independent variables	X1	Infant mortality (per 1000 births)
	X2	Literacy %
	X3	Phones (per 1000)
	X4	Birthrate
	X5	Deathrate
Dependent variable	y	GDP ($ per capita)

Description:

1. **Infant mortality rate**: Infant mortality is the death of an infant before his or her first birthday. The infant mortality rate is the number of infant deaths for every 1,000 live births.
2. **Literacy**: The percentage of population aged 15 years and over who can both read and write with understanding a short simple statement on his/her everyday life. Generally, 'literacy' also encompasses 'numeracy', the ability to make simple arithmetic calculations.
3. **Phones** (per 1000): The number of people in a population who have the access to a phone for every 1000 people. This feature is an indicator of the number of people who are privileged enough to have access to technology.
4. **Birthrate:** The number of live births per thousand of population per year.
5. **Deathrate:** The ratio of deaths to the population of a particular area or during a particular period of time, usually calculated as the number of deaths per one thousand people per year.
6. **GDP ($ per capita):** Per capita gross domestic product (GDP) is a metric that breaks down a country's economic output per person and is calculated by dividing the GDP of a country by its population.

EMPIRICAL ANALYSIS

OBJECTIVE:

To perform linear regression analysis on the given dataset of 195 countries to:

- Check and establish the OLS assumptions.
- Determine the weights of each of the feature and thereby describe their respective impact.

DATA SOURCE:

Secondary data has been collected for all the variables from the latest available database of all countries and the required features for our analysis were picked out from the data.

https://www.kaggle.com/fernandol/countries-of-the-world

METHODOLOGY:

Multiple linear regression using OLS estimators under the Classical linear Regression Model. Python using Jupyter Notebook interface was utilized to effectively carry out the analysis. The libraries used for the analysis were mainly statsmodel and sklearn.

MULTIPLE LINEAR REGRESSION

Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. The variable that we want to predict is known as the dependent variable, while the variables we use to predict the value of the dependent variable are known as independent or explanatory variables. A multiple linear regression model can be mathematically expressed as:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon$$

- Y: the predicted value of the dependent variable
- β0: the y-intercept/bias (value of y when all other parameters are set to 0)
- β1X1: the regression coefficient (B1) of the first independent variable (X1)
- βnXn: the regression coefficient of the last independent variable
- Ɛ: model error

In our analysis, we have 5 independent variables and one dependent variable, GDP. Therefore, the equation can be formulated as:

$$GDP = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Where X_1, X_2, X_3, X_4, X_5 are infant mortality, literacy, phones (per 1000), birthrate, deathrate respectively. Regression analysis aims at finding the optimal value of the intercept, β_0 and the coefficients β_1,…, β_5.

Assumptions of a Multiple Linear Regression Model

1. Linearity

The first assumption of multiple linear regression is that there is a linear relationship between the dependent variable and each of the independent variables.

Test for Linearity:

The best way to check the linear relationships is to create scatterplots and then visually inspect the scatterplots for linearity. The initial scatter plot of our data appeared to show a non-linear relation. To overcome with problem, we deployed log transformation. Plotting a scatter plot again showed that all the independent variables are linearly related to the dependent variable.

2. No Multicollinearity

Yet another assumption is that the data should not show multicollinearity, which occurs when the independent variables are highly correlated with each other. When independent variables show multicollinearity, there will be problems figuring out the specific variable that contributes to the variance in the dependent variable.

Test for Multicollinearity:

The best method to test for the assumption is the Variance Inflation Factor (VIF) method. 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. Values of VIF that exceed 10 are often regarded as indicating multicollinearity.

The initial VIF values showed values beyond the optimal range. To mitigate this issue, we use the process of standardization. Standardization refers to the process of subtracting the mean and dividing by the standard deviation. Centering the variables and standardizing them will both reduce the multicollinearity. However, standardizing changes the interpretation of the coefficients. So, we centered or subtracted the respective mean from the features in this case. Subsequently, the VIF values were in the optimal range of below 10. Thus, our model has no multicollinearity.

3. No Autocorrelation

Autocorrelation refers to the degree of correlation between the values of the same variables across different observations in the data. The model assumes that the observations should be independent of one another. Simply put, the model assumes that the values of residuals are independent.

Test for Autocorrelation:

To check for autocorrelation, we use the Durbin-Watson test. A value of 2.0 means there is no autocorrelation detected in the sample. The Durbin-Watson statistic for our dataset is 1.883 which indicated the absence of autocorrelation in our model.

4. No Heteroscedasticity / Homoscedasticity

Multiple linear regression assumes that the amount of error in the residuals is similar at each point of the linear model. This scenario is known as homoscedasticity. Heteroscedasticity refers to situations where the variance of the residuals is unequal over a range of measured values.

Test for Heteroscedasticity

We will be using Goldfeld Quandt test to look for heteroscedasticity. Under this test, we consider the following hypothesis:

- Null hypothesis, H0: Homoscedasticity or Absence of heteroscedasticity
- Alternate hypothesis, H1: Presence of heteroscedasticity

For any hypothesis test, the decision rule is:

- If p-value < level of significance (alpha); then null hypothesis is rejected.
- If p-value > level of significance (alpha); then we fail to reject the null hypothesis.

The p value of our model was 0.42, which is greater than the level of significance 0.05. Therefore, we fail to reject the null hypothesis. Hence, we conclude that our model is homoscedastic or has no heteroscedasticity.

Regression Results

After running a linear regression model using sklearn in python, we devised the following equation for our analysis.
The model had an R squared value of 0.80.

Multiple linear regression model:

$$GDP = 8.52 - 0.011\mathrm{X}_1 + 0.002\mathrm{X}_2 + 0.002\mathrm{X}_3 - 0.02\mathrm{X}_4 + 0.02\mathrm{X}_5$$

Where X_1, X_2, X_3, X_4, X_5 are infant mortality rate, literacy, phones (per 1000), birthrate, deathrate respectively.

CONCLUSION

To recapitulate, we deployed a multiple linear regression model on our 'countries of the world' dataset which had five independent features namely, infant mortality rate, literacy, phones (per 1000), birthrate, deathrate respectively. The dependent variable for our analysis was 'GDP'. For the study, we took the data of 195 random countries.

The initial aim of our analogy was to check and handle the four basic assumptions of regression: Linearity, No Multicollinearity, No Autocorrelation, Homoscedasticity. We successfully established these assumptions using various techniques in python.

Furthermore, we built a linear regression model using sklearn and found out the optimal coefficients for each feature as well as the intercept value. This model was further used to predict values in the test dataset and check the accuracy by calculating the residuals and the percentage difference for each of the predicted values.

From the regression model it is evident that infant mortality rate and birthrate are negatively proportional to GDP as their respective coefficients are negative. Meanwhile, features like literacy, phones (per 1000) and deathrate are positively related to GDP.