A Comprehensive Study on the Factors Impacting the GDP (per capita) of Major Economies around the Globe using Regression Analysis

Nimisha Bhide
Student, University of Mumbai
B.E. Computer Engineering
Vidyalankar Institute of Technology, Mumbai, India

Saurabh Khanolkar
Student, University of Mumbai
B.E Computer Engineering
Vidyalankar Institute of Technology, Mumbai, India

ABSTRACT

Financial Architecture aims at sustainability of an Economy. This is done by ensuring a consistent growth rate. GDP is a strong indicator of the growth of an economy. A Higher GDP of an economy reflects a robust growth. This leads to the definition of GDP (per capita). This study focuses on the GDP (per capita) as an indicator of a nation’s prosperity. The ratio of the GDP of an economy to its population is termed as the GDP (per capita). This study considers GDP (per capita) as a function of 17 factors. Further on, out of these 17 factors, 5 of the most statistically significant factors are identified using the Backward Elimination Algorithm. Thus, a statistically significant regression model is designed and the impact of each of the 5 factors on the GDP (per capita) is gauged. It was found that the combination of the aforementioned 5 statistically significant variables could explain 83% of the variance in the GDP (per capita) of the economies. The F statistic increased from 51.13(before applying Backward Elimination Algorithm); to 168.6 (after the application of the Algorithm) and hence, signifying the increase in the overall significance of the model. The authors firmly believe that that this study will form a foundation to the higher level policy making in the future.

General Terms

Multiple Regression, Accuracy.

Keywords

Multiple Regression, Backward Elimination, GDP (per capita), Regression Analysis, Correlation, Hypothesis testing.

1. INTRODUCTION

Financial architecture broad term. It refers to the framework and series of measures that are considered necessary to prevent future economic crises. Further on, it helps manage these crises when they occur. Financial Architecture refers to the structures, practices and rules which are designed in order to overcome the influence of crisis on the economy. Financial Architecture aims at sustainability of an Economy. This is done by ensuring a consistent growth rate. GDP is the indicator of the growth of an economy.[10] Higher GDP of an economy reflects a rosy picture as it portrays a better position of the economy. But this isn’t a true representation of the prosperity of a nation. GDP (per capita) is a metric that truly represents the purchasing power per capita of the economy.[6] Hence, GDP (per capita) forms a very important metric of the purchasing power of the nation’s citizens.[1] There are certain macro factors operating in the economic environment that will influence the GDP (per capita). Financial crises are a disruption or sudden change in the activities of operating environment which have a significantly negative impact on economic developments. Such negative impact of crisis can be mitigated to a certain extent by identifying the factors and by analysing the early signals indicated by these factors operating in the environment.

2. OBJECTIVES AND HYPOTHESIS OF THE STUDY

2.1 The main objectives of the study are

- Identifying the relationship between selected variables and GDP (per capita) of an Economy.
- 2. Identifying the factors that most significantly impact the target variable i.e. GDP per capita of an Economy.[11]
- 3. Develop the most statistically significant model using the significant factors identified.

2.2 Hypothesis of the Study

Hypothesis for testing the overall significance of the model

2.2.1 H0 (Null Hypothesis)

None of the selected variables is a significant predictor of the target variable i.e. GDP (per capita) of the economy

2.2.2 Ha (Alternate Hypothesis)

At least one of the selected variables is a significant predictor of the Target variable i.e. GDP (per capita) of the economy.

2.3 Hypothesis for feature selection

2.3.1H0 (Null Hypothesis)

The feature under consideration is a not a significant indicator of the GDP (per capita); when all other variables are included in the model

2.3.2Ha (Alternate Hypothesis):

The feature under consideration is a significant indicator of the GDP(per capita); when all other variables are included in the model.

2.4 Significance level (α) selected

The significance level for this study is selected to be 0.05

3. RESEARCH DESIGN AND METHODOLOGY

Exploratory research design is adopted in the present study. The study seeks to extract information about the influence and relationship between GDP (per capita) and selected variables of an Economy.

3.1 Data description and Sources

The data is collected by using secondary sources relating to the selected variables. Following are the 17 factors the study considers:

1) Population: The number of people living in the country whose GDP is being calculated.

2)Area: The total area of land in the country in square miles.

3)Population Density: The number of people living per square mile area.

4)Coastline: It is a ratio of coast to total area.

5)Net Migration: The net migration experienced by the economy.

6)Infant mortality: The number of deaths under one year of age occurring among the live births in a given geographical area during a given year, per 1,000 live births occurring among the population of the given geographical area during the same year..

7)Literacy: Percentage literacy.

8)Phones (per 1000): Phones per 1000 in the population.

9)Crops: Crops grown in the economy.

10)Industry: People involved in the industry sector.

11)Birth rate: The number of births per thousand of population per year.

12)Death rate: The number of deaths per thousand of population per year.

13)Agriculture: People involved in the Agricultural sector.

14)Arable: The fertile piece of land present among the total area covered by the economy.

15)Services: People in the service sector.

16)Other: The other parameters that influence the GDP of a country.

17)Climate-label: The climate experienced by the economy.

3.2 Techniques used in analysis

The study attempts to examine the relationship between GDP (per capita) and selected variables of economies world-wide using Regression Analysis.[4] More specifically, the study aims to develop a statistically significant model which cherry picks the most statistically significant factors that have a significant impact on the GDP (per capita). This study uses the “Backward Elimination Algorithm” for feature selection of the significant factors.[1] Further on, the significance of the entire model is judged by examining the F statistic.[12] The accuracy of the final model is examined using the adjusted-R2 metric.

3.3 Initial Multivariate correlational Analysis

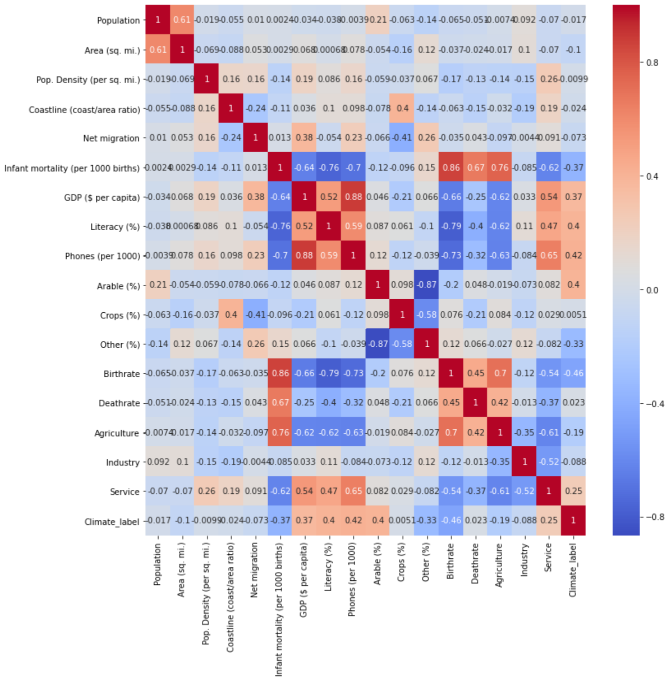


Figure 4-a: Multivariate Correlation

The multivariate correlational analysis suggests that the target variable GDP (per capita) is strongly correlated with the following explanatory variables:

3.3.1 Infant mortality

- Correlation coefficient: -0.64
- This implies that “infant mortality” shows a linear negative relationship with GDP (per capita)

3.3.2 Phones per 1000

- Correlation coefficient: 0.88
- This implies that “phones per 1000”

  shows a strong linear positive relationship with GDP (per capita)

3.3.3Birth rate

- Correlation coefficient: -0.66
- This implies that “Birth Rate” shows a linear negative relationship with GDP (per capita)

3.3.4 Agriculture

- Correlation coefficient: -0.62
- This implies that Agriculture shows a linear negative relationship with GDP (per capita)

Further on, some notable correlations observed are:

1)Infant mortality and

- Agriculture (positively correlated)
- Birth rate (strongly positively correlated)
- Phones (per 1000) (negatively correlated)
- Literacy (negatively correlated)

2) literacy and:

- Agriculture (negatively correlated)
- Birth rate (negatively correlated)

3) Phones (per 1000) and:

- Agriculture (negatively correlated)
- Birth rate (negatively correlated)

4)Birth rate and:

- Agriculture (positively correlated)

3.4 Conditions for Multiple linear regression

The following conditions were checked for before Regression Analysis was performed:

3.4.1 Linearity

A Linear relationship between the explanatory variables and the target variable. It exhibits a simple non-trivial relationship.[7] This assumption is violated when the points on the plot cannot be represented with a straight line.[5]

3.4.2 Normality and homoscedasticity

Normality assumes that the error terms are normally distributed. If this is not the case then the central limit theorem can be used. Homoscedasticity, refers to errors having equal variance.[3]

3.4.3 No multicollinearity

Multi collinearity is a statistical phenomenon. If two or more explanatory variables in a multiple regression model are highly correlated multicollinearity occurs. One variable can be linearly predicted from the others with a non-trivial degree of accuracy. This situation may result in coefficient estimates changing erratically in response to small changes in the model or the data [2]. One solution for this is to drop one of the variables. The Backward Elimination Algorithm can handle multicollinearity with ease.

3.5 Backward Elimination Algorithm

The study begins with 17 independent variables that may have an impact on the target variable. The aim is to figure out which of the factors have a more significant impact on the target variable than the rest of the factors[13]. For this, the Backward Elimination Algorithm is used. The significance level under consideration in this study is 0.05. The Algorithm is as Follows:

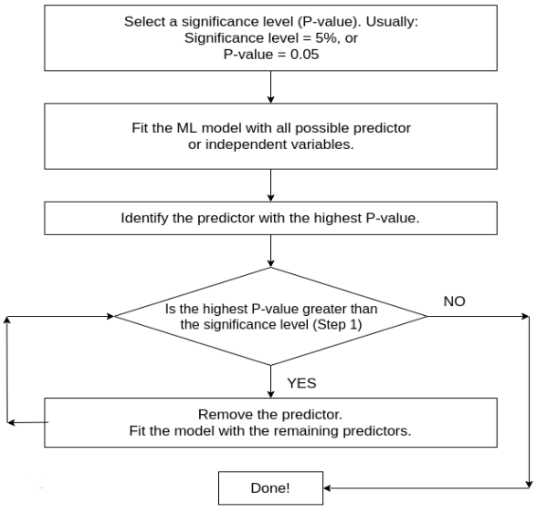
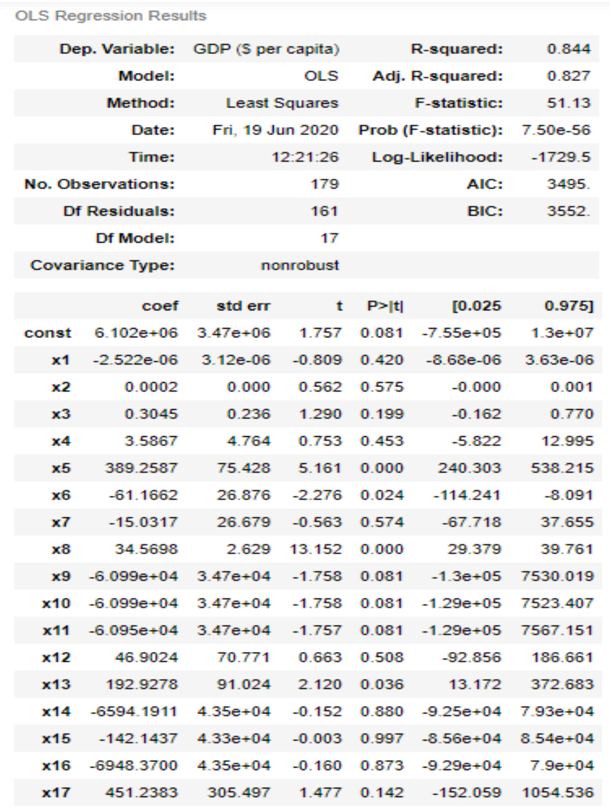

Select a significance level (P-value). Usually:
Significance level = 5%, or
P-value = 0.05
Fit the ML model with all possible predictor or independent variables.
Identify the predictor with the highest P-value.
Is the highest P-value greater than the significance level (Step 1)
NO
YES
Remove the predictor.
Fit the model with the remaining predictors.
Done!

Figure 4-b: Backward Elimination Process

This study uses **F statistic** as the measure of overall significance of the model. The higher the F statistic is, the more significant the model as a whole is. The p-value corresponding to the F-statistic indicates whether the model is significant or not.

3.6 Application of the Algorithm to the Data

Initially all the 17 independent variables are fit to the model. The following regression output is obtained:



OLS Regression Results

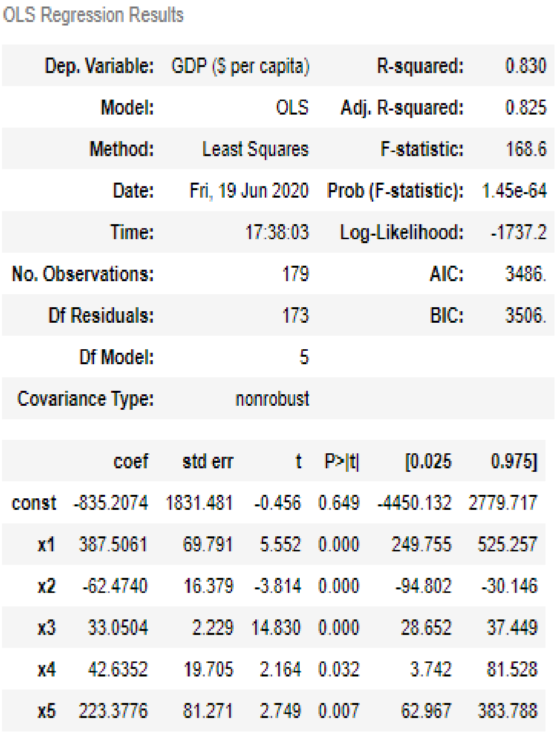
Dep. Variable:	GDP ($ per capita)	R-squared:	0.844
Model:	OLS	Adj. R-squared:	0.827
Method:	Least Squares	F-statistic:	51.13
Date:	Fri, 19 Jun 2020	Prob (F-statistic):	7.50e-56
Time:	12:21:26	Log-Likelihood:	-1729.5
No. Observations:	179	AIC:	3495.
Df Residuals:	161	BIC:	3552.
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.102e+06	3.47e+06	1.757	0.081	-7.55e+05	1.3e+07
x1	-2.522e-06	3.12e-06	-0.809	0.420	-8.68e-06	3.63e-06
x2	0.0002	0.000	0.562	0.575	-0.000	0.001
x3	0.3045	0.236	1.290	0.199	-0.162	0.770
x4	3.5867	4.764	0.753	0.453	-5.822	12.995
x5	389.2587	75.428	5.161	0.000	240.303	538.215
x6	-61.1662	26.876	-2.276	0.024	-114.241	-8.091
x7	-15.0317	26.679	-0.563	0.574	-67.718	37.655
x8	34.5698	2.629	13.152	0.000	29.379	39.761
x9	-6.099e+04	3.47e+04	-1.758	0.081	-1.3e+05	7530.019
x10	-6.099e+04	3.47e+04	-1.758	0.081	-1.29e+05	7523.407
x11	-6.095e+04	3.47e+04	-1.757	0.081	-1.29e+05	7567.151
x12	46.9024	70.771	0.663	0.508	-92.856	186.661
x13	192.9278	91.024	2.120	0.036	13.172	372.683
x14	-6594.1911	4.35e+04	-0.152	0.880	-9.25e+04	7.93e+04
x15	-142.1437	4.33e+04	-0.003	0.997	-8.56e+04	8.54e+04
x16	-6948.3700	4.35e+04	-0.160	0.873	-9.29e+04	7.9e+04
x17	451.2383	305.497	1.477	0.142	-152.059	1054.536

Figure 4-c: OLS Regression Results-I

The p-value of the model as a whole is 7.5e-56 i.e. 7.5*10^-56. The p-value is lesser than the significance level set for the study. This implies that the Null hypothesis used for testing the overall significance of the model can be rejected. Hence, At least one of the selected variables is a significant predictor of the Target variable i.e. GDP (per capita) of the economy. Further on, note that the value of the **F statistic is 51.13**.

Using the **Backward elimination Algorithm**, a **parsimonious** model is reached. The final model is as follows:



OLS Regression Results

Dep. Variable:	GDP ($ per capita)	R-squared:	0.830
Model:	OLS	Adj. R-squared:	0.825
Method:	Least Squares	F-statistic:	168.6
Date:	Fri, 19 Jun 2020	Prob (F-statistic):	1.45e-64
Time:	17:38:03	Log-Likelihood:	-1737.2
No. Observations:	179	AIC:	3486.
Df Residuals:	173	BIC:	3506.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-835.2074	1831.481	-0.456	0.649	-4450.132	2779.717
x1	387.5061	69.791	5.552	0.000	249.755	525.257
x2	-62.4740	16.379	-3.814	0.000	-94.802	-30.146
x3	33.0504	2.229	14.830	0.000	28.652	37.449
x4	42.6352	19.705	2.164	0.032	3.742	81.528
x5	223.3776	81.271	2.749	0.007	62.967	383.788

Figure 4-d: OLS Regression Results-II

Note that the value of the **F statistic is 168.6.** This implies that the parsimonious model achieved, is way more significant as compared to the model the study started off with (F statistic : 51.13). In the final model achieved, all the factors have an associated p-value less than the significance level of 0.05.

It was found that the factors that have a significant impact on the GDP (per capita) of an economy are:

- X1 = Net Migration …(NM)
- X2 = Infant mortality (per 1000)…(IM)
- X3 = Phones (per 1000) …(PH)
- X4 = Literacy (%) …(LI)
- X5 = Death rate …(DR)

3.7 The Final model:

A linear model is given by:

$Y = \beta0 + \beta1\ (X1) + \beta2(X2) + \beta3(X3) + \ldots + \beta n(Xn)$

$\beta0$ = constant

$\beta1,2,3,\ldots,n$ = coefficients of predictors

The final model includes the five aforementioned predictors and is represented as:

GDP (per capita) = -835.2074 + 387.5061(NM) - 62.4740(IM)+33.0504(PH) + 42.6352(LI) + 223.3776(DR)

4. RESULT AND ANALYSIS

4.1 Analysing the F statistic:

F statistic is the measure of overall significance of the model. The higher the F statistic is, the more significant the model as a whole is. The value of the F statistic was noted and analysed at the end of each iteration of the Backward Elimination Algorithm.[9] The following diagram represents the F statistic over the 13 iterations of the Backward Elimination Algorithm:

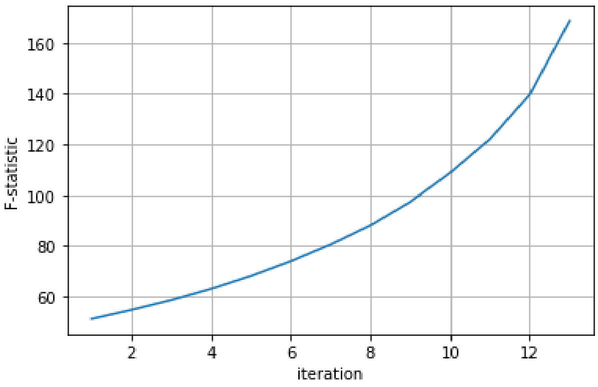


Figure 5: F-statistics

This implies that the **overall significance of the model increases** with elimination of every non-significant factor from the initial model using the Backward Elimination Algorithm.

4.2 Accuracy of the parsimonious model:

The accuracy of the final model is measured using the metric: R2(R squared). R2 signifies the percent of variability in the target variable that is explained by a set of independent variables. The R2 of the final model is 0.830. This implies that 83% of the total variance in the dependent variable can be explained by the explanatory variables. Further on, the Adjusted R2 of the model, which is the true representation of the accuracy is 0.825.

4.3 Interpretation of the final model:

4.3.1Net migration:

All else remaining constant, for a unit's increase in Net migration, the GDP (per capita) exhibits an increase of 387.5 units. Note that a positive net migration implies an inward flux of migrants into the country and vice versa. This implies that the GDP (per capita) exhibits an increase as the people migrating into the country increases.

4.3.2Infant mortality (per 1000):

All else remaining constant, for a unit's increase in the infant mortality, the GDP (per capita) exhibits a decrease of 62.47 units.

4.3.3 Phones (per 1000):

All else remaining constant, for a unit's increase in the phones (per 1000), the GDP (per capita) exhibits an increase of 33.05 units.

4.3.4 Literacy:

All else remaining constant, for a unit's increase in the percent literacy, the GDP (per capita) exhibits an increase of 42.63 units.

4.3.5 Death rate:

All else remaining constant, for a unit's increase in the death rate, the GDP (per capita) exhibits an increase of 223.37 units.

5. CONCLUSIONS

The policy makers have to respond when early signals about the economy are received before reaching a critical situation. It becomes very difficult to address the problem. The study revealed that Net Migration, Infant Mortality, Number of Phones owned by the population, Literacy and death rate are better indicators of the prosperity of an economy. Increase in the Net migration, phones, literacy, death Rate have a positive impact on the GDP (per capita) whereas, an increase in the infant mortality has a negative impact on the nations GDP.

The aforementioned factors can serve as early indicators of the GDP (per capita) in the near future. This can in turn help the policy makers make better decisions.

6. REFERENCES

[1] Divya K. and Devi V., 2014. A Study on Predictors of GDP: Early Signals. Procedia Economics and Finance, 11, pp.375-382.

[2] Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. The Review of Economic and Statistics, 92-107.

[3] Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. Econometrica: Journal of the Econometric Society, 1287-1294

[4] Wang, H. and Rhee, W., 1995. An algorithm for estimating the parameters in multiple linear regression model with linear constraints. Computers & Industrial Engineering, 28(4), pp.813-821.

[5] Turóczy, Z. and Marian, L., 2012. Multiple Regression Analysis of Performance Indicators in the Ceramic Industry. Procedia Economics and Finance, 3, pp.509-514.

[6] TY - JOUR, Maity, Bipasha and Chatterjee, B.,2012/01/01 Forecasting GDP Growth Rates of India.: An Empirical Study.Int. J. Econom. Manage. Sci.,1

[7] FENG, R., 2019. Effect of Physical Education on Employment Ability of University Students Based on Multiple Linear Regression Analysis. DEStech Transactions on Economics, Business and Management, (emba).

[8] Karamazova, E., Jusufi Zenku, T. and Trifunov, Z., 2017. Analysing and Comparing the Final Grade in Mathematics by Linear Regression Using Excel and SPSS. International Journal of Mathematics Trends and Technology, 52(5), pp.334-344.

[9] Kumar, S. and Muhuri, P., 2019. A novel GDP prediction technique based on transfer learning using CO2 emission dataset. Applied Energy, 253, p.113476.

[10] Saqib, N., 2013. The Effect of Exchange Rate Fluctuation on Trade Balance: Empirical Evidence from Saudi Arab Economy. SSRN Electronic Journal.

[11] Lehmann, R. and Wohlrabe, K., 2015. Forecasting GDP at the Regional Level with Many Predictors. German Economic Review, 16(2), pp.226-254.