Analysis of the Factors Affecting the GDP of several countries

Project Report-1

(Group 8)

MBA652A – Statistical Modelling for Business Analytics

Submitted By:

Devansh Jindal (231140007)
Suraj Chaudhary (231140025)
Amogh Gupta (231140005)
Saransh Sharma (231290012)
Aishwarya Awasthi (22125010

Submitted To:

Prof. Devlina Chatterjee
IME Department
IIT Kanpur

Table of Contents

Introduction

Gross domestic product (GDP) is the total monetary or market value of all the finished goods and services produced within a country's borders in a specific time period. As a broad measure of overall domestic production, it functions as a comprehensive scorecard of a given country's economic health. The calculation of a country's GDP encompasses all private and public consumption, government outlays, investments, additions to private inventories, paid-in construction costs, and the foreign balance of trade. Exports are added to the value and imports are subtracted. Since exports plays important role in improving GDP, we have considered export facilitating Indicators. We have conducted a study to determine those factors that are actually impacting the GDP per capita. which will provide a clear picture and help in decision making. Though the numbers of factors which cause a skewed ratio are many, we have limited our study to just four factors that are Information and Communication Technology, Physical Infrastructure, Business Environment and Border and transport efficiency.

Objective

The objective of the study is to find out what impact does the factors "Information and Communication Technology", "Physical Infrastructure", "Business Environment" and "Border and transport efficiency", could have on the GDP per capita. in several industries.

Methodology

We have decided to adopt regression analysis to perform our study as it is a tried and tested way to estimate what impact does a variable (or a set of variables) creates upon the variable of interest.

Data Description and Source

This data has been taken from World Bank. It contains the export facilitating Indicators like Information and Communication

Technology, Physical Infrastructure, Business Environment and Border and transport efficiency of 112 countries from the year 2004 to 2007. 101 countries are developing and the rest are developed. Export Facilitating Indicators are considered because exports play a huge role in GDP of any country. Link for the data has been provided in the reference section.

Steps in Regression Analysis:

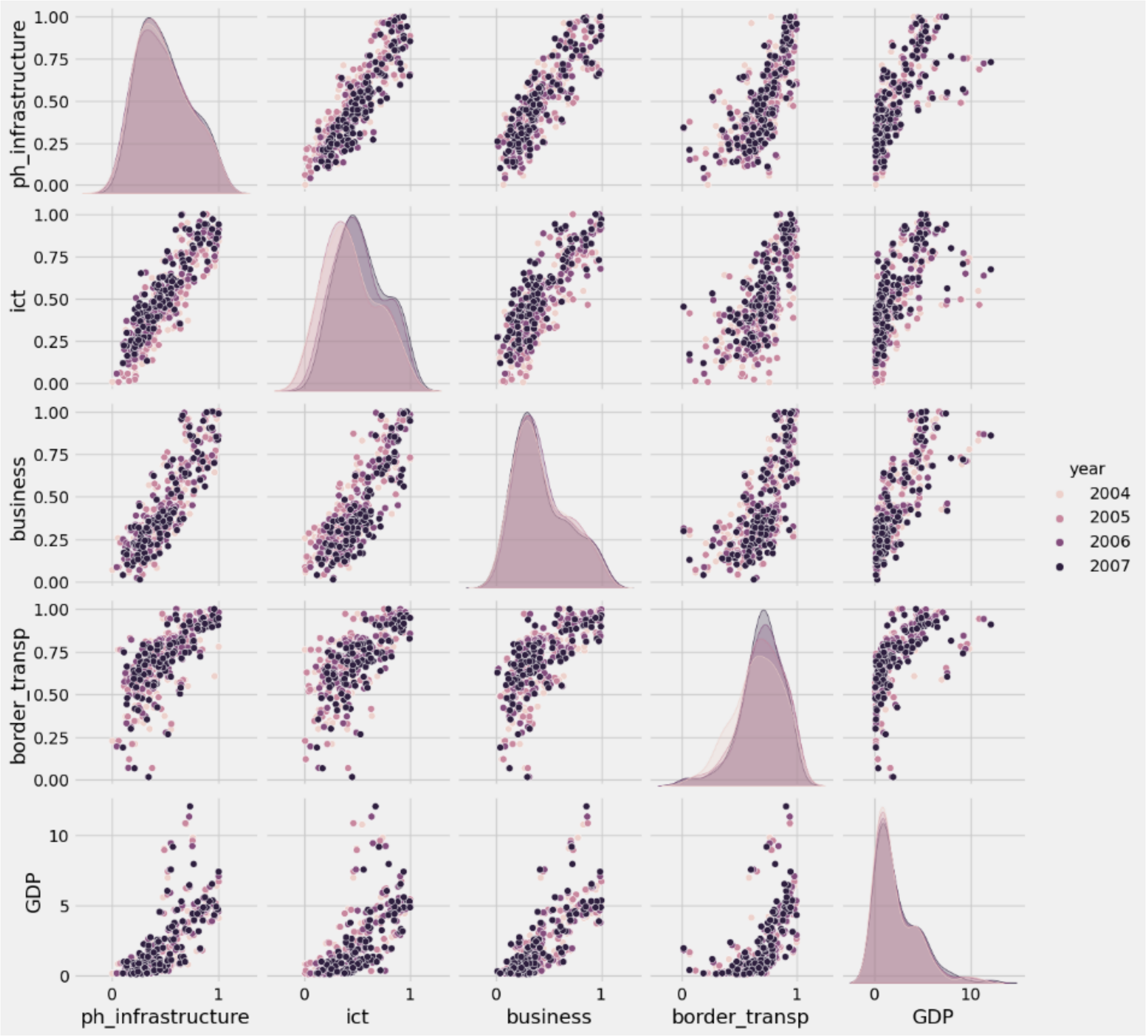
a. Selection of Dependent and Independent Variables
   - Our variable of interest was Gross Domestic Product per capita. (GDP per capita.) i.e., a dependent variable.
   - Explanatory or Independent variables in our study are mentioned here below:
      i. **ph_infrastructure**: measures the level of development and quality of ports, airports, roads, and rail infrastructure.
      ii. **ict**: Information and communications technology (ICT) is interpreted as the extent to which an economy uses information and communications technology to improve efficiency, and productivity as well as to reduce transaction costs. It contains indicators on the availability, use, absorption, and government prioritization of ICT.
      iii. **business**: measures the level of development of regulations and transparency. It is built on indicators of irregular payments, favouritism, government transparency, and measures to combat corruption.
      iv. **border_transp**: Aims at quantifying the level of efficiency of customs and domestic transport that is reflected in the time, cost, and number of documents necessary for export and import procedures.

b. Performing descriptive statistical analysis of our data set along with some simple graphical analysis (Scatter Plots, Box Plots etc.) in order to obtain meaningful summary of our dataset and its measures (mean, median, quartile etc.). We have divided GDP per capita. by 10000 to scale it down.

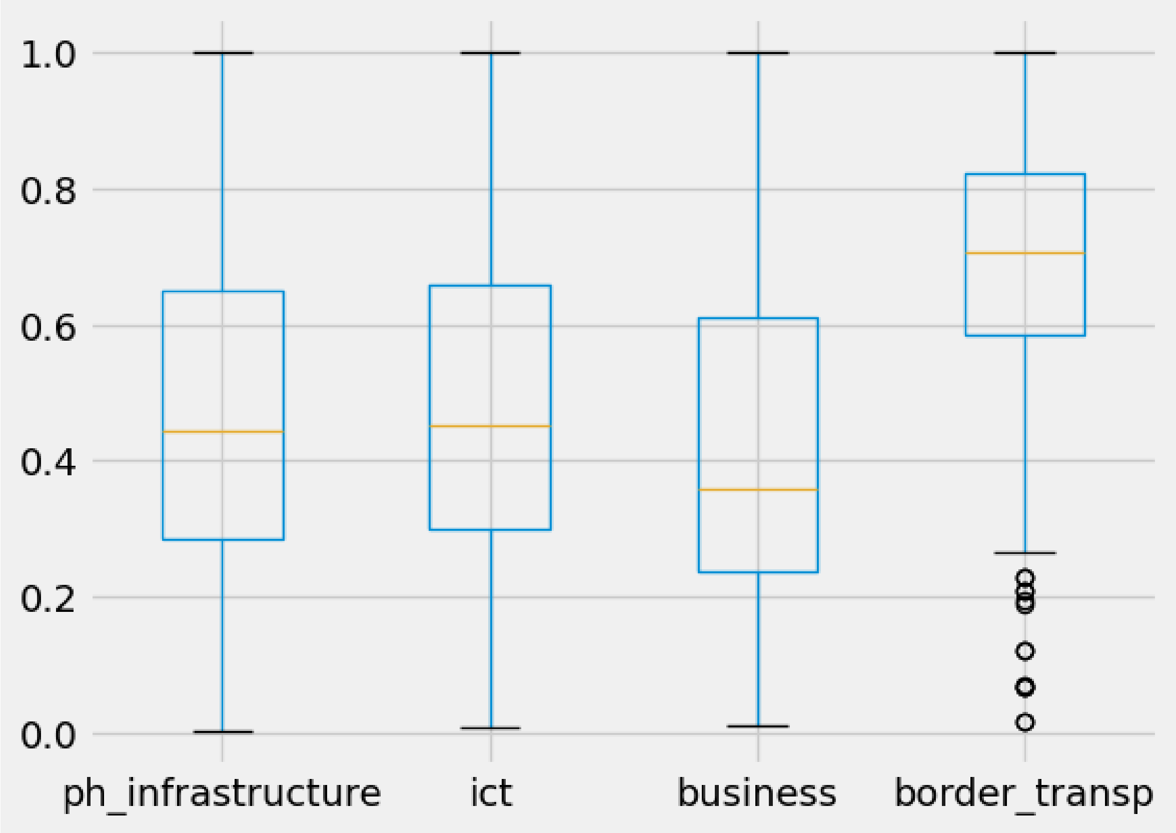
Descriptive Statistics of dataset:

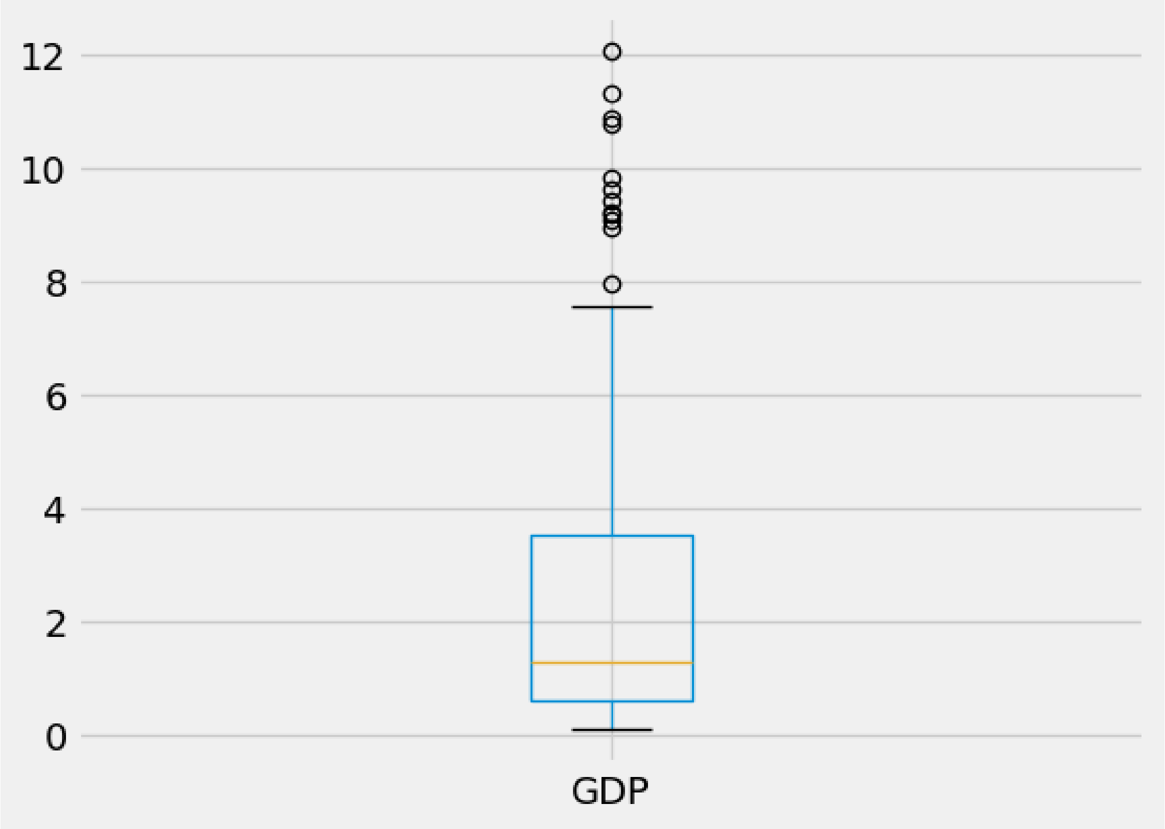
	Min. value	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Max. value**
GDP	0.0782	0.5942	1.292	2.211	3.53	12.065
ph_infrastructure	0.00115	0.2844	0.444	0.480	0.650	1
ict	0.00802	0.299	0.453	0.484	0.659	1
business	0.0123	0.238	0.359	0.430	0.611	1
border_transp	0.0159	0.585	0.705	0.689	0.822	1

Scatter Plots:



Boxplots:


1.0
0.8
0.6
0.4
0.2
0.0
ph_infrastructure
ict
business
border_transp


12
10
8
6
4
2
0
GDP

Observations:

i. Descriptive analysis of data provided us the simple summary of our data; however, for better understanding we've decided to perform some graphical analysis too.

ii. From "Scatter Plots" we have observed that relationship between dependent and independent variables was linear to large extent, it is for the same reason we've decided to perform "Multi Linear Regression Analysis" for our study.

iii. From "Scatter Plots" Linear relationship was observed among independent variables too. To check for the level of correlation among independent variables, it is now necessary to generate a correlation matrix for the same.

iv. From "Boxplots", we've detected the presence of outliers, we've decided to remove outliers. Also, small outliers are taken care of by using robust regression method.

c. Generation of Correlation matrix to check for interrelationship between the independent variables and to check for multicollinearity.

	ph_infrastructure	**ict**	**business**	**border_transp**	**GDP**
ph_infrastructure	1	0.875	0.859	0.684	0.706
Ict		1	0.836	0.683	0.665
business			1	0.681	0.774
border_transp				1	0.611
GDP					1

From the above correlation matrix, it is clear that there is a strong correlation between all pair of independent variables. However, correlation does not necessarily mean causality. Also, we've decided to check for multicollinearity in our regression model too.

d. White’s Test to check for the heteroskedasticity.

Null Hypothesis H0: Homoscedasticity.

Alternative hypothesis H1: Heteroskedasticity.

	Test Statistic	**P value**	**comments**
ph_infrastructure	15.952437	2.087693e-07	Reject the null hypothesis: Heteroskedasticity is present
ict	9.813778	6.814245e-05	Reject the null hypothesis: Heteroskedasticity is present
business	12.124404	7.571364e-06	Reject the null hypothesis: Heteroskedasticity is present
border_transp	12.245451	6.752348e-06	Reject the null hypothesis: Heteroskedasticity is present

As there is a presence of heteroskedasticity in all the independent variables, it is now required to obtain heteroskedastic standard robust error while performing regression analysis using these variables.

e. To understand the incremental explaining power of each of the independent variable, regression analysis was carried out in a forward stepwise manner, considering only one independent variable in Model 1, all combination of two variables in Model 2, considering all combination of three independent variables in Model 3 and finally considering all four variables:

	ph_infrastructure	**ict**	**business**	**border_transp**
	Model 1	Model 2	Model 3	Model 4
Adjusted R^2	0.490	0.433	0.592	0.423

	ph_infrastructure + ict	**ph_infrastructure + business**	**ph_infrastructure + border_transp**	**ict + business**	**ict + border_transp**	**business + border_transp**
	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Adjusted R^2	0.499	0.598	0.540	0.593	0.502	0.617

	ph_infrastructure + ict + business	**ph_infrastructure + ict + border_transp**	**ph_infrastructure + business + border_transp**	**ict + business + border_transp**	**ph_infrastructure + ict + business + border_transp**
	Model 11	Model 12	Model 13	Model 14	Model 15
Adjusted R^2	0.598	0.541	0.617	0.616	0.618

Models 10, 13,14 and 15 are giving relatively high adjusted R^2.

f. Test to check for multicollinearity was performed on model 10, 13,14 and 15:

Independent Variable	**VIF (value)**
business	1.95
border_transp	1.95

Independent Variable	**VIF (value)**
border_transp	2.105
ph_infrastructure	3.957
business	3.975

Independent Variable	**VIF (value)**
border_transp	2.148
business	3.506
ict	3.525

Independent Variable	**VIF (value)**
border_transp	2.184
business	4.335
ict	4.800
ph_infrastructure	5.386

As the VIF value is significantly smaller than 10 in every case, So, multicollinearity will not be problematic in our regression model.

Since correlation between ict and ph_infrastructure is high, we have decided to drop ict because dropping ict from model 15 does not bring a significant change in adjusted R^2 (model 13).

We can also drop ph_infrastructure from model 13 as the adjusted R^2 is not affected. But ph_infrastructure is important factor of hard infrastructure; we will not drop it.

g. Regression Analysis: After performing regression on different combinations of independent variables, the best and simple model obtained is as follows:

	coef.	**Robust std err**	**t**	**P>|t|**	**95% confidence interval**	
intercept	-2.240	0.177	-7.286	0.000	-2.845	-1.636
ph_infrastructure	0.801	0.384	1.420	0.156	-0.308	1.911
business	5.011	0.349	9.282	0.000	3.950	6.072
border_transp	2.700	0.357	4.696	0.000	1.570	3.831

Adjusted $R^2 = 0.617$

$R^2 = 0.620$

h. Omitted variable Bias:

- level of education
- population growth and density
- economic geography

Conclusion

Final regression equation for our analysis is based on two independent variables “business” and “border_transp” and is given here below:

GDP = -2.240 + 5.011 (business) + 2.700 (border_transp) + 0.801(ph_infrastructure)

It has been concluded for above equation that:

- “business” and “border_transp” are significantly contributing towards GDP.
- “ict” and “ph_infrastructure” is highly correlated with each other as well as with the other two variables. So, we have omitted ict because adjusted R^2 was not getting significantly affected.
- “ph_infrastructure” can also be omitted as it is also not affecting adjusted R^2. But it is an important indicator of Hard Infrastructure.

- This indicates that GDP per capita. and trade facilitators have a close relationship with each other. But for finding a linear regression and understanding the causal relation of GDP per capita. ph_infrastructure, business and border transport are considered.

References

- Dataset for GDP is taken from: https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD
- Dataset for export facilitators is taken from: world bank , https://doi.org/10.48529/tp5f-zg09
- ALBERTO PORTUGAL-PEREZ and JOHN S. WILSON * The World Bank, Washington (2011): Export Performance and Trade Facilitation Reform: Hard and Soft Infrastructure
- Stock, J. H., & Watson, M. W. (2015). Introduction to econometrics.