JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Mapping Correlation between GDP and Poverty rate of India using Linear Regression

Saumya Gupta[1], Pradeep Rai[2]

[1]Computer Science and Engineering, PSIT College of Engineering, Abdul Kalam Technical University, Kanpur, India
[2]Computer Science and Engineering, PSIT College of Engineering, Abdul Kalam Technical University, Kanpur, India

**Corresponding Author: saumyagupta431997@gmail.com*

Available online at: www.ijcseonline.org

18/May/2018, Published: 31/May/2018

***Abstract*—**We aim to project the impact of the Gross Domestic Product of India on the overall poverty rate of the country through the trailing years using data science. The correlation between GDP and Poverty rates has been modelled for the years 1981-2015. On getting a high correlation, we have used Linear Regression in order to train a model corresponding to the World development Indicators (a world-bank dataset) and found out their individual contributions towards the GDP of the country. The results found during the research are immensely helpful to define the major contributors of the current economic conditions of India. Also, these results can be further formulated to predict the poverty rates of the country.

***Keywords*—**GDP(Gross Domestic Product), Poverty rates, Data Science, Pearson's correlation, Linear regression

I. INTRODUCTION

Being in the top 10 list of largest areas, populations and democracies of the world, India is one of the rising economies. India's development stature has at times been referred to as "a paradox of wealth and poverty". There have been lengthy discussions as to whether or not the Gross domestic product varies inversely with the poverty rate of the country.

The World development Indicators dataset, formulated by the World Bank and published on Kaggle is used for the research. Last updated on 14-08-2017, this dataset contains information about the development indicators of 264 countries. The data for India is extracted and studied further. 32 indicators were filtered out based on their context similarities (finance and economy) and their consistent availability throughout the years.

To accumulate the labels for the study, the GDP World Bank data is used. Last updated on 16-02-2018, this dataset contains 14 csv files containing statistics of 14 social issues. Out of these, GDP by country data is chosen for further study.

The content of the paper is organized as follows, Section I contains the introduction of the problem in hand and the sources of the data. Section II contains a list of similar researches done which guided us to formulate our results. Section III contains the basic methodology involved and introduces the algorithm used i.e. Linear Regression. Section IV contains a detailed description of the datasets used for our research. Section V describes results and discussion, shows the graphical analysis of the study and derives mathematical constants. Section VI concludes the research work with future directions.

II. RELATED WORK

The research being a blend of two different domains – Data Science and Social Science, we found a plethora of work done on similar domains. Some of the most significant ones are listed here.
Angus Deaton in his paper –" Price Indexes, Inequality, and the Measurement of World Poverty" [1] studies the economic inequalities corresponding to the purchasing power parity (PPP) price indexes from the International Comparison Project.
Sanjay G. Reddy in his paper - "Counting the poor: the truth about world poverty statistics" [2] too takes a deep dive into understanding the trends of world poverty.
Human Development Reports of UNDP [3] of a series of years were referenced in order to determine the major development indicators. "Poverty Lines in Theory and Practice" [4] provided an insight to the methods being used to measure poverty rates since years and how appropriate

those methods have proved out to be. The immense usefulness of Linear Regression in analysis was proved in the papers – “Author Age Prediction from Text using Linear Regression” [5] and “Regression Models Based on Log-incremental Payments” [6].

III. Methodology

Pearson’s correlation is used in order to determine the correlation between GDP and Poverty, keeping in consideration the data for the years 1981-2015. Pearson’s correlation coefficient is the constant that measures the statistical relationship, or association, between two continuous variables. Revolving around the covariance concept, it is known to be the best method of measuring the association between variables of a similar context. It gives information about the magnitude as well as the nature of their association.

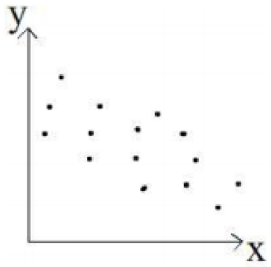


Fig. 1.1 Negative correlation

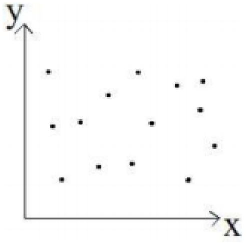


Fig. 1.2 No correlation

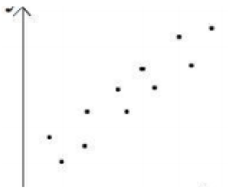


Fig 1.3 Positive correlation

The pandas.DataFrame.corr method is used in order to compute the correlation.

In order to find out individual correlation coefficients for each indicator with GDP, a linear regression model is trained that involves a total of 32 features. These features are trained for the years 1960 – 2016.

Linear regression:

Linear regression attempts to model the relationship between two variables by formulating a linear equation to define the relationship in data. One variable is considered to be an independent variable, and the other is considered to be a dependent variable.

Least Squares Regression:

This method calculates the line of best fit for the observed data by minimizing the sum of the squares of the perpendicular deviations from each data point to the line. Because the deviations are first squared, then summed, there are no cancellations between positive and negative values.

A linear regression line has an equation of the form $\boldsymbol{Y = a + bX}$, where $\boldsymbol{X}$ is the independent variable (in this case, the individual indicator) and $\boldsymbol{Y}$ is the dependent variable(in this case, GDP). The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept.

IV. DATASET DESCRIPTION

In our work we have used World Development Indicators Dataset[7] in order to find year-wise values of poverty indicators for India. The data set consists of 6 tables namely-

1. WDISeries
2. WDICountry
3. WDIData
4. WDIFootNote
5. WDICountry-Series
6. WDISeries-Time

Also, we used The GDP World Bank dataset[8] The GDP World Bank data consisted the GDP (Gross Domestic Product) of all the countries for the same time frame i.e., 1962 to 2014. It had following tables:

1. Account At a Financial Institution Male 15 Adults
2. Agricultural Machinery Per Unit of Arable Land
3. Adult Population Literate
4. ATMMachines Per 100000 Adults
5. Births Attended by Skilled Health Staff of Total
6. Domestic Credit To Private Sector
7. GDP by Country
8. GDP MetaData

9. Legal Rights Strength Index
10. MetaData Country
11. Population Per Country
12. public education expenditure as share of gdp
13. Rural Population of Total Population
14. Women Making Informed Choices to Reproductive HealthCare

We used the GDP by Country table to record the year-wise GDP for India.
Due to the unavailability of cumulated information regarding the poverty rates for previous years, we created this data through the information given on the websites – Statista[9] and Ieconomics[10] .

V. Results and Discussion

The following graph was plotted for year v/s GDP –

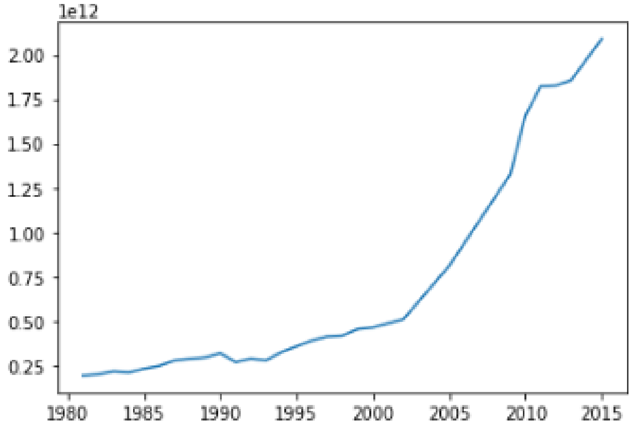


Fig 1.4 Graph between Years (X-axis) and scaled GDP (Y-axis)

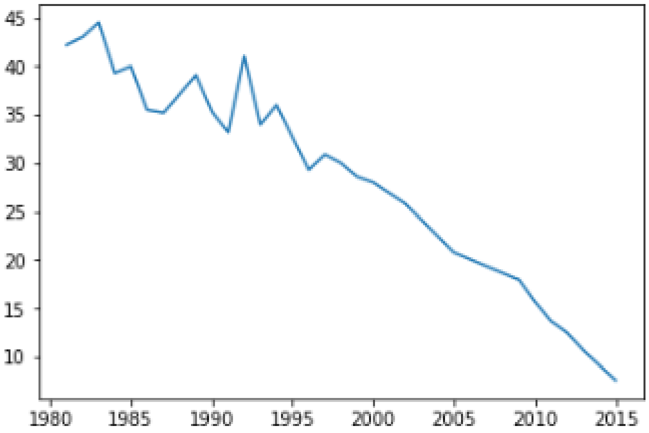
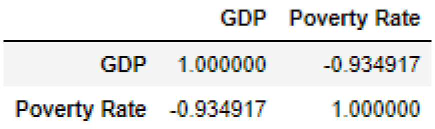


Fig 1.5 Graph between Year (X-axis) and scaled Poverty rate (Y-axis)



	GDP	Poverty Rate
GDP	1.000000	-0.934917
Poverty Rate	-0.934917	1.000000

Fig 1.6 Pearson’s correlation constant between GDP and poverty rate

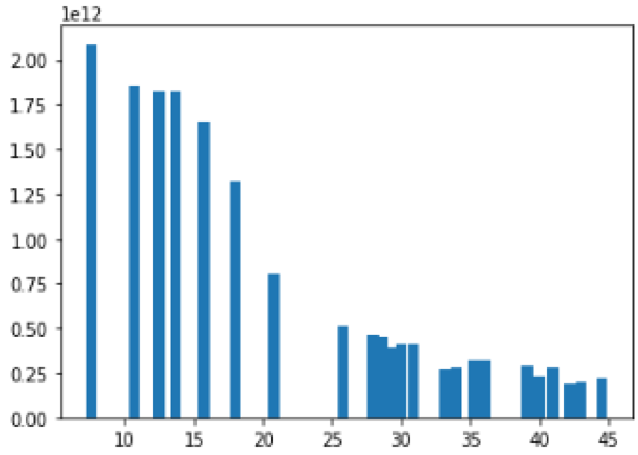


Fig 1.7 Bar graph between Poverty rate (X-axis) and GDP (Y-axis)

Correlation coefficients of individual indicators with the GD P of the respective year were calculated using a linear regress ion model trained with an accuracy score of 80.142368% These indicators, with their correlation coefficient against the GDP are listed as under-

1. NY.ADJ.NNTY.CD : Adjusted net national income (current US$).
   Coefficient: -7.197928e+00

2. NY.ADJ.NNTY.PC.CD: Adjusted net national income per capita (current US$)
   Coefficient: 8.907367e+09

3. NV.AGR.TOTL.ZS : Agriculture, value added (% of GDP)
   Coefficient : -8.284191e+10

4. FM.LBL.BMNY.GD.ZS : Broad money (% of GDP)
   Coefficient : -3.813806e+10

5. GC.DOD.TOTL.GD.ZS: Central government debt, total (% of GDP)
   Coefficient: -2.088255e+10

6. BN.CAB.XOKA.GD.ZS: Current account balance (% of GDP)
Coefficient: -4.337594e+09

7. FS.AST.CGOV.GD.ZS: Claims on central government, etc. (% GDP)
Coefficient: 8.341490e+11

8. NY.GDP.DISC.CN: Discrepancy in expenditure estimate of GDP (current LCU)
Coefficient: 4.582214e-02

9. FS.AST.DOMS.GD.ZS: Domestic credit provided by financial sector (% of GDP)
Coefficient- 8.106877e+11

10. FS.AST.PRVT.GD.ZS: Domestic credit to private sector (% of GDP)
Coefficient-4.372537e+11

11. FD.AST.PRVT.GD.ZS: Domestic credit to private sector by banks (% of GDP)
Coefficient- 4.372537e+11

12. GC.XPN.TOTL.GD.ZS: Expense (% of GDP)
Coefficient0-1.950827e+11

13. NE.EXP.GNFS.ZS: Exports of goods and services (% of GDP)
Coefficient-1.867089e+10

14. BX.KLT.DINV.WD.GD.ZS: Foreign direct investment, net inflows (% of GDP)
Coefficient- 3.233170e+11

15. BM.KLT.DINV.WD.GD.ZS: Foreign direct investment, net outflows (% of GDP)
Coefficient- 7.933767e+10

16. NE.CON.GOVT.ZS: General government final consumption expenditure (% of GDP)
Coefficient- 916751e+10

17. SE.XPD.TOTL.GD.ZS: Government expenditure on education, total (% of GDP)
Coefficient- 8.860038e+10

18. NE.GDI.TOTL.ZS: Gross capital formation (% of GDP)
Coefficient- 2.003623e+09

19. NY.GDS.TOTL.ZS: Gross domestic savings (% of GDP)
Coefficient- 1.757666e+10

20. NE.GDI.FTOT.ZS: Gross fixed capital formation (% of GDP)
Coefficient- 1.041529e+11

21. NE.DAB.TOTL.ZS: Gross national expenditure (% of GDP)
Coefficient : -1.557237e+10

22. NY.GNS.ICTR.ZS: Gross savings (% of GDP)
Coefficient : -2.574429e+10

23. NE.IMP.GNFS.ZS: Imports of goods and services (% of GDP)
Coefficient : 3.097854e+09

24. NY.GDP.DEFL.KD.ZG: Inflation, GDP deflator (annual %)
Coefficient : 5.323560e+09

25. NV.IND.MANF.ZS: Manufacturing, value added (% of GDP)
Coefficient: -1.098730e+11

26. CM.MKT.LCAP.GD.ZS: Market capitalization of listed domestic companies (% of GDP)
Coefficient : -5.250514e+09

27. GC.AST.TOTL.GD.ZS: Net acquisition of financial assets (% of GDP)
Coefficient : 2.921575e+11

28. GC.NFN.TOTL.GD.ZS : Net investment in nonfinancial assets (% of GDP)
Coefficient : -6.273402e+11

29. GC.NLD.TOTL.GD.ZS : Net lending (+) / net borrowing (-) (% of GDP)
Coefficient : -2.111217e+11

30. GC.REV.XGRT.GD.ZS : Revenue, excluding grants (% of GDP)
Coefficient : 2.410309e+11

31. GC.TAX.TOTL.GD.ZS : Tax revenue (% of GDP)
Coefficient : 2.811015e+10

32. NE.TRD.GNFS.ZS : Trade (% of GDP)
Coefficient : 2.176865e+10

VI. CONCLUSION

From the derived results, it is evident that surpassing numerous allegations that GDP might not be an accurate measure of poverty, GDP actually is a very promising determinant of poverty showing a high Pearson's correlation factor of -0.934917 for India. This clearly suggests that with increasing GDP, poverty has shown steep decrement in its figures.

The strongest contribution to India's GDP is given by - Claims on central government(%GDP) with the highest positive correlation coefficient of 8.341490e+11.

(Claims on central government, i.e. FS.AST.CGOV.GD.ZS include loans to central government institutions and net deposits. This means keeping all other factors constant, 1 unit change in FS.AST.CGOV.GD.ZS would reflect to 8.341490e+11 units of change in GDP.)

These results are immensely helpful to define the major contributors of the current economic conditions of India. Also, these results can be further formulated to predict the poverty rates of the country

REFERENCES

[1] Deaton, Angus. 2010. "Price Indexes, Inequality, and the Measurement of World Poverty." American Economic Review, 100 (1): 5-34.DOI: 10.1257/aer.100.1.5

[2] Sanjay G. Reddy, "Counting the poor: the truth about world poverty statistics"

[3] Human Development Reports, United Nations development program.

[4] Martin Ravallion, "Poverty Lines in Theory and Practice", Georgetown University

[5] Dong Nguyen Noah A. Smith Carolyn P. Rose, "Author Age Prediction from Text using Linear Regression", Language Technologies Institute Carnegie Mellon University, Pittsburgh, PA 15213, USA

[6] Christofides, S., Regression Models Based on Log-incremental Payments, Claims Reserving Manual, 1990.2, Institute of Actuaries, London.

[7] World Development Indicators | A World Bank data published by Kaggle.

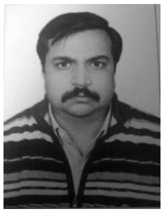
[8] GDP World Bank Data | A world bank data published by Kaggle

[9] Statista – The statistics portal for market data, market research and market studies.

[10] Ieconomics | Search and visualization of economic indicators.

Authors' Profile

Saumya Gupta is pursing Bachelor of Technology in Computer Science and Engineering from PSIT College of Engineering, Kanpur. Her major research work includes machine learning, data science, big data, deep learning and social sciences. She is currently working on similar researches and studying the data correlation between the GDP and poverty rates of Japan and USA.



Pradeep Rai, BTech, MTech in Computer Science and Engineering. He is currently working as a Head of the Computer Science and Engineering Department at PSIT College of Engineering, Kanpur. He has published more than 5 research papers in reputed international journals that have more than 100 citations online. His main research work focuses on digital image processing, machine learning, data mining and deep web. He has 14 years of teaching experience and 8 years of research experience. Some of his previous research papers are namely, 'A survey of clustering techniques', 'A review of MANET's security aspects and challenges, Comparison of data mining techniques for forecasting diabetes mellitus', 'Identifying Cyber black holes (deep web)', etc.