Statistical and Predictive Analysis of R&D In India

Jalem Raj Rohit, Desidi Siva Prakash
CoE of Systems Science,
Indian Institute of Technology, Jodhpur

Abstract

This research involves statistical analysis of the R&D data of India which includes the number of technicians, researchers, grants, number of research articles published and number of patent applications and identifying a trend and classification of the same depending upon the regression analysis. Machine Learning algorithms are implemented on the data for better and intelligent analysis. This is an attempt to learn and gain various insights of the R&D department of India from the data available and be able to tell a very interesting story.

Index Terms—Statistics, Regression, Machine Learning, Education, R&D.

I. INTRODUCTION

The R&D department of a country is very important for the growth and scientific development of the country. The number of researchers, the number of technicians, the published and the patent applications are the factors which drive the R&D department of any country. The research and Development in a country play a very significant role in the technological advancement of a country. So, proper analysis of the R&D is very necessary. Understanding the trends of various data measures in the R&D data is important for getting a proper and clear idea about the advancements and improvements in the industry.

Many statistical techniques exist for analyzing the data including regression analysis, Deep Learning approach, etc. So, proper selection of the analysis technique is very important. Machine Learning has some great concepts with which the regression analysis can be done effectively. Proper visualization is very important for comparative and statistical analysis.

II. DETAILS ABOUT DATA

The Data obtained is from an open-source Data repository. The dataset contains the previous 32 years of data about the percentage expenditure with respect to the GDP of a country, the number of researchers, the number of technicians, the number of research/scholarly articles published and also the number of patent applications. The data is in the excel and the csv formats. The datasets of the R&D departments of about 100+ countries are available and used for comparative analysis.

III. FEATURES OF DATA

The data primarily consists of 5 features, basing on which the analyses are done. They are:

- Research and Development Expenditure – This data says the amount of expenditure spent on Research and Development as a percentage out of the country's GDP. Expenditures for research and development are current and capital expenditures (both public and private) on creative work undertaken systematically to increase knowledge, including knowledge of humanity, culture, and society, and the use of knowledge for new applications. R&D covers basic research, applied research, and experimental development.
- Researchers in the R&D department -- Researchers in R&D are professionals engaged in the conception or creation of new knowledge, products, processes, methods, or systems and in the management of the projects concerned. Postgraduate PhD students (ISCED97 level 6) engaged in R&D are included.
- Technicians in R&D -- Technicians in R&D and equivalent staff are people whose main tasks require technical knowledge and experience in engineering, physical and life sciences (technicians), or social sciences and humanities (equivalent staff). They participate in R&D by performing scientific and technical tasks involving the application of concepts and operational methods, normally under the supervision of researchers.

- Scientific and Technical article journals -- Scientific and technical journal articles refer to the number of scientific and engineering articles published in the following fields: physics, biology, chemistry, mathematics, clinical medicine, biomedical research, engineering and technology, and earth and space sciences.

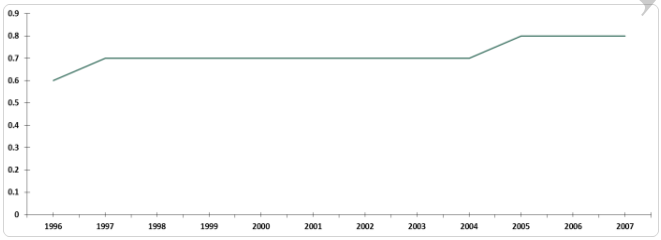
Patent Applications -- Patent applications are worldwide patent applications filed through the Patent Cooperation Treaty procedure or with a national patent office.

IV. INDIAN R&D DATA IS HIGHLY NON-UNIFORM

Through a detailed time-series analysis of the data of the India R&D Department, it can be concluded that the statistics of the R&D department is highly non-uniform with quite a lot of steep rises and dips. The necessary explanation of the time-series graphical analysis of all the five features is given below.

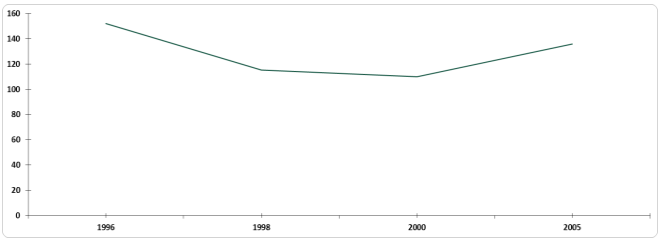
1. R&D Expenditure (% of GDP)

The R&D Expenditure of India has been stably increasing from 1996 to 2000; where it experienced a sudden drop from almost 0.75% in 2000 to about 0.7% in 2003; which can be considered as a steep drop.



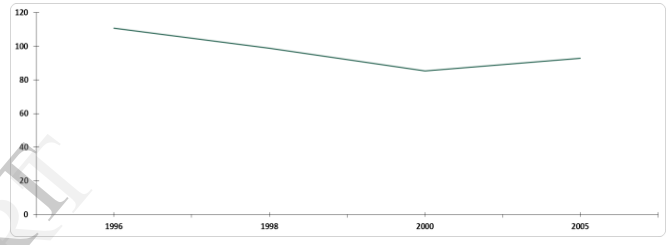
2. Researchers in R&D

The number of researchers also has seen a very steep drop from about 152 per million people in 1996 to 115.4 in 1998. The value have again improved to 135.8 in 2005. Both can be considered as very steep drops and also very steep rises; keeping in mind the high fluctuations. The respective changes are -24.07% and 23.45% respectively.



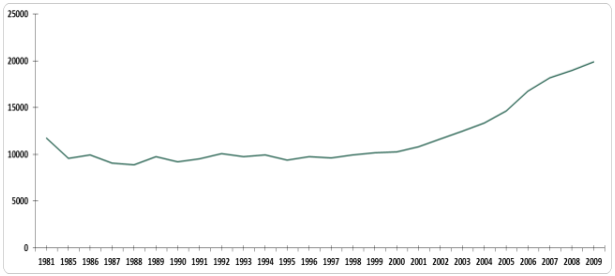
3. Technicians in R&D

The number of technicians per million people has taken a very steep drop from about 110.7 in 1996 to 85.4 in 2000; which can be considered a very awkward result in the data. The depreciation and the appreciation rates are close to 13.55% and 10.76% respectively; which are very steep values for the likes of a department.



4. Scientific and technical journal articles

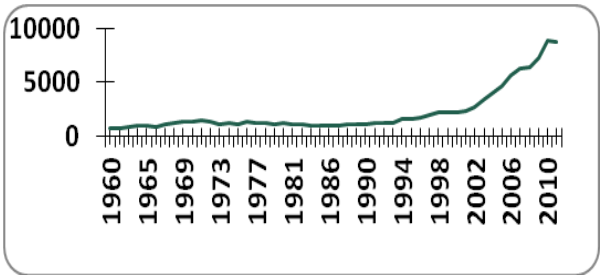
The number of technical journal articles released/published by the R&D department of India has been on a constant rise since 2000. Before the year 2000, there has been very unstable with a lot of minor fluctuations. The increase rate is approximately around 5% since 2000; which can be considered as healthy.



5. Patent Applications

The number of patent applications filed by the R&D department of India has also been on a constant rise since 2000. But this is not a smooth increase rate as compared to the previous data, since

the increase rates are around 25%; which can be considered as very steep.



V. REASONS FOR ABNORMAL BEHAVIOUR IN 2000-2001

From the above visualizations, it can be clearly stated that all the data obtained showed an abnormal behavior in the region 2000-2001. By proper analysis of the financial sector happenings of India in the time-span of 2000-2001; it can be clearly understood that India has undergone a very steep drop in the economic indicators during the above mentioned financial year. So, the expenditure has taken a steep drop till 2000. Immediately after the financial crisis, Indian markets have shown a very appreciable growth or improvement which has resulted in the improvement of the Expenditure data and also the remaining four features as observed. From the analysis of the expenditure data, it would be very logical to assume depreciation in the remaining features too. The steepness cannot be accurately determined; but the depreciation can be predicted through the previous visualizations behavior.

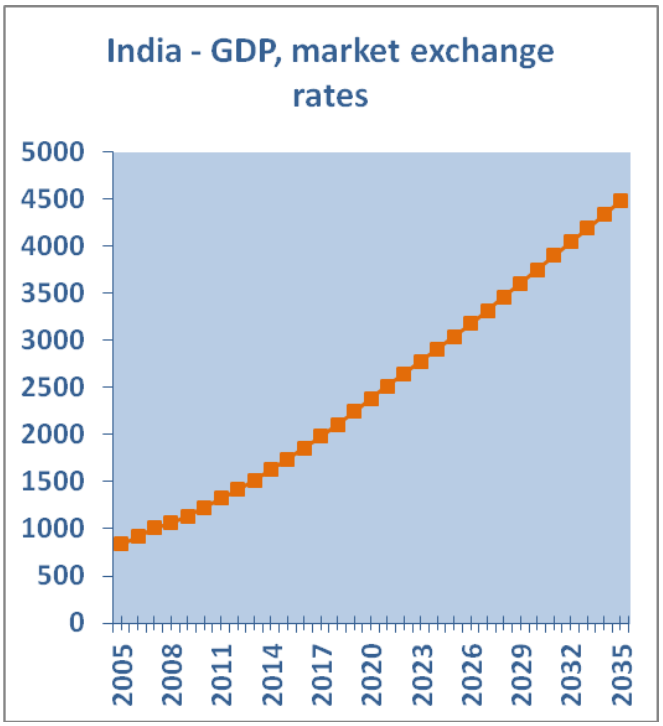
But an interesting point to observe is that, the rest of the 4 visualizations and data do not show any depreciation in the near post- 2000 region, unlike the data of the GDP Expenditure, which is very interesting. By careful analysis, it can be concluded that the data of the remaining four features show a steady growth from the year 2000. That was the time when India has started to feel the impact of the financial crisis. The data-sets have shown a steady decline till 2000; and then started to shoot up; which can be considered healthy in terms of national development. 23.45%, 8.63%, 7.98%, 18% are the average improvement rates of the features. So, it can be stated that the R&D department have shown very healthy signs of development during the post-financial crisis period; even though the expenditure have taken a slump for a short period of time.

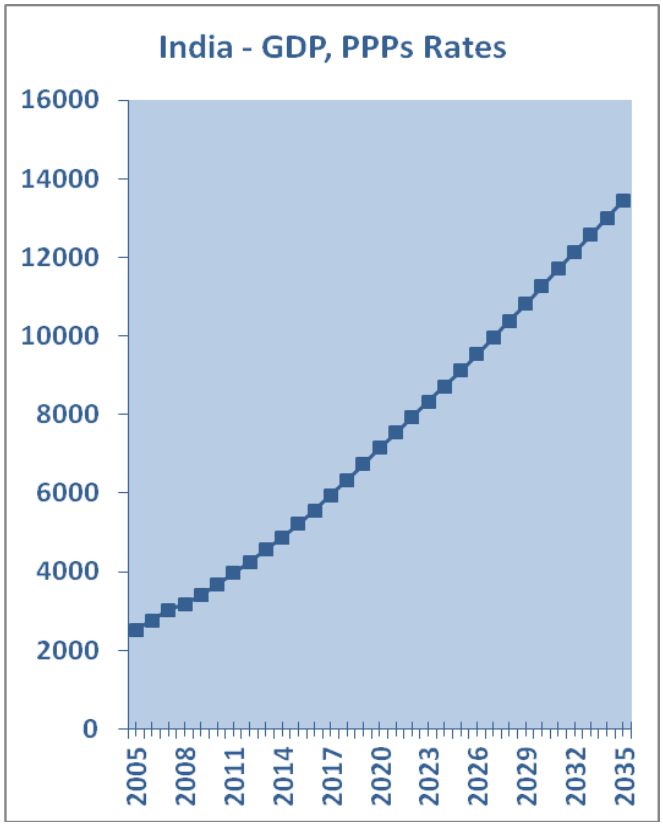
VI. FORECAST OF INDIA's GDP

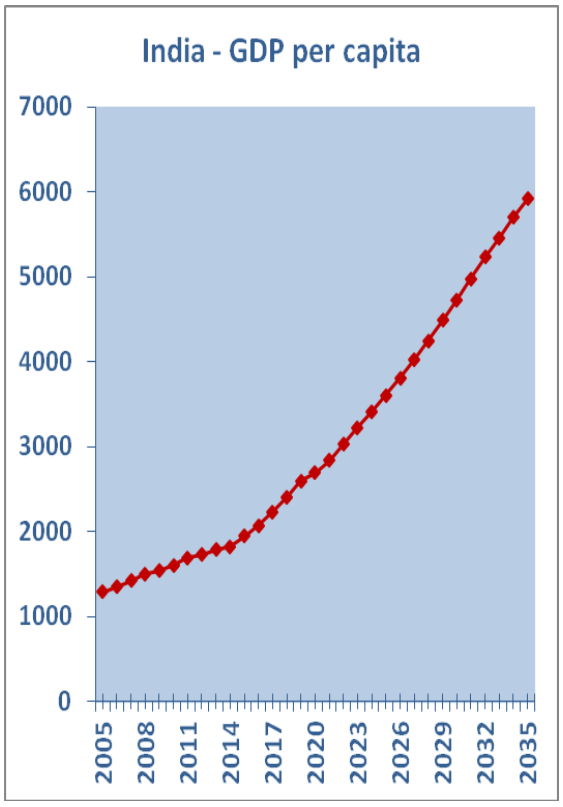
From external data sources, the data of India's previous 20+ years GDP data has been obtained, The forecast has been done on the data using the technique of Logistic Regression. The reason of choosing Logistic Regression; over linear regression is primarily the shape of the data. As the data is non-linear in nature; the Linear Regression approach would increase the mean squared error.

The constant values of the Logistic Regression formula are selected by applying the Gradient Descent Algorithm on the training data; for minimizing the cost function. (The entire approach and code is present in the project repository.)

The results which have been obtained are very encouraging in perspective of the Indian Economy. The resulting visualization is shown below. The expected GDP is constantly increasing.







VII. SIMILAR FORECASTS

Similar trends are shown by the remaining four features. This can also be inferred from the correlations observed from the viualizations of the features. The correlation of all the features during the post 2011 financial period is almost 1. So, a similar trend can be obtained for the remaining data too.

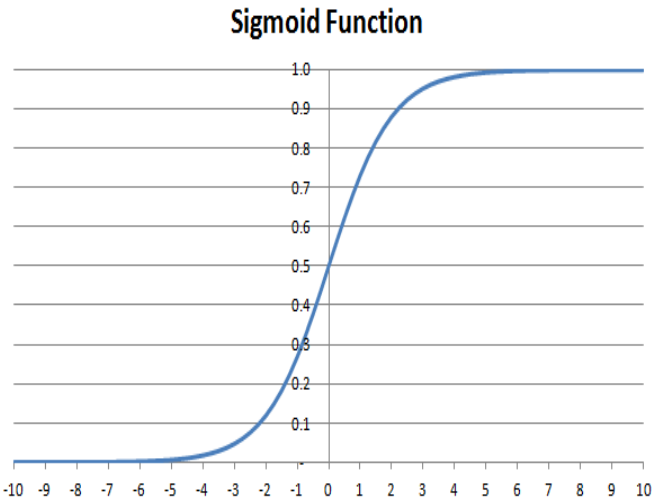
According to the Human Development Index (HDI); India is forecasted to cross China and the United States of America by 2020. The fore-cast is also done using the Logistic Regression approach; which a similar optimization technique, that is the Gradient Descent Algorithm.

The visualizations of the HDI statistical analyses are present in the repository.

VIII. MATHEMATICAL FORMULAE USED

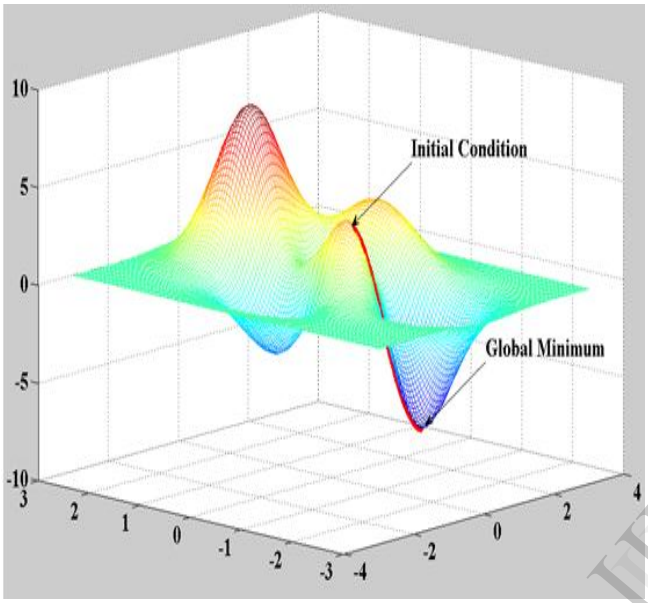
The mathematical formulae used during the analysis of the R&D Data are mentioned below.

The type of function used in the Logistic Regression analysis if the Sigmoid function, which has a logistic shape and is asymptotic to 1 and 0.



The above image is scaled for better resolution. The sigmoid function is symmetric.

The optimization algorithm used for the optimization of the error function is the Gradient Descent Algorithm. The general intuition of this algorithm is that it slowly decreases the value of theta by decreasing the gradient of the error function iteratively. When the minimum point (global minimum) is reached; the algorithm stops and returns the value of theta; which is the optimal value. So, by using the value of theta obtained through this algorithm, the error(squared) is getting minimized.



IX. SOFTWARE AND TECHNOLOGY USED

The following software and technology is used in the analysis of the data and during the work on this paper.

- OCTAVE and MATLAB – Octave and Matrix Laboratory are the software used for the simulations of the error functions and the implementation of the Gradient Descent Algorithm. The learning rate assumed in the algorithm is 0.01.
- R – R is a statistical computing language; used in the statistical analysis of the data obtained. The visualizations are obtained using the ggplot2 package in the R language.
- QGIS – The geographical visualizations are done using the open-source project QGIS.
- Python – The Python programming language is used to deploy the sigmoid function into the Logistic Regression.

X. REFERENCES

The following bibliography is used in the analysis:

[1] Indian Economy – Endemic Crisis, Prof. Arvind.

[2] The code and the necessary visualizations are present in the repository at www.github.com/Dawny33

[3] The data obtained is completely open-source.

ANNEXURE:

The code for the Sigmoid function:

In Matlab:-

```
function g = sigmoid(z)

g = zeros(size(z));
one = ones((size(z)));
g = one ./ (one + e .^ (-z));
end
```

The code for the Gradient Descent algorithm

In Matlab:-

```
function [J, grad] =costFunctionReg(theta,X, y, lambda)
m = length(y); % number of training examples
n = size(X,2); % number of features
J = 0;
grad = zeros(size(theta));
one = ones(m,1);

g = sigmoid(X * theta);
grad = 1/m * X' * (sigmoid(X * theta) - y) + lambda/m * theta;
grad(1) = 1/m * X(:,1)' * (g - y); %dont add lamda/m *theta to grad0

J = 1/m * ( ((log(g))' * -y) - ((one - y)' * (log((one - g))))) + lambda/(2*m) * sum(theta(2:n,1).^2);
end
```