IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9617

INDIAN LITERACY RATE: A STUDY

Likhitha S[1] and Deepa Yogesh Kamat[2]

[1,2]Department of Statistics, Nrupathunga University Government Science College, Bengaluru-560001, India

Abstract: A census is the procedure of systematically acquiring and recording information about the members of a given population. The term is used mostly in connection with national population and housing censuses; other common censuses include agriculture, business, and traffic censuses. Here we are focussing on literacy rate, any one above age 7 who can read and write in any language was considered a literate. In censuses before 1991, children below the age 5 were treated as illiterates. In this paper, we are estimating the population and literacy rate for the year 2021 using 2011 census data. We have used cluster analysis, simple regression, logistic curve method, multiple regression and R programming as the tools to estimate the population and literacy rate of 2021.

Key words: Census, literacy rate, regression, cluster analysis, R programming

1. INTRODUCTION

A census is the procedure of systematically acquiring and recording information about the members of a given population. The term is used mostly in connection with national population and housing censuses; other common censuses include agriculture, business, and traffic censuses. The decennial Census of India has been conducted 15 times, as of 2011. While it has been conducted every 10 years, beginning in 1872, the first complete census was conducted in the year 1881. Post 1949, it has been conducted by the Registrar General and Census Commissioner of India, under the Ministry of Home Affairs, Government of India. All the census since 1951 are conducted under 1948 Census of India Act.

Literacy: Any one above age 7 who can read and write in any language was considered a literate. In censuses before 1991, children below the age 5 were treated as illiterates. The literacy rate taking the entire population into account is termed as "crude literacy rate", and taking the population from age 7 and above into account is termed as "effective literacy rate". Females constitute about 50% of country's human resource. But lack of education snatches their chance to be a part of the progress and development of India. Education is regarded as a key instrument for the empowerment of women. Education changes their worldview, improves their chances of employment, facilitates their participation in public life, and also influences their fertility. Although considerable progress has been made with regard to literacy and education, the overall picture still remains unfavourable to women. Some common causes for the lack of education amongst girl child and women are:

i) Poverty is the root cause of many problems in India and also of low female literacy rate. More than one-third of the population in India is living below the poverty line. Though the government is putting efforts to make the primary education free but still parents are not ready to send their girls to school.

ii) Accessibility to schools - In most of the rural areas, lack of easy accessibility to school is another reason for low female literacy rate. Parents do not prefer to send girls to schools if these are located at a far distance from their village or home.

iii) Lack of adequate school facilities - some of the schools are really in pathetic conditions and do not have even basic facilities.

But in spite of all these reasons, National Women's Parliament wants women of all ages to understand and realize that education can actually end the vivacious cycle of poverty, their misfortune, so that they can live a life with pride.

Our government has introduced various schemes to ensure higher literacy rate, like "Shiksha Sahayog Yojana", "Sarva Shiksha Abhiyan ","Saakshar Bharat", "Kanya Saaksharta Protsahan Yojna", "Kasturba Gandhi Balika Vidyalaya Yojna", etc. But unfortunately, due to lack of awareness people are unable to take advantage of those schemes.

Sex ratio: Sex ratio is used to describe the number of females per 1000 males. Sex ratio is a valuable source for finding the population of women and what is the ratio of women to that of men. Since decades, India has seen a decrease in the sex ratio. The major cause of the decrease of the female birth ratio in India is considered to be the violent treatments meted out to the girl child at the time of the birth. The sex ratio in India was almost normal during the phase of the years of independence, but thereafter it started showing gradual signs of decrease. Though the sex ratio in India has gone through commendable signs of improvement in the past 10 years, there are still some states where the sex ratio is still low and is a cause of concern for the NGO organisations. Some facts related to the sex ratio in India follows, the main cause of the decline of the sex ration in India is due to the biased attitude which is meted out to the women. The main cause of this gender bias is inadequate education.

Data Source Website: https://www.kaggle.com.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9617

Original census data released (and owned) by the Registrar General and Census Commissioner of India under the Ministry of Home Affairs, Government of India. http://censusindia.gov.in/

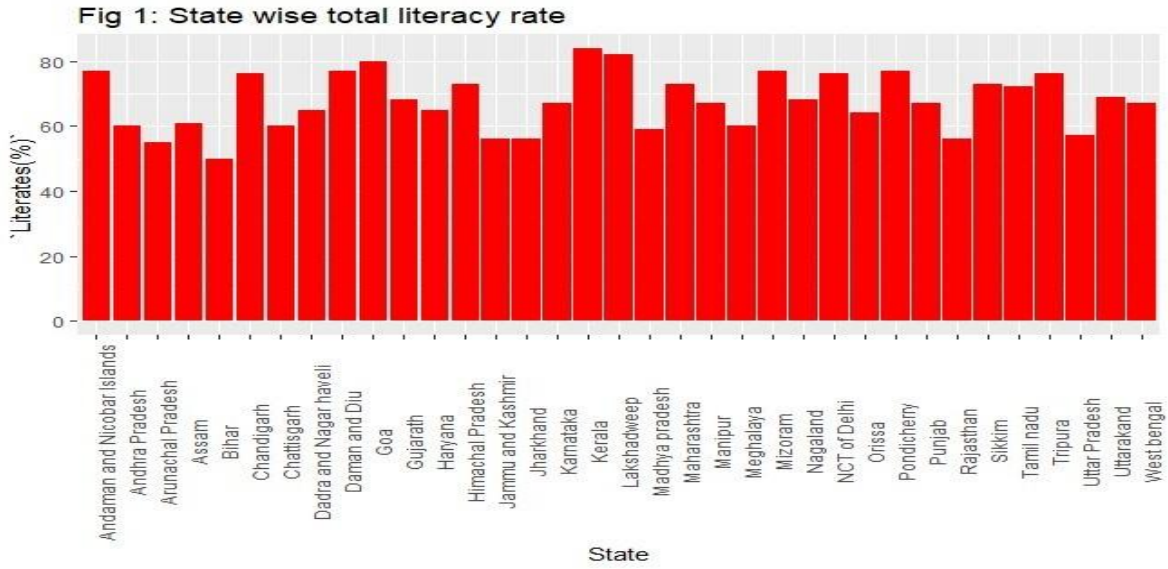
In section 2, we discuss the objectives, in the Section 3, analysis of the data and interpretation of the results are given. Section 4 gives the references.

2. Objectives: i) To study the literacy rate, female literacy rate in India.
ii) To predict population for the year 2021.
iii) To forecast literacy rate.

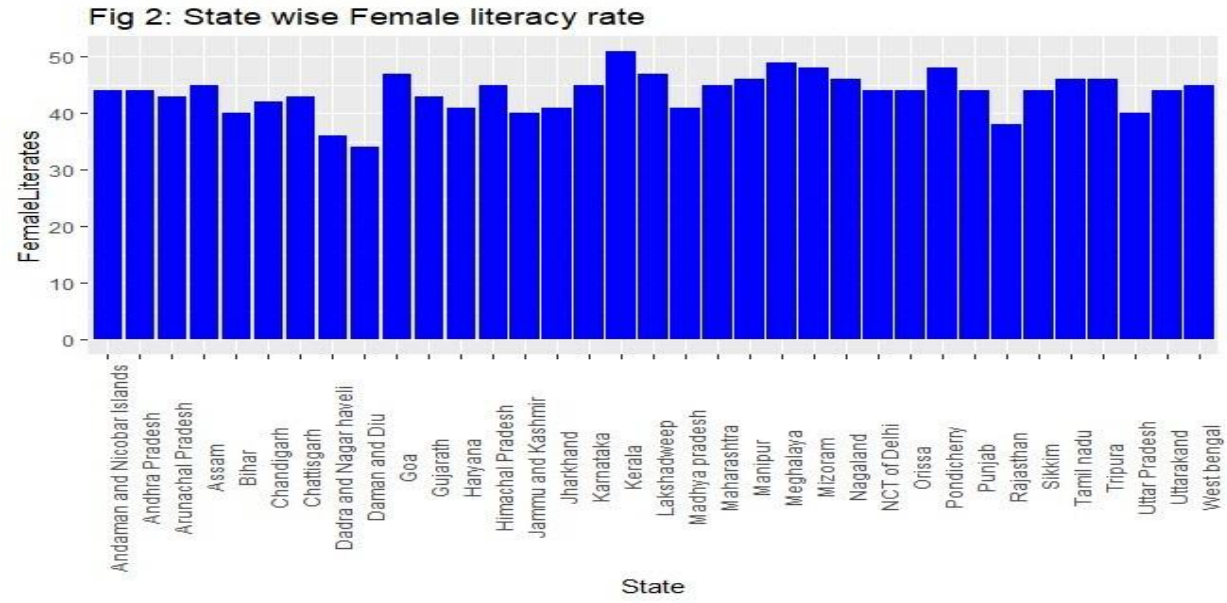
3. Analysis and Interpretation:

In the data, we have considered the States, total literacy rate, female literacy rate and sex ratio, total population, male population, female population. We have considered percentages of total literacy rate and female literacy.

From Fig 1, we can see that Kerala has highest literacy rate and where as it is less in Bihar.



From Fig 2, we can see that Kerala has highest female literacy rate and where as it is less in Daman and Diu.



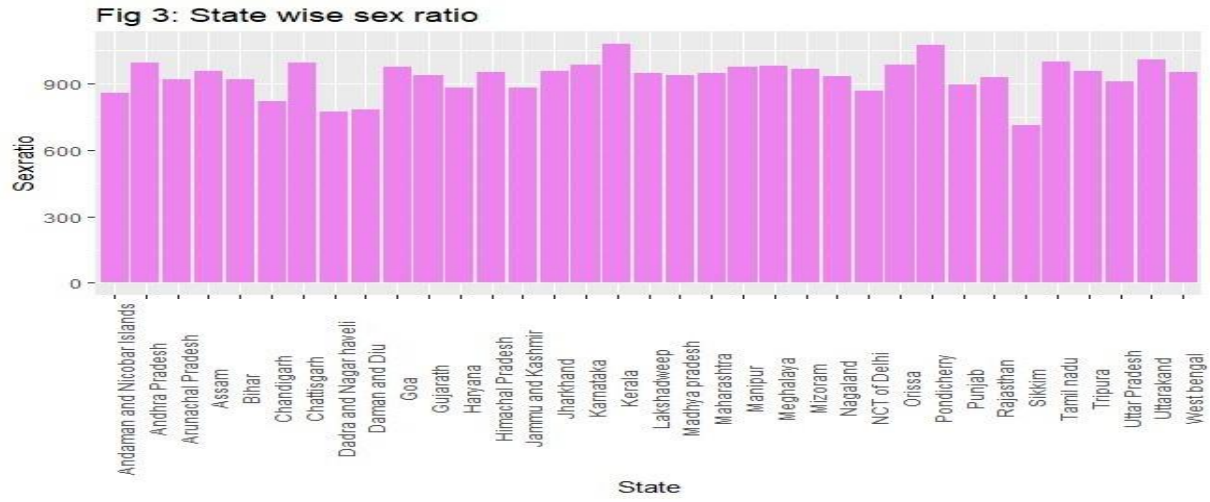
From Fig 3, we can see that Sex ratio is more in Kerala and less in Sikkim.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9617



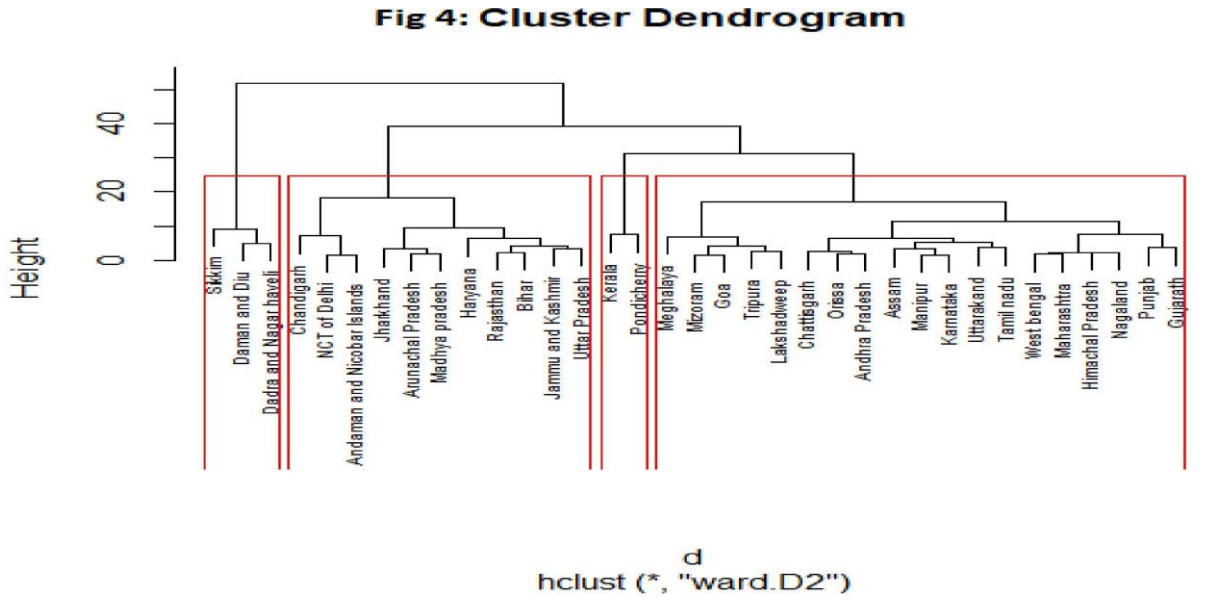
Cluster Analysis

Cluster analysis or clustering is the procedure of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). The objectives of cluster analysis is to assign observations to groups ("clusters") so that observations within each group are similar to one another with respect to variables or attributes of interest, and the groups themselves stand apart from one another. In other words, the objective is to divide the observations into homogeneous and distinct groups. In contrast to the classification problem where each observation is known to belong to one of a number of groups and the objective is to predict the group to which a new observation belongs, cluster analysis seeks to discover the number and composition of the groups. Cluster analysis is also used to group variables into homogeneous and distinct groups.

Hierarchical clustering

Hierarchical methods of cluster analysis start with each object in a cluster of its own and then continually join clusters together, until there is only one cluster consisting of all the objects. Clusters are joined on the basis of 'shortest distance' between clusters. Alternatively, we can start with one cluster of all the objects and then split this cluster into more and more clusters. The first method is the most common among the two. The graph which represents hierarchical clustering is called 'dendrogram".

Dendrogram: To check how many clusters we need to carry out the analysis, we have generated dendrogram. Dendrogram is a main graphical tool for looking at a hierarchical cluster solution. This is a tree-like display that lists the objects which are clustered along x-axis, and the distance at which the cluster was formed along the y-axis.



IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9617

By looking at the dendrogram we have considered the 4 clusters. Then k-means clustering is used.

K- means clustering:

This is a non-hierarchical clustering method. K-means clustering covers a group of techniques that find clusters by optimizing various criteria. K-means clustering needs the number of clusters to be stated. Knowing the number of clusters to choose could be difficult. The dendrogram from a hierarchical method used on the data might indicate the number to choose. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Table 1: Cluster analysis

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Size	2	18	3	12
Sex Ratio	1078	969	757	898
Literates(%)	80.5	69.1	71.8	62.7
Female Literates(%)	49.7	45.4	37.8	41.4

By looking at the above table, we conclude that

i) Cluster 1 is the smallest cluster consisting of 2 states. It has highest sex ratio. And also highest total literacy rate and female literacy rate.

ii) Cluster 2 is the largest cluster consisting of 18 states.

iii) Cluster 3 is the second smallest cluster consisting of 3 states. It has lowest sex ratio and female literacy rate.

iv) Cluster 4 is the second largest cluster consisting of 12 states. It has lowest total literacy rate.

Regression Analysis

The correlation between female literacy rate and sex ratio is 0.6431768 and the p-value is $3.084 * 10^{-5}$. Since the correlation coefficient between female literacy rate and sex ratio is significant we have fitted the simple linear regression model. Simple linear regression: It is a linear regression model with a single explanatory variable. That is, it concerns two-dimensional sample points with one independent variable and one dependent variable. It is a linear function that, as accurately as possible, predicts the dependent variable values as a function of the independent variables. The simple linear regression model is given by, $Y = \beta_0 + \beta_1 X + \epsilon$

where, Y is the dependent variable, X is the independent variable, β_0 is the intercept and β_1 is the slope.

In our study, Y= Sex ratio, and X=female literacy rate.

$\hat{y} = 325.39 + 13.92 * x$

i) p-value: $3.084 * 10^{-5}$. Hence it is statistically significant.

ii) $R^2 = 0.4136764$. Hence 41% of total variation in sex ratio is explained by variation in female literacy rate.

iii) Adjusted $R^2 = 0.395909$. Hence 39.6% of total variation in sex ratio is explained by variation in female literacy rate.

Scatter plot: Scatter plot is a 2 dimensional data visualization that uses dots to represent the values obtained for 2 different variables. One plotted along the x-axis and the other plotted along the y-axis. When we to show the relationship between 2 variables, scatterplots are used.

Fig 6 shows that there is a linear relationship between the female literacy and sex ratio.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

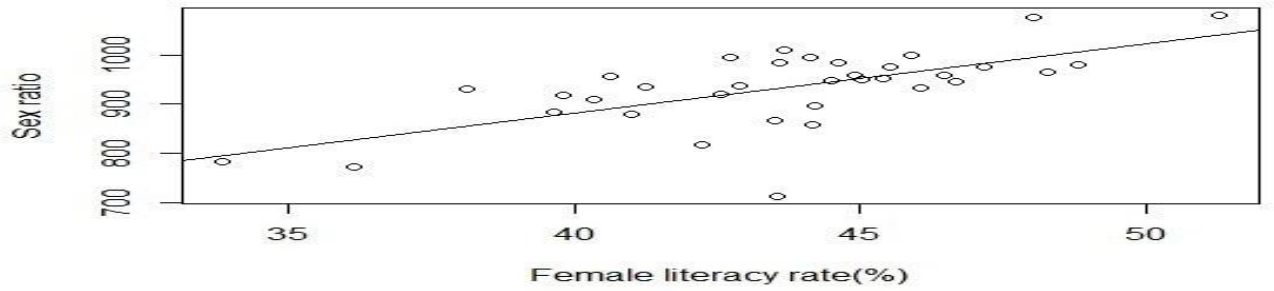
IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

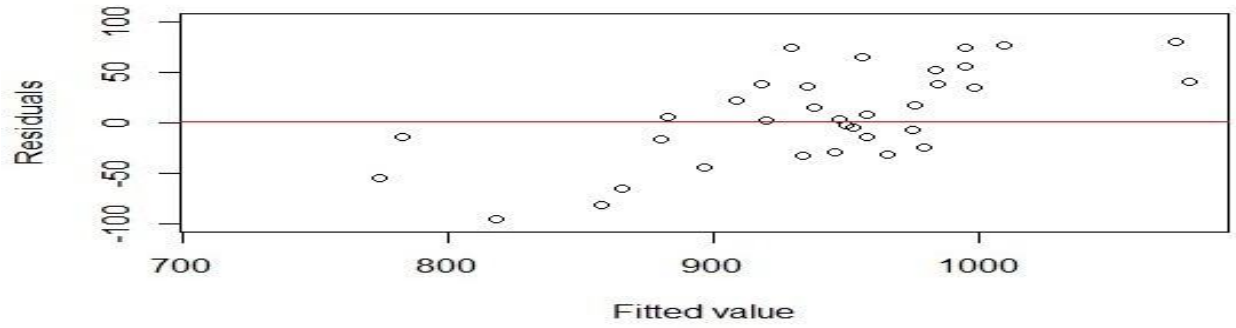
DOI: 10.17148/IARJSET.2022.9617

Fig 6: Scatter plot



Residual plot: A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

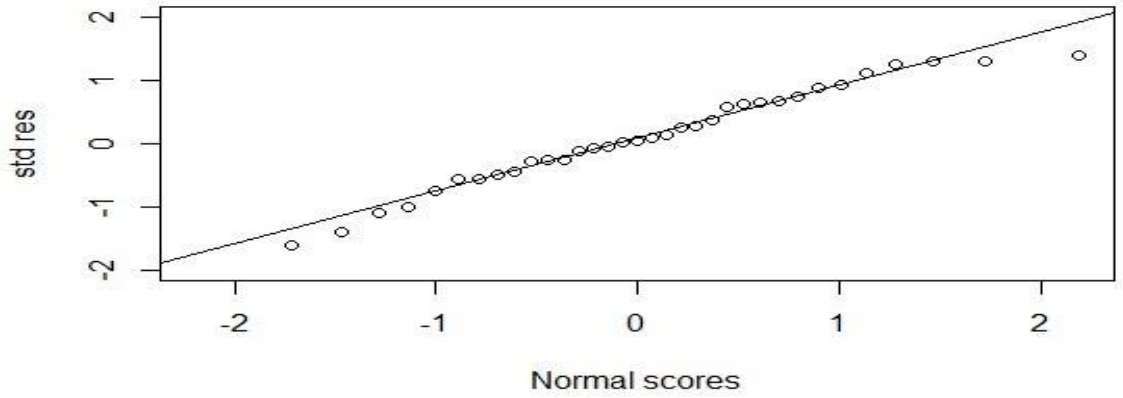
Fig 7: Residual plot



Since the plot does not follow any particular pattern, there is no model inadequacy.

Normal Probability Plot: It is used to check whether or not a data set is approximately normally distributed. From Fig 8, we can say that, the data follows normal distribution.

Fig 8: Normal prob plot



Population Forecasting: Forecasting is the expected outcome in the near future. Projection gives a clear picture of what future will look like, based on the past and assumptions about the future. Projection acts as an indicator of change over time, helps in evaluating impact of any policy or program by comparing the feature before and after the occurrence of an evet. The demographic projections are based on the basis of projected populations for particular area and for particular period. The present and past population are obtained from census population records. After collecting these population figures, Logistic curve method is used to predict the population.

Logistic Curve Method: This method is used when the growth rate of population due to births, deaths and migrations takes place under the normal situation and it is not subjected to any other changes like epidemic, war, earthquake or any natural disaster, etc. The population follows the growth curve characteristics of living things within limited space and economic opportunity. If the population is plotted with respect to time, the curve so obtained under normal condition

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9617

looks like S-shaped curve and is known as logistic curve. A mathematical solution for the logistic curve, which can be represented by an autocatalytic first order equation, is given by $\log_e\left(\frac{P_S-P}{P}\right) - \log_e\left(\frac{P_S-P_0}{P_0}\right) = -k.P_s.t$

where, P= Population at any time t , P_s= Saturation population, P_0= Base population, t= time in years k= constant.

After solving we get, $P = \frac{P_S}{1+m\log_e^{-1}(n.t)}$ where, m and n are constants, given by, $m = \frac{P_S-P_0}{P_0}$, $n = -K.P_s$

If only three pairs of charactristic values P_0, P_1, P_2 at times $t = t_0$, t_1 and $t_2 = 2t_1$ extending over the past records are chosen, the saturation population P_s and constant m and n can be obtained by the following equations, $Ps = \frac{2P_0P_1P_2-P_1^2(P_0+P_2)}{P_0P_2-P_1^2}$, $m = \frac{P_S-P_0}{P_0}$, $n = \frac{2.3}{t_1}\log_{10}\frac{P_0(P_S-P_1)}{P_1(P_S-P_0)}$

• Total population for the years 1991, 2001 and 2011 is given below:

Table 2: Total population

1991	846420000
2001	1028610000
2011	1210190000

Predicted total population for the year 2021 is 1379462280 (i.e., 1379 million.)

• Total male population for the years 1991, 2001 and 2011 is given below:

Predicted male population for the year 2021 is 708164458 (i.e. 708 million).

Table 3: Total male population

	Male population
1991	439360000
2001	532160000
2011	623720000

• Total female population for the years 1991, 2001 and 2011 is given below:

Predicted male population for the year 2021 is 671316664 (i.e. 671 million).

Table 4: Total female population

	Female population
1991	407060000
2001	496450000
2011	586470000

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

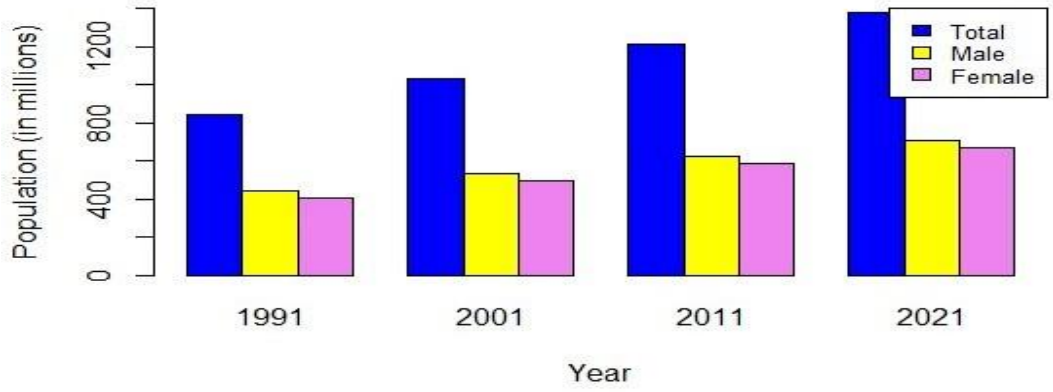
IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9617

Fig 9: Population in India



Literacy Projection: Literacy is the ability to read and write in any language. Since literacy plays a vital role for education and development, early formulation of literacy goals are required. Multiple linear regression is used to forecast the literacy rate here.

Multiple Regression: Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable. The variables we are using to predict the value of the dependent variable are called the independent variables.

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, where Y is dependent variable, X_1 and X_2 are the independent variables, β_0 is the intercept, β_1 and β_2 are the regression coefficients for X.

In our study, literacy rate is the dependent variable, and the independent variables are total male population and total female population.

$\hat{y}$= -1.3881332+ (0.3907186*x_1)-(0.2860505*x_2).

Table 4: Actual and predicted literacy rate

Year	Actual Literacy rate(%)	Predicted Literacy rate(%)
1951	18.32	20.89640
1961	28.31	26.11599
1971	34.45	34.04669
1981	43.56	42.29488
1991	52.21	53.83827
2001	65.38	64.52691
2011	74.04	74.55084
2021	_	84.66256

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9617

Fig 10: Literacy rate in India

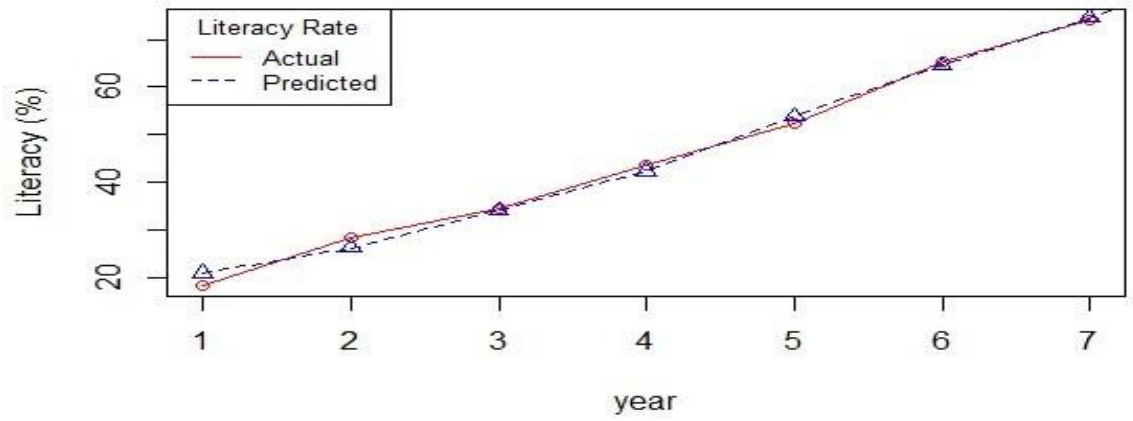
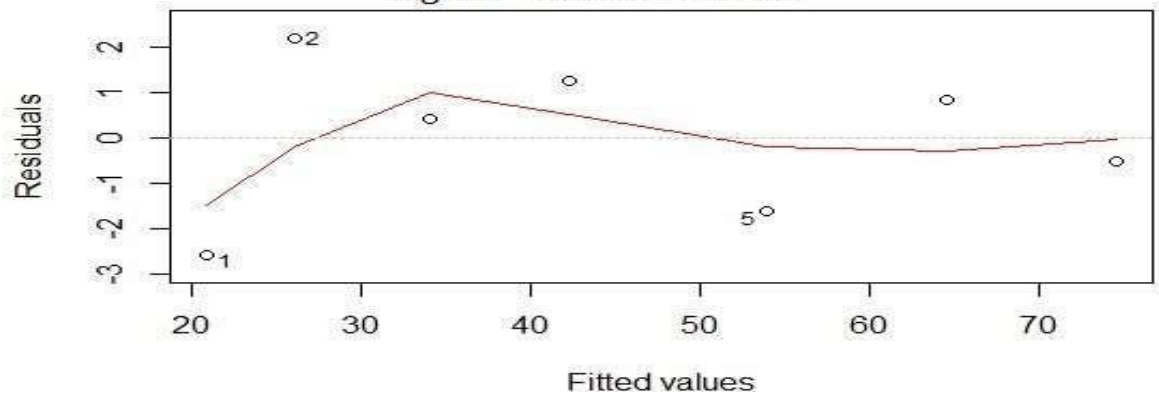


Fig 11: Residuals vs Fitted



Since the plot (Fig 11) does not follow any particular pattern, there is no model inadequacy.

Fig 12: Normal Q-Q

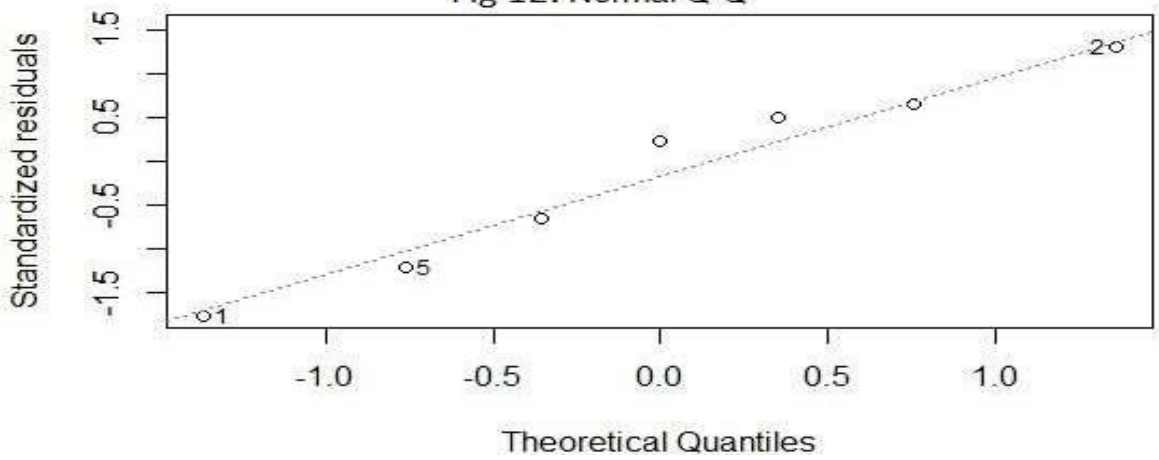


Fig 12 shows that the residuals follows normal distribution.

CONCLUSION:

In the country like India the growing population prediction is very much needed and important to predict the demands and supply, business growth, medical measures, real estates, etc. Henceforth, India trying best to make the country literacy rate towards cent percent, by creating various schemes across the states and county, attracting children and adults towards the educational systems. So, this project is small try towards the methodology of population and its literacy increment by considering various important aspects presented.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9617

ACKNOWLEDGEMENT:

The authors also thank **Dr Praveena A S**, Assistant Professor of Statistics, Manasagangothri, Mysuru for her valuable guidance which led to the improvement of the manuscript.

REFERENCES

1. https://www.kaggle.com
2. Zelterman D (2015) Applied Multivariate Statistics with R, Springer.
3. Dekker M , Regression analysis and its Application- A Data Oriented Approach.
4. Poston.D.L & Micklin.M Handbook of population, Kluwer.
5. NPTEL IIT Kharagpur Web Course.
6. Agarwal B.L. Basic Statistics (2005) New Age International Publishers.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License