[[1]](#footnote-1)

Analysis of Indian Suicide Data for Prediction of Number of Suicides Given Features of a Population

Selva Priyanka S, Sudeep Galgali, Selva Priya S*,* Shashank BR,

*Dept. of CSE, M. S. Ramaiah Institute of Technology*

*Abstract* — We present a study aimed at discovering the prime factors that affect the rate of suicides in certain regions of India. The Indian government maintains a database of the reported suicide cases in each region of India; this is made public for the purpose of data analytics. This data is classified under various different sub categories. From our study we were able to identify the prime sub categories that contribute significantly to the suicide rates.

**Three datasets were used in our study, the marital status of the population, the Educational level of the population and the Primary Census, all the data was from the year 2011. These datasets were subject to pre-processing to obtain it in the required format. Pearson correlation was computed to determine the strength of the features on the suicide rates and then linear regression was used to develop a model for prediction of suicide rates.**

**The results obtained were quite remarkable, 9 features were found to have a significant linear relationship with suicide rates. The prediction model developed using these 9 features gave us a linear fit close to 96% and prediction accuracy of 94%.**

*Index Terms* — Linear Regression, Pearson Correlation, Suicides, Suicide Predictions

# INTRODUCTION

Every year about 800,000 commit suicides all over the world [1], of these 135,000 (17%) are residents of India, [2] a nation with 17.5% of world population.  Between 1987 to 2007, the suicide rate increased from 7.9 to 10.3 per 100,000,[3] with higher suicide rates in southern and eastern states of India.[4] In 2012, Tamil Nadu (12.5% of all suicides), Maharashtra (11.9%) and West Bengal (11.0%) had the highest proportion of suicides[2] Among large population states, Tamil Nadu and Kerala had the highest suicide rates per 100,000 people in 2012. The male to female suicide ratio has been about 2:1. [2]

Estimates for number of suicides in India vary. For example, one study projected 187,000 suicides in India in 2010, [5] while official data by the Government of India claims 134,600 suicides in 2010. [2]

According to [WHO](https://en.wikipedia.org/wiki/World_Health_Organization) data, the [age standardized](https://en.wikipedia.org/wiki/Age_adjustment) suicide rate in India is 16.4 per 100,000 for women (6th highest in the world) and 25.8 for men (ranking 22th). [6]

In this paper an attempt has been made to identify the causes of suicides. This is done by calculating how correlated features of a state are with the number of suicides. The features include marital, professional and educational status of the state’s population. After this the significant features are used to predict number of suicides.

The dataset creation process is explained in Section II followed by the modeling algorithm in Section III. The simulation results are presented and discussed in Section IV and finally the conclusions are indicated in Section V.

# Dataset creation

The aim of dataset creation is to prepare a single dataset for analysis which is sourced by multiple datasets. The list of datasets used is

* 12 Marital status of population datasets for each state in 2011[7]
* 12 Educational level of population datasets for each state in 2011[7]
* 12 Primary census abstract datasets for each state in 2011[7]

The 12 states considered are

* Andhra Pradesh
* Bihar
* Gujarat
* Karnataka
* Kerala
* Madhya Pradesh
* Maharashtra
* Punjab
* Rajasthan
* Tamil Nadu
* Uttar Pradesh
* West Bengal

The analysis dataset is created in which each state has the following features

1. Total Persons
2. Total Males
3. Total Females
4. Marital Status - Never Married Persons
5. Marital Status - Never Married Males
6. Marital Status - Never Married Females
7. Marital Status - Currently Married Persons
8. Marital Status - Currently Married Males
9. Marital Status - Currently Married Females
10. Marital Status - Widowed Persons
11. Marital Status - Widowed Males
12. Marital Status - Widowed Females
13. Marital Status - Separated Persons
14. Marital Status - Separated Males
15. Marital Status - Separated Females
16. Marital Status - Divorced Persons
17. Marital Status - Divorced Males
18. Marital Status - Divorced Females
19. Illiterate - Persons
20. Illiterate - Males
21. Illiterate - Females
22. Literate - Persons
23. Literate - Males
24. Literate - Females
25. Educational level - Literate without educational level - Persons
26. Educational level - Literate without educational level - Males
27. Educational level - Literate without educational level - Females
28. Educational level - Below Primary - Persons
29. Educational level - Below Primary - Males
30. Educational level - Below Primary - Females
31. Educational level - Primary - Persons
32. Educational level - Primary - Males
33. Educational level - Primary - Females
34. Educational level - Middle - Persons
35. Educational level - Middle - Males
36. Educational level - Middle - Females
37. Educational level - Matric/Secondary - Persons
38. Educational level - Matric/Secondary - Males
39. Educational level - Matric/Secondary - Females
40. Educational level - Higher secondary/Intermediate/Pre-University/Senior secondary - Persons
41. Educational level - Higher secondary/Intermediate/Pre-University/Senior secondary - Males
42. Educational level - Higher secondary/Intermediate/Pre-University/Senior secondary - Females
43. Educational level - Non-technical diploma or certificate not equal to degree - Persons
44. Educational level - Non-technical diploma or certificate not equal to degree - Males
45. Educational level - Non-technical diploma or certificate not equal to degree - Females
46. Educational level - Technical diploma or certificate not equal to degree - Persons
47. Educational level - Technical diploma or certificate not equal to degree - Males
48. Educational level - Technical diploma or certificate not equal to degree - Females
49. Educational level - Graduate & above - Persons
50. Educational level - Graduate & above - Males Educational level - Graduate & above - Females
51. Total Worker Population Person
52. Total Worker Population Male
53. Total Worker Population Female
54. Main Working Population Person
55. Main Working Population Male
56. Main Working Population Female
57. Main Cultivator Population Person
58. Main Cultivator Population Male
59. Main Cultivator Population Female
60. Main Agricultural Laborers Population Person
61. Main Agricultural Laborers Population Male
62. Main Agricultural Laborers Population Female
63. Main Household Industries Population Person
64. Main Household Industries Population Male
65. Main Household Industries Population Female
66. Non Working Population Person
67. Non Working Population Male
68. Non Working Population Female
69. Number of Suicides

These features are considered only for the part of the population aged between 15 and 29 due to a WHO study that proves this is a crucial age for suicide.

# Pearson correlation and MODELING

This section introduces the modeling techniques used for the analysis. This includes Pearson correlation and Regression modeling.

## Pearson correlation

Correlation is a technique to measure how two sets of data are related, Pearson correlation also called as the Pearson Product Moment correlation particularly describes the linear relationship [8]. The correlation coefficient always lies between -1 and 1. A value of -1 indicates that there exists a perfectly negative relationship i.e. If we have two variables X (independent) and Y (dependent), Y is expected to yield a line with a negative slope when plotted with respect to X. Conversely a value of 1 indicates a perfect positive relationship. A value of 0 indicates no correlation implying that Y is independent of the variation in X.

We have calculated the Pearson correlations for all the 69 features with the number of suicides and selected only those for modeling which have a value greater than 0.5.

## Regression Modeling

Linear Regression is the statistical method of modeling the relationship between a dependent variable and one or more independent variables [9]. In the case of a single independent variable the process is termed Simple Linear Regression and in the case of multiple independent variables it is called multivariate linear regression. The variables are modeled by (1).

*y = f(x) + c* (1)

Where y is the dependent variable and x = x1, x2, x3…..xn are the independent variables and n is the number of independent variables.

# SIMULATION AND RESULTS

## Pearson Correlation

The correlations of the 9 significant features with the number of suicides are tabulated in Table 1.

Table 1: Pearson correlations of the 6 significant features

|  |  |
| --- | --- |
| **Feature** | **Correlation co-efficient** |
| Marital Status – Widowed Females | 0.5025 |
| Marital Status - Separated Persons | 0.7209 |
| Marital Status - Separated Females | 0.8618 |
| Marital Status – Divorced Females | 0.5558 |
| Educational level - Technical diploma or certificate not equal to degree - Persons | 0.5842 |
| Educational Level- Technical Diploma or Certificate not equal to Degree - Males | 0.6265 |
| Main Working Population Female | 0.6429 |
| Main Agricultural Laborers Population Female | 0.5914 |
| Main Household Industries Population Female | 0.5452 |

From Table 1 we can identify the main sub categories of the population that influence the suicide rates in certain regions of India as

* Separated and widowed Females have a very high correlation with the number of suicides, which prompts us to believe that this is the most important factor in predicting the suicide rates.
* The next in line is the farmer, the main agricultural laborers population.
* Men with a technical diploma or certificate not equal to degrees.
* People who own or work in household industries.
* People with a technical diploma or certificate not equal to degrees.
* Working female population

## Linear Regression Modeling

A linear model is developed for the significant features to predict number of suicides The Classical Ordinary Least Squares method is used to estimate the number of suicides.

In order to evaluate a multivariate regression model a test can be carried out known as the F-test. The F-test is basically used to compare an intercept only model with the specified model. It involves setting up a null hypothesis which states that ‘The coefficients of all independent variables involved in the model are equal to zero’. The alternate hypothesis is that ‘At least one of the independent variable involved in the model has a non-zero coefficient’. Following this, a number of steps are carried out in the F-test.

* Calculation of test statistic (denoted by F) assuming the null hypothesis is true as given by (2).
* Calculation of 95% confidence interval I for degrees of freedom using an F-table or statistical software.
* Acceptance of the null hypothesis if F ∈ I; reject if F ∉ I.
* Calculation of the p-value using statistical algorithms

*F=*  (2)

The results are summarized in Table 2.

Table 2: Summary of results

|  |  |
| --- | --- |
| R-squared | 0.961 |
| Percentage Fit | 96% |
| Percentage Accuracy | 94.1363% |
| Ftest Pvalue | 0.00697 |
| RMS Error | 543 |

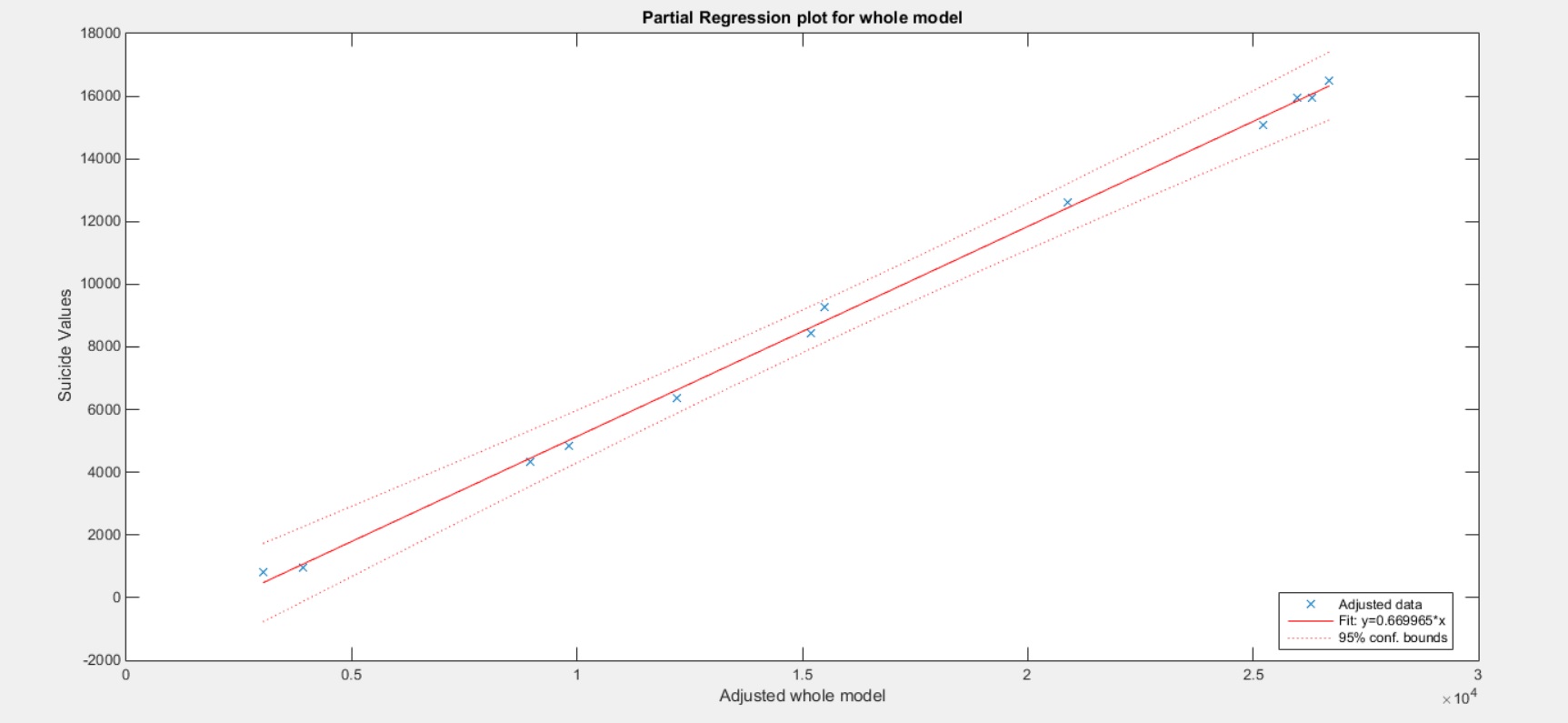


Figure : Partial Regression Plot of Model

# CONCLUSION

The results obtained are extraordinary; a percentage accuracy of 99.83% is exceedingly high. The partial regression plot shows that all the features lie within 95% confidence bounds. From this we can conclude that the rate of suicides depends mainly on the identified factors of the total population. If given the total number of separated females, farmers, men with no degrees or just a technical diploma, people working on household industries, total number of people without degrees or with just a technical diploma, people working in household industries, divorced females and people without jobs of a particular state, we can predict with a great accuracy the number of suicides for that state.

# FUTURE WORK

We would like to conduct such a study considering the remaining states and find what group of the population has a significant relationship on the number of suicides. We also want to consider features other than the population.

Acknowledgment

We would like to thank Dr. K G Srinivasa for giving us this opportunity to perform our study and also for his valuable inputs. We would also like to thank the faculty of MSRIT for their continued support.

References

1. Suicide prevention (SUPRE) World Health Organization (2012)
2. Suicides in India The Registrar General of India, Government of India (2012)
3. Vijaykumar L. (2007), Suicide and its prevention: The urgent need in India, Indian J Psychiatry;49:81-84,
4. Polgreen, Lydia (March 30, 2010). "Suicides, Some for Separatist Cause, Jolt India". The New York Times.
5. Patel, V.; Ramasundarahettige, C.; Vijayakumar, L.; Thakur, J. S.; Gajalakshmi, V.; Gururaj, G.; Suraweera, W.; Jha, P. (2012). "Suicide mortality in India: A nationally representative survey". The Lancet 379 (9834): 2343. doi:10.1016/S0140-6736(12)60606-0.
6. Suicide Rates - Data by country. World Health Organization 2012. Retrieved 30 November 2015.
7. Census India 2011
8. David A. Freedman, “Statistical Models: Theory and Practice”, Cambridge University Press. P. 26, 2009.Richard Taylor, “Interpretation of the correlation coefficient:A Basic Review” ,JDMS 1:35-39, January/February 1990.
9. Hayashi, Fumio (2000). Econometrics. Princeton University Press.ISBN-0-691-01018-8

.

1. [↑](#footnote-ref-1)