A Mathematical Essay on Linear Regression*

Assignment 1

Hari Shankar SJ
Dept. of Electrical Engineering
IIT Madras
Chennai, India
ee17b011@smail.iitm.ac.in

***Abstract*—This document is a essay on Regression model called Linear Regression In real world scenarios prediction using ML algorithms is very much needed. Apart from that there are scenarios where with a given set of data one would like to find if there is any relation between the data. Using Linear regression one can prove if there is a linear relation between the variables. For each the variables has to divided into two classes of Dependent and Independent. Also if they don't have a linear relation we can guess the possibility of Non-Linear relation. Finally we solve one such problem where we have to find if the Cancer incidence and Death are related with the socio-economic status of the citizens of USA.**

I. Introduction

Linear Regression was first found by Sir Francis Galton, an English statistician in 1894. It is used to find is there is a linear relationship between the Dependent and Independent Variable which are chosen by the person with the domain knowledge. It can be used in various domains like medicine and policy making,etc.

II. Linear Regression

A. Concepts of the Linear Regression

In linear Regression we fit the two classed of Data in to a linear equation. To identify the parameters of the equations by minimizing a loss function defined by us. Commonly used loss function is the least square error.

B. Mathematical Representation of Linear Model

We will represent out independent Variables as x_n where n is denoted that it is the n^{th} data point. let y_n and $y'n$ be our predicted and ground truth values of dependent Variable for x_n. The Linear equation in which we try to fit the data which one dependent and one independent variable can be represented as follows,

$$y_n(x) = ax_n + b \tag{1}$$

Here a and b care called the coefficients and intercepts. The loss function $L(y', (a, b)$ is represented as,

$$L(y', (a, b)) = \frac{1}{N}\sum_{n=N}^{N}(y'_n - (ax_n + b))^2 \tag{2}$$

Identify applicable funding agency here. If none, delete this.

Thus the a and b are given by,

$$argmin_{a,b}L(y', (a, b)) \tag{3}$$

C. Multi Variant Linear Regression

The Independent Variables also called as features could be more than one in usual cases. whee the variables are turned into Matrices and Vectors.

$$Y = AX + B \tag{4}$$

The solution is got by equating the derivative of Loss function to 0.

III. Application example of LR

A. Problem to solve

An American NGO want to find if the Socio-economic status of people affect the Incident and Deaths due to Cancer in America. We were given the following Data for each County. M and F indicated the gender.
Independent Data

- *All poverty, M poverty, F poverty* - The raw poverty Data in each county.
- *M with, M without, F with, F without, All with, All without* - The no of people who are with and without medical insurance
- *Med Income, Med Income Black, Med Income Asian, Med Income Nat Am, Hispasian* - The median income of different races in that area.

Dependent Data

- *Incidence Rate, Avg Ann Incidence Rate* - The Age adjusted Incidence Rate per 100,000 peoples and Annual Average Rate of Incidence.
- *Mortality Rate, Avg Deaths* - The Age Adjusted Death Rate per 100,000 people and the average Death Rate due to cancer.
- *Recent Trend Incd, Recent Trend Deaths* - The trend in change of Incidence rate and death rate observed in 5 years.

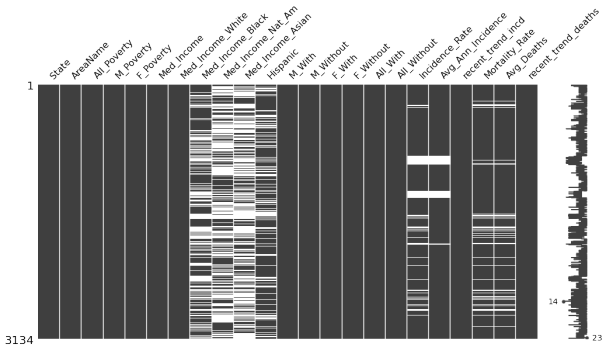

Fig. 1. Missing Matrix

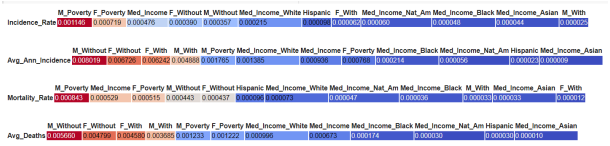

Fig. 3. Coefficients of LR

RMSE for the predicted values:

	Incidence_Rate	Avg_Ann_Incidence	Mortality_Rate	Avg_Deaths
RMSE	14.530005	47.563618	11.686365	31.489834

Fig. 4. RMSE values

B. Missing Data

Fig 1 shows the Matrix of Data that are missing in each column. First lets focus on the missing values in the dependent values. We observe that there is an high correlation between the 4 dependent data. We first replace the columns data cells which has '3 or fewer' number of avg annual Incidence to 2, so can the Iterative Imputer which uses iterative regression to fill the empty cells of the the columns. Similarly we repeat the same for the dependent Data set.

C. Method to prove the relation

To check if they have a linear relation we first split the data set into train and test data set in the ration of 8:2. The splitting is done randomly. We first train the linear regression (Normalized) model to the Train data set and we test it on the tests data. The prediction we got are shown in the Fig 2. Note - the data with 'All' tags are removed since they can be represented as the sum of M and F. The coefficients of

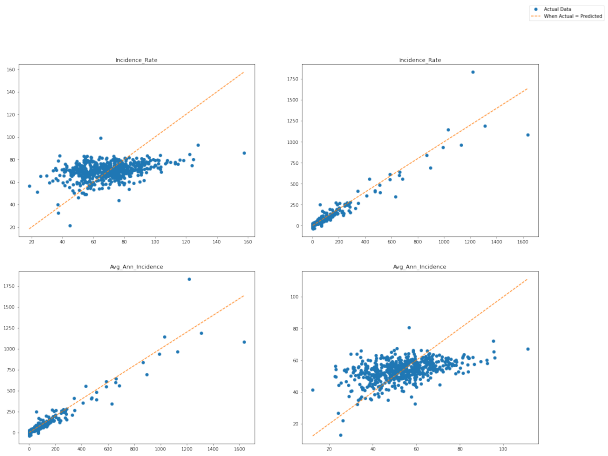

Fig. 2. Plot of Predicted Results

each feature is tabulated and we observe the features related to Poverty, Medical Insurance have high weightage. Also the correlation found before is also high among this variables as seen in fig 5.

After we find the important features, we try to predict the recent trends using classification with linear models. We follow the similar approach and the (train,test) split has equal ratios of each labels. Note- The rising trends data were dropped because it is observed in less than 2 percent cases. The results of the classification is shown below. Accuracy of Indicence Prdiction = 93.15 and Accuracy of Death Prdiction = 77.86.

With this quantitative data we can say the variables have a linear relation between them.

D. Visualizing the Data to see the Relation

Some plots between the variables are shown to see the relation visually.

CONCLUSION

We have proved that the results of Linear Regression are in the lines of expectations of the NGO. We can try for a better models, the actual relation could be more complex one. The results may change due to the assumptions of missing data. but since the assumption made doesn't vary the values more from ground truth we have a decent confidence level in this result.

REFERENCES

Imperative Imputer
https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeIm
Missingno
https://github.com/ResidentMario/missingnoSomething Linky
Seaborn Pairplot
https://seaborn.pydata.org/generated/seaborn.pairplot.html

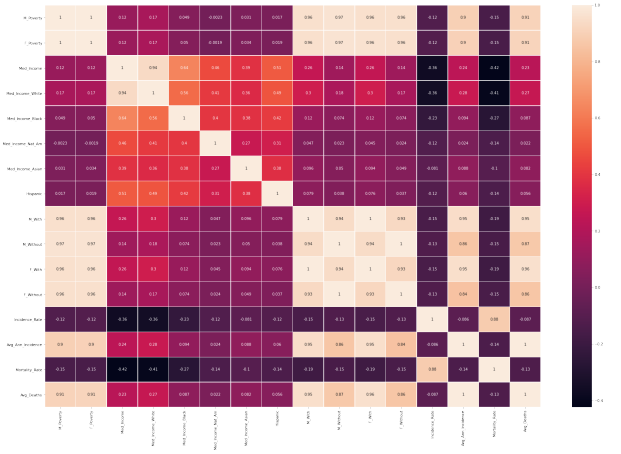

Fig. 5. Correlation Value

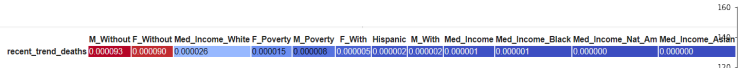

Fig. 6. The coefficient of LR for Trend in Deaths

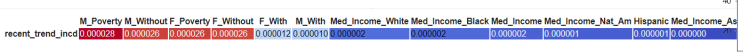

Fig. 7. The coefficient of LR for Trend in Incidence

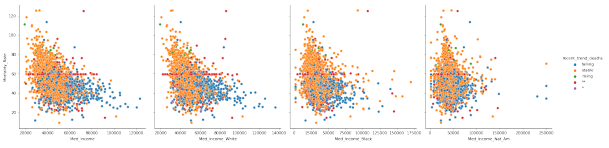

Fig. 8.

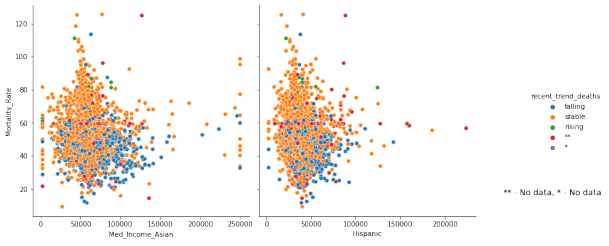

Fig. 9.

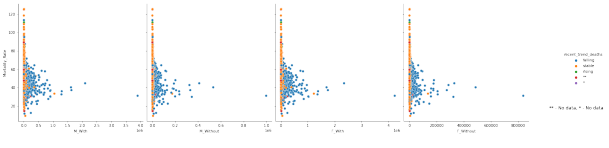

Fig. 10.

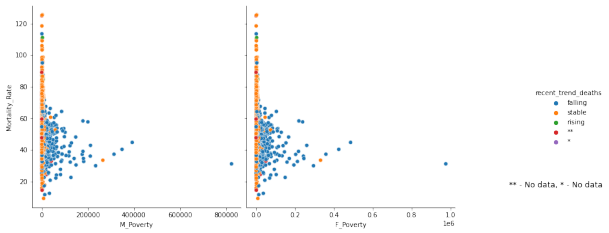

Fig. 11.

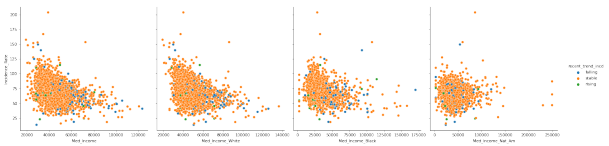

Fig. 12.

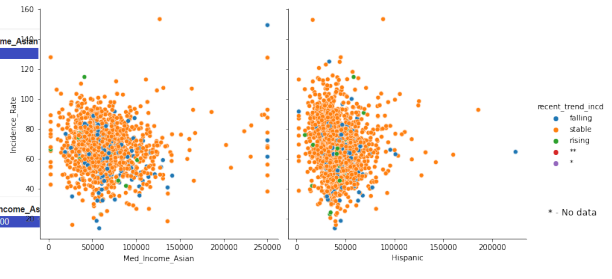

Fig. 13.

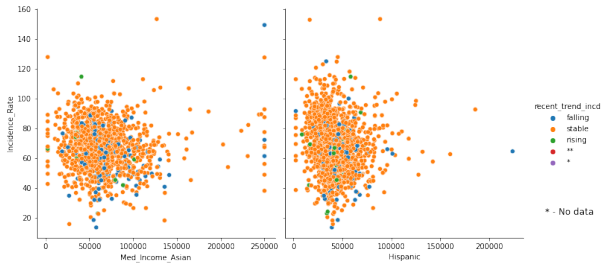

Fig. 14.

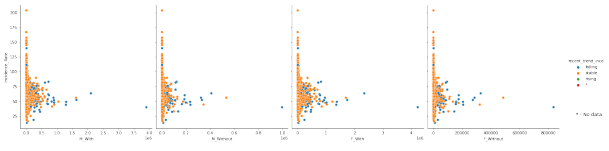

Fig. 15.

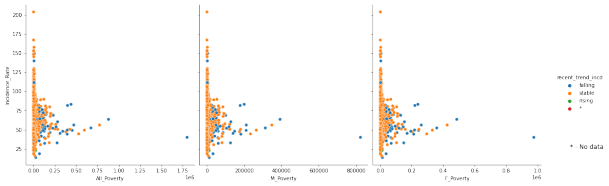

Fig. 16.