Titanic Disaster
Linear Regression

Devcharan K
Indian Institute of Technology, Madras

***Abstract*—Linear regression refers to the mathematical technique of fitting given data to a function of a certain type. It is best known for fitting straight lines. In this paper, data was analyzed using Linear Regression techniques to find if there is any correlation between socio-economic status and cancer incidence and mortality of the US population.**

I. Introduction

Linear Regression deals with the numerical measures to express the relationship between two variables. The relationship between variables can be either strong or weak or even inverse. For example, if we have two attributes to consider, say temperature and ice cream sales. A vendor will sell more ice creams on a hot day than on cold days. So naturally, there will be a linear relationship between temperature and ice cream sales. This relationship would look like a straight line with a positive slope with x axis being temperature and y axis being revenue.

The data is fitted onto the curve $y = mx + c$. Here x is called the independent or predictor variable and y is called the dependent or response variable. The fitted curve can then be used to predict how the attributes will behave at points we don't have data for yet. This is a feature of inferential statistics. The example given in the previous paragraph takes into consideration 2 attributes. We, as humans, cannot visualize more than three such attributes to perform a regression analysis, but with the use of modern techniques regression analysis can be performed on an n-dimensional space with n attributes.

In this paper the data of the US population about income, socio-economic status and cancer incidence is analyzed. The goal is to try to find a correlation between socio-economic status and cancer incidence and mortality. The goal of this paper is to find both quantitative and visual evidence that the non-profit can take and use to further their mission.

In Section 2 the Least Squares Method of Linear Regression and its applications is explained. In Section 3 the problem and the method solve it is shown. Section 4 is an Exploratory analysis of the cleaned data-set with the usage of Python packages namely Pandas, NumPy, SciPy, Seaborn, Matplotlib, Plotly and visualizations of the same. Finally the inferences of the analysis is presented in Section 5.

II. Linear Regression

Regression is a category of Supervised Learning. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model. Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable *causes* the other (for example, higher GRE score does not *cause* higher CGPA), but that there is some significant association between the two variables. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values. Say $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ are n observations from an experiment. We are interested in finding a curve

$$y = f(x) \tag{1}$$

closely fitting the given data of size 'n'. Now at $x = x_1$ while the observed value of y is y_1, the expected value of y from the curve is $f(x_1)$. The residual is defined as $e_1 = y_1 - f(x_1)$. Likewise, the residuals of all other points can be defined as $e_1, e_2, e_3, \ldots, e_n$. Some residuals may be positive and some may be negative. We would like to find a curve fitting the given data such that the residual at any x_i is as small as possible.

In other words, we want a curve fitting the data such that the sum of the squares of all the residuals (say E) is minimum. Thus we consider

$$\sum_{i=1}^{n} e_i^2 \tag{2}$$

and find the best curve (1) that minimizes (2).

III. The Problem

The data on Income, Medical Insurance, Cancer incidence and mortality of the US population in the year 2015 is

available . The problem is to determine if there is any correlation between socio-economic status and cancer incidence and mortality. The data is sorted by area and also by ethnicity, economic status(above or below poverty line), medical insurance.

The analysis begins by cleaning the data by removing any unusual data values like asterisks in a column of integer values etc. All these data points are converted to NaN, ie. Not a Number. The missing data points need to be filled. So the mean, median and mode of each column is taken and the missing data points are filled with one of these quantities depending on the distribution

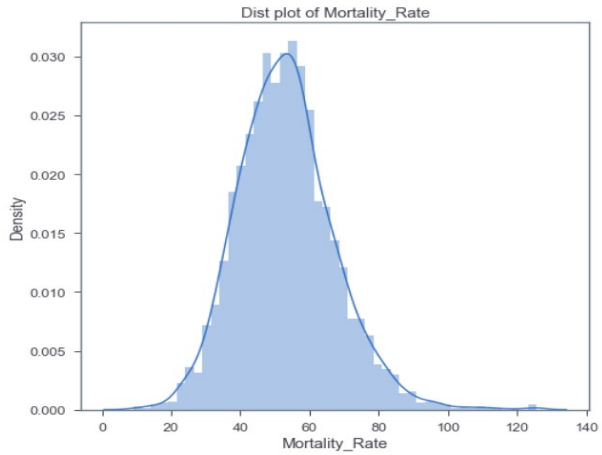

Fig. 1. Distribution of mortality rate

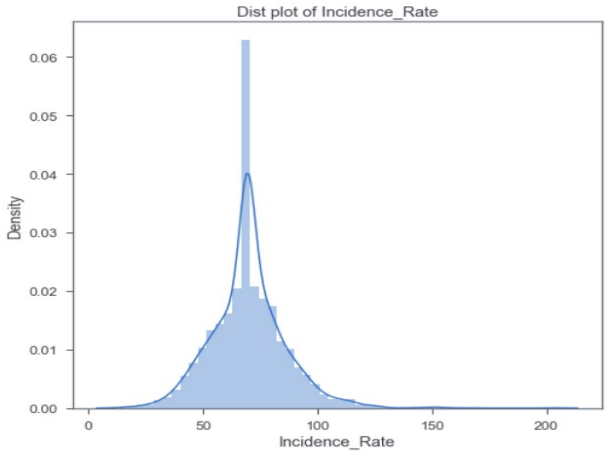

Fig. 2. Distribution of Cancer Incidence

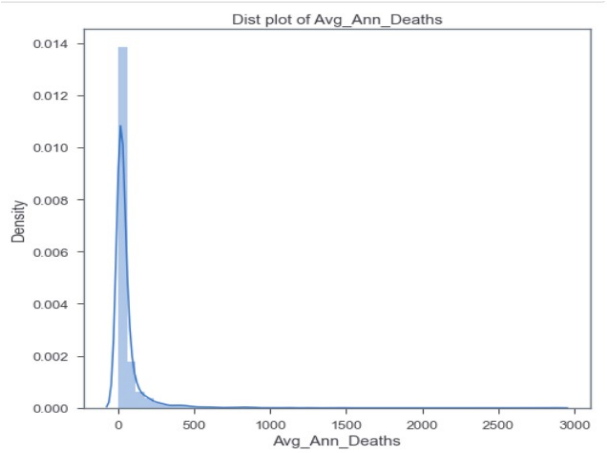

Fig. 3. Distribution of Annual Deaths

For the attributes Incidence rate, Mortality rate and Average annual deaths, the mean, median, mode and standard deviation etc. are calculated to plot the distribution curves. The plots (From Fig(1) to Fig(3)) reveal that the data of 'Average annual deaths' is very skewed to the right. There are a large number of outliers. The mean values cannot be used in this case because the outliers will have a significant impact on the mean values. For symmetric distributions the mean can be used to fill the missing data values but in this case median should be used as the distribution is not symmetric. The attribute 'Recent Trend' is a categorical column. It has three possible values namely 'rising', 'stable' and 'falling'.In this case mode should be used to fill the missing data points. Once the NULL values and the invalid data points are taken care of, the analysis of the data can be done.

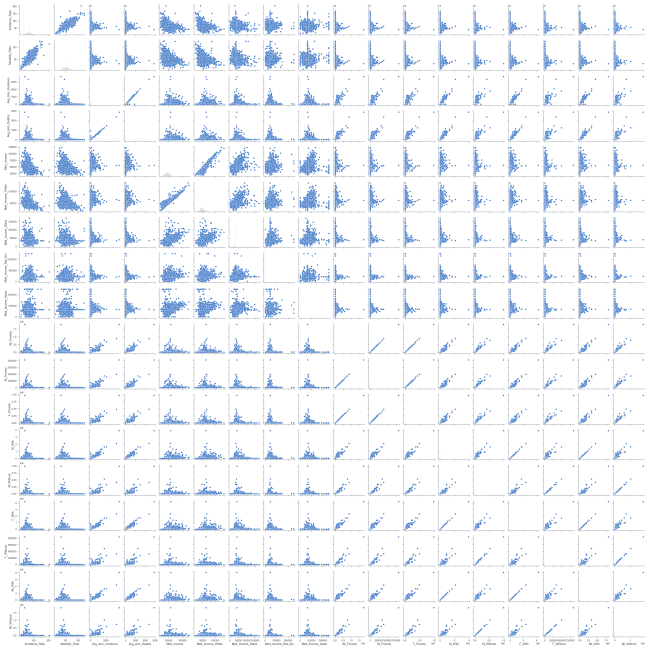

Fig. 4. Pair-wise plot

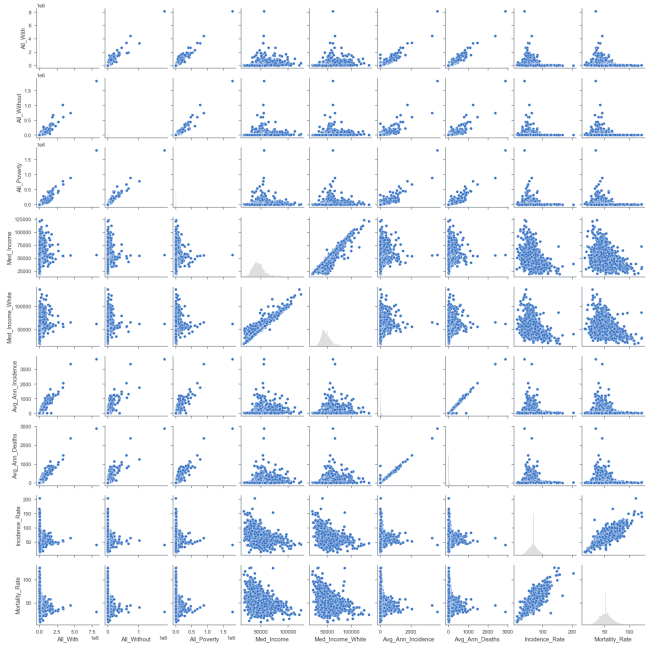

Fig. 5. Pair-wise plot with selected attributes

IV. Exploratory analysis

The goal is to find a correlation between socioeconomic and cancer incidence and mortality. The attributes namely, M with, M without, F With, F Without, All With, All Without come under socio-economic criteria. They have to be compared against the attributes Incidence Rate, Mortality Rate, Avg Ann Incidence, Avg Ann Deaths and recent trend. So first those columns are separated into different sets of data. A pairplot is created using the seaborn package. This command creates a pair-wise plots of all the socio-economic parameters vs the Incidence parameters. This plot gives an idea of which attributes are correlated to each other the most. Some attributes show very high correlation and others very low. For example, as shown in Fig(4), the Males without medical insurance and All without medical insurance are very highly correlated. So one of these can be used in our exploratory analysis and the results of the analysis can be applied to the highly correlated attributes of the chosen attribute. The attribute pairs that have fairly high correlation(0.8-0.99) could also have important implications in the analysis.

Some of the correlations as seen above are not useful for the analysis. For example the correlation between Female poverty and All poverty does not give any extra meaningful information. The females in an area are a subset of the total number of people in the area, and hence the correlation. So, the umber of columns considered for the pairwise plot can be reduced and plotted again with the relevant attributes only as in Fig(5).

A. Regression

Plotting a regression plot between two attributes shows the best fit line for the given data points. The regression plot command is applied to different pairs of attributes to see the plot and draw inferences.

B. Average Annual Death

As seen in Fig(6)., the regression plot of Annual death vs All poverty is rising. This is as expected because more poverty in an area would mean the average annual death rate is higher. This is the same way for the plots of Average Annual death vs All with, All without. As seen in Fig(9)., the regression plot of Average Annual death vs Med Income is rising very slowly. We can infer that Annual death is not affected by the income of the population as much as it is by issues like poverty and insurance availability.

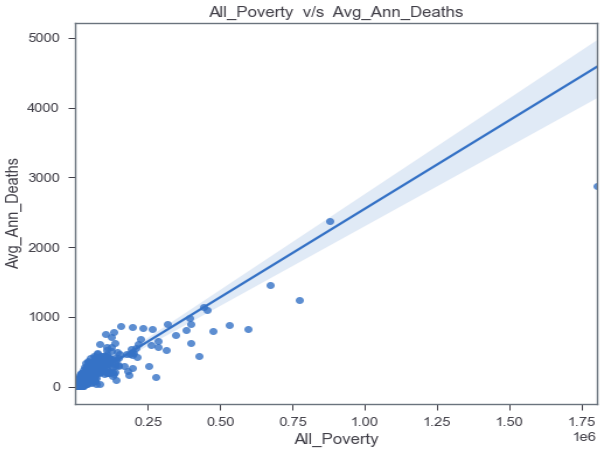

Fig. 6. Avg Annual Death vs All Poverty

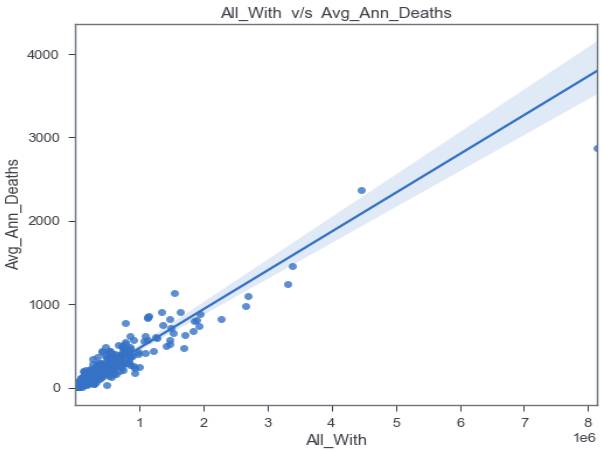

Fig. 7. Avg Annual Death vs All with Insurance

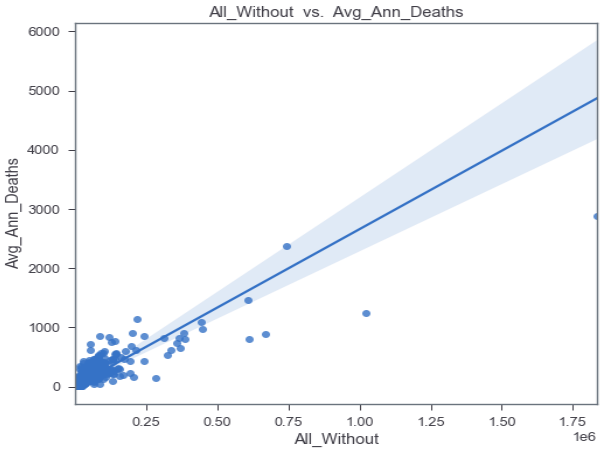

Fig. 8. Avg Annual Death vs all without Insurance

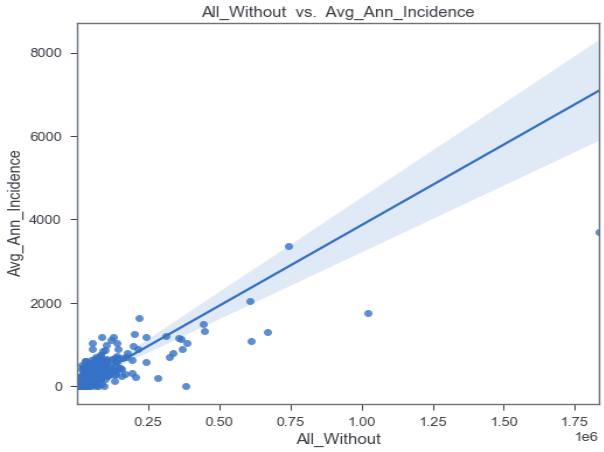

Fig. 11. Avg Annual Incidence vs All without Insurance

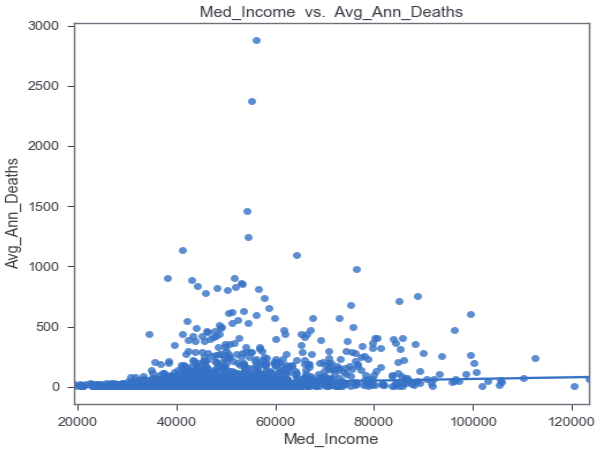

Fig. 9. Avg Annual Death vs Med Income

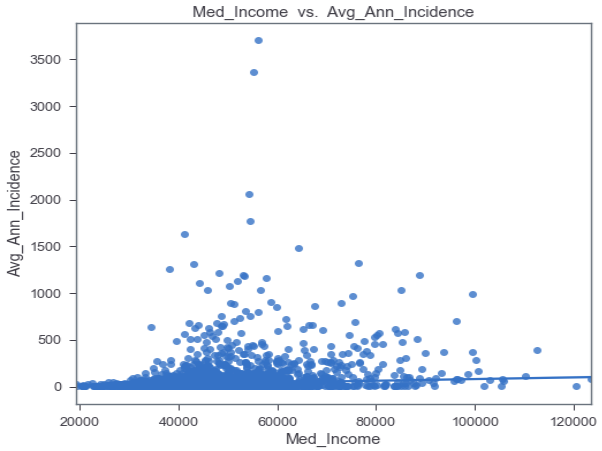

Fig. 12. Avg Annual Incidence vs Med Income

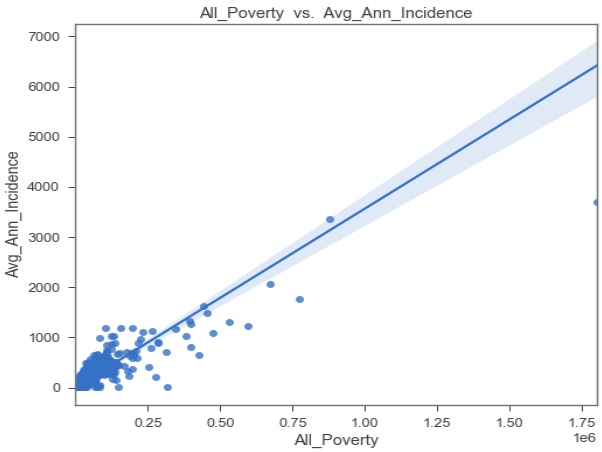

Fig. 10. Avg Annual Incidence vs All Poverty

C. Average Annual Incidence

As seen in Fig(10)., the regression plot of Average Annual Incidence vs All poverty is rising. This is the same reason as seen with Average Annual deaths.

Once again, Fig(12). shows that the regression plot of Annual incidence vs Med Income has a fairly level curve. Once again the inference is that the Annual Incidence rate is not affected by the income of the population as much as it is by the other considered factors.

D. Mortality Rate

In Fig(13)., the Mortality rate vs All poverty curve has a negative slope which could be due to the smaller sample size of higher poverty data points. A similar pattern is seen with Mortality Rate vs All With and All Without. In Fig(14)., the Mortality rate vs Med Income curve has a negative slope. As the Income becomes higher, the mortality rate reduces.

The Mortality rate vs Native American, Hispanic and Asian regression plot shows a fairly level curve even for higher

incomes. The Mortality rate vs White and black regression plot shows a curve with a negative slope. So the mortality rate reduces as income is higher.

This shows that the racial,social and economic factors all have different effects on the mortality rate, Annual cancer incidence rate and the Annual death rate.

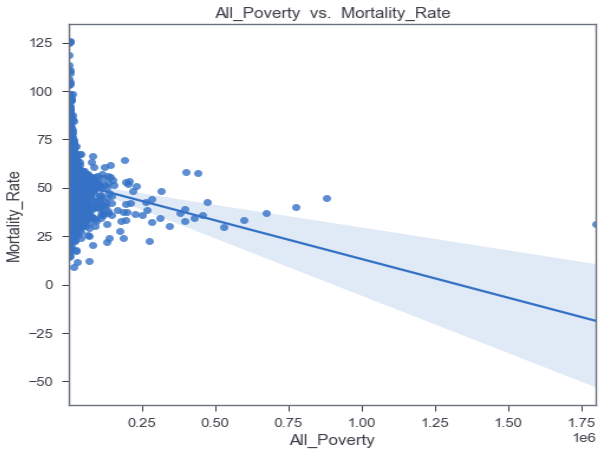

Fig. 13. Mortality rate vs All Poverty

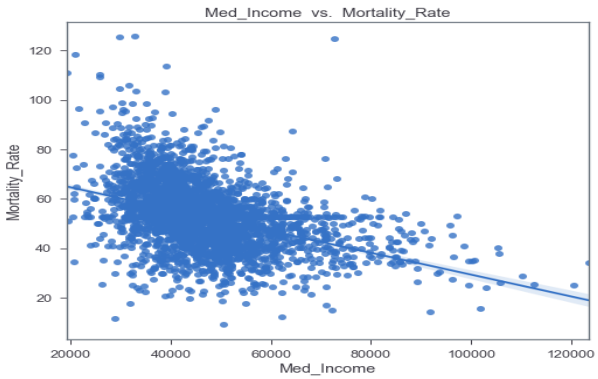

Fig. 14. Mortality rate vs Med Income

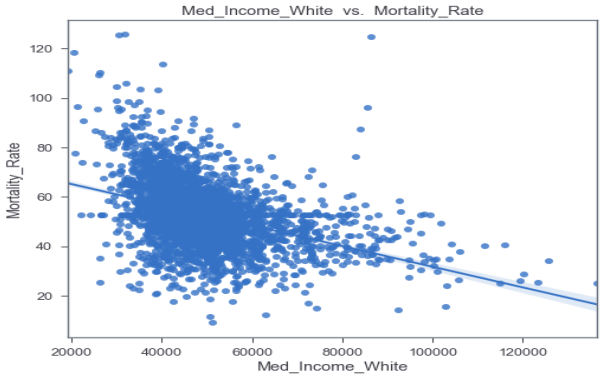

Fig. 15. Mortality rate vs Med Income White

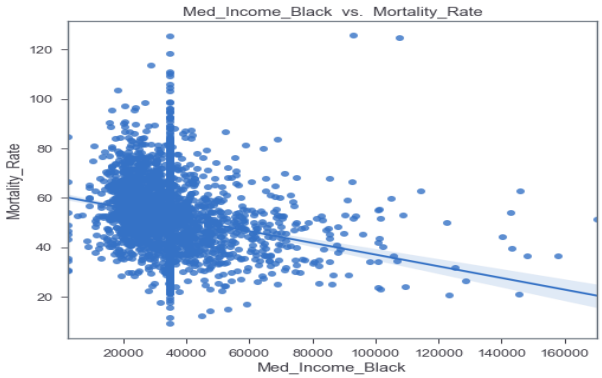

Fig. 16. Mortality rate vs Med Income Black

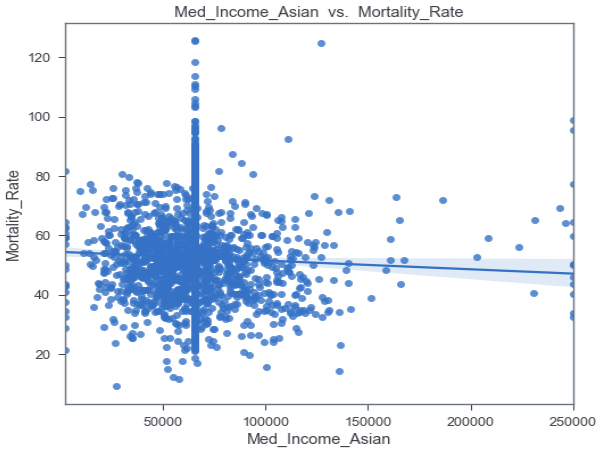

Fig. 17. Mortality rate vs Med Income Asian

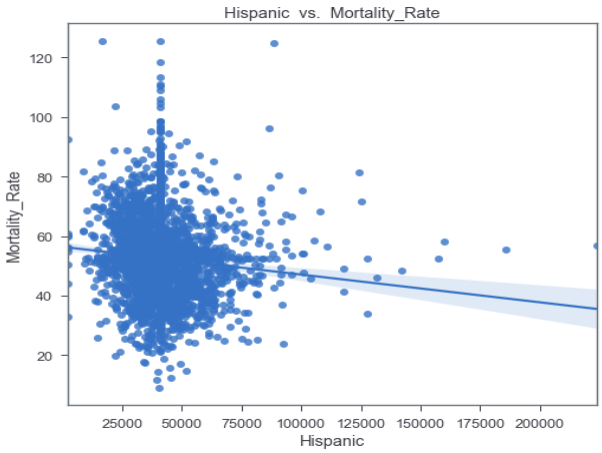

Fig. 18. Mortality rate vs Med Income Hispanic

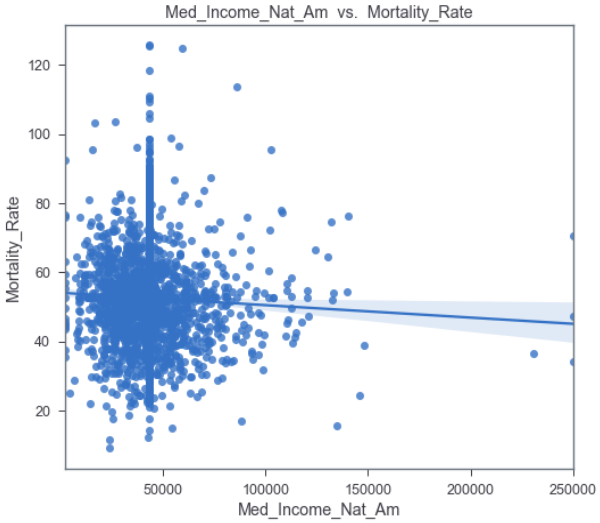

Fig. 19. Mortality rate vs Med Income Native American

V. Inferences

The Mortality rate, Death rate and Incidence rate are affected by socio-economic factors in different ways. White and Black people generally tend to see lower mortality rates as their income increases, but higher income does not seem to affect the mortality rate of Hispanics, Asians and Native Americans to the same degree. The Mortality rate does not reduce at the same rate for the latter three groups.

Conclusion

Linear Regression is a widely used technique in many branches of science and technology. It is a core concept in Machine Learning and Data Science. In this paper the techniques of statistics and Linear Regression have been used to arrive at meaningful results which can be used by non profit organizations to further their mission of equal medical treatment of all individual irrespective of their socio-economic status.

References

[1] ”Linear Regression” Wikipedia [Accessed on 17 September] Available: https://en.m.wikipedia.org/wiki/Linear_regression
[2] ”Covariance and correlation” [Accessed 17 September 2021] Available: https://en.m.wikipedia.org/wiki/Covariance_and_correlation
[3] Trevor Hastie, Robert Tibshirani, Jerome Friedman, ”The Elements of Statistical Learning Data Mining, Inference, and Prediction”, 2017
[4] M. Huang, ”Theory and Implementation of linear regression,” 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), 2020, pp. 210-217, doi: 10.1109/CVIDL51233.2020.00-99.
[5] The Basics : Linear Regression [Accessed on 17 September] Available: https://towardsdatascience.com/the-basics-linear-regression-2fc9f5124687
[6] Linear Regression notes ,Yale University [Accessed on 17 September] Available: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm
[7] EE4708 Regression, Dr. Kaushik Mitra and Dr. Ramakrishna Pasumarthy