Assignment 1 Version 2: A mathematical essay on linear regression

S, Karthik
ME18B149 - IDDD Data Science
Indian Institute of Technology Madras
Chennai, India
me18b149@smail.iitm.ac.in

***Abstract*—This document is an overview of the concept of linear regression. It gives a general idea of the topic, along with the mathematical aspects, as well as an application involving linear regression based on data from a real-world problem.**

***Index Terms*—Linear Regression, Machine Learning, Data Science**

I. Introduction

Supervised learning is a machine learning task that learns a model or a function to give an output to every possible input value based on previous data at various instants. Linear Regression is a type of supervised learning where the output is assumed to be a linear function of the input parameters.

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric. In a single input model, this function is a line. In a multi input model, this function represents a plane or a hyperplane.

The specific problem discussed is the effect of socioeconomic status on cancer incidence and mortality using the available data for each district in the US. Linear regression techniques will be used to analyze the data which is provided by the US Government.

This paper is a case study through which the principles of linear regression are implemented. It aims to examine whether low income groups are at a greater risk of being diagnosed and dying from cancer. The results and the conclusion of this analysis will be helpful for a non-profit organization with lobbying and fundraising whose mission is to advocate for better health outcomes for low income populations in the US.

II. Linear regression

A. Model Structure

Linear regression is a regression task in which the model being learnt is assumed to be a linear function of parameters. The most general form of the linear regression model is:

$$f(\mathbf{x},\boldsymbol{\theta}) = \hat{y} = \theta_0 + \Sigma_1^n \theta_j x_j = \mathbf{x}^{\boldsymbol{T}}\boldsymbol{\theta} \tag{1}$$

where $\hat{y}$ is the predicted output for a given input, $\mathbf{x} = (x_0, x_1, ..., x_n)^T$ is a n-dimensional input feature vector with $x_0 = 1$ (x_0 is not an actual input, it is written along with the inputs for convenience), $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_n)^T$ are the model parameters and n is the order of the linear model.

Equation (1) gives the function for a single data point. Expanding it to multiple data points, we obtain the following equation

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\theta} \tag{2}$$

where $\boldsymbol{X}$ is the matrix with the inputs being it's rows, $\hat{\boldsymbol{y}}$ being the vector of all the outputs.

B. Cost Function

Cost Function is a measure of how wrong the model is in terms of its ability to estimate the relation between inputs and outputs. There exist different types of cost functions and the most popular among them is the Mean Squared Error (MSE) between the predicted and observed outputs. The MSE cost function for a dataset with N samples is given by

$$J(\boldsymbol{\theta}) = \frac{1}{2N}\Sigma_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{3}$$

where $\hat{y}_i$ is the predicted output and y_i is the observed or given output.

C. Solution Objective

Given a dataset of ($\mathbf{X}$ and outputs ($\boldsymbol{y}$), the objective of Linear Regression is to find a function i.e, $\boldsymbol{\theta}$ that minimizes the cost function. Mathematically,

$$minimize_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{4}$$

D. Analytical Approach

If the matrix $\boldsymbol{X}$ is invertible, then a direct solution can be obtained as follows

$$\boldsymbol{\theta} = \boldsymbol{X}^{-1}\boldsymbol{y} \tag{5}$$

Equation (5) need not work for all the cases. It will not work if the matrix $\boldsymbol{X}$ is not invertible, or if the set of equations do not have a solution.

If $\boldsymbol{X}$ is not a square matrix, we can pre-multiply Equation (2) by $\boldsymbol{X}^T$.

$$\boldsymbol{X}^T\boldsymbol{y} = (\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\theta}$$
$$\boldsymbol{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{6}$$

Equation (6) is known as the normal equation which gives the solution of a line or a plane that passes through all the data points involved in the matrix.

However, it is clearly known that for most of the real world cases, it is impossible to have a linear function that perfectly passes through all the data points. Also, as the number of data points in a real world problem are large numbers, matrix inversion becomes a very computationally expensive task. Hence, this solution is very restrictive in nature.

E. Gradient Descent

Gradient descent is an optimization algorithm used to minimize the cost function by moving in the direction opposite to that of the gradient of the cost function in an iterative manner.

Unlike the previous solution, this method will always give a result and doesn't have any condition on $\boldsymbol{X}$. It is also more computationally efficient with slight variations in how the data is used (For example, data is processed in batches in Mini-Batch Gradient descent)

The mathematical implementation is given as follows where

Algorithm 1 Gradient Descent

```
$\boldsymbol{\theta_0}$ = random_init()
for $i = 0, 1, 2, \ldots, num\_iterations - 1$
    $g^i \leftarrow \nabla J(\boldsymbol{\theta_i})$
    $\boldsymbol{\theta^{i+1}} \leftarrow \boldsymbol{\theta^i} - \alpha g^i$
end
```

α is known as the learning rate which can be appropriately set by the user.

Some of the major choices that have to be made for implementing gradient descent are

1) Parameter Initialization: Ideally, one should choose $\boldsymbol{\theta_0}$ to be close to the minimum. Since in most cases, it is difficult to make a guess of such a choice, they are initialized to zero. A common workaround for this problem is to run the algorithm several times from widely dispersed starting points, and then pick the best result.

2) Stopping Criteria: Stopping criteria is a condition which determines that the gradient descent algorithm has converged. Ideally we want the algorithm to stop when the minimum has been reached. There are multiple ways of defining a stopping criteria 1. Setting the maximum number of iterations apriori 2. Stopping when the change in cost function is less than a certain threshold 3. Stopping when the change in parameters is less than a certain threshold 4. Stopping when the change in gradient is less than a certain threshold

3) Learning Rate: There is a tradeoff in selecting the value of learning rate. If the value is set high, then the gradients can blow up which might lead to skipping the minimum. If it is set to a very low value, then it will take a lot of iterations to converge. A viable solution is to adjust the learning rate dynamically while having a decay factor when it is near the minimum.

F. Grid Search

Grid search method is a brute force method for optimization where a domain is chosen and the function value is plotted at all the locations and the minimum is found out. Clearly, this means that we need to have a reasonable guess of the minima region beforehand. Also, it it computationally very expensive. An application of this can be during fine tuning where the minima is already found using other methods and the neighbourhood can be searched for better results.

III. The Problem

A. Overview

I am a data scientist employed by a consulting company. Our consulting firm has been hired by a nonprofit organization whose mission is to advocate for better health outcomes for low income populations in the United States. We've been asked to examine whether low income groups are at greater risk for being diagnosed and dying from cancer. If successful, the analysis will help the nonprofit with lobbying and fundraising.

B. Goals and Objectives

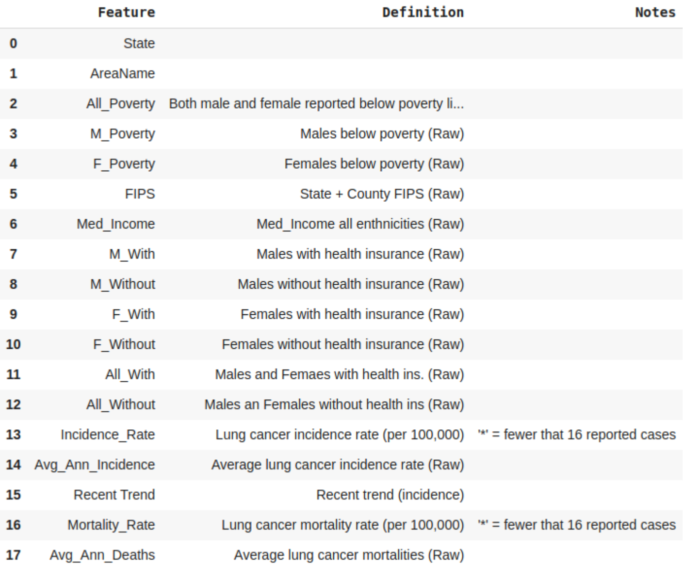
Demonstrate whether or not cancer incidence and mortality are correlated with socioeconomic status. Provide both quantitative and visual evidence that the nonprofit can take and use to further their mission. 4 steps 1. Gather, Clean and prepare data 2. Exploratory analysis 3. Statistical model 4. Visualizations

C. Reading the data

We will be using python throughout this assignment. The dataset is presented using a file named merged_data.csv. which can be read directly using the read_csv command from the pandas package and stored into a dataFrame (see Code **??**).

IV. Data preparation and cleaning

The dataset contains countywise details of various features like population in poverty, mortality rate. The complete details of all the attributes are shown in Fig 1.



	Feature	Definition	Notes
0	State		
1	AreaName		
2	All_Poverty	Both male and female reported below poverty li...	
3	M_Poverty	Males below poverty (Raw)	
4	F_Poverty	Females below poverty (Raw)	
5	FIPS	State + County FIPS (Raw)	
6	Med_Income	Med_Income all enthnicities (Raw)	
7	M_With	Males with health insurance (Raw)	
8	M_Without	Males without health insurance (Raw)	
9	F_With	Females with health insurance (Raw)	
10	F_Without	Females without health insurance (Raw)	
11	All_With	Males and Femaes with health ins. (Raw)	
12	All_Without	Males an Females without health ins (Raw)	
13	Incidence_Rate	Lung cancer incidence rate (per 100,000)	'*' = fewer that 16 reported cases
14	Avg_Ann_Incidence	Average lung cancer incidence rate (Raw)	
15	Recent Trend	Recent trend (incidence)	
16	Mortality_Rate	Lung cancer mortality rate (per 100,000)	'*' = fewer that 16 reported cases
17	Avg_Ann_Deaths	Average lung cancer mortalities (Raw)	

Fig. 1. Exhaustive list of all the attributes in the given dataset

Our first task is to analyze the data and filter out the attributes that are not required or redundant in this analysis.

Let us first check the number of unfilled values in each of the parameters. Table I gives the number of empty cells under every feature.

TABLE I
EMPTY VALUES UNDER EACH VARIABLE

Index	Feature	Empty_Values
0	State	0
1	AreaName	0
2	All_Poverty	0
3	M_Poverty	0
4	F_Poverty	0
5	FIPS	0
6	Med_Income	1
7	Med_Income_White	2
8	Med_Income_Black	1210
9	Med_Income_Nat_Am	1660
10	Med_Income_Asian	1757
11	Hispanic	681
12	M_With	0
13	M_Without	0
14	F_With	0
15	F_Without	0
16	All_With	0
17	All_Without	0
18	fips_x	0
19	Incidence_Rate	0
20	Avg_Ann_Incidence	0
21	recent_trend	0
22	fips_y	0
23	Mortality_Rate	0
24	Avg_Ann_Deaths	0

1) Unreported data: It is observed that the median income for individual ethnicities have been unreported at a huge number of counties. Also, since our analysis only covers socio-economic status from a geographical region, the above data is not considered for the model.

2) Unwanted data: Genderwise split of data is not necessary in our analysis as it is about the effects on the entire population. Recent trend is not used in our analysis as there is already a quantified value of the incidence and the mortality rates.

From Fig 1, Mortality_Rate and Incidence_Rate have some entries as * which means the number of reported cases is 16 or less.

A. Mortality_Rate

Two possible decisions can be made in a broad scale.

- The counties corresponding to the starred entries can be removed. The disadvantage of this is that valuable data might be lost.
- These values can be imputed using the existing data in a reasonable manner. The risk of this is that there is a high chance of the data being skewed towards being incorrect.

To determine the course of action, the plan was made which was to check the population of the unreported counties and compare the mortality rates to that of counties with similar population that have reported. Population is not directly a column in the dataset. Fortunately, there are 2 columns **All_With, All_Without** which gives the population with and without health insurance. Adding them up will give the total population. A new column named **"Population"** was created in the dataset.

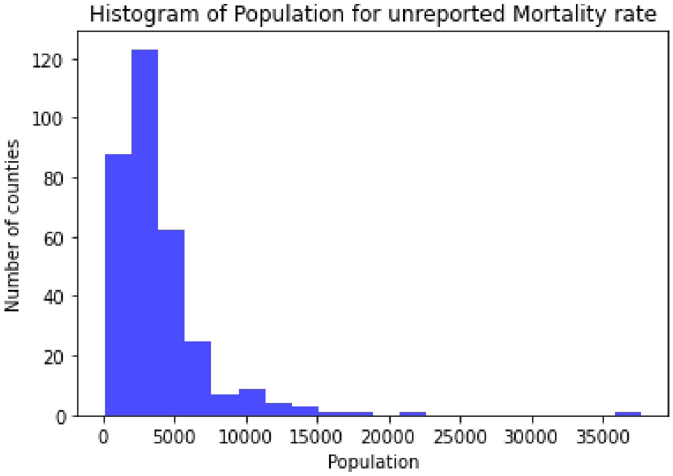


Fig. 2. Distribution of population of counties with unreported deaths

Fig 2 shows the distribution of the population among the unreported counties. It is clear that the population of the unreported counties are in the range of **0 to 10000**. If we plot the distribution of Mortality Rate for all the reported counties with population from 0 to 10000, we get Fig **??** The following analysis was done only for those counties with population between 0 and 10000.

- Number of counties which reported deaths: 411
  Mean mortality rate among reported counties: 53.19
  Median of the above: 52.6
- Number of counties with unreported deaths: 308

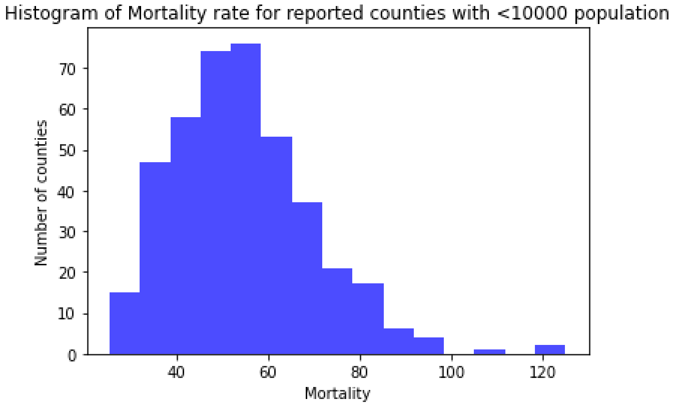


Fig. 3. Distribution of mortality rate of reported counties with population between 0 and 100000

From Fig 1, the unreported counties have a mortality rate of less than 16. However, from the above analysis, it is clear that the distribution is centered around 52 deaths per 100,000 which is completely opposite of the counties with reported deaths. This means there is a very high possibility that the counties which did not report failed to collect data and/or under-reported the deaths. Hence, we can conclude that it is best if we eliminate the counties which did not report mortality rate.

To check the above, we can plot the distribution of mortality rate for the counties with population between 0 and 10000 (Fig 3). We can see that the distribution is fairly normal in this case. Introducing 308 unreported cases as less than 16 will most likely skew the distribution which is not ideal. Hence, we have substantial data to show that the counties with unreported deaths have most likely failed to track the cases.

Final Decision: The counties with Mortality Rate unreported will be excluded from the analysis.

B. Incidence Rate

The counties with unreported mortality rates have been removed. From the remaining, there are counties which have unreported incidences due to privacy/legislative reasons. Since the Mortality rate data is clean, eliminating these counties too will result in significant loss of valuable data which is not desired as mortality rate is the primary variable with the incidence rate being secondary.

1) Approach 1: Logically speaking, there should be a positive correlation between the incidence rate and the mortality rate. If the correlation coefficient is higher than 0.9, we can conclude that incidence rate and mortality rate are near-perfectly correlated and use the mortality rates to predict incidence rate for the missing counties.

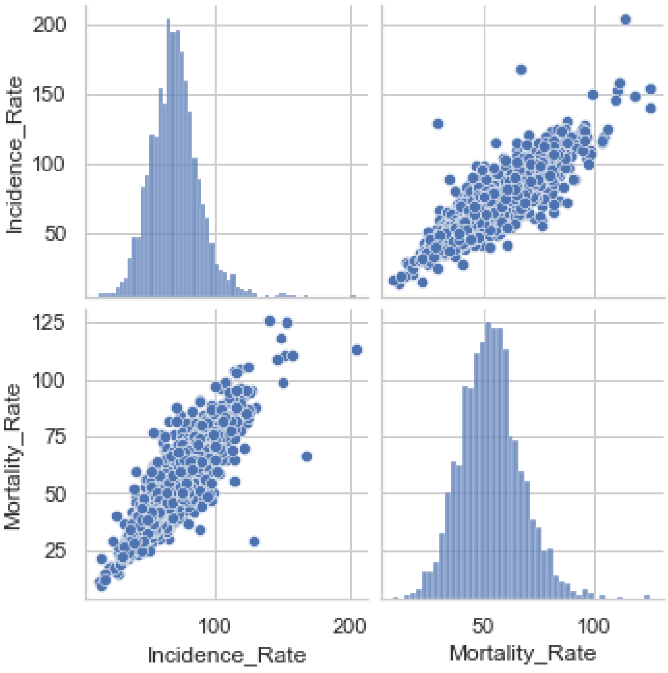


Fig. 4. Pairplot between Mortality Rate and Incidence Rate

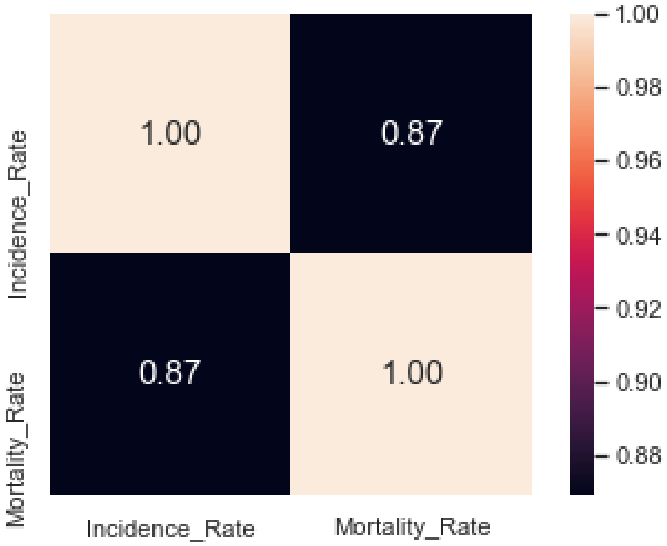


Fig. 5. Correlation heatmap between Mortality Rate and Incidence Rate

Using the available data, a pair plot and a heatmap was plotted for the above 2 variables. The correlation coefficient came out to be 0.8 which means they are positively correlated but not strong enough to proceed with the above proposed solution.

2) ***Final Approach****:* Plotting the distribution of population for the unreported counties (Fig 6), we see that most of the counties have population less than 25000. Hence, a good estimate would be to assign the median of the incidence rate from the reported counties under the same population bracket. From the distribution shown in Fig 7, the median is 67.3 which can be assigned to the missing incidence rate values.

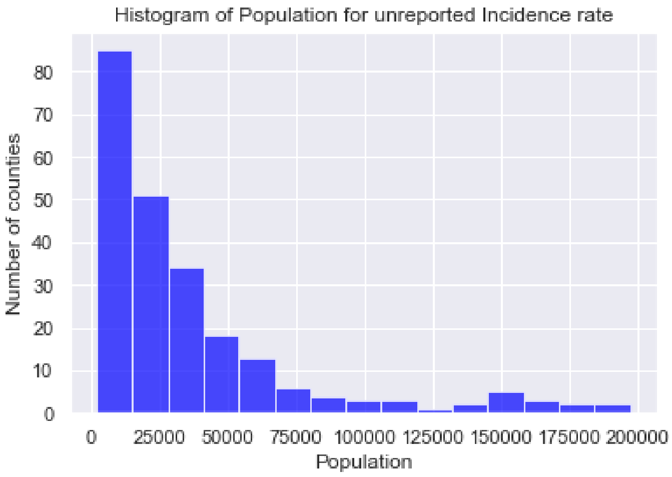


Fig. 6. Distribution of population of counties with unreported Incidence Rate

C. Redundant data

Since Incidence and Mortality Rates already exist in the dataset, the values of Average Annual incidence and deaths become redundant in the analysis. For others, a heatmap was plotted containing the correlation coefficients (Fig 8).

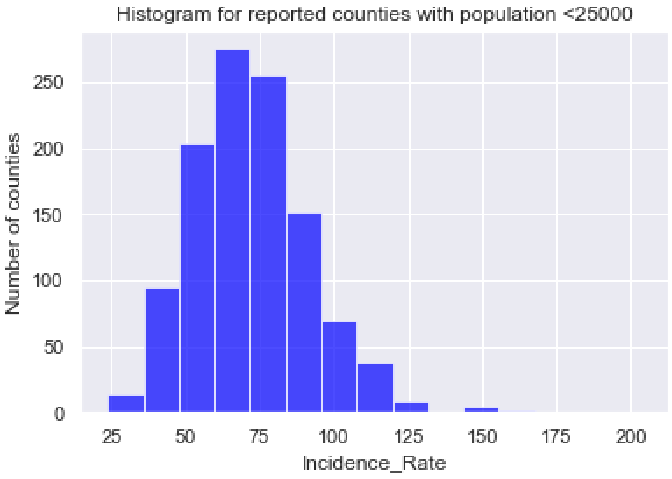


Fig. 7. Distribution of Incidence Rate for counties with population less than 25000

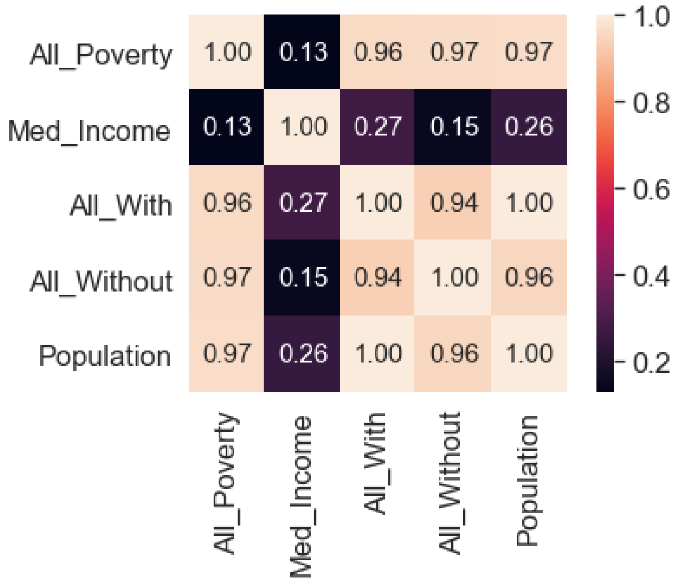


Fig. 8. Correlation heatmap between parameters(per 100,000)

From the plot, we can see that **All_Poverty, All_With, All_Without and Population** are very highly correlated. 3 out of the 4 are redundant. This is also very misleading as it is reasonable that as population increases the number of people in poverty and the others also increase. It is the fraction of the population that we have to actually analyze. Changing all the counts per 100000 to counts per capita, we get Fig 9

All_With and All_Without are complimentary as expected so we can remove the latter for redundancy. Others seem to be either weakly/moderately correlated which means there are no other redundant features.

With this, the data is clean and ready for model creation and analysis.

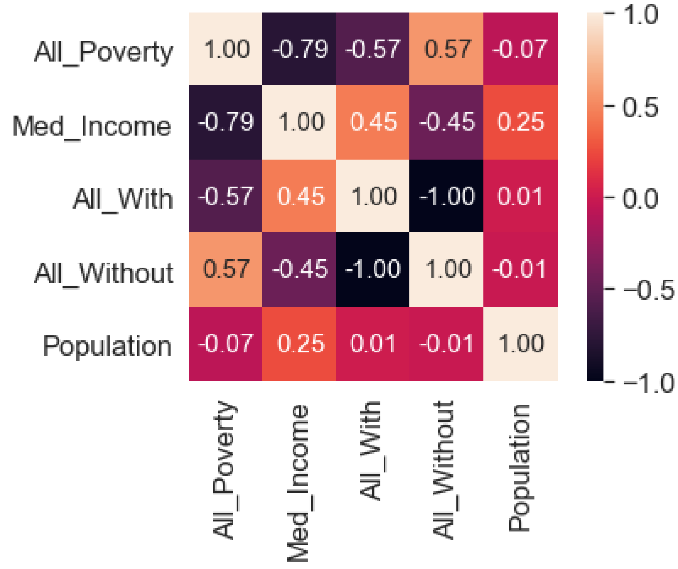


Fig. 9. Correlation heatmap between parameters(per capita)

V. Statistical Model

With the data being ready, the next step is to create a model between the independent variables and the dependent variable. LinearRegression feature from scikit-learn package was used to implement all the models. For every model, the coefficients and the $R^2 score$ is determined to observe the patterns and trends. All the data was normalized to prevent extreme values of the coefficients.

A. Incidence rate as the dependent variable

The input variables in this case are **'All_Poverty', 'Med_Income', 'All_With', 'Population'**.

1) Model 1: This is a basic linear regression model without any extra features. The results of the model are shown in Table II

TABLE II
Model 1 - Incidence Rate

Parameter	Coefficient
All_Poverty	-0.0128
Med_Income	-18.9396
All_With	-0.0018
Population	-0.0
Constant	0.0143
R^2 **score**	0.15

Looking at the R^2 value, it seems like this is a very weak model which is reasonable as we can say that the incidence is not very dependent on the socio-economic status. The reason being the cost of getting diagnosed is not very high unlike the treatment. A key takeaway from this is that the coefficient of Median Income is negative and also a very high value. This shows how strongly the income levels affect the incidence rate. One more interesting observation is that the coefficient of population is very low compared to others. This can be explained by the fact that since all the counting variables are in the per capita scale, they themselves include the population aspect and population as a variable becomes redundant in this case.

2) Model 2: Ridge regression is implemented in this model to prevent the coefficients from reaching an exceedingly high value. Table III shows the parameters with alpha set as 0.1.

TABLE III
MODEL 2 - INCIDENCE RATE

Parameter	Coefficient
All_Poverty	0.009
Med_Income	-0.0016
All_With	-0.0021
Population	-0.0
Constant	0.0043
R^2 **score**	0.08

To check the effect of alpha, a plot was drawn for R^2 value for various values of alpha (Fig 10). It is clear from the plot that the best model is when there is no regularization.

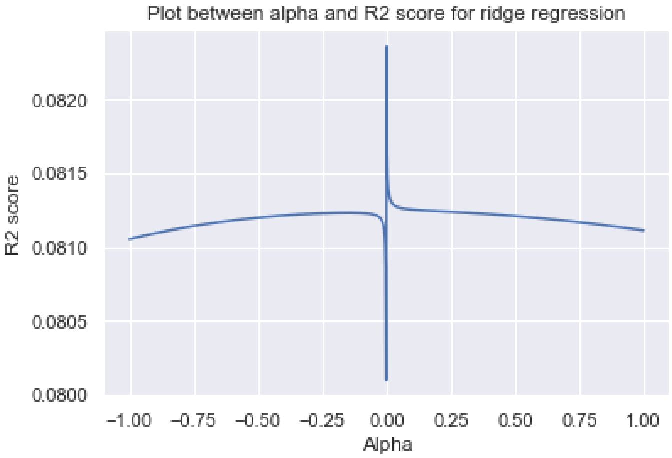


Fig. 10. Plot showing the variation of R2 value with alpha

B. Mortality rate as the dependent variable

The input variables in this case are **'All_Poverty', 'Med_Income', 'All_With' and 'Incidence_Rate**. Population is removed in these models as it was concluded that it is a redundant feature. Logically speaking, the incidence rate should definitely have an influence over the number of deaths. The influence of the same was monitored using the observations from the following models.

*1) Model 1 - **Without Incidence Rate**:* Similar to last section, a simple linear regression model was chosen in this case. The results of the model are shown in Table IV.

The results are similar to that of Model 1 of the Incidence model where the Median Income heavily affects the mortality rate.

TABLE IV
MODEL 1 - MORTALITY RATE

Parameter	Coefficient
All_Poverty	-0.0115
Med_Income	-16.7785
All_With	-0.0032
Constant	0.0135
R^2 **score**	0.14

TABLE V
MODEL 2 - MORTALITY RATE

Parameter	Coefficient
All_Poverty	0.0006
Med_Income	0.2657
All_With	-0.0022
Incidence_Rate	0.7706
Constant	0.0018
R^2 **score**	0.94

*2) Model 2 - **With Incidence Rate**:* Including the incidence rate as one of the independent variables changes the table drastically as shown in Table V.

The value of R^2 is very high which means this is a very predictable model. However, the coefficients of All_Poverty and Med_Income have changed sign which means their effects are changed to the opposite direction compared to Model 1.

*3) Model 3 - **Without other features**:* This model was implemented to check how removing other features affect the coefficient of the incidence rate

TABLE VI
MODEL 3 - MORTALITY RATE

Parameter	Coefficient
Incidence_Rate	0.773
Constant	-0.0
R^2 **score**	0.94

There are no significant changes in both the coefficient and the R^2 value between both model 2 and 3. This is due to the influence of Incidence_rate. From Table II, we can see that Incident rate itself is heavily influenced by the median income. Hence, in this model, it becomes redundant.

VI. MODEL VISUALIZATION

Fig 11 to Fig 15 shows the plots between the actual and the fitted values.

Some of the observations that can be confidently made based on the above analysis and visualizations are:

- Incidence Rate dominates the influence over all the other parameters on the Mortality Rate
- Of the other variables, the income affects both the incidence and the mortality rate the most. For every 1% decrease in median income, the incidence rate increases by around 19% and mortality rate by around 16%. Hence, the lower income sections are at a high risk of getting diagnosed with cancer.

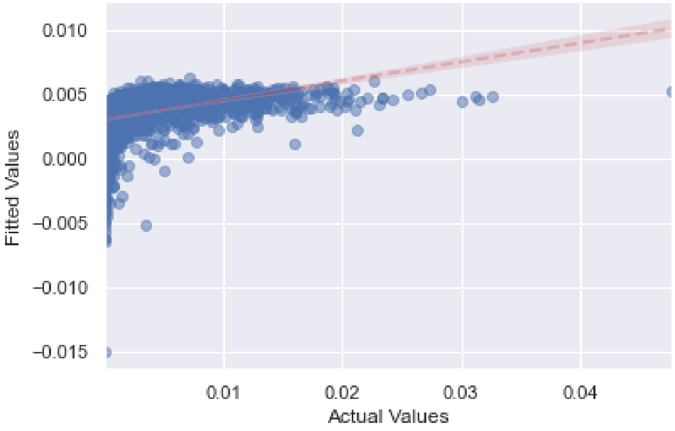


Fig. 11. Model 1 - Incidence Rate

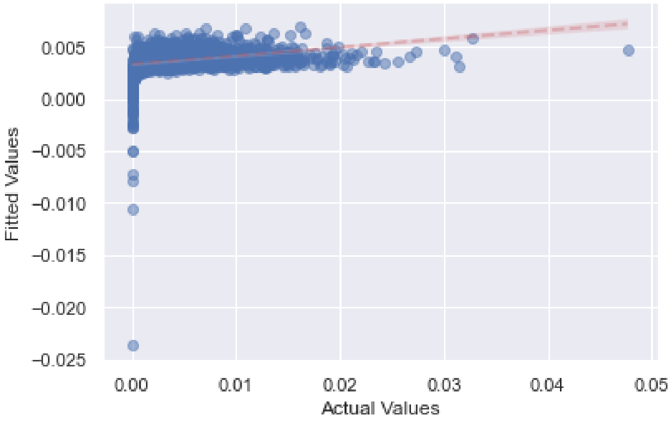


Fig. 12. Model 2 - Incidence Rate

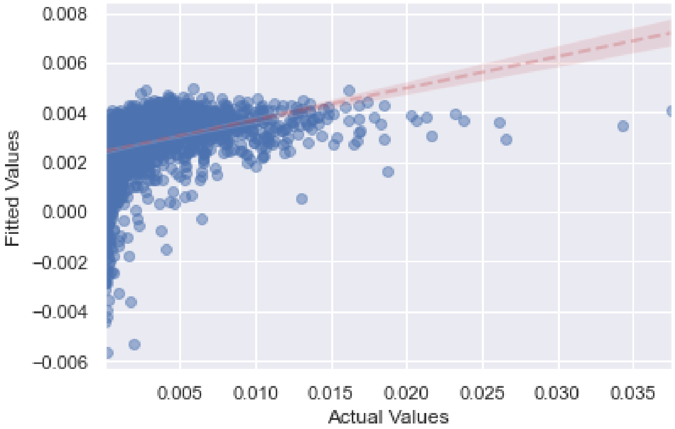


Fig. 13. Model 1 - Mortality Rate

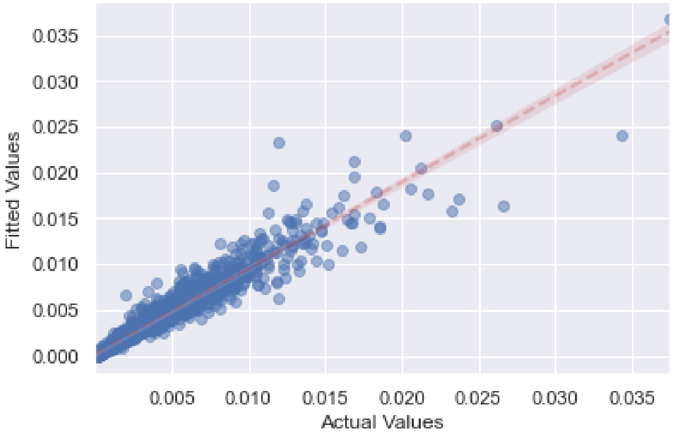


Fig. 14. Model 2 - Mortality Rate

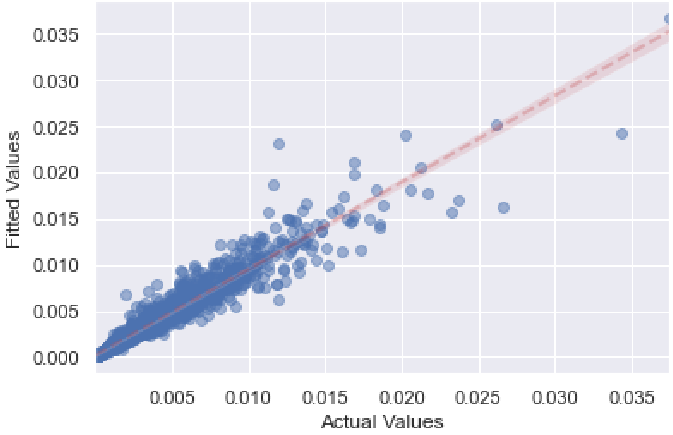


Fig. 15. Model 3 - Mortality Rate

- Poverty level per capita too has a similar effect on both the rates. However, the degree of influence is very small compared to income.
- From the visualizations, models with incidence rate as an independent variable seem to be highly accurate compared to other models. This is logical because cancer deaths happen only to those people who are diagnosed.

VII. Conclusion

The influence of socio-economic factors on the rates of cancer diagnosis and mortality were analyzed in great detail using techniques of linear regression. It was found out that low income populations are at a higher risk of getting diagnosed. The NGO should focus on it's most part on the counties which have a high incidence rate combined with a low income since mortality is most likely due to shortage of funds for the low income groups and in turn they are the ones which are likely to get diagnosed with the same,

References

[1] Müller, Andreas and Sarah Guido, "Introduction to machine learning with Python: a guide for data scientists", 2016.
[2] Understanding Linear Regression, Towards Data Science.
[3] Dr. Kaushik Mitra, Dr. Ramkrishna Pasumarthy, "Regression", EE4708: Data Analytics Laboratory - Week 3