Assignment-1: A Mathematical Essay on Linear Regression

Shania Mitra
Chemical Engineering
Indian Institute of Technology Madras
Chennai, India
ch18b067@smail.iitm.ac.in

***Abstract*—In this study, we examine the effect of low income on cancer diagnosis and treatment among populations in the United States. We demonstrate the correlation of cancer incidence and mortality with socioeconomic status, and provide both quantitative and visual evidence for the same.**

***Index Terms*—Linear Regression, Visualization, Cancer Diagnosis**

I. Introduction

Cancer survival is known to vary by socio-economic group. A report published by the International Agency for Research on Cancer (IARC) in 1997 indicated that lower SES tends to have higher cancer incidence and poorer cancer survival than higher SES in both developed and less-developed countries. In this study, we use data from CDC's National Program of Cancer Registries Cancer Surveillance System (NPCR-CSS) as well Incidence data provided by the SEER Program to probe these correlations further using linear regression models.

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One or more variables are considered to be explanatory variables, and one is considered to be a dependent variable. 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.

In this study, we use linear regression to explore the correlation between cancer incidence and mortality rates and socio-economic and racial factors. We begin by gathering, cleaning and preparing the data, following which we perform exploratory analysis. Finally, we build statistical models and perform visualizations to provide quantitative and visual evidence of the relationships observed. In the next section, we highlight the key principles underlying Linear Regression. In section 3, we discuss the insight and observations drawn from the data and the models. Finally, in section 4 we outline the salient features of the study and present further avenues of possible investigation.

II. Linear Regression

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- Linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables
- Linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables

A. Formulation

Given a data set $\{y_i,\, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors x is linear. This relationship is modeled through a disturbance term or error variable an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors.

The model takes the form
$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

The following are the major assumptions made by standard linear regression models with standard estimation techniques (e.g. ordinary least squares):

- Weak exogeneity: This means that the predictor variables x can be treated as fixed values, rather than random variables
- Linearity: This means that the mean of the response variable is a linear combination of the parameters (regression coefficients) and the predictor variables.

- Constant variance (a.k.a. homoscedasticity): This means that the variance of the errors does not depend on the values of the predictor variables.
- Independence of errors: This assumes that the errors of the response variables are uncorrelated with each other.

B. Estimation methods

A large number of procedures have been developed for parameter estimation and inference in linear regression. These methods differ in computational simplicity of algorithms, presence of a closed-form solution, robustness with respect to heavy-tailed distributions, and theoretical assumptions needed to validate desirable statistical properties such as consistency and asymptotic efficiency. Some common methods among others include:

- Least Squares Estimation
- Maximum Likelihood Estimation
- Bayesian Estimation

C. Least Squares Estimation

Assuming that the independent variable is $\overrightarrow{x_i} = \left[x_1^i, x_2^i, \ldots, x_m^i\right]$ and the model's parameters are $\vec{\beta} = [\beta_0, \beta_1, \ldots, \beta_m]$, then the model's prediction would be

$$y_i \approx \beta_0 + \sum_{j=1}^{m} \beta_j \times x_j^i$$

If $\overrightarrow{x_i}$ is extended to $\overrightarrow{x_i} = \left[1, x_1^i, x_2^i, \ldots, x_m^i\right]$ then y_i would become a dot product of the parameter and the independent variable, i.e.

$$y_i \approx \sum_{j=0}^{m} \beta_j \times x_j^i = \vec{\beta} \cdot \overrightarrow{x_i}$$

In the least-squares setting, the optimum parameter is defined as such that minimizes the sum of mean squared loss:

$$\overrightarrow{\hat{\beta}} = \arg\min_{\vec{\beta}} L(D, \vec{\beta}) = \arg\min_{\vec{\beta}} \sum_{i=1}^{n} \left(\vec{\beta} \cdot \overrightarrow{x_i} - y_i\right)^2$$

Now putting the independent and dependent variables in matrices X and Y respectively, the loss function can be rewritten as:

$$\begin{aligned} L(D, \vec{\beta}) &= \|X\vec{\beta} - Y\|^2 \\ &= (X\vec{\beta} - Y)^\top (X\vec{\beta} - Y) \\ &= Y^\top Y - Y^\top X\vec{\beta} - \vec{\beta}^\top X^\top Y + \vec{\beta}^\top X^\top X\vec{\beta} \end{aligned}$$

As the loss is convex the optimum solution lies at gradient zero. The gradient of the loss function is (using Denominator layout convention):

$$\begin{aligned} \frac{\partial L(D, \vec{\beta})}{\partial \vec{\beta}} &= \frac{\partial \left(Y^\top Y - Y^\top X\vec{\beta} - \vec{\beta}^\top X^\top Y + \vec{\beta}^\top X^\top X\vec{\beta}\right)}{\partial \vec{\beta}} \\ &= -2X^\top Y + 2X^\top X\vec{\beta} \end{aligned}$$

Setting the gradient to zero produces the optimum parameter:

$$\begin{aligned} -2X^\top Y + 2X^\top X\vec{\beta} &= 0 \\ \Rightarrow X^\top Y &= X^\top X\vec{\beta} \\ \Rightarrow \overrightarrow{\hat{\beta}} &= \left(X^\top X\right)^{-1} X^\top Y \end{aligned}$$

III. The Problem

In this study, our aim is to examine whether low income groups are at greater risk for being diagnosed and dying from cancer to help a nonprofit with lobbying and fundraising. For this, firstly we demonstrate whether or not cancer incidence and mortality are correlated with socioeconomic status and secondly, we provide both quantitative and visual evidence that the nonprofit can take and use to further their mission.

A. Data Preparation

1) Existing features in the data set: The merged dataset, used in this study, consists of 25 columns and 3134 samples, which are areas in various states. Interpretation of the features is as follows:

1) State: State of respective area, totally 51 in number
2) Area: Name of area in which sampling is done
3) All_Poverty: Number of people of both genders below the poverty line. Similarly, M_poverty is for males, F_Poverty is for females
4) FIPS: Zipcode of the area
5) Med_Income: Median Income of all ethnic groups in the area
6) Med_income_White, Black, etc: Median Income of a particular ethnic group in the area
7) All_With: Number of individuals having an insurance in the area; Along these lines, All_Without, Without_Male, With_Male, Without_Female, With_Female are defined
8) Incidence_Rate: Number of cancer cases detected per 100,000 people in the area
9) Mortality_Rate: Number of mortalities per 100,000 people

To impute the missing values, two approaches are considered:

1) State wise medians are filled in in place of the empty values such that the distribution does not get distorted, as would be the case if we imputed using means
2) Missing Values are dropped

Multiple values in the columns had noisy characters such as '**', '*', these were replaced with missing values and '3 or fewer' was replaced with 3. Some numbers were followed by a '#' which was removed. In case of '* since we cannot be sure of how many cases it is reliably (below 16), these points are not considered.

Finally, columns 'FIPS', 'fips_x','fips_y' were dropped assuming pincode holds no relation with incidence or mortality rates.

State	0.000000
AreaName	0.000000
All_Poverty	0.000000
M_Poverty	0.000000
F_Poverty	0.000000
FIPS	0.000000
Med_Income	0.031908
Med_Income_White	0.063816
Med_Income_Black	38.608807
Med_Income_Nat_Am	52.967454
Med_Income_Asian	56.062540
Hispanic	21.729419
M_With	0.000000
M_Without	0.000000
F_With	0.000000
F_Without	0.000000
All_With	0.000000
All_Without	0.000000
fips_x	0.000000
Incidence_Rate	13.401404
Avg_Ann_Incidence	6.668794
recent_trend	14.901085
fips_y	0.000000
Mortality_Rate	10.370134
Avg_Ann_Deaths	10.370134

Fig. 1. Percentage of missing values in each of the columns

	All_Poverty	Med_Income	Mortality_Rate
mean	1.522966e+04	46819.837855	53.188537
min	1.000000e+01	19328.000000	9.200000
50%	4.294000e+03	45075.000000	52.600000
max	1.800265e+06	123453.000000	125.600000

Fig. 2. Mean, Mininum, Median and Maximum of 3 features from the dataset

In figure 2, we list a sample of mean, median, mininum and maximum of 3 features as an example of the measures considered in this study

2) Additional features:

1) Total Number of Males: Obtained by summing number of males insured and number of males not insured, similarly we obtain the total number of females (It was verified that number of males with insurance + number of males without insurance + number of females with insurance + number of females without insurance == all with insurance + all without insurance)
2) Total Population: The total population of the area is obtained by summing number of males and females
3) Female Ratio: Number of females divided by the total number of people, similarly we can obtain ratio of males in the area
4) Female Poverty Ratio: Ratio of poor females among total number of females, similarly Male Poverty Ratio can be obtained
5) Poverty Ratio: Number of poor people divided by total number of people in that area
6) Female Insurance Ratio: Number of females insured divided by total number of females, similarly, we can obtain male insurance ratio
7) Number of Groups above Median: This is a number ranging from 0 to 5 indicating the number of ethnic groups having median income above the median income of the entire population
8) For the purpose of visualisation, many of the above features have been bucketed by splitting across the median/mean:
   a) Insurance Ratios: The male and female insurance ratios are split across the median and combined to form 4 categories
      i) Low Female, Low Male: both genders in the area have insurance ratios lower than the median
      ii) Low Female, High Male
      iii) High Female, Low Male
      iv) High Female, High Male
   b) Female Population: Female Ratio is binarised (Low, High) by splitting across the median
   c) Poverty Category: Poverty Ratio and Median Income are split across the mean and combined to form 4 categories
      i) Most Poor: This category implies most people in the area are poor since poverty ratio is high and median income is low (1298 samples)
      ii) Many Poor, Few Very Rich: This category implies that a majority of the people are poor (high poverty ratio) but very few are very rich, biasing the mean (mean 'median_income' is high) (1237 samples)
      iii) Few Very Poor, Many Rich: This category consists of areas with low poverty ratios and low mean income (463 samples)
      iv) Many Rich: This category consists of areas with low poverty ratios and high mean income (135 samples)

B. Exploratory Analysis

In this section we look at the trends between various features:

In Figure 3 we see the correlation matrix forming clusters of features, indicating that groups of features are highly linearly related to others in that group. From this we can already see that median income and mortality rate have a negative moderate correlation coefficient, indicating that as the median income increases, mortality rates decrease.

In figure 4 we plot the median income distributions for each of the ethnic groups. We observe that blacks have the lowest mean 'median income' while asians have the highest mean 'median income'. Further, native americans and Asians have a spread out distribution with higher variance while whites, blacks and hispanics have lower variance in income, indicating that most blacks and hispanics are payed low incomes, hinting at racial discrimination.

In figure 5 we plot the mean mortality rate against the mean incidence rate of each of the 51 states. There is strong evidence of a linear relationship confirming that as the rate of incidence increases, mortality rates also increase.

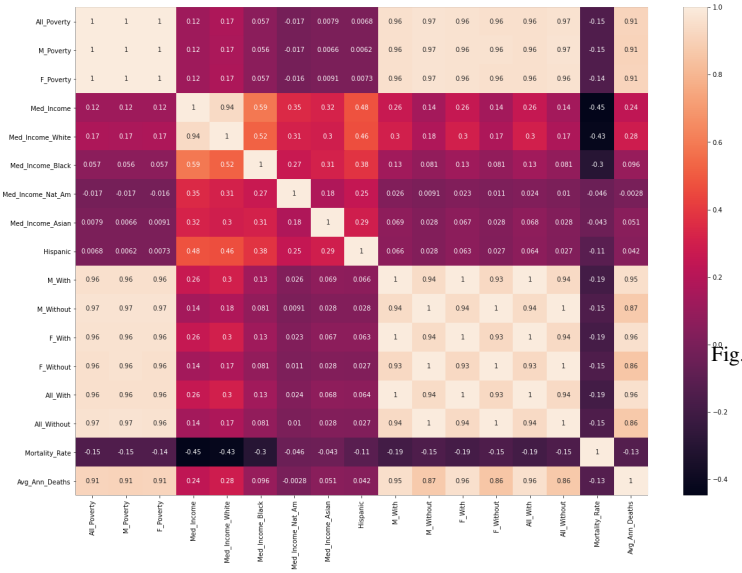

Fig. 3. Pearson Correlation Coefficient among all numeric features

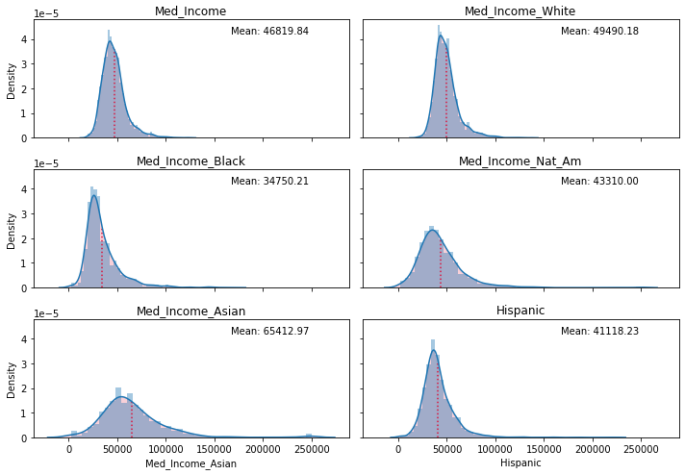

Fig. 4. Distribution of Median Incomes in all areas (a) all ethnic groups (b) Whites (c) Blacks (d) Native Americans (e) Asians (f) Hispanics

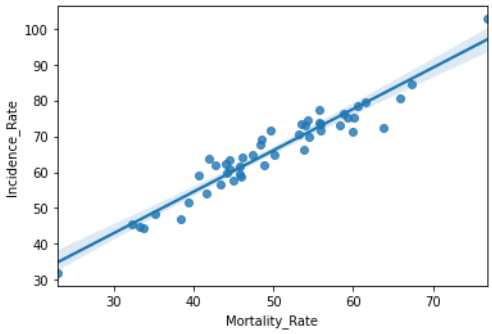

Fig. 5. Mean Incidence Rate vs Mean Mortality Rates for the 51 states

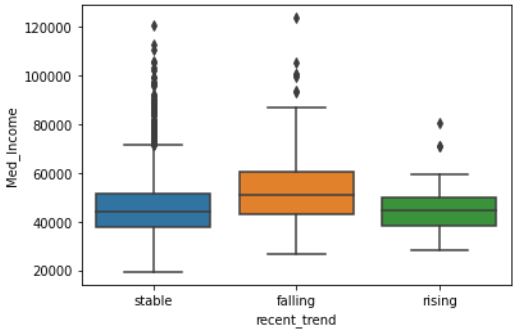

Fig. 6. Recent trend in number of cases vs Median Income in each area

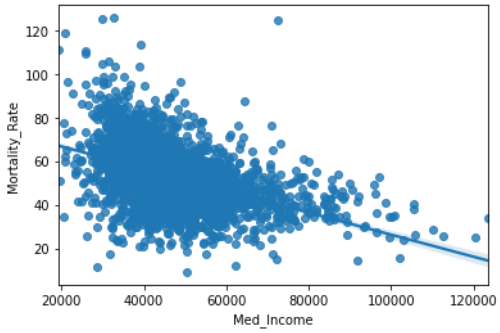

Fig. 7. Plot of Mortality Rate vs Median Income

In figure 6, the trend of rise/fall of cases is plotted against median income. It is clearly seen that in regions with a falling trend in cases, the median "median_income" is higher than the median in other regions, suggesting that better income may help people afford better healthcare and facilities, leading to a falling trend in cases.

In figure 7, it is seen that there is a huge cluster of points at low median incomes, however, at high median incomes we can clearly see that there is trend of decreasing mortality rate with income.

In figure 8 we observe that areas with more ethnic groups earning above the median salary are more likely to subscribe to an insurance. This insurance subscription, as can be seen in figure 13, does not, however, translate to a lower incidence rate or a lower mortality rate

In figure 9 it is seen that among all ethnic groups, sub-groups with higher median income tend to show a falling trend in cases. This plot further shows us that among all ethnic groups, blacks have the lowest median income even among the areas with falling cases.

In figure 10, it is clearly visible that on an average females are more poverty-stricken than males. From figure 11, we observe that regions where poor females are more in number as compared to poor males, incidence and mortality rates are higher. Further, from figure 12 it can be inferred that regions with higher female population (above the median female population) have higher mean and median incidence and mortality rates. These three plots help us deduce that

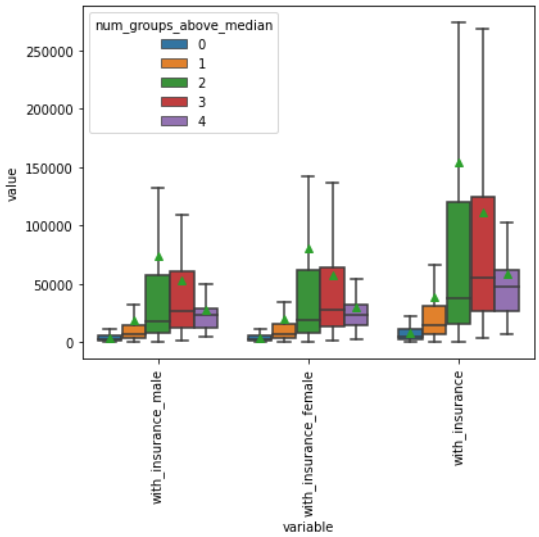

Fig. 8. Plot of number of males/females/all with/without insurance in regions grouped by number of ethnic groups having income above median income

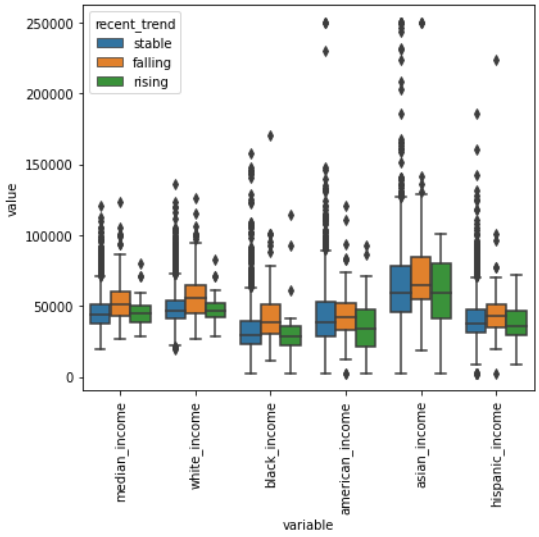

Fig. 9. Plot of median income of various ethnic groups in regions with stable, falling and rising trend in number of cases

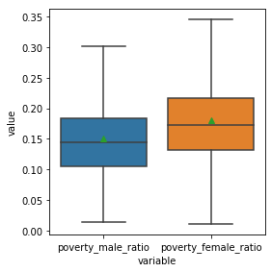

Fig. 10. Distribution of poverty ratios among males and females in all the areas

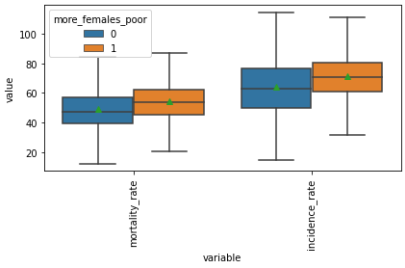

Fig. 11. Plot of Incidence and Mortality Rates in areas with a greater ratio of poor females in orange and greater ratio of poor males in blue

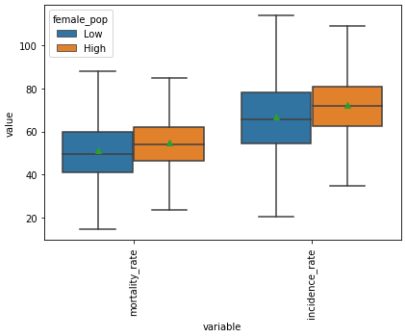

Fig. 12. Plot of Mortality and Incidence Rates for regions with low and high female population

females, who tend to be poorer, are more prone to cancer and must be made aware of the disease and its treatment.

In figure 13, we observe that median mortalities and incidence rates are unaffected by rates of insurance among males and females.

As explained in section A, areas are grouped into 4 regions according to the ratio of poor people and the median income of the area. In figure 14 we confirm that regions that have mostly poor people (low median income and high ratio of poor people) have the highest mortality and incidence rates, again suggesting that socioeconoic status and cancer incidence/mortality have direct linkage.

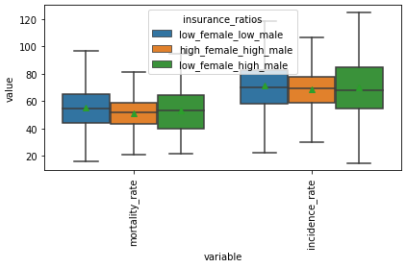

Fig. 13. Plot of Mortality and Incidence Rates among areas with low male and female insurance ratios in blue, high female and high male insurance ratios in orange and low female and high male insurance ratios in green. It must be noted that no samples with high female and low male insurance ratios exist.

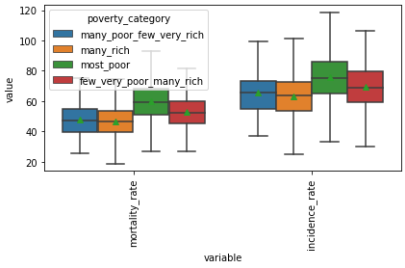

Fig. 14. Incidence and Mortality Rates for each poverty category

Dataset	Incidence Rate	Mortality Rate
Train	0.0434	0.0480
Test	0.05064	0.04514

TABLE I
TRAIN AND TEST R^2 VALUES OF LINEAR MODEL USING NUM_POVERTY ONLY

C. Statistical Models and Visualizations

To build the linear regression models, the data was scaled to the range of 0 to 1 by dividing by the maximum value of each feature. Further a train-test split of 80-20 was followed.

1) Modelling Incidence and Mortality Rates as a function of Number of poor people only: From table I and figure 15 it is clearly seen that the Num_poverty alone does not explain incidence and mortality rates enough and additional information is required since both train and test performances are poor.

2) Modelling Incidence and Mortality Rates as a function of Number of poor people and Median Income only: From table II and figure 16 it is seen that inclusion of median_income among explantory variables significatly improves the performance, however train and test performance still remain poor.

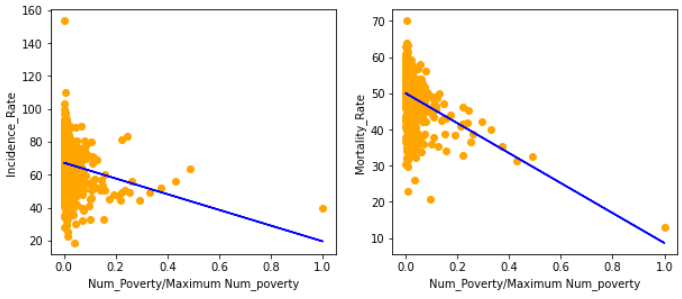

Fig. 15. Normalized num_poverty vs Incidence and Mortality rates. The blue line signifies the learned model while the orange points signify the samples in the entire data

Dataset	Incidence Rate	Mortality Rate
Train	0.1748	0.2121
Test	0.2073	0.3043

TABLE II
TRAIN AND TEST R^2 VALUES OF LINEAR MODEL USING NUM_POVERTY AND MEDIAN_INCOME

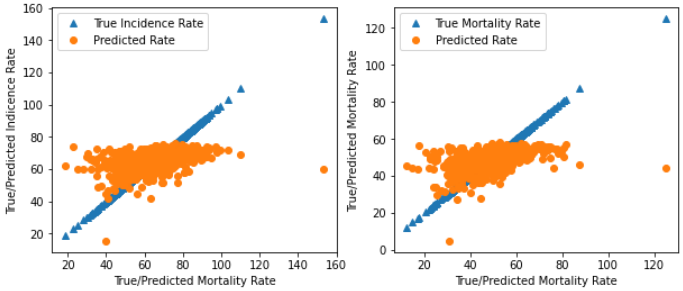

Fig. 16. Plot of predicted incidence/mortality rates vs true incidence/mortality rates. The blue triangles signify the ideal outcome, where true and predicted outcomes are the same, while the orange circles show us the predicted values for the given true values, for the entire dataset

Dataset	Incidence Rate	Mortality Rate
Train	0.9680	0.9600
Test	0.3083	0.2991

TABLE III
TRAIN AND TEST R^2 VALUES OF LINEAR MODEL USING ALL ENGINEERED FEATURES AFTER DROPPING ONES CAUSING MULTICOLLINEARITY

3) Modelling Incidence and Mortality Rates as a function of engineered features: In this case, all new features listed in A along with the existing ones are processed and fed to the the final model. As part of the processing:

1) All features with pearson correlation coefficient > 0.8 are dropped
2) The remaining features are fed to the model and hypothesis tests on the coefficients are conducted to see if they are significantly above 0, (i.e., null hypothesis: coefficient of feature = 0, alternate hypothesis: coefficient of feature <> 0). Features with p-value > 0.1 are dropped (i.e., confidence interval is taken to be 90%). In this step, 5 features: 'black_income', 'asian_income_more', 'hispanic_income_more', 'num_groups_above_median', 'more_females_insured' are dropped

From table III and figure 17 it is seen that inclusion of newly created features radically improves train performance and significantly improves test performace suggesting that the new features contain significant amount of information. However, the vast difference between train and test R2 scores

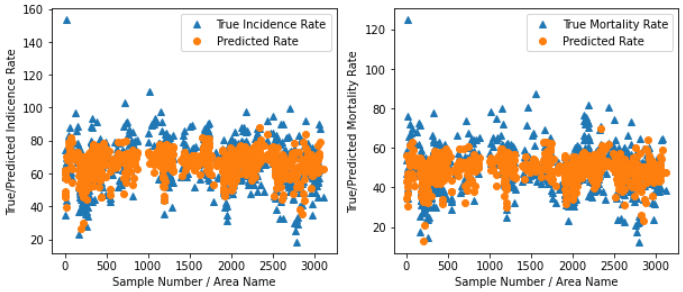

Fig. 17. Plot of true y-value (blue) and predicted y-value (orange) sample by sample

TABLE IV

Dataset	Incidence Rate	Mortality Rate
Train	0.5799	0.5669
Test	-0.1318	-0.8895

TABLE V
TRAIN AND TEST R^2 VALUES OF LINEAR MODEL USING POLYNOMIAL OF EXISITNG FEATURES OF DEGREE 2

suggest that overfitting takes place.

4) Modelling Incidence and Mortality Rates as a function of all features using Polynomial Regression: To build the linear model, all features with correlation coefficient > 0.8 to prevent multicollinearity and all features including interaction features of degree 2 are created and fed to the model.

From table IV and figure 18 we observe that the train performance upon inclusion of the polynomial features improves from the univariate (with num_poverty) and bivariate (with num_poverty, median_income) cases significantly, but the test performance is the poorest of all cases, suggesting that drastic overfitting takes place. In figure 18, only the test set is plotted to observe the poor performance, and it can be seen that the predicted values are highly scattered and far away from the true values.

From all these models it can be said that there is some correlation between socioeconomic factors and incidence/mortality rate (since, upon inclusion of those factors prediction capability of the model improves) however it is not very strong since test performance is not satisfactory in each of the cases.

IV. CONCLUSIONS

In this study, we observed that although not high, there is some correlation between socio-economic status and incidence and mortality (from the models as well as the exploratory analysis). Further, females are poorer and more prone to cancer incidence and mortality and hence the non-profit organization must create more awareness among them about the disease and must make efforts in the direction of creating more employment among women and strive for equitable pay among those that already work. To improve the performance of models, in future, non-linear models may be explored.

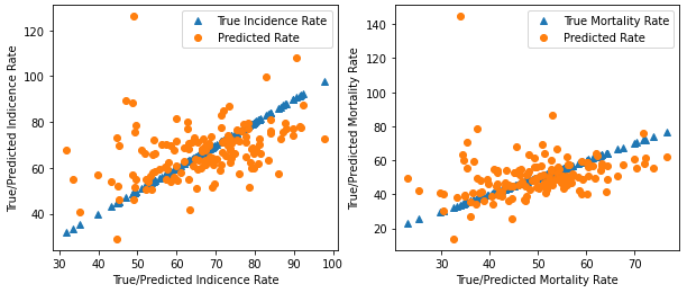

Fig. 18. Plot of predicted incidence/mortality rates vs true incidence/mortality rates. The blue triangles signify the ideal outcome, where true and predicted outcomes with polynomial features are the same, while the orange circles show us the predicted values for the given true values, for the test set ONLY

REFERENCES

[1] https://en.wikipedia.org/wiki/Linear_regression
[2] https://ioair.link/7djh8f
[3] https://www.annalsofoncology.org/article/S0923-7534(19)40327-X/fulltext
[4] https://link.springer.com/chapter/10.1007/978-981-15-1831-7_4
[5] https://towardsdatascience.com/laymans-introduction-to-linear-regression-8b334a3dab09
[6] https://www.statsmodels.org/stable/examples/notebooks/generated/predict.html
[7] https://aegis4048.github.io/mutiple_linear_regression_and_visualization_in_python
[8] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.