Assignment 1: a mathematical essay on linear regression

Gautham Govind A
Dept. of Electrical Engineering
Indian Institute of Technology Madras
ee19b022@smail.iitm.ac.in

***Abstract*—The objective of this assignment is to explore the mathematical formalism behind linear regression and then to use it in a real-life application. In this assignment, as a real-life application, linear regression is used to formally identify the relationship between socioeconomic status and cancer incidence, mortality rates. Linear regression is implemented using Python. The analysis enables us to arrive at the conclusion that the socioeconomic status does indeed have an impact on the cancer incidence. mortality rates.**

***Index Terms*—linear regression, python, visualization**

I. Introduction

Cancer is one of the most pressing problems faced by society today. While there exists methods to prevent as well as cure cancer, question remains as to whether everyone has easy access to it. It is thus necessary to systematically analyse the impact the socioeconomic status of an individual has on their chance of cancer incidence/ mortality. The goal here is therefore to make use of data to see how cancer incidence/mortality rates of a person is influenced by their income/societal status.

With the technological advancements we have made, it is now possible to make use of data analytic tools to derive insights from data. By aggregating, cleaning and analysing large amounts of data, it is possible to draw inferences regarding how some parameters may influence others. Linear Regression is one such model which can be used for building linear relationships between variables. The attempt is to make use of Linear Regression to see how various factors affect cancer incidence/mortality of a population.

In this work, we make use of Python for performing the necessary data manipulations. Various libraries in Python are used for cleaning the data, visualizing the data and analyzing the data. Namely, the libraries pandas, seaborn and scikit-learn are extensively used. Jupyter notebook environment is used to easily integrate concise explanation for the code along with the code itself.

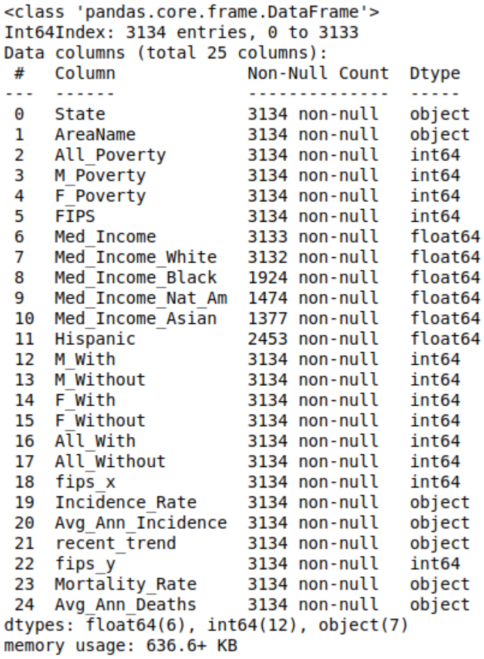
After a careful analysis of the available data, and using linear regression, we arrive at the conclusion that the socioeconomic status of an individual does have a significant impact on his chance of cancer incidence/ mortality. More specifically, we find that poverty rate in general has a positive correlation with cancer incidence/ mortality, meaning as poverty rate increases there is a higher chance for cancer incidence/ mortality in that population. We also find that the median income has a negative correlation with cancer incidence/ mortality. Thus, it can be said that low-income groups are at a greater risk of suffering from cancer and it is imperative to address the issue to ensure well being of all members of the society.

Section II gives an overview of the various techniques used for data cleaning and an initial exploratory analysis. A lot of insights can be gained just by making qualitative observations from the given data. Section III gives a short description of the mathematical formalism behind linear regression. section IV sates the various results that are obtained by applying linear regression in this particular case. Section V gives a summary of the major conclusions drawn from the analysis.

II. Exploratory Data Analysis

In this section, we describe the process of data cleaning and visualization.

The given dataset consists of 3134 rows and 25 columns. A brief overview of the dataset is presented in Figure 1.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3134 entries, 0 to 3133
Data columns (total 25 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   State               3134 non-null   object
 1   AreaName            3134 non-null   object
 2   All_Poverty         3134 non-null   int64
 3   M_Poverty           3134 non-null   int64
 4   F_Poverty           3134 non-null   int64
 5   FIPS                3134 non-null   int64
 6   Med_Income          3133 non-null   float64
 7   Med_Income_White    3132 non-null   float64
 8   Med_Income_Black    1924 non-null   float64
 9   Med_Income_Nat_Am   1474 non-null   float64
 10  Med_Income_Asian    1377 non-null   float64
 11  Hispanic            2453 non-null   float64
 12  M_With              3134 non-null   int64
 13  M_Without           3134 non-null   int64
 14  F_With              3134 non-null   int64
 15  F_Without           3134 non-null   int64
 16  All_With            3134 non-null   int64
 17  All_Without         3134 non-null   int64
 18  fips_x              3134 non-null   int64
 19  Incidence_Rate      3134 non-null   object
 20  Avg_Ann_Incidence   3134 non-null   object
 21  recent_trend        3134 non-null   object
 22  fips_y              3134 non-null   int64
 23  Mortality_Rate      3134 non-null   object
 24  Avg_Ann_Deaths      3134 non-null   object
dtypes: float64(6), int64(12), object(7)
memory usage: 636.6+ KB
```

Fig. 1. Summary of the raw dataset

Our first task is to weed out columns which are not of relevance. Since our objective is to make inferences for the whole population and not to make inferences for various states/areas separately, discerning between states/areas is unnecessary. This renders columns State, AreaName, FIPS, fips_x and fips_y. Hence, we drop these columns.

Next, we observe that the columns Incidence_Rate, Avg_Ann_Incidence, Mortality_Rate and Avg_Ann_Deaths are all marked as having ”object” datatype. This is an indication that not all values in these columns are numerical. This will present a problem during mathematical analyses. Hence, it is essential to examine these columns and try to make all values numerical or categorical.

On examination, we find the that there are some non-numerical entries in each of these columns. The type of entry and method adopted to convert it to a numerical value are discussed below:

- * - This indicates that the actual value is very low. It seems reasonable to replace these entries using 0. However, this results in the concentration of a lot of values at 0 and severely impedes the performance of regression models. Hence, these rows were removed.
- entries ending with # - This is just an error in formatting. The problem can be resolved by simply removing the # symbol.
- _ and __ - These indicate lack of information. Although there are imputation strategies available to deal with such data, we do not apply them here. This is because of two reasons:
  - Missing data is present in variables like Incidence_Rate which is what we would like to predict from other parameters. In this sense, these are like target variables and hence imputation seems non-ideal.
  - The rows with missing data for these variables accounts for only about 6% of the total number of rows. Since this is a small fraction, there isn’t really a lot of loss of data.

After performing all these operations, we are left with a cleaned dataset, whose summary is presented in Figure 2.

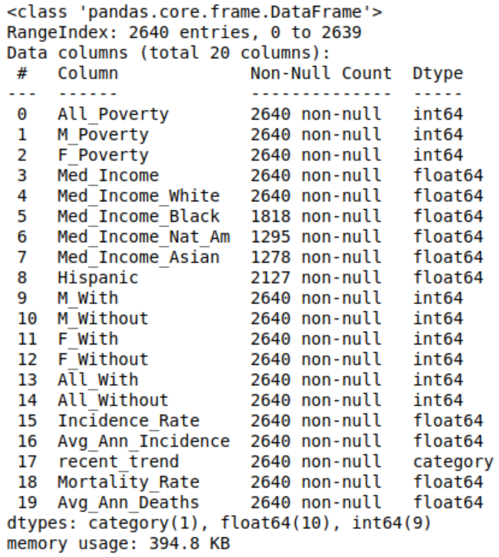
At this point, we also make an important observation regarding the variables Incidence_Rate, Avg_Ann_Incidence, Mortality_Rate and Avg_Ann_Deaths. Avg_Ann quantities represent the **numbers for the entire population** , whereas Incidence_Rate and Mortality_Rate represent the **numbers normalized using the total population.** Since what is of relevance to us is the normalized rate, we shall be **considering only Incidence_Rate and Mortality_Rate in all further analysis.**

Our objective is to see the impact of the following two factors on cancer incidence and mortality:

1) Economic status
2) Social status

A. Economic status

From the available dataset, we make the following observations:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2640 entries, 0 to 2639
Data columns (total 20 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   All_Poverty        2640 non-null   int64
 1   M_Poverty          2640 non-null   int64
 2   F_Poverty          2640 non-null   int64
 3   Med_Income         2640 non-null   float64
 4   Med_Income_White   2640 non-null   float64
 5   Med_Income_Black   1818 non-null   float64
 6   Med_Income_Nat_Am  1295 non-null   float64
 7   Med_Income_Asian   1278 non-null   float64
 8   Hispanic           2127 non-null   float64
 9   M_With             2640 non-null   int64
 10  M_Without          2640 non-null   int64
 11  F_With             2640 non-null   int64
 12  F_Without          2640 non-null   int64
 13  All_With           2640 non-null   int64
 14  All_Without        2640 non-null   int64
 15  Incidence_Rate     2640 non-null   float64
 16  Avg_Ann_Incidence  2640 non-null   float64
 17  recent_trend       2640 non-null   category
 18  Mortality_Rate     2640 non-null   float64
 19  Avg_Ann_Deaths     2640 non-null   float64
dtypes: category(1), float64(10), int64(9)
memory usage: 394.8 KB
```

Fig. 2. Summary of the processed dataset

- The variables we have with respect to the economic status are the following:
  - Number of individuals below poverty line
  - Median income of individuals
  - Number of individuals who have health insurance
- We expect a **positive correlation between incidence/mortality rate and poverty rate, whereas a negative correlation between median income and incidence/mortality rate.**
- In the case of **health insurance**, we **do not expect it to have much impact on the incidence rate** itself, whereas we expect the number of individuals who are insured (normalized with population) to have some **negative correlation with mortality rate.**

On calculating, we observe that **correlation between All_Poverty, which is the number of individuals below poverty line, and Incidence_Rate is -0.027 which is almost 0!** Does this mean there is no correlation between the two? The important fact to note here is that All_Poverty is the total number of individuals below the poverty line, i.e. it has not been normalized with total population, whereas Incidence_Rate is already normalized. Hence, it is **necessary to normalize this quantity as well with respect to total population.**

We can approximate the total population as the sum of all people with health insurance and all people without health insurance. We also calculate male and female population separately. We add these quantities as separate columns and then compute the **Poverty rate**, which is the number of individuals below poverty lane normalized using population, separately for males and females as well as for the entire population.

After calculating the poverty rates, we calculate the correlation between poverty rate (across all individuals) and incidence rate, mortality rate. We observe correlation values of **0.259** and **0.285**, which indicate **significant positive correlation.** To visualize these observations, we make scatter plots for incidence rate and mortality rate against poverty rate. These plots are presented in Figure 3 and Figure 4.

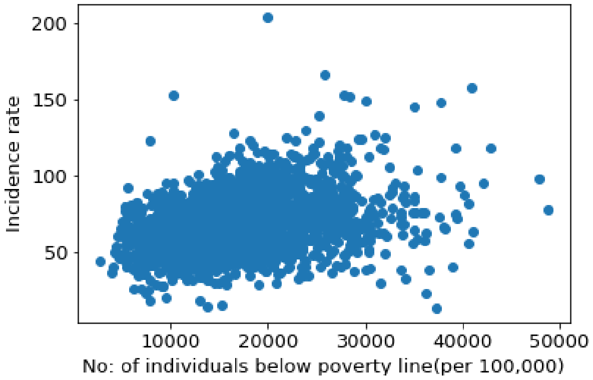


Fig. 3. Cancer incidence v/s Poverty

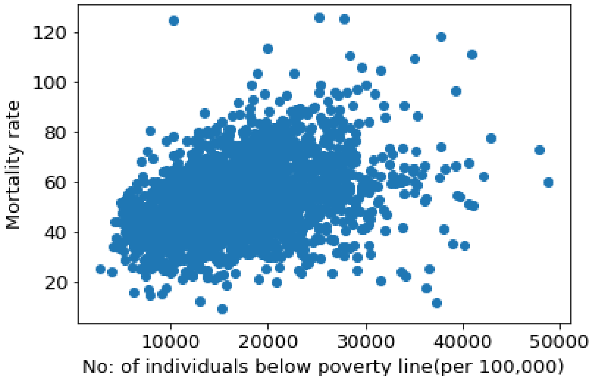


Fig. 4. Mortality v/s Poverty

From the plots, it is clear that cancer incidence/mortality increases, in general, with increase in poverty.

Next, we observe the relationship between median income and cancer incidence/mortality rate. On calculating, we find that the correlation values are **-0.256** and **-0.276** which indicate **significant negative correlation.** To visualize these observations, we make scatter plots for incidence rate and mortality rate against median income. These plots are presented in Figure 5 and Figure 6.

From the plots, it is clear that cancer incidence/mortality decreases, in general, with increase in income.

From the correlation values and plots, it can be qualitatively concluded that as poverty increases/ income decreases, there is a higher chance for cancer incidence/mortality in general.

We will now consider the impact of having a health insurance on cancer incidence/ mortality. On calculating, we observe correlation values of **-0.038** and **-0.124** for incidence and mortality respectively. It seems like there **isn't any correlation at all between cancer incidence and having a health insurance.** This is to be expected because a health insurance is of use only in the treatment of a disease; not so much for its prevention. Unlike the case of cancer incidence, there seems to be a **significant negative correlation between having a health insurance and mortality rate.** This is illustrated in Figure 7. Again, this is to be expected since having a health insurance provides an incentive to individuals to seek appropriate treatment without considering financial limitations.

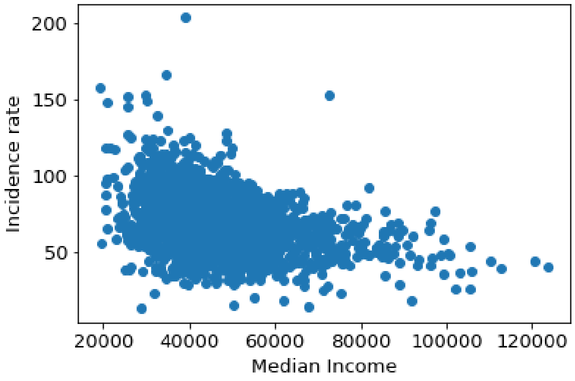


Fig. 5. Cancer incidence v/s Income

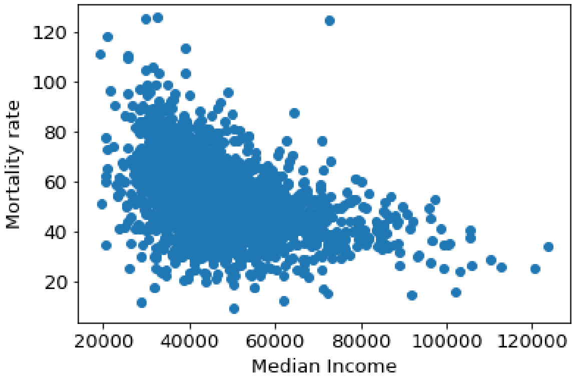


Fig. 6. Mortality v/s Income

B. Social status

Our target is to associate social status with cancer incidence/ mortality. Towards this, we shall consider the following two broad social classification criterion:

- Gender
- Ethnicity

The major challenge we face in this case is that we **do not have direct access to incidence/mortality rate for population subsections separately.** For instance, we do not know the separate count of cancer incidence of male individuals and cancer incidence of female individuals. This prohibits any

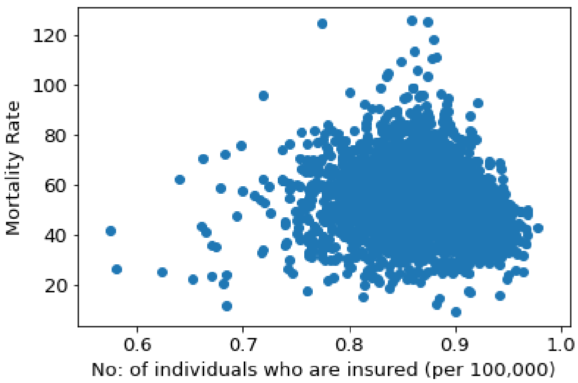


Fig. 7. Mortality v/s Insurance

direct comparison. As a result, it is necessary to make an **indirect comparison** based on the available parameters.

We first consider gender. For gender, we consider the poverty rate for males and females separately. Since we have already concluded from earlier analysis that poverty rate has a positive correlation with both incidence rate and mortality rate, **if we can conclude that one section has a higher poverty rate in general, we can conclude that this section suffers from higher incidence/mortality as well.**

To see how the poverty rates of the two sections compare, we make use of a box plot. The box plot is shown in Figure 8. From the plot, we can **qualitatively see that females have higher poverty rate, in general.** To make this more quantitative, consider Figure 9. **Clearly, mean and median of poverty rate of females are higher than males.** Hence, from this, it can be estimated that female population in general has higher poverty rate and **consequently higher values for cancer incidence and mortality rate.**

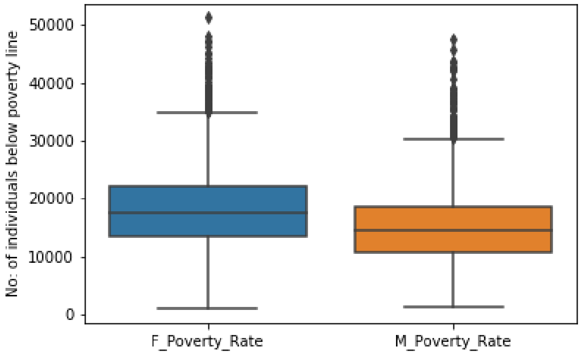
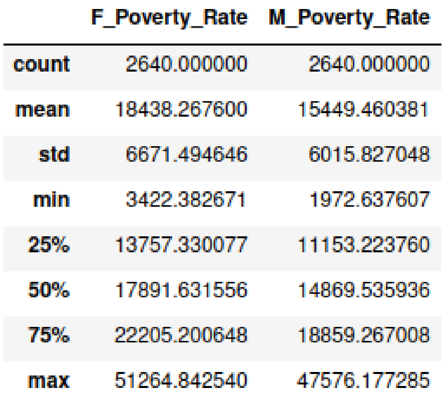


Fig. 8. Gender v/s Poverty

Next, we explore the relationship between ethnicity and cancer incidence/mortality. In this case, the parameter available to us in median income. As we have seen before, a lower median income would correspond to higher cancer incidence/mortality.



	F_Poverty_Rate	M_Poverty_Rate
count	2640.000000	2640.000000
mean	18438.267600	15449.460381
std	6671.494646	6015.827048
min	3422.382671	1972.637607
25%	13757.330077	11153.223760
50%	17891.631556	14869.535936
75%	22205.200648	18859.267008
max	51264.842540	47576.177285

Fig. 9. Statistics for poverty distribution

To see how the median income varies among different ethnic groups, we again make use of a box plot. The plot is given in Figure 11. It can be seen that the median income of different sections are different, with some sections having higher values and others having lower. A quantitative description is given in Figure 12. It can be seen that the mean of the median income follow the following order for ethnic groups: Asians > Whites > Native Americans > Hispanics > Blacks. Hence, from this, it can be estimated that cancer incidence/mortality rate also follows the same order for ethnic groups.

It must be noted that in all the analysis we have done in this subsection, we have attempted to correlate social status with economic indicators like poverty and income. This approach, though insightful, cannot guarantee a completely accurate analysis, since for this we will need information regarding incidence/mortality rate of each section of the population separately.

C. Incidence v/s Mortality

Finally, for the sake of completion, we explore the relationship between cancer incidence and mortality. This is fairly obvious, as we expect a high positive correlation between the two. Visual evidence is presented in Figure 10. We observe a correlation of **0.867** It is also worthwhile to note here that in general as one increases the other also increases with the only exception being the case of insurance rate, which has already been discussed.

III. Model: Linear Regression

In this section, we will give a brief overview of the mathematical formalism behind the linear regression model.

Regression is a statistical technique where both the dependent and independent variable takes continuous values and a model is fitted on the explanatory variables. Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data.

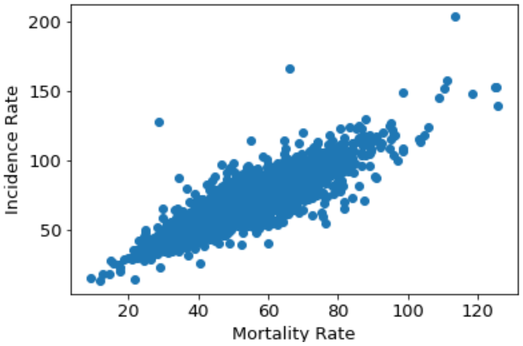


Fig. 10. Incidence v/s Mortality

It must be noted that in general, linear regression is used for creating a linear model of existing data and then using this to predict values for scenarios in which the target variables is unknown. However, in our particular use case, rather than attempting to do this, we will make the model and then use it on the same data; the idea is to see how well the model models the proposed relationship between the variables. A general mathematical description of linear regression is presented below.

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 ϵ — an unobserved random variable that adds ”noise” to the linear relationship between the dependent variable and regressors. Thus the model takes the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\mathsf{T}\boldsymbol\beta + \varepsilon_i,$$

for $i = 1, \ldots, n$ and where T denotes the transpose, so that $\mathbf{x_i^T}\boldsymbol\beta$ is the inner product between vectors $\mathbf{x_i}$ and $\boldsymbol\beta$.

Often these n equations are stacked together and written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol\beta + \boldsymbol\varepsilon,$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

The goal, in general ,is to compute the vector $\boldsymbol\beta$ when given $\mathbf{y}$ and $\mathbf{X}$. For the case of simple linear regression, also known as Ordinary Least Squares (OLS), there exists a closed form solution for the following optimization problem, in which we find $\hat{\boldsymbol\beta}$ such that:

$$\hat{\boldsymbol\beta} = \underset{\boldsymbol\beta}{\arg\min}\, ||\mathbf{X}\boldsymbol\beta - \mathbf{y}||^2$$

The closed form solution is given by:

$$\hat{\boldsymbol\beta} = (\mathbf{X^TX})^{\mathbf{-1}}\mathbf{X^Ty}$$

After computing a solution, it is necessary to have some sort of benchmark to measure the quality of the solution. In our use case, we would like to know how well the chosen parameter represents the target variable. Towards this, we make use of the following two metrics:

1) **Mean Squared Error (MSE)**: MSE is the mean of squares of the error terms, i.e., MSE is given by:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2$$

where y_i is the actual value and $\hat{y_i}$ is the predicted value. Note that this is an absolute metric, meaning the value of MSE cannot be compared across different datasets but can be used for comparison among different models of the same problem.

2) **Coefficient of determination** (R^2): R^2 is the proportion of the variation in the dependent variable that is predictable from the independent variable(s). It is computed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y_i})^2}{\sum_{i=1}^{N}(y_i - \bar{y_i})^2}$$

where y_i is the actual value, $\hat{y_i}$ is the predicted value and $\bar{y_i}$ is the sample mean given by:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Unlike MSE, R^2 is a relative measure and hence is somewhat data agnostic. Typically, R^2 values lies in between 0 and 1, with higher values signifying a better fit.

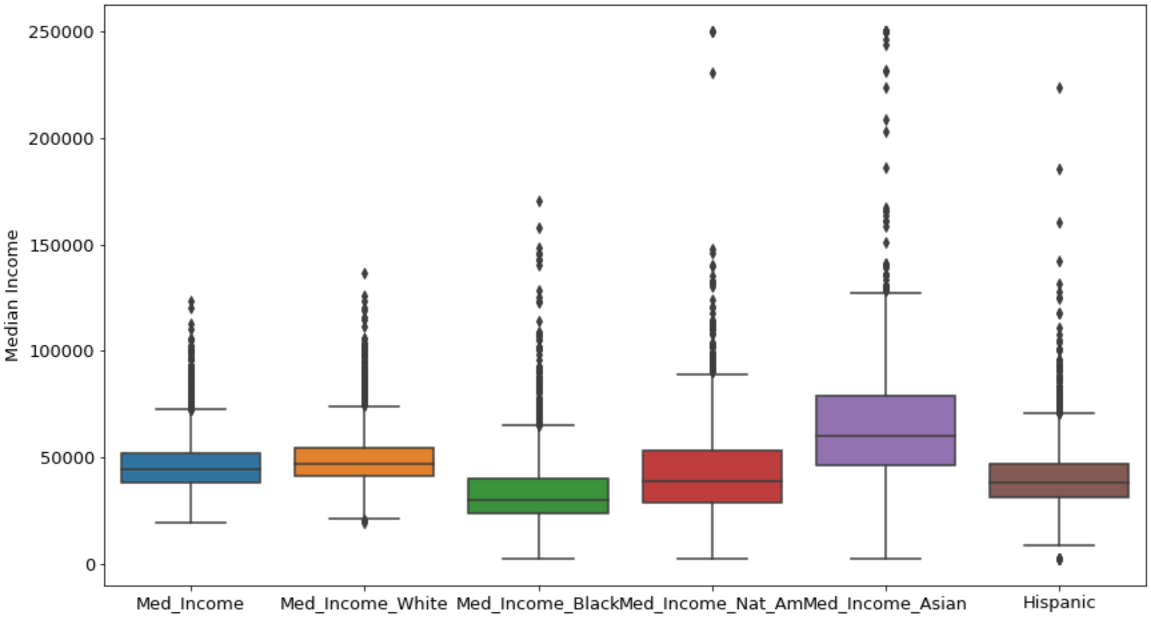
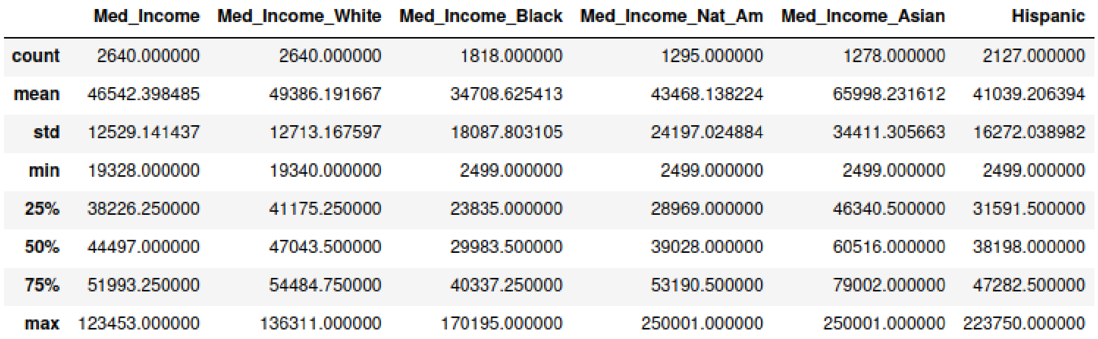


Fig. 11. Ethnicity v/s Median income



	Med_Income	Med_Income_White	Med_Income_Black	Med_Income_Nat_Am	Med_Income_Asian	Hispanic
count	2640.000000	2640.000000	1818.000000	1295.000000	1278.000000	2127.000000
mean	46542.398485	49386.191667	34708.625413	43468.138224	65998.231612	41039.206394
std	12529.141437	12713.167597	18087.803105	24197.024884	34411.305663	16272.038982
min	19328.000000	19340.000000	2499.000000	2499.000000	2499.000000	2499.000000
25%	38226.250000	41175.250000	23835.000000	28969.000000	46340.500000	31591.500000
50%	44497.000000	47043.500000	29983.500000	39028.000000	60516.000000	38198.000000
75%	51993.250000	54484.750000	40337.250000	53190.500000	79002.000000	47282.500000
max	123453.000000	136311.000000	170195.000000	250001.000000	250001.000000	223750.000000

Fig. 12. Statistics for income distribution

IV. Modelling

In this section, we discuss the application the linear regression model to our problem.

Our primary goal is to identify how cancer incidence and mortality rates are affected by income. More specifically, we would like to see if low-income groups are more prone to cancer and subsequent fatality. The two main parameters we can consider for analysing the income of a particular group are poverty rate and median income.

We employ linear regression keeping cancer incidence rate and mortality as the target variables. We consider the following four models, each taking different input variables:

1) Model 1: Poverty rate as the only input variable
2) Model 2: Median income as the only input variable
3) Model 3: Both poverty rate and median income as input variables
4) Model 4: Polynomial features generated from poverty rate and median income as input variables

A. Model 1

We obtain the plot shown in Figure 13 for incidence rate and plot shown in Figure 14 for mortality rate. The values of bench-marking parameters are summarised in Table I.

B. Model 2

We obtain the plot shown in Figure 15 for incidence rate and plot shown in Figure 16 for mortality rate. The values of

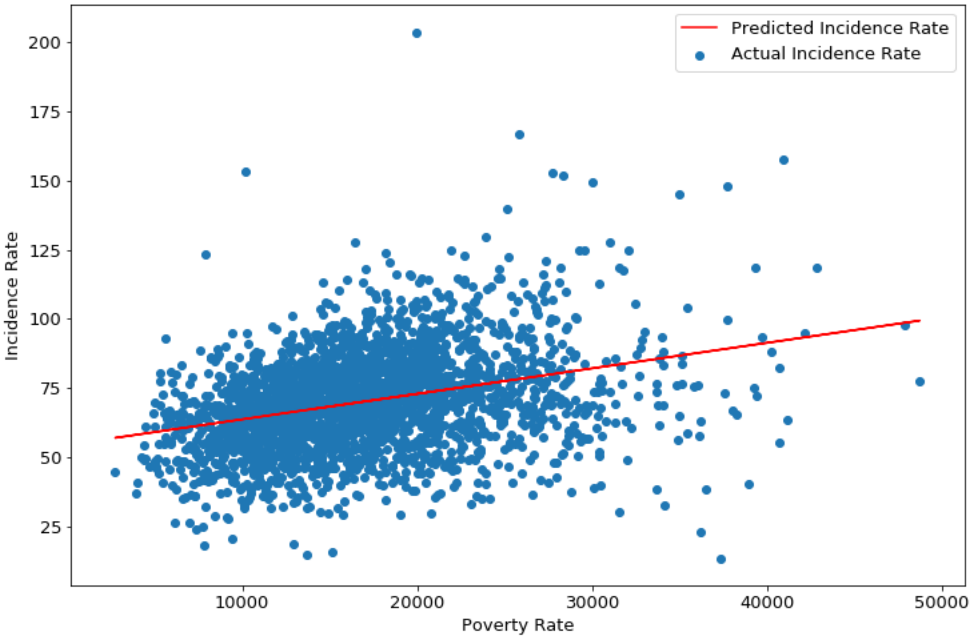


Fig. 13. Incidence prediction from poverty rate

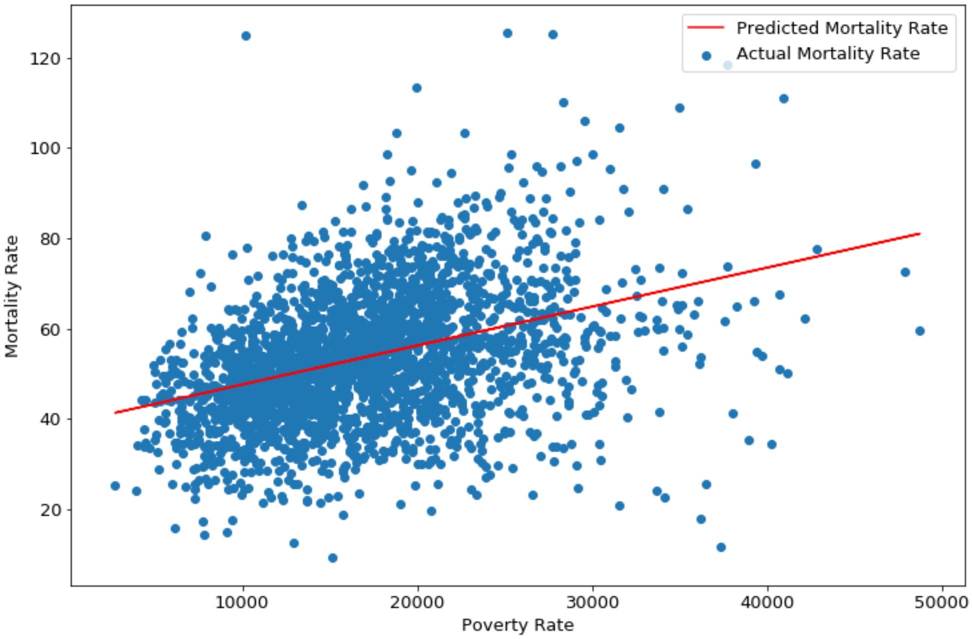


Fig. 14. Mortality prediction from poverty rate

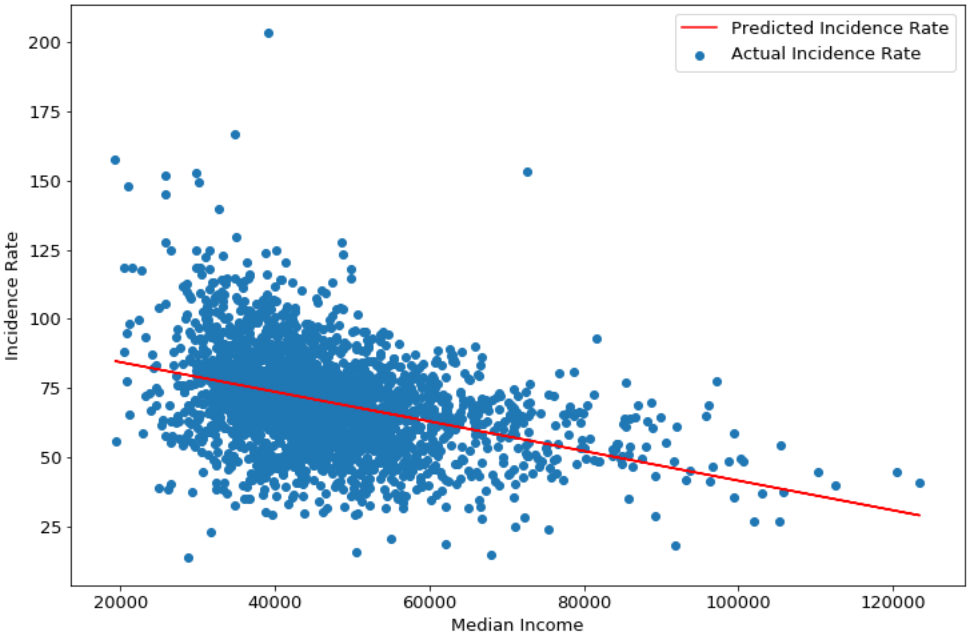


Fig. 15. Incidence prediction from median income

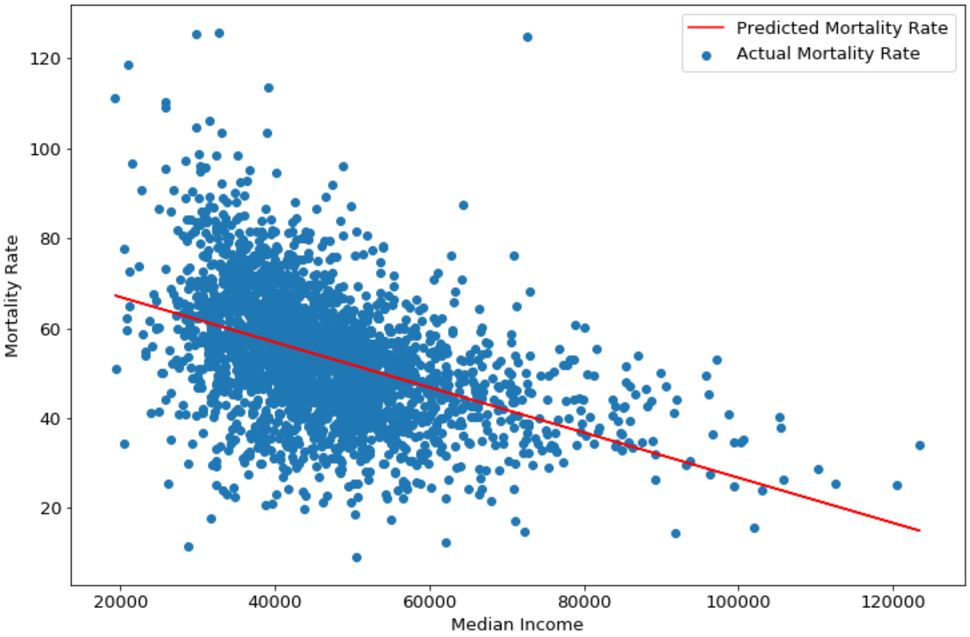


Fig. 16. Mortality prediction from median income

TABLE I
BENCHMARKS FOR MODEL 1 (ONLY POVERTY RATE USED)

Type	Incidence Rate	Mortality Rate
MSE	279.397	169.968
R^2	0.107	0.147

TABLE II
BENCHMARKS FOR MODEL 2 (ONLY MEDIAN INCOME USED)

Type	Incidence Rate	Mortality Rate
MSE	267.633	159.726
R^2	0.144	0.198

bench-marking parameters are summarised in Table II.

It can be seen that we obtain **better fit, in general by using median income instead of poverty rate.** This can be seen in terms of lower MSE values and higher R^2 scores.

C. Model 3

In this case, we give both poverty rate and median income as input paramters. Since, we have two input parameters, it is not possible to visualize the plot in 2D. We can still however obtain the bench-marking parameters. The values of bench-marking parameters are summarised in Table III.

It can be seen that we get a better fit than model 2, though only by a marginal amount, again **signifying that majority of the contribution is from the median income parameter.**

D. Model 4

In this case, we generate a polynomial of degree 6 from the combination of parameters poverty rate and median income. Again, since we have multiple input parameters, it is not possible to visualize the plot. We can still however obtain the bench-marking parameters. The values of bench-marking parameters are summarised in Table IV.

It can be seen that we get a **better fit than model 3, by a significant amount.** This makes sense intuitively because it is not necessary that cancer incidence/mortality is a linear function of poverty rate and median income, but rather can depend on higher powers of these parameters.

TABLE III
BENCHMARKS FOR MODEL 3 (INCOME + POVERTY)

Type	Incidence Rate	Mortality Rate
MSE	267.032	159.186
R^2	0.146	0.201

TABLE IV
BENCHMARKS FOR MODEL 3 (INCOME + POVERTY)

Type	Incidence Rate	Mortality Rate
MSE	249.674	146.292
R^2	0.202	0.266

V. CONCLUSIONS

From the four models we have built, it is **unambiguously clear that cancer incidence and mortality are very much impacted by the socioeconomic status.** By using a mathematical model, it has been formally proved that there exists a **positive correlation** between poverty rate and cancer incidence/mortality whereas there exists a negative correlation between median income and cancer incidence/mortality. The strength of the correlation has also been **quantified through bench-marking parameters, namely MSE and $\mathbf{R^2}$.**

It has also been inferred that **gender and ethnicity play a role in determining the economic status** of an individual as evidenced by the poverty rate and median income distributions. It has also been observed that individuals **not having a health insurance are at a greater risk of suffering death from cancer as compared to individuals who are insured.**

From the above listed conclusions, it is abundantly clear that **making health-related policies specifically directed at the low-income sections of the population** is the need of the hour. Focus should be on providing accessible and affordable health insurance as well as treatment to all sections of the population, irrespective of ethnicity and gender.

VI. AVENUES FOR FURTHER RESEARCH

Although inferences were made regarding how social status affects cancer incidence/mortality, since separate values for cancer incidence/mortality rates for different ethnic/gender sections were not available, direct predictions were not made. Rather, predictions were made indirectly using parameters like median income and poverty. If data can be collected separately for different ethnic/gender groups, a lot more insights can be drawn. Also, there are probably several other parameters which influence cancer incidence/ mortality. A detailed study of these parameters may also be a worthile avenue for further exploration.

REFERENCES

[1] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., 2nd ed., 2019.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[3] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[4] Wes McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference* (Stéfan van der Walt and Jarrod Millman, eds.), pp. 56 – 61, 2010.

[5] "Linear regression." https://en.wikipedia.org/wiki/Linear_regression.