A Mathematical Essay on Linear Regression

Aman Kumar
(EE21B013)
Dept. of Electrical Engineering
Indian Institute of Technology Madras
ee21b013@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 remains one of the most critical challenges confronting society today. Despite the availability of preventive measures and treatments, questions persist about the accessibility of these options for all individuals. Therefore, it is crucial to systematically evaluate how an individual's socioeconomic status influences their likelihood of developing or succumbing to cancer. This study aims to utilize data to examine how a person's cancer incidence and mortality rates are affected by their income and social standing.

With advancements in technology, we now have access to sophisticated data analytics tools that can help us extract meaningful insights from large datasets. By aggregating, cleaning, and analyzing vast amounts of data, we can uncover relationships between various factors. Linear Regression is one such analytical model that allows us to establish linear relationships between variables. In this study, Linear Regression is applied to explore how different factors impact cancer incidence and mortality within a population.

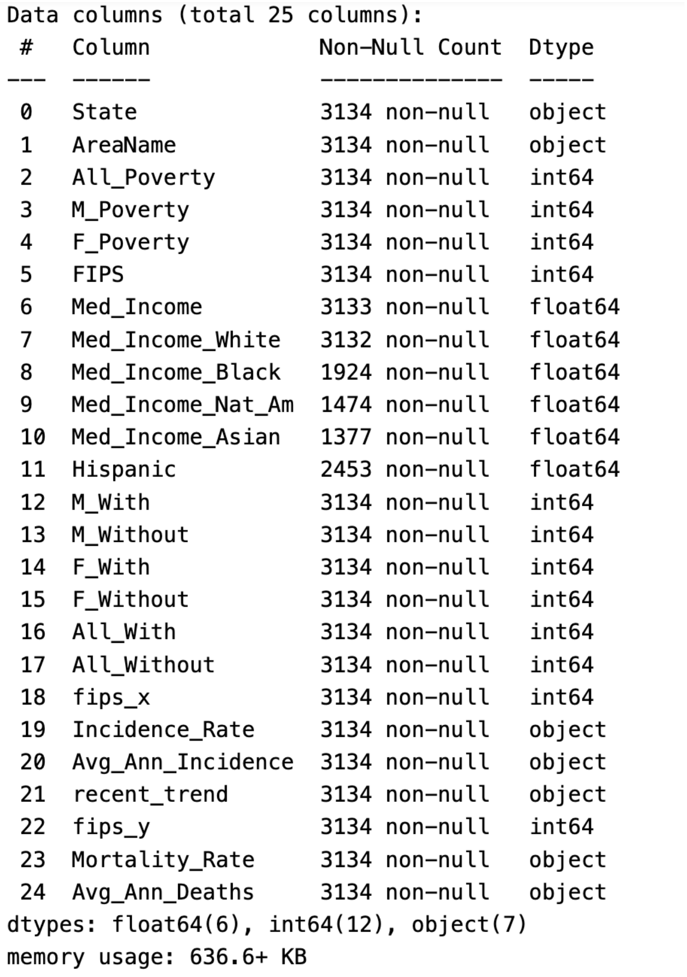
In this analysis, Python is employed to carry out the necessary data manipulations. Several Python libraries, including pandas, seaborn, and scikit-learn, are extensively utilized for data cleaning, visualization, and analysis. The Jupyter notebook environment is used to seamlessly integrate explanatory text alongside the code, facilitating a clear and concise presentation of the work.

After conducting a thorough analysis of the available data using linear regression, we conclude that an individual's socioeconomic status significantly affects their risk of cancer incidence and mortality. Specifically, we find a positive correlation between poverty rates and cancer outcomes, indicating that higher poverty rates are associated with increased cancer incidence and mortality. Conversely, there is a negative correlation between median income and cancer outcomes, suggesting that lower-income groups are at a higher risk of developing and dying from cancer. These findings underscore the need to address socioeconomic disparities to improve the health and well-being of all members of society.

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.



```
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 eliminate columns that are not relevant. Since our goal is to make inferences about the entire population rather than specific states or areas, distinguishing between states or areas is unnecessary. As a result, the columns State, AreaName, FIPS, fips x, and fips y become irrelevant, and we drop them.

Next, we notice that the columns Incidence Rate, Avg Ann Incidence, Mortality Rate, and Avg Ann Deaths are all classified as having an ”object” datatype. This suggests that not all values in these columns are numerical, which could pose a problem during mathematical analyses. Therefore, it is crucial to review these columns and attempt to convert all values to either numerical or categorical form.

Upon review, we discover that there are some non-numerical entries in each of these columns. The type of entry and the method used to convert it to a numerical value are discussed below:

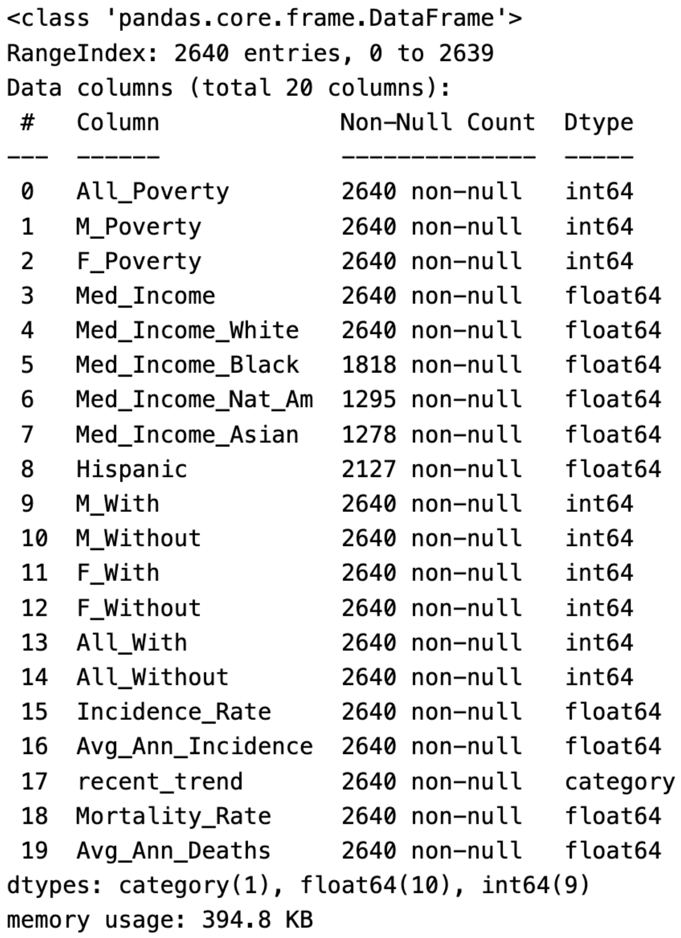
- * - This indicates that the true value is extremely low, so substituting these entries with 0 might seem appropriate. However, this approach results in a high concentration of 0s, which severely impacts the effectiveness of regression models. Consequently, these rows were excluded.
- Entries ending with # - This is just an error in formatting. The problem can be resolved by simply removing the # symbol.
- I removed the columns `Med_Income_Black`, `Med_Income_Nat_Am`, `Med_Income_Asian`, and `Hispanic` from the dataset because they contained a high proportion of missing values, with approximately 25% of the data being NULL.
- “and” - These indicate a 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 account 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 completing these operations, we are left with a cleaned dataset, summarized 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



```
<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

A. Economic Status

From the available dataset, we make the following observations:

- The variables we have with respect to 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.

Upon calculation, we observed that the correlation between `All_Poverty`, which represents the number of individuals below the poverty line, and the `Incidence_Rate` is -0.13—almost negligible! Does this imply that there is no correlation between the two? An important consideration here is that All_Poverty is the total number of individuals below the poverty line, meaning it has not been normalized relative to the total population, whereas the Incidence_Rate is already normalized. Therefore, it is essential to normalize All_Poverty with respect to the total population.

We approximate the total population as the sum of all individuals with and without health insurance. Additionally, we calculate the male and female populations separately. These quantities are added as separate columns, and we then compute the Poverty_Rate—the number of individuals below the poverty line normalized by the population—separately for males, females, and the entire population.

After calculating the poverty rates, we compute the correlation between the Poverty_Rate (across all individuals) and the Incidence_Rate and Mortality_Rate. We observe correlation values of 0.32 and 0.38, respectively, indicating a significant positive correlation. To visualize these findings, we present scatter plots for Incidence_Rate and Mortality_Rate against Poverty_Rate in Figure 3 and Figure 4.

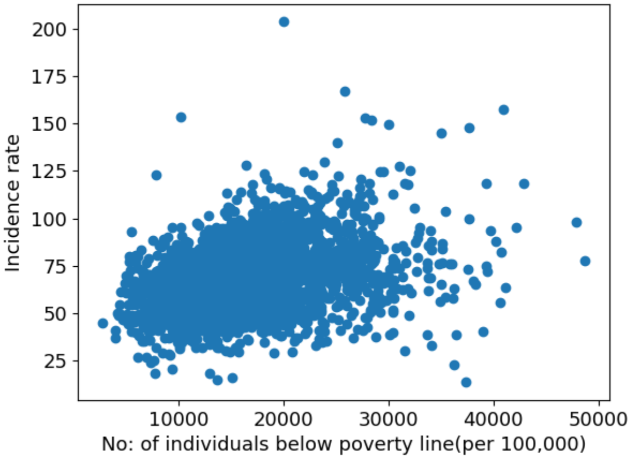


Fig. 3: Cancer incidence vs Poverty

From the plots, it is evident that cancer incidence and mortality generally increase with rising poverty levels.

Next, we examine the relationship between median income and cancer incidence and mortality rates. Upon calculation, we find correlation values of -0.37 and -0.44, respectively, indicating a significant negative correlation. To visualize these observations, we present scatter plots for Incidence_Rate and Mortality_Rate against Median_Income in Figure 5 and Figure 6.

From the plots, it is clear that **cancer incidence and mortality decrease, in general, with an increase in income**.

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

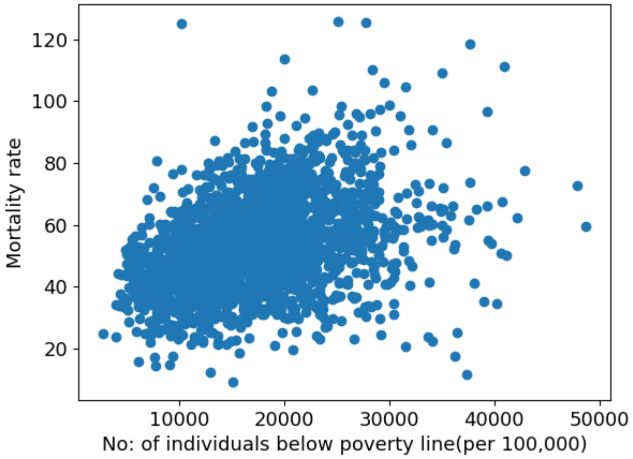


Fig. 4: Mortality vs Poverty

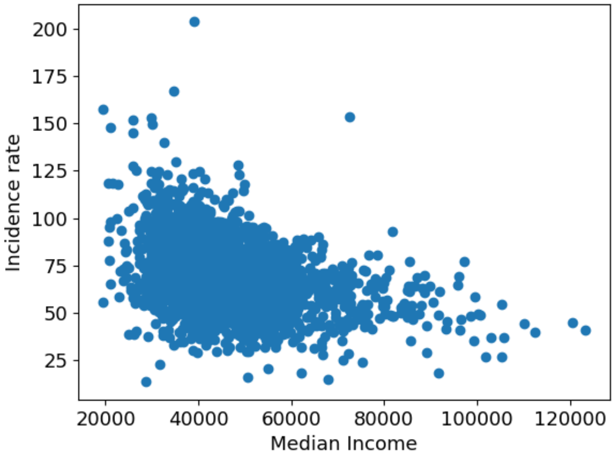


Fig. 5: Cancer incidence vs Income

We will now consider the impact of having health insurance on cancer incidence and mortality. On calculation, we observe correlation values of -0.038 and -0.124 for incidence and mortality, respectively. It appears that **there is minimal correlation between cancer incidence and having health insurance**. This is expected because health insurance is primarily useful for the treatment of a disease rather than its prevention. However, unlike the case of cancer incidence, there seems to be a **significant negative correlation between having health insurance and the mortality rate**. This is illustrated in Figure 7. This correlation is anticipated as having health insurance encourages individuals to seek appropriate treatment without financial constraints.

B. Social Status

Our goal is to associate social status with cancer incidence and mortality. To achieve this, we will consider the following two broad criteria for social classification:

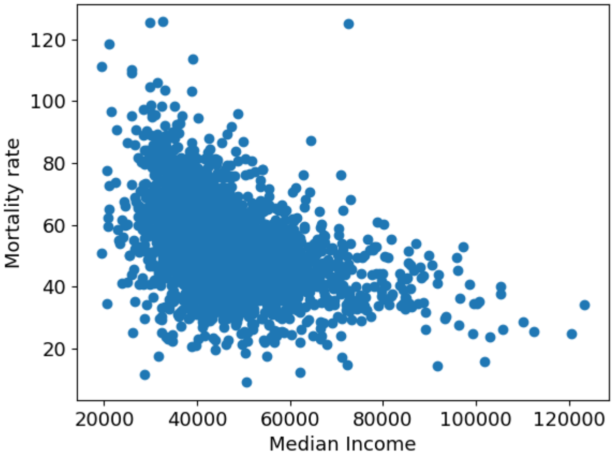


Fig. 6: Mortality vs Income

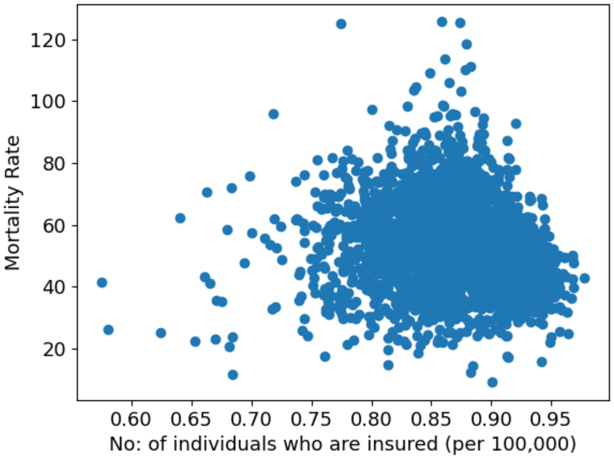


Fig. 7: Mortality vs Insurance

- Gender
- Ethnicity

The major challenge we face is that we do not have direct access to incidence and mortality rates for population subsections separately. For example, we lack data on the separate counts of cancer incidence for male and female individuals, which prevents any direct comparison. Consequently, we must rely on indirect comparisons based on the available parameters.

First, we consider gender. For gender, we analyze the poverty rates for males and females separately. Given our earlier conclusion that poverty rate is positively correlated with both incidence rate and mortality rate, if one gender group is observed to have a higher poverty rate, we can infer that this group likely experiences higher incidence and mortality rates as well. To compare the poverty rates between genders, we use a **box plot**, shown in Figure 8. The plot qualitatively indicates that females generally have a higher poverty rate. For a more quantitative assessment, refer to Figure 9. The **mean** and **median** poverty rates for females are clearly higher than those for males. Therefore, it can be estimated that the female population, in general, has a higher poverty rate and, consequently, higher **cancer incidence** and **mortality rates**.

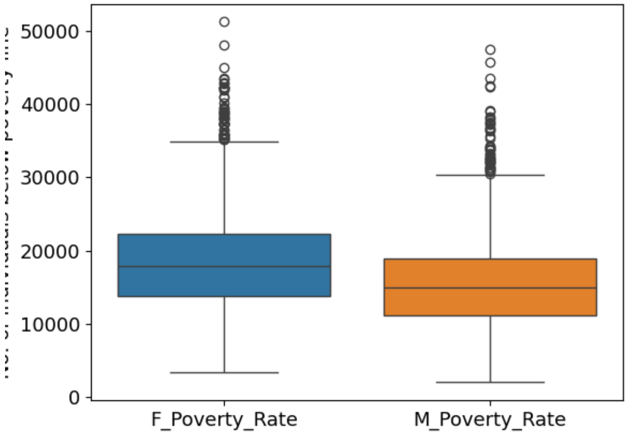
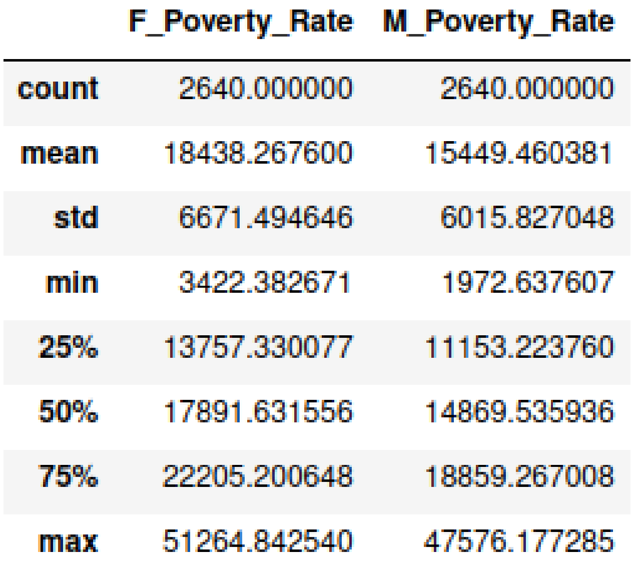


Fig. 8: gender v/s poverty



	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

It must be noted that in all the analyses conducted in this subsection, we have attempted to correlate **social status** with **economic indicators** such as **poverty** and **income**. Although this approach is insightful, it cannot guarantee a completely accurate analysis, as it lacks information regarding **incidence** and **mortality rates** for each subsection of the population separately.

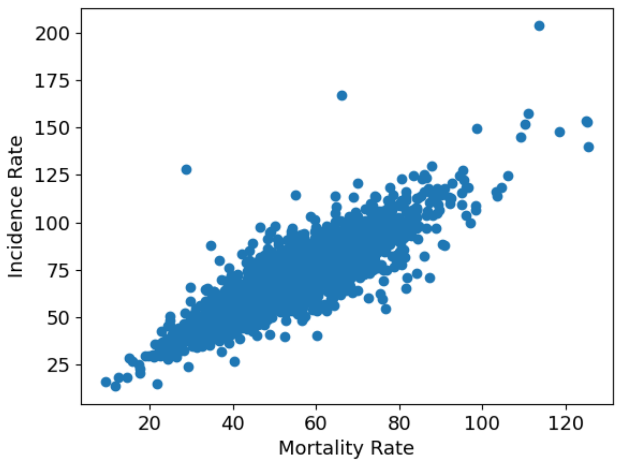


Fig. 10: Incidence v/s Mortality

C. *Incidence vs Mortality*

Finally, for the sake of completeness, we explore the relationship between **cancer incidence** and **mortality**. This relationship 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 noteworthy 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 provide a brief overview of the **mathematical formalism** behind the **linear regression model**.

Regression is a statistical technique where both the **dependent** and **independent variables** take continuous values, and a model is fitted to the **explanatory variables**. **Linear regression** is a linear approach for modeling 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.

It must be noted that, in general, linear regression is used to create a linear model of existing data and then use this model to predict values for scenarios where the **target variable** is unknown. However, in our specific use case, rather than predicting unknown values, we will **fit the model to the same data** to evaluate how well the model represents 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:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i = x_i^T \beta + \epsilon_i$$

for $i = 1, \ldots, n$, where T denotes the transpose, so that $x_i^T \beta$ is the inner product between vectors x_i and β. Often these n equations are stacked together and written in matrix notation as:

$$y = X\beta + \epsilon,$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

The goal, in general, is to compute the vector β when given y and 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 $\hat{\beta}$ such that:

$$\hat{\beta} = \text{argmin}_\beta \|X\beta - y\|^2$$

The closed-form solution is given by:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

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:

- **Mean Squared Error (MSE)**: MSE is the mean of the squares of the error terms, 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 MSE is an **absolute metric**, meaning its value cannot be compared across different datasets but can be used for comparison among different models of the same problem.
- **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 is somewhat data agnostic. Typically, R^2 values lie between 0 and 1, with higher values signifying a better fit.

IV. Modeling

V. Application of the Linear Regression Model

In this section, we discuss the application of 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 aim to determine if **low-income groups** are more prone to cancer and subsequent fatality. The two main parameters for analyzing the income of a particular group are **poverty rate** and **median income**.

We employ linear regression with **cancer incidence rate** and **mortality rate** as the target variables. We consider the following four models, each incorporating 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

B. Model 1

We obtain the plot shown in Figure 11 for **incidence rate** and the plot shown in Figure 12 for **mortality rate**. The values of **benchmarking parameters** are summarized in Table I.

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

C. Model 2

We obtain the plot shown in Figure 13 for **incidence rate** and the plot shown in Figure 14 for **mortality rate**. The values of **benchmarking parameters** are summarized in Table II.

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

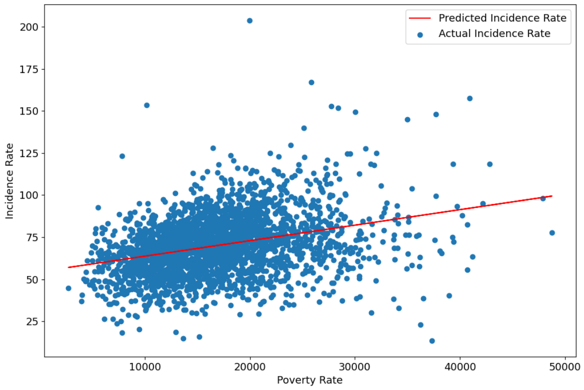


Fig. 11: Incidence prediction from poverty rate

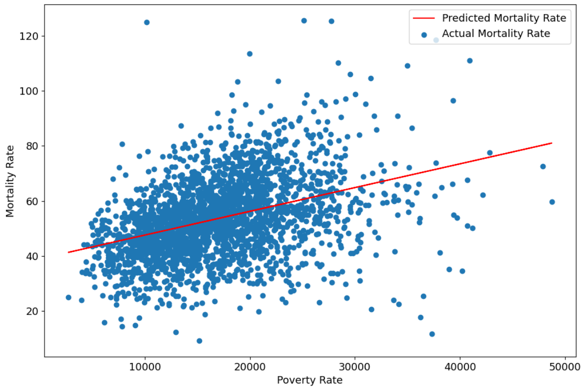


Fig. 12: Mortality prediction from poverty rate

D. Model 3

In this case, we use both **poverty rate** and **median income** as input parameters. Since we have **two input parameters**, it is not possible to visualize the plot in **2D**. We can still, however, obtain the **benchmarking parameters**. The values of these parameters are summarized in Table III.

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

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 **benchmarking parameters**. The values of these parameters are summarized in Table IV.

It can be seen that we achieve 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.

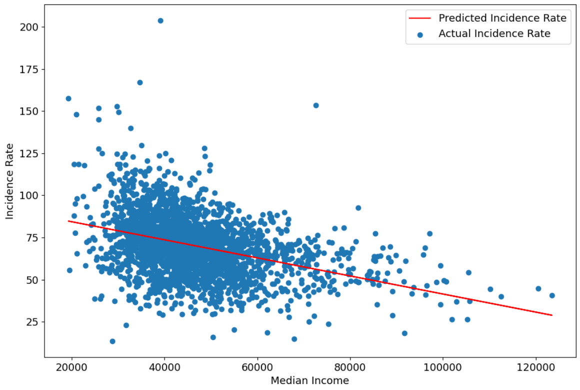


Fig. 13: Incidence prediction from median income

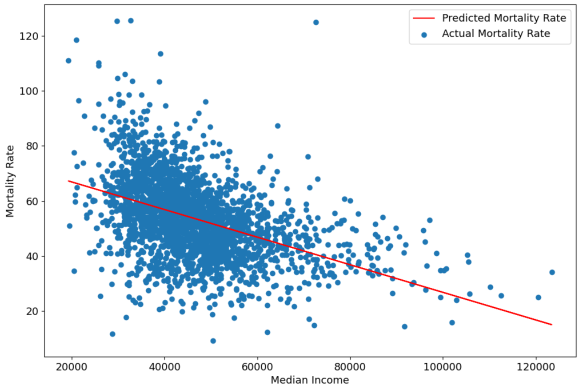


Fig. 14: Mortality prediction from median income

VI. Conclusions

VII. Conclusions

From the four models we have built, it is **unambiguously clear** that **cancer incidence** and **mortality** are significantly impacted by **socioeconomic status**. By using a **mathematical model**, it has been formally proven that there exists a **positive correlation** between **poverty rate** and cancer incidence/mortality, whereas there is a **negative correlation** between **median income** and cancer incidence/mortality. The strength of these correlations has also been quantified through benchmarking parameters, namely **MSE** and **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. Additionally, it has been observed that individuals without **health insurance** are at a higher risk of suffering death from cancer compared to individuals who are insured.

From the above conclusions, it is **abundantly clear** that there is an urgent need to develop health-related policies specifically targeted at the **low-income sections** of the population. The focus should be on providing **accessible and affordable health insurance** and treatment to all sections of the population, irrespective of **ethnicity** and **gender**.

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

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 4 (Income + Poverty Polynomial)

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

VIII. 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 that influence cancer incidence/mortality. A detailed study of these parameters may also be a worthwhile 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 et al., "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 (Stefan van der Walt and Jarrod Millman, eds.)*, pp. 56 – 61, 2010.

[5] "Linear regression." [Online]. Available: https://en.wikipedia.org/wiki/Linear$_{r}egression$