DAL 2023-ASSIGNMENT 1

Kaja Vikas
Departmentof Engineering Design
Indian Institute of Technology, Madras
Chennai, India
ed20b026@smail.iitm.ac.in

***Abstract*—This paper presents a comprehensive analysis of a global dataset that encompasses socioeconomic indicators, health insurance coverage data as well as cancer incidence and mortality rates in diverse regions and populations. To extract meaningful insights from this dataset, we utilize the data analysis technique of Linear Regression. Prior to applying the model, we employ data visualization techniques to gain an understanding of the dataset's characteristics and trends. Moreover, we intricately explore the pivotal process of feature selection, an essential step in preparing the dataset for the subsequent application of the Linear Regression model. We also delve into the mathematical aspects and applications of linear regression. Upon establishing the Linear Regression model, we conducted a dataset analysis, yielding valuable insights that illuminate the correlations between socioeconomic factors and cancer rates. These results offer essential contributions to the understanding of this critical public health issue and provide a foundation for future research and policymaking in this domain.**

I. Introduction

Cancer is a formidable global health challenge, affecting millions of lives each year. Despite significant advancements in medical science and treatments, the prevalence of cancer remains a pressing concern for healthcare professionals, policymakers, and society as a whole. The burden of cancer is not distributed uniformly, and disparities in cancer incidence and mortality rates have been observed across different regions and demographic groups. To effectively combat cancer and reduce these disparities, it is imperative to understand the intricate relationship between cancer rates and socioeconomic factors. The dataset at hand, comprising various socioeconomic and cancer-related variables, provides a valuable resource for investigating this relationship. It includes information on poverty rates, income levels, demographics, and cancer-related statistics across different areas. This dataset is a rich source of information that allows us to explore how socioeconomic factors might influence cancer incidence and mortality rates. By utilizing linear regression analysis, we aim to uncover visual and quantitative relationships between these variables, shedding light on the role of income, poverty, and demographic factors in cancer outcomes.

Linear regression is a powerful statistical technique that serves as a fundamental tool in data analysis. At its core, linear regression aims to establish a clear and quantifiable relationship between independent variables (predictors or features) and a dependent variable (target variable). It operates under the fundamental assumption that this relationship is linear in nature, meaning that changes in the independent variables are directly proportional to changes in the dependent variable. By modeling the linear associations between various socioeconomic factors (independent variables) and cancer incidence or mortality rates (dependent variables), we can quantitatively assess the impact of these factors on cancer outcomes. This method enables us to uncover valuable insights into how changes in income, poverty rates, and other socioeconomic variables relate to changes in cancer rates.

In deploying our linear regression model to investigate the correlation between socioeconomic factors and cancer rates, we have undertaken several critical steps to ensure a robust and effective analysis. Our initial step involved the preparation of the dataset, where we focused on data cleaning. During this phase, we meticulously handled missing values, eliminated unnecessary columns, and ensured data consistency. This meticulous data preparation process was crucial in ensuring that our dataset was reliable and primed for analysis. To identify the most influential predictors of cancer rates, we leveraged correlation matrices and data visualization techniques. These powerful tools allowed us to gain valuable insights into the interplay between various socioeconomic factors and cancer incidence/mortality rates. Through the correlation matrix, we visualized the relationships between different variables, providing us with an initial understanding of which factors might be significant drivers of cancer rates. Subsequently, with our linear regression model deployed on carefully selected features, we unveiled the significance of specific attributes in shaping cancer rates, particularly in relation to Incidence Rate (IR) and Mortality Rate (MR). By visualizing the importance of these features through bar graphs, we transformed abstract coefficients into tangible insights, shedding light on the driving forces behind cancer incidence and mortality.

Upon uncovering the intricate relationships between poverty rates, health insurance coverage, ethnicity, median incomes, and their influence on cancer incidence and mortality rates, a range of strategic actions and applications can be employed to address public health challenges. The results can guide the formulation of public health policies. Resource allocation can be optimized by prioritizing areas with higher cancer rates and specific socioeconomic risk factors for additional funding, screening programs, and support services. Moreover, health education campaigns can be designed to cater to communities with distinct risk profiles, ensuring that awareness and preventive measures are culturally sensitive and effective. The findings can also guide future research endeavors to delve

deeper into the mechanisms underpinning these relationships, potentially yielding novel insights for cancer prevention and treatment. Equity initiatives can be advanced to address health disparities, targeting social determinants of health, improving healthcare access, and advocating policies to alleviate poverty and bolster health insurance coverage.

In this section of Introduction, we outline the structure and contents of the subsequent sections. Section 2 delves into the details of linear regression, providing a comprehensive explanation of the algorithm. Section 3 focuses on Dataset Handling and exploration, elucidating the procedures and techniques employed in data preprocessing and exploration. Section 4 illustrates the application of linear regression on the dataset. Section 5 is the conclusion and future work, summarizes the findings and discusses potential directions for future research. Finally, Section 6 lists the references used throughout the paper.

II. Linear Regression in Detail

Linear regression is a widely used statistical method for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors or features).

A. Target Variable and Independent Variables

1) Target Variable (Dependent Variable): The target variable is the main focus of the analysis. In our study, it's represented by cancer incidence and mortality rates. These rates quantify the prevalence and impact of cancer in various contexts.

2) Independent Variables (Predictors or Features): Independent variables are the factors under investigation, such as poverty rates, health insurance coverage, ethnicity, and median incomes. They are used to explain changes in the target variable. In our analysis, these independent variables serve as inputs to our linear regression models.

B. Principles of Linear Regression

Linear regression is a fundamental statistical technique used to model the relationship between a dependent variable (target) and one or more independent variables (predictors or features). It relies on the following mathematical expression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

Here:

- Y is the dependent variable, representing the outcome we seek to predict (e.g., cancer incidence or mortality rate).
- $X_1, X_2, \ldots, X_n$ are the independent variables, which can be numeric or categorical factors (e.g., poverty rates, health insurance coverage, ethnicity, median incomes).
- β_0 is the intercept, and $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients, representing the impact of each independent variable on the dependent variable.
- ϵ denotes the error term, capturing unexplained variability.

C. Assumptions of Linear Regression

Linear regression relies on critical assumptions, including linearity (the relationship is linear), independence of errors (residuals are not correlated), homoscedasticity (constant variance of residuals), and normally distributed errors.

D. Objective Function in Linear Regression

The objective function in linear regression is to minimize the sum of squared differences between predicted and actual values, known as the Least Squares Objective Function:

$$\text{Minimize } \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{ni}))^2$$

This function guides model training to determine the β coefficients that yield the best-fit linear equation.

E. Limitations of Linear Regression

Linear regression has limitations, including its assumption of linearity, sensitivity to outliers, multicollinearity concerns, and limited ability to capture complex non-linear relationships. These limitations should be considered when interpreting results and deciding if linear regression is appropriate for a given analysis.

III. Dataset Handling and Exploration

In this study, our primary objective is to explore the correlation between socioeconomic factors and cancer incidence and mortality rates. To achieve this aim, we have thoughtfully selected a subset of columns from the given dataset for further analysis and data cleaning. The chosen columns are as follows: *All_Poverty*, *M_Poverty*, *F_Poverty*, *Med_Income*, *Med_Income_White*, *Med_Income_Black*, *Med_Income_Nat_Am*, *Med_Income_Asian*, *M_With*, *M_Without*, *F_With*, *F_Without*, *All_With*, *All_Without*, *Incidence_Rate*, and *Mortality_Rate*.

These columns have been selected based on their relevance to our problem statement. *All_Poverty*, *M_Poverty*, and *F_Poverty* represent poverty rates, which are important socioeconomic factors that can impact cancer outcomes. *Med_Income*, *Med_Income_White*, *Med_Income_Black*, *Med_Income_Nat_Am*, and *Med_Income_Asian* provide information on median incomes among different racial and ethnic groups, which can help us investigate disparities in cancer rates. *M_With*, *M_Without*, *F_With*, *F_Without*, *All_With*, and *All_Without* represent health insurance coverage, which is another critical factor affecting cancer detection and treatment.

Furthermore, *Incidence_Rate* and *Mortality_Rate* are our primary outcome variables, reflecting cancer incidence and mortality rates, respectively. These columns are essential for assessing the impact of socioeconomic factors on cancer outcomes.

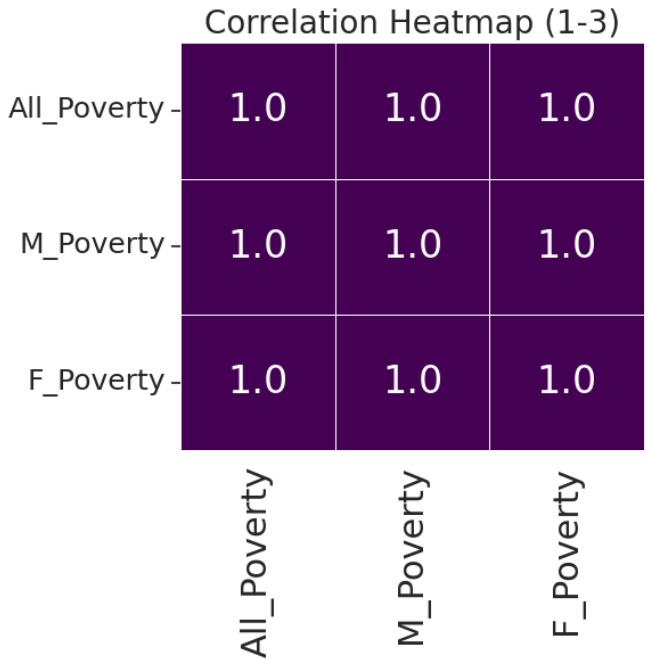
In the data cleaning process, we addressed several issues to ensure the quality and reliability of our dataset. First, we identified and removed any entries where *Incidence_Rate* had

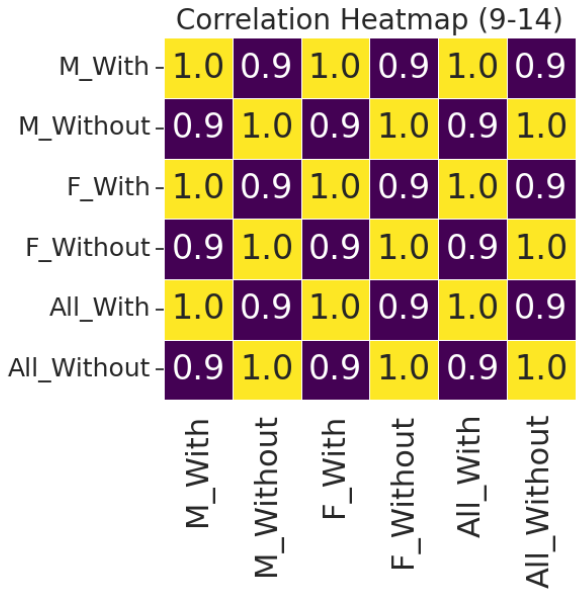
a '#' character at the end, converting these values to floating-point numbers. Additionally, we replaced certain characters ('_', '__', '*') with NaN values in the entire dataset, making it consistent and more suitable for analysis. Entries with '*' in *Incidence_Rate* and *Mortality_Rate* were replaced with a value of 16, where '*' represents fewer than 16 cases. Since 16 is nearly smaller compared to the mean of incident rate and mortality rate, we can consider assigning 16 to the '*' in incident rate and mortality rate columns. Finally, we converted all the data into float values.

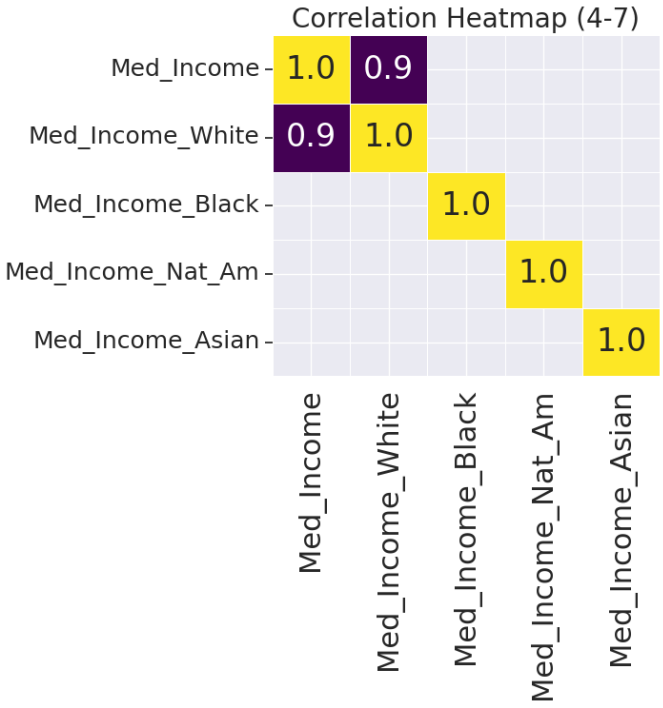
To handle missing values, we imputed them with the respective column means. This approach ensures that our dataset is complete and ready for further statistical analysis. Deleting rows with missing values can result in a significant loss of data, especially when there are missing values in multiple columns. By using the column mean as a replacement for missing values, you ensure that the central tendency of the data is maintained. This approach prevents distortion of the data distribution, which can happen if you use a single constant value (e.g., 0) for all missing values.

These data cleaning steps were crucial to obtain reliable results in our investigation of the correlation between socioeconomic factors and cancer incidence and mortality rates.

A correlation matrix is a statistical tool used to quantify and summarize the degree of linear relationship or association between pairs of variables within a dataset. It is a square matrix where the diagonal elements represent the correlation of each variable with itself (which is always 1), and the off-diagonal elements represent the correlations between pairs of variables. Based on the observed correlations, we have identified key features for inclusion in our linear regression model. Specifically, *All Poverty*, *M Poverty*, and *F Poverty* exhibit strong positive correlations among themselves, and as such, we have chosen *All Poverty* as a representative feature for our model.

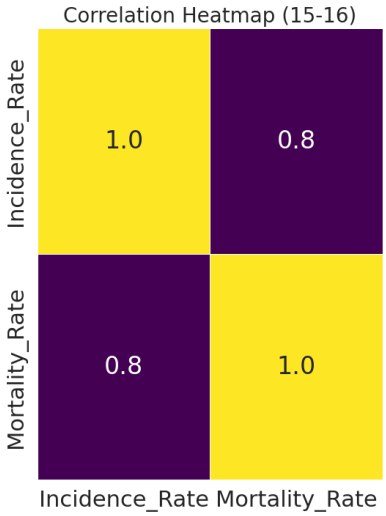






Also, the variables related to health insurance coverage, namely *M With*, *F With*, *F Without*, *All with*, and *All Without*, display strong positive correlations. Consequently, we have selected *All with ins* and *All Without Ins* to represent this group of features in our linear regression model.

Regarding median income across different racial groups, such as *Med Income*, *Med Income White*, *Med Income Black*, *Med Income Nat Am*, and *Med Income Asian*, we have observed that they do not exhibit high correlations with each



other. Hence, we consider all of these variables as individual features for our linear regression model.

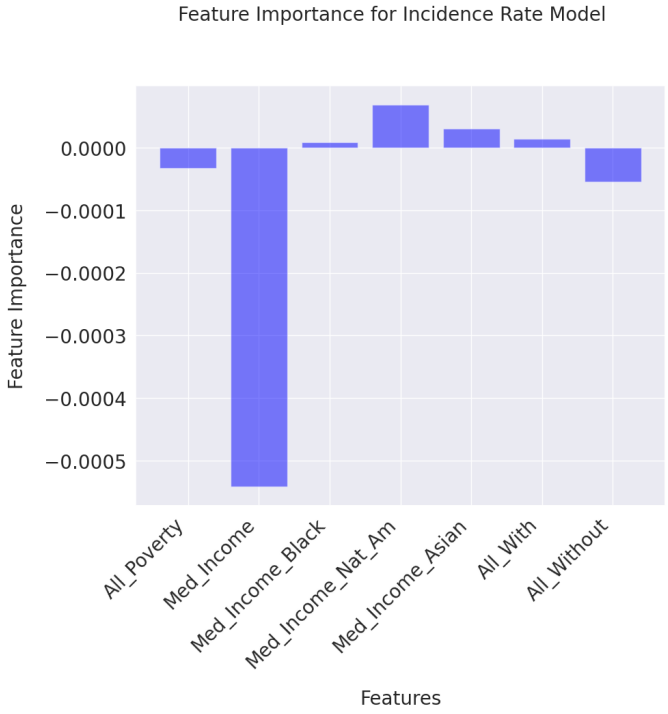
Furthermore, it is worth noting that *Incident Rate* and *Mortality Rate* demonstrate a correlation with each other, as evident from the correlation matrices. This selection of features is based on their respective correlations, and it aims to ensure that our model captures the relevant socioeconomic factors and health-related variables that may influence cancer incidence and mortality rates.

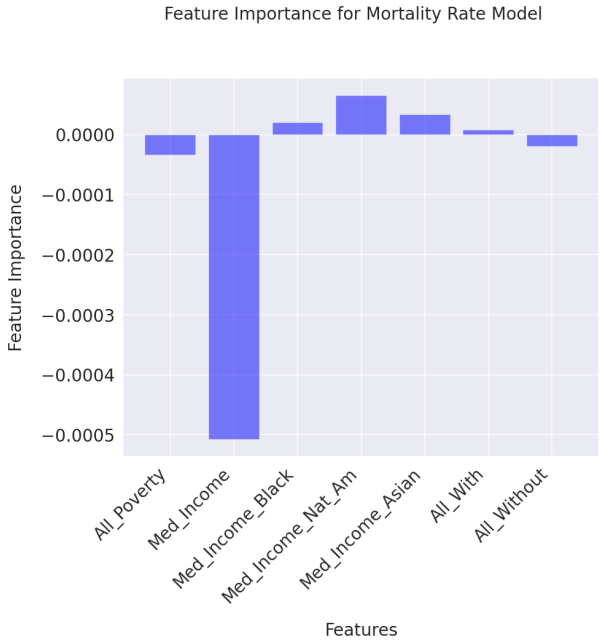
IV. Applying Linear Regression

In this analysis, we applied linear regression models to investigate the relationship between socioeconomic factors and cancer incidence and mortality rates. Two distinct linear regression models were created, one for predicting Incidence Rate (IR) and another for predicting Mortality Rate (MR). We used the scikit-learn library in Python, a powerful tool for machine learning and data analysis. Specifically, we imported the Linear Regression class for building the regression models, train_test_split for dividing our dataset into training and testing sets, and mean_squared_error for evaluating model performance.

To begin, we split our data into training and testing sets, allocating 70% for training and 30% for testing. This separation ensures that our models are trained on one subset and tested on another, allowing us to assess their generalization capabilities. After setting up our data, we created two distinct linear regression models: one aimed at predicting Incidence Rate (IR) and the other focused on predicting Mortality Rate (MR).

The evaluation of the models is crucial to assess their accuracy. We employed the mean squared error (MSE) as the evaluation metric for both models. The MSE measures the average squared difference between predicted values and actual values. Lower MSE values indicate better model performance. For our specific models, the Mean Squared Error for the Incidence Rate model was approximately 15.03, and for the Mortality Rate model, it was approximately 11.76. These values serve as indicators of how well our linear regression models approximate the observed cancer incidence and mortality rates based on the selected socioeconomic features. In summary, this application of linear regression models using scikit-learn libraries provides valuable insights into understanding the impact of socioeconomic factors on cancer incidence and mortality rates, aiding in the identification of significant predictors in the context of public health and healthcare policy.





These bar graphs visually represent the feature importance for our linear regression models predicting Incidence Rate (IR)

and Mortality Rate (MR). Each bar corresponds to a specific feature from the dataset, and the height of the bar indicates the magnitude of the coefficient assigned to that feature in the respective model. Features with taller bars are considered more influential in predicting the target variable. so from bar graphs median income has largest influence on incident rate and mortality rate. Both male and female below poverty line also has some influence on incident rate and mortality rate. people without medical insurance also had slight affect on incident rate and mortality rate.

V. Conclusions and Future Work

In conclusion, our analysis using linear regression models has revealed valuable insights into the factors influencing cancer incidence rate (IR) and mortality rate (MR). Median income, especially across diverse racial and ethnic groups, emerged as a pivotal factor significantly impacting both IR and MR. Moreover, poverty rates among males and females below the poverty line also play a substantial role in predicting cancer outcomes. Additionally, the presence of medical insurance appears to have a modest yet noteworthy influence on IR and MR. These findings emphasize the importance of addressing socioeconomic disparities and healthcare access to improve cancer prevention and treatment outcomes. In the future, we aim to expand our research beyond linear regression, exploring more complex models such as multinomial regression to unveil intricate relationships and patterns within the dataset. By considering a broader range of variables and adopting advanced modeling techniques, we strive to enhance our predictive capabilities and gain a deeper understanding of the multifaceted factors affecting cancer incidence and mortality rates. This research paves the way for comprehensive interventions and policies aimed at reducing disparities and fostering equitable cancer care for all.

Future work entails not only the exploration of multinomial regression but also the implementation of machine learning algorithms such as decision trees, random forests, and support vector machines. These models can uncover non-linear relationships and intricate interactions among variables, providing a more holistic understanding of cancer outcomes. Moreover, geographical analysis and spatial modeling techniques can be employed to assess regional disparities and their impact on cancer incidence and mortality. Additionally, the integration of lifestyle and behavioral factors into predictive models can further refine our ability to identify high-risk populations and tailor preventive strategies. Overall, our research opens avenues for more sophisticated analyses and data-driven interventions that can contribute to the advancement of public health initiatives and the reduction of cancer disparities on a broader scale.

References

[1] https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/

[2] https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.

[3] https://www.geeksforgeeks.org/ml-linear-regression/