Data Analytics Lab: Assignment-1
A Mathematical Essay on Linear Regression

Arjav Singh
Metallurgical and Materials Engineering
Indian Institute of Technology Madras
Chennai, India
mm20b007@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 between cancer incidence, mortality, and socioeconomic status and provide quantitative and visual evidence.**

***Index Terms*—Introduction, Linear Regression, Problem, Conclusion**

I. Introduction

Cancer survival disparities linked to socioeconomic groups are well-established. A 1997 report from the International Agency for Research on Cancer (IARC) highlighted that lower socioeconomic status (SES) is associated with higher cancer incidence and worse survival rates in developed and less-developed countries. Our study employs data from the CDC's National Program of Cancer Registries Cancer Surveillance System (NPCR-CSS) and SEER Program's Incidence data to investigate these relationships using linear regression models.

Linear regression, a technique to model the relationship between variables, fits a linear equation to observed data. It designates certain variables as explanatory and one as the dependent variable. The prevalent least-squares method computes the best-fitting line by minimizing the sum of squared deviations between data points and the line.

Here, we utilize linear regression to investigate links between cancer incidence, mortality rates, socioeconomic factors, and race. We initially gather, refine, and prepare data, followed by exploratory analysis. Subsequently, we construct statistical models and visualizations to offer quantitative and visual confirmation of our findings. The subsequent section outlines fundamental Linear Regression principles, while section 3 presents insights from data and models. Finally, section 4 summarizes the study's key aspects and suggests potential future research directions.

II. Linear Regression

Linear regression is a statistical method of estimating the relationship between a scalar response and independent and dependent variables (also called dependent variables 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 form of analysis estimates the coefficients of the linear equation involving one or more independent variables that best predict the value of the dependent variable. Linear Regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.

Linear Regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Because linear regression is a long-established statistical procedure, the properties of linear regression models are well understood and can be trained very quickly.

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

- Linear regression is used for prediction by fitting a model to observed data, enabling future response predictions even when explanatory variables change.
- Linear regression explains how the response variable's variation relates to explanatory variables, identifying strengths and redundancies in their relationships.

A. Formulation

Given a dataset $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n $\mathbf{x}$ p size, a linear regression model assumes a linear relationship between the dependent variable y and the vector of regressors $\mathbf{x}$. This relationship is modeled through a disturbance term or error variable ε — an unobserved random variable that adds "noise" to the dependent and regressors' linear relationship. Thus, the model takes the form:

$$y_i = \theta_0 + \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\theta} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where T denotes the transpose so that $\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\theta}$ is the inner product between vectors $\mathbf{x}_i$ and $\boldsymbol{\theta}$. Often, these n equations are stacked together and written in matrix notation as:

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

where $\mathbf{y}$ is a column vector, $\mathbf{X}$ is a matrix, $\boldsymbol{\theta}$ is a column vector of coefficients, and $\boldsymbol{\varepsilon}$ is a column vector of errors.

The following are the major assumptions made by standard linear regression models with standard estimation techniques:

- **Weak exogeneity**. This basically means that predictor variables $\mathbf{X}$ 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/homoscedasticity**. This implies that the variability or variance of the errors is independent of predictor variable values. Consequently, the response variability remains consistent across differing response magnitudes for fixed predictor values.
- **Independence of errors**. This assumes that the errors of the response variables are uncorrelated with each other.

B. Cost Function

A cost function measures how wrong the model is in terms of its ability to estimate the relation between inputs and outputs. Different types of cost functions exist, and the most popular among them is the Squared Error between the predicted and observed outputs. The formula for the mean squared error for a dataset with N samples is given by:

$$J(\theta) = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

where $\hat{y}_i$ is the predicted output, and y_i is the observed or given output. The objective is to find θ that minimizes the cost function J.

C. Parameter estimation using Least Square

Many 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, etc. One of the most common methods is the Least Square Estimation.

Assuming that the independent variable is $\vec{\mathbf{x}_i} = [x_{i1}, x_{i2}, \ldots, x_{im}]$, and the model's parameters are $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots, \theta_m]$, then the model's prediction would be:

$$y_i \approx \theta_0 + \sum_{j=1}^{m} \theta_j \times x_{ij}$$

If $\mathbf{x}_i$ is extended to $\vec{\mathbf{x}}_i = [1, x_{i1}, x_{i2}, \ldots, x_{im}]$, then y_i would become a dot product of the parameter and the independent variable, i.e.

$$y_i \approx \sum_{j=1}^{m} \theta_j \times x_{ij} \approx \boldsymbol{\theta} \cdot \vec{\mathbf{x}}_i$$

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

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} L(D, \boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (\boldsymbol{\theta} \cdot \vec{\mathbf{x}}_i - y_i)^2$$

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

$$L(D, \boldsymbol{\theta}) = \|\mathbf{X}\boldsymbol{\theta} - \mathbf{Y}\|^2$$

$$= (\mathbf{X}\boldsymbol{\theta} - \mathbf{Y})^T (\mathbf{X}\boldsymbol{\theta} - \mathbf{Y})$$

$$= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}$$

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

$$\frac{\partial L(D, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}$$

Setting the gradient to zero produces the optimum parameter:

$$-2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} = 0$$

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\theta}$$

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

D. Metrics for model evaluation

1) **R-squared value**: This value ranges from 0 to 1. Value '1' indicates that the predictor perfectly accounts for all the variation in Y. Value '0' indicates that predictor 'x' accounts for no variation in 'y'.
   - **Regression sum of squares (SSR)**. This gives information about how far the estimated regression line is from the average of the actual output.
   $$Error = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2$$
   - **Sum of Squared error (SSE)**. How much the data varies around the regression line (predicted values).
   $$Error = \sum_{i=1}^{n} (y - \hat{y})^2$$
   - **Total sum of squares (TSS)**. This tells how much the data point moves around the mean.
   $$Error = \sum_{i=1}^{n} (y - \bar{y})^2$$
   $$R^2 = 1 - \frac{SSE}{SSTO}$$
2) **Null-Hypothesis and P-value**: The p-value attributed to each term assesses the null hypothesis, implying the coefficient's equality to zero, thus indicating no impact. A p-value below 0.05 indicates grounds for rejecting the null hypothesis. In simpler terms, a predictor with a low p-value is likely to significantly enhance your model, as alterations in the predictor correspond to shifts in the response variable. On the contrary, a higher (insignificant) p-value implies a lack of connection between predictor alterations and response variations.

III. Problem

We are faced with a task involving testing the hypothesis that cancer incidence and mortality are linked to socioeconomic status. The metrics for socioeconomic status are provided through Poverty, Income, and Insurance data. The dependent variables for testing the hypothesis encompass Incidence rate, Average incidence rate, mortality rate, and average death rate. For this, we will demonstrate how well cancer incidence and mortality correlate with socioeconomic status by providing quantitative and visual evidence.

A. Data pre-processing

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:

- State: State of the respective area, total 51.
- Area: Name of area in which sampling is done.
- All Poverty: Number of people of both genders below the poverty line. Similarly, M poverty is for males, and F Poverty is for females.
- FIPS: Zipcode of the area.
- Med Income: Median Income of all ethnic groups in the area.
- All With: Number of individuals having insurance in the area; Along these lines, All Without, Without Male, With Male, Without Female, With Female are defined.
- Incidence Rate: Number of cancer cases detected per 100,000 people in the area.
- Mortality Rate: Number of mortalities per 100,000 people

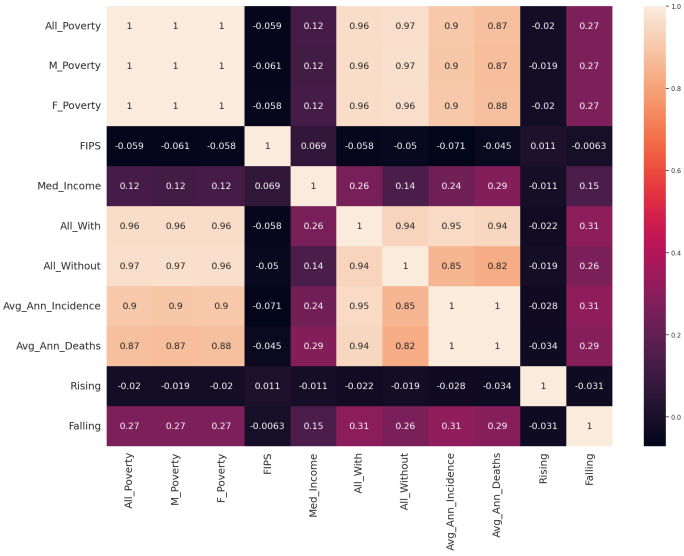

Fig. 1. Pearson Correlation Coefficient among all numeric features.

In Figure 1, 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 the median income increases, and mortality rates decrease.

We begin by examining the impact of social status on median income, focusing on the only available social data. We create a distribution plot (Figure 2) illustrating the median incomes of various communities in different states. Preliminary observations reveal a trend where Asians tend to have higher incomes, while incomes for Black individuals are generally lower. These distinct income distributions across social groups and states suggest that median income could be a valid

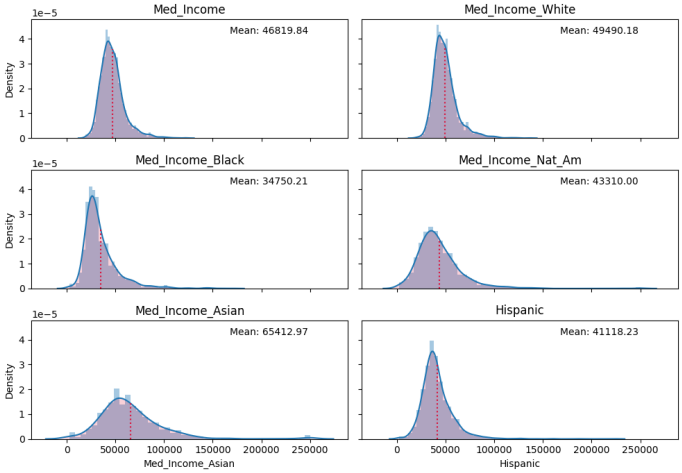

Fig. 2. Median income by state.

factor in determining Average Annual Incidence and Deaths. Consequently, we can infer that social status is also likely to be a relevant factor.

Next, we assess the presence of null values in each column, revealing that columns containing Median Income data for social groups such as Native Americans, Asians, Blacks, and Hispanics have a substantial number of missing values, ranging from 20% to 55%. To address this issue, we opt to remove these columns. Additionally, we identify columns with non-numeric values, including Incidence Rate, Average Annual Incidence, Recent Trend, Mortality Rate, and Average Annual Deaths.

An important observation is that the independent columns are not population-normalized and lack population data. Consequently, it is prudent to train our model using Average Annual Incidence and Average Annual Deaths rather than Incidence Rate and Mortality Rate, as the latter are merely normalized versions of the average values. Consequently, we drop these two dependent columns.

Upon evaluating the Average Annual Deaths column data, we notice that 325 data points are marked with an asterisk. These asterisks denote data suppressed due to confidentiality when fewer than 16 cases were reported. We must address this missing data, but we will evaluate the other columns before proceeding.

In the Average Annual Incidence column, we encounter three types of non-numeric data: '3 or fewer,' ' ', and ' '. These instances are relatively small compared to the total dataset. We replace '3 or fewer' with 3 and the other instances with null values to address this.

Furthermore, we employ feature extraction to create two new columns, indicating whether the recent trend is rising or falling, and subsequently, we drop the original recent trend column. With some data points now containing null values, we must handle these gaps before training our model. Several methods exist, including removing rows with missing data, imputing missing values with the median within each state,

or utilizing a model capable of handling missing data. In this paper, we choose the latter approach.

B. Visualization

Our analysis begins with creating scatter plots illustrating the relationship between Average Annual Incidence and Average Annual Deaths (Figure 3). These plots reveal a strong correlation between the two variables, suggesting that testing on one of these variables would yield similar results for the other. This assumption can be reasonably made. Subsequently, we generate a pair plot including all poverty-related and median income columns. This analysis confirms the presence of a significant correlation, as supported by a correlation heat map (Figure 4), among the poverty-related columns—namely, All Poverty, M Poverty, and F Poverty. This high multicollinearity issue prompts us to address it by removing the M Poverty and F Poverty columns.

Following the same procedure, we repeat this analysis for the insurance-related columns and drop the M With, M Without, F With, and F Without columns. Moving forward, we create pair plots (Figure 5) for the remaining columns—All Poverty, Median Income, All With, and All Without. These visualizations indicate a substantial correlation between All Poverty and the Insurance columns (Figures 6, 7, 8, 9). However, we will address this issue in subsequent steps. Additionally, we examine scatter plots illustrating the relationships between the independent and dependent columns.

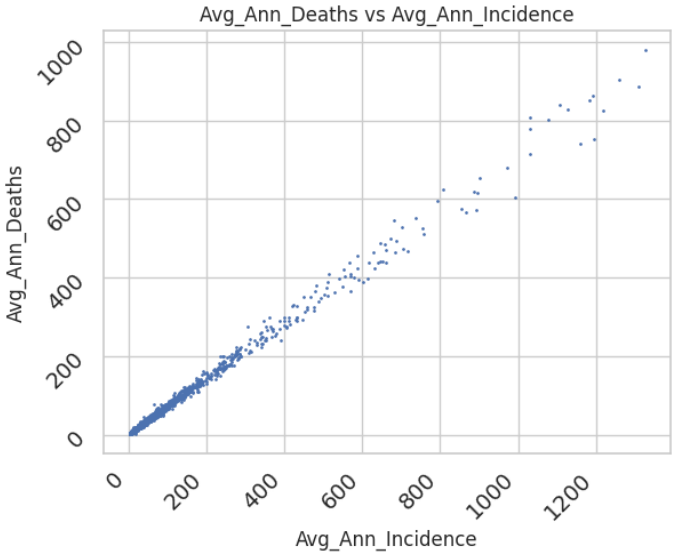

Fig. 3. Average Annual Death vs Average Annual Incidence.

C. Statistical Linear Regression Modelling

We employed the Statsmodels library in Python to construct linear regression models, offering the advantage of obtaining significance values for the regression coefficients. Initially, we created two separate models for the dependent variables, Average Annual Deaths and Average Annual Incidence, with the independent variables being All Poverty, Median Income, All

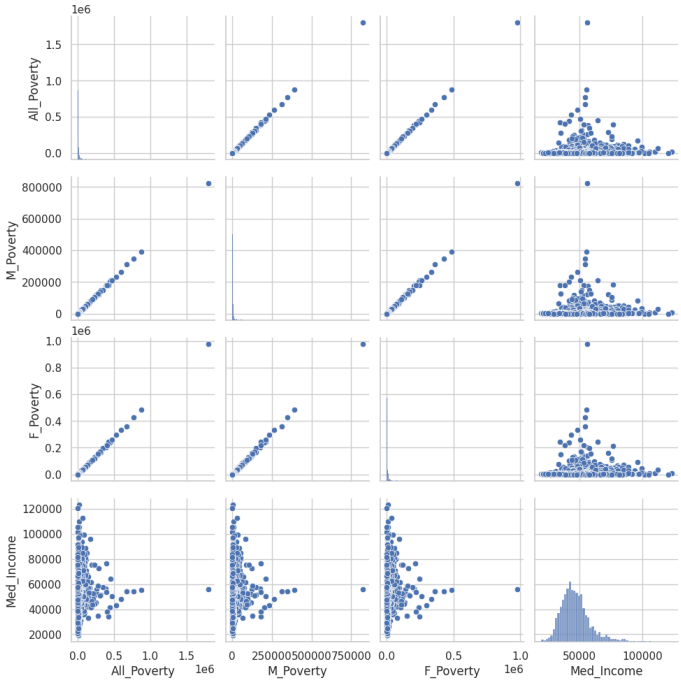

Fig. 4. Pairplots of Poverty vs Median Income.

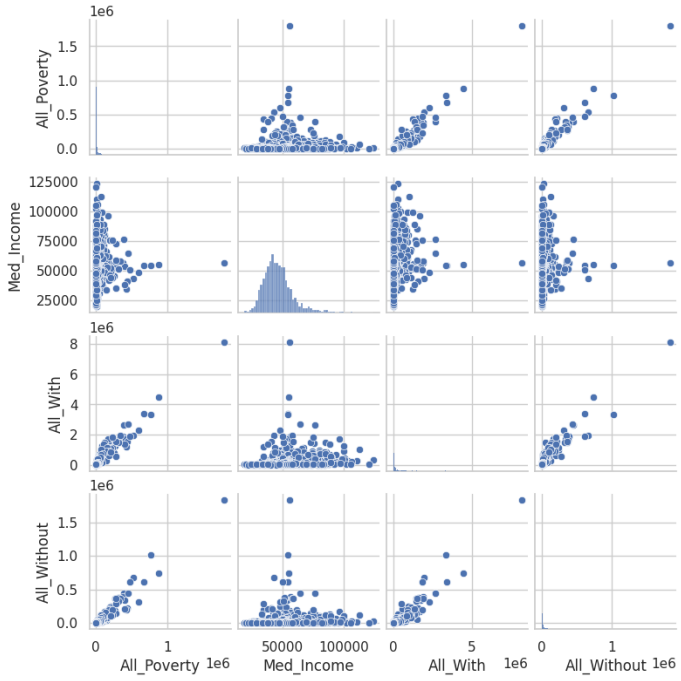

Fig. 5. Pairplots of Poverty vs Median Income.

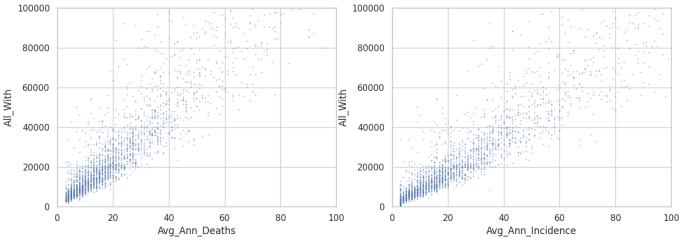

Fig. 6. All with insurance vs Average annual Death and Average annual Incidence.

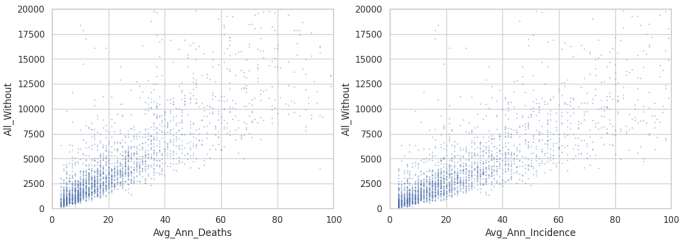

Fig. 7. All without insurance vs Average annual Death and Average annual Incidence.

With, All Without, Rising, and Falling. Both models yielded an Adjusted R-squared value of 0.922, and all coefficients were statistically significant except for Rising, which had a P-value exceeding 0.05 (for a 95% confidence level). It's worth noting that the Statsmodels library automatically handled null values in the dataset.

Next, we assessed whether our model met the assumptions for linear regression. We began by calculating the Variance Inflation Factor (VIF) for all independent variables to check

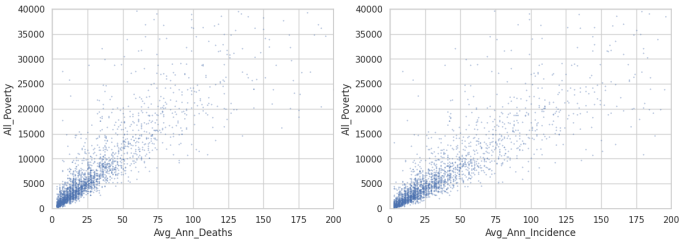

Fig. 8. All Poverty vs Average annual Death and Average annual Incidence.

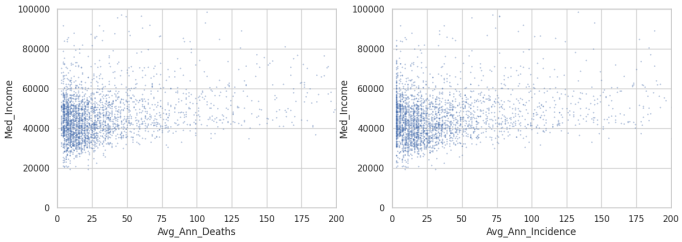

Fig. 9. Median Income vs Average annual Death and Average annual Incidence.

for multicollinearity. The first model revealed that All Poverty had the highest VIF, approximately 25.7, which could lead to statistically insignificant coefficients. Consequently, we iteratively removed variables until all VIF values were below the threshold of 5.

Subsequently, we examined the normality of the residuals. By plotting histograms of our data alongside normal and t distributions (Figure 10) and creating QQ plots (Figure 11) for both distributions, we observed that the QQ plot for the normal distribution did not closely align with the straight line at the ends, deviating upwards on the right and downwards on the left. This indicated that the residuals had heavier tails than a normal distribution. To address this, we explored the QQ plot for the t distribution, which exhibited a much better fit, confirming that the residuals followed a t distribution. The presence of fatter tails in the distribution suggested the presence of more outliers.

Next, we turn our attention to assessing heteroscedasticity. Initially, we visualize this by creating a regression plot (Figure 12) and consistently find a strong Pearson R coefficient, indicating a high correlation.

Subsequently, we delve into analyzing the data further. We construct a Lowess curve (Figure 12) and a scatter plot for our residuals (Figure 14) to scrutinize any discernible trends. We notice that the Lowess curve consistently falls below the y=0 line for lower values, suggesting our model tends to overpredict these values. Conversely, for most of the residuals, they cluster above the y=0 line, indicating that our model tends to underpredict these values.

Finally, we proceed to visualize the residuals, looking for any patterns in the changing variance across different model values. What we observe is a distinctive cone-shaped pattern in the residual plot, a common indication of heteroscedasticity. This phenomenon suggests that as the fitted values increase, the variance in the residuals also increases, signifying a notable presence of heteroscedasticity in our model.

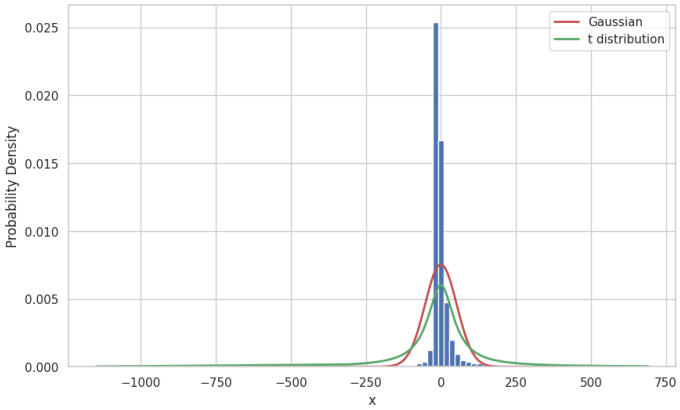

Fig. 10. Probability Density with common distributions.

IV. Conclusion

After constructing a statistical model that incorporates various features as proxies for socioeconomic status and treats

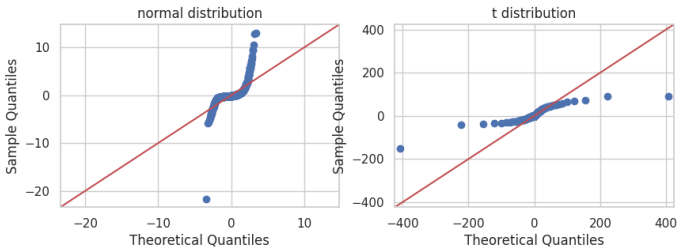

Fig. 11. QQ Plots.

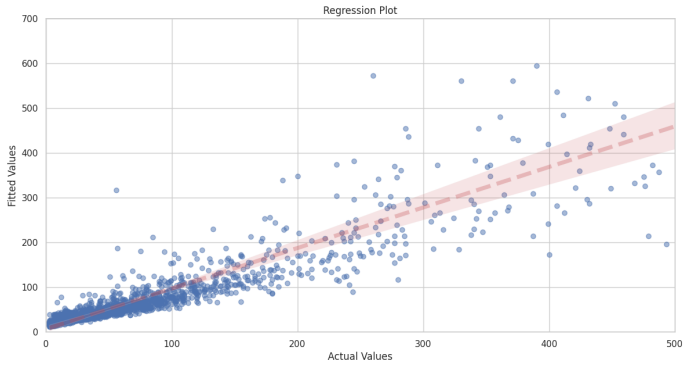

Fig. 12. Regression Plot.

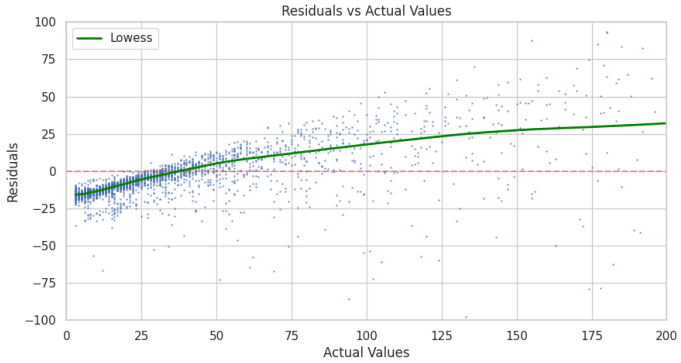

Fig. 13. Residual vs Actual Values.

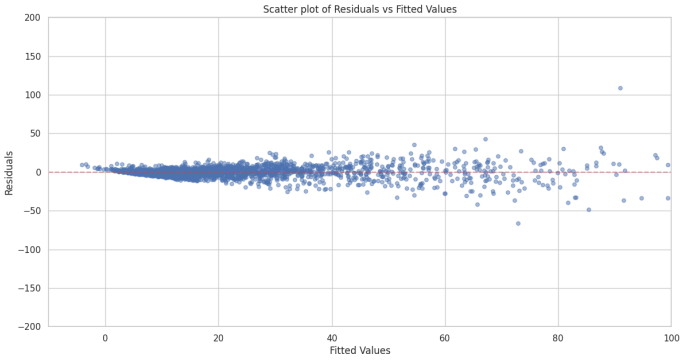

Fig. 14. Heteroscadasticity.

Average Incidence and Deaths as the dependent variables, we observed significant correlations among these features. Notably, factors such as Poverty, Median Income, and Insurance, which serve as economic indicators, displayed substantial relationships. Additionally, our analysis of Median Income data across different communities and states revealed varying mean values, suggesting that Median Income plays a crucial role in determining mortality rates. Furthermore, the disparities in Median Income among different social groups support the notion that socioeconomic factors are indeed significant. Consequently, we accept our hypothesis that there exists a correlation between socioeconomic status and cancer incidence as well as mortality.

While assessing the assumptions of Linear Regression, we identified and addressed issues such as multicollinearity (which has been resolved), non-normality (with residuals aligning better with a t-distribution), and heteroscedasticity (indicated by a cone-shaped residual plot). To further enhance our model, potential future improvements could involve the identification and treatment of outliers. Additionally, investigating the root causes of variable variance could be beneficial. Some potential solutions may include employing weighted regression models or applying transformations to the dependent variable. One such transformation could involve normalizing all independent variables with population data and then utilizing Mortality Rate and Death Rate for modeling purposes, thereby eliminating the population's effect on the data.

References

[1] Social Determinants of Health in Non-communicable Diseases. [Online]. Available: https://link.springer.com/book/10.1007/978-981-15-1831-7

[2] IBM. "Linear Regression." Available online: https://www.ibm.com/topics/linear-regression.

[3] Wikipedia. "Linear Regression." Available online: https://en.wikipedia.org/wiki/Linear_regression.

[4] M. Rutecki, "Regression Models Evaluation Metrics," Kaggle, [Online]. Available: https://www.kaggle.com/code/marcinrutecki/regression-models-evaluation-metrics.

[5] Wikipedia. "P-value." [Online]. Available: https://en.wikipedia.org/wiki/P-value.

[6] Wikipedia, "Student's t-distribution," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Student%27s_t-distribution.

[7] "7 Ways to Handle Missing Values in Machine Learning," *Towards Data Science*, 2023. [Online]. Available: https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e.

[8] Minitab Blog, "How to Interpret Regression Analysis Results: P-values and Coefficients," Minitab Blog, [Online]. Available: https://blog.minitab.com/en/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients.

[9] GeeksforGeeks, "Detecting Multicollinearity with VIF (Python)," GeeksforGeeks, [Online]. Available: https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/.

[10] Jim Frost, "Understanding Heteroscedasticity in Regression," Statistics by Jim, [Online]. Available: https://statisticsbyjim.com/regression/heteroscedasticity-regression/.

[11] Stack Exchange, "How to Interpret This Shape of QQ Plot of Standardized Residuals," Cross Validated (Stack Exchange), [Online]. Available: https://stats.stackexchange.com/questions/481413/how-to-interpret-this-shape-of-qq-plot-of-standardized-residuals.