DAL Assignment 1: A Mathematical Essay On Linear Regression

C Vamshi Krishna Reddy
CH20B112
EE5708: Data Analytics Laboratory
IIT Madras, TamilNadu, India
ch20b112@smail.iitm.ac.in

***Abstract*—This document serves as a model for showcasing the application of Linear Regression in assessing whether individuals from low-income backgrounds face an elevated risk of cancer diagnosis and mortality. Its aim is to support a nonprofit organization's efforts to advocate for improved health outcomes among low-income populations in the United States.**

***Index Terms*—regression, linear regression, statistical modeling**

I. Introduction

Machine learning, a subset of both artificial intelligence (AI) and computer science, is dedicated to leveraging data and algorithms to replicate human learning processes. It excels at forecasting future outcomes by drawing insights from historical data or filling gaps in existing datasets.The rapidly advancing field of data science relies significantly on machine learning. Generally, machine learning algorithms can be classified into three main categories: supervised, unsupervised, and reinforcement learning. In the context of this paper, we employ Linear Regression, a technique falling within the supervised category.

Linear regression, a widely-used mathematical model, offers simplicity and interpretability in predicting outcomes, finding applications in diverse fields from environmental and biological sciences to social sciences and business. This technique assumes a linear relationship between input variables and an output variable, functioning both as a machine learning and statistical algorithm. In this discussion, we delve into the concepts of linear regression, emphasizing its goal of identifying the best-fitting line that minimizes prediction errors. These errors, defined as the deviations between data points and the regression line, are typically addressed through the method of least squares, often referred to as ordinary least square regression.

This paper aims to use Linear regression in explaining whether the social and economic background of a person influences his incidence and mortality rates of cancer in United States is affect.

The paper provides a detailed overview of linear regression by solving a real world problem and insights about the relation between variables like income, poverty, number of insurance citizens and number of cancer cases, incidence rate of cancer in given population.

II. Linear Regression

Regression is a widely employed application in machine learning, especially in supervised settings. It essentially helps us understand how certain independent factors connect to a particular outcome or result. In simpler terms, it's like a tool used in machine learning to make predictions, particularly for continuous results. The core idea is to teach algorithms how different independent variables are linked to an outcome or dependent variable. Once the model grasps this relationship, it becomes capable of predicting what might happen when faced with new, unexpected data or completing missing parts in existing data.

Linear regression is one of the most commonly used statistical approaches in regression models. In this, the model assumes a linear relationship between the input and output variables. Additionally, it assumes that observations are independent of each other, that the errors or discrepancies follow a normal distribution, that input variables are unrelated to output variables, and that there is minimal to no overlap or redundancy among the input variables.

A. Mathematical Equation of Linear Regression

Linear regression can be expressed through a mathematical equation given below:

$$Y(X, W) = W^T X \tag{1}$$

Here X refers to a set of input features, and W refers to the weights corresponding to individual features and Y refers to the Output variable (or) result.

B. Cost Function

The cost function acts as a measure of how much the model's predictions deviate from actual outcomes. Usually, the Mean Squared Error is normalized and adopted as a cost function in the context of linear regression.

$$[CostFunction(J) : (1/n)(\sum_{i=1}^{n}(yi - yp)^2)] \tag{2}$$

C. Cost Function optimization

Gradient descent can be used in estimating the model's parameters by optimizing cost function.The gradient of the cost function is obtained by differentiating the cost function with respect to the separate weights and combining them into a single vector. Differentiating each of the *wi* we obtain:

$$\frac{\partial}{\partial wi}J(w) = \sum_{i=1}^{n}(\frac{w^T X - yi}{m})xi \tag{3}$$

D. Evaluation Model

R-squared is employed as a metric to assess the goodness of fit of a model to the provided data. Mathematically, it can be represented as:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_{i=1}^{n}(yi - \hat{y}_i)}{\sum_{i=1}^{n}(yi - \bar{y})} \tag{4}$$

III. The Problem

The objective is to determine if individuals with incomes below the average face a higher risk of cancer development and mortality compared to those with higher incomes. We possess data on lung cancer incidence and mortality rates categorized by region. Our focus is on uncovering potential connections between factors such as the poverty rate, median income within different population segments, and the number of people with health insurance. The process initiates with data cleansing and then exploratory data analysis. Subsequently, we seek to establish correlations between these factors and the response variables, namely, incidence and mortality rates.

A. Data Cleaning

The given dataset has a large number of missing values. They should be eliminated or modified using common methods such as mean estimation-based filling, distribution-based filling, etc. Certain features like AreaName, FIPS, fips x, fips y which do not add any additional value to the problem statement were removed, and columns like "State" were turned into categorical features using Label Encoder. The new column total population was derived from the "All with" and "All without" columns and used to normalize columns like "All poverty" and "All with." to calculate "Poverty proportion", "proportion of insurance" columns. Now the median income,Poverty proportion,Proportion of insurance are used as independent variables and "Mortality Rate", "Incidence Rate" are considered as dependent variables to assess the situation for a region.

From figure 1 it is evident that the count of null-values are almost half percent which was incredibly high in the race-specific median income columns. As a result, it was decided not to include these columns in the final linear model because they might have an impact on the goodness of fit. The remaining columns which have fewer null values are filled with mean values.

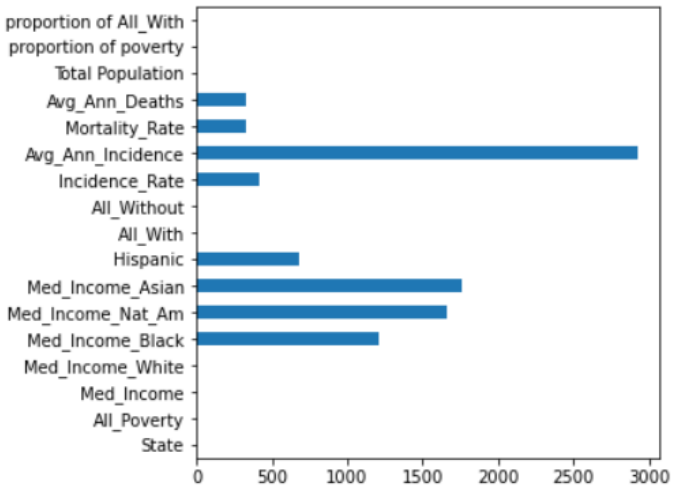

Fig. 1. Bar Graph distribution of Null Values.

B. Exploratory Data Analysis and Visualisations

In the dataset, we are provided with columns All with, All without which refers to number of individuals with and without insurance. These are dependent on overall population of the area so it is converted into proportion of people with insurance before comparing with incidence rate and mortality rate. Similarly, we have converted number of poor to proportion of poor people by normalizing with All with + All without (i.e. total population). and female population of the place in the columns reflecting the men and women with and without insurance.

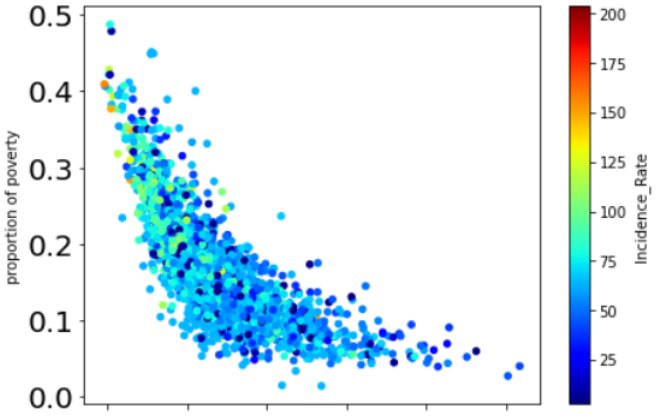

Fig. 2. Proportion of Poverty vs Incidence Rate.

Figure 2 shows a scatter plot between proportion of poverty and Median Income. An important point to note is that the features are highly correlated. Proportion of poverty and Median Income have a very correlation of -0.75. From Figure 2 and Figure 3, The scatter plot between Median Income and Proportion of poverty shows qualitatively that proportion of poverty is a good indicator of Mortality and Incidence rates. Counties with a higher value of proportion of poverty have

higher mortality rates and those with a lower values have lower mortality rates.

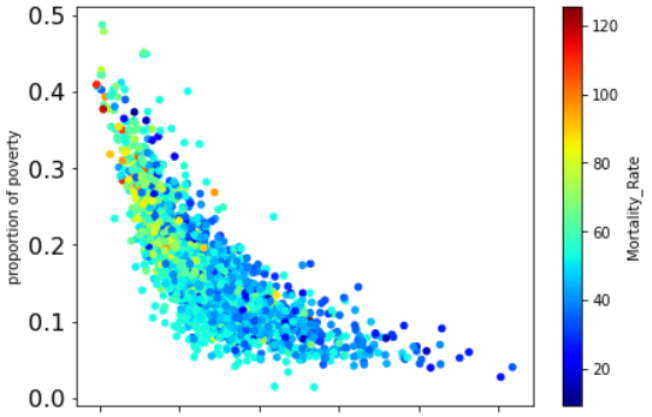

Fig. 3. Proportion of Poverty vs Mortality Rate.

C. Correlation Heat Map

The heat map is plotted using a clean data set to identify the correlation between different features with the Incidence and Mortality Rates. From this heat map of correlation, we can see that both incident rate and mortality rate are primarily correlated with median income, proportion of poor, and proportion of people with insurance. The incidence rate of cancer is higher in areas with lower median incomes (negative correlation) and higher percentages of the population living in poverty (positive correlation). It is apparent that regions with and without robust economies experience have different cancer incidence rates. According to a correlation map between poverty and median income, persons who are economically disadvantaged have a higher mortality risk from cancer than those who are wealthy. There is a tiny but favorable correlation between the death rate and the number of insurance holders.

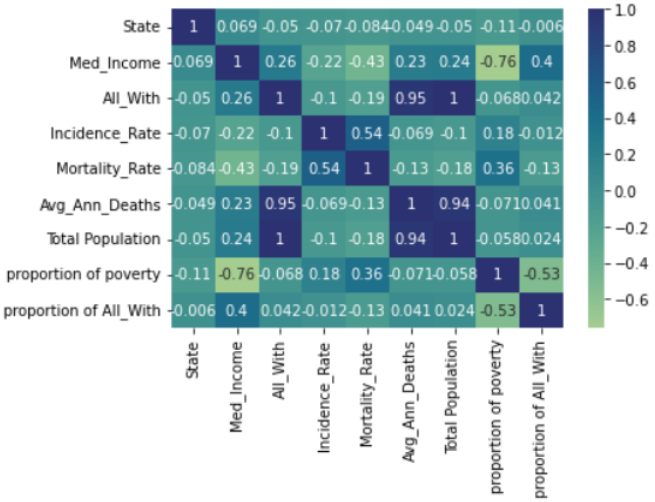

Fig. 4. Correlation Heat Map.

Running a linear regression model on mortality rate with median income, proportion of poverty and all with (insurance) and another model on Incidence Rate with median income, proportion of poverty and all with (insurance) we obtain below coefficients.

TABLE I
MORTALITY RATE

Coefficient of features			
Median Income	***proportion of poverty***	***All With***	***Intercept***
-3.43e-04	2.39e+01	-4.77e-06	65.72

TABLE II
INCIDENCE RATE

Coefficient of features			
Median Income	***proportion of poverty***	***All With***	***Intercept***
-3.72e-04	9.93	-4.63e-06	79.51

From these tables, it is clear that both Mortality Rate and Incidence Rate are inversely proportional to the median income and directly proportional to poverty percentage. Percentage of people with insurance has significantly low impact on both mortality rate and incidence rate as its coefficient is of the order 10^{-6}.

IV. CONCLUSION

The correlation matrix clearly indicates that individuals with lower incomes are more likely to experience cancer development and mortality compared to those in more stable financial circumstances. This conclusion is drawn from the negative correlation observed between incidence rates and median income. Through linear regression analysis, we were able to quantitatively establish relationships between incidence and death rates with median income, poverty levels, and the number of individuals with health insurance. However, it was not feasible to do the same for race-specific disparities due to the limited availability of data. In summary, low-income individuals exhibit higher cancer incidence and mortality rates, as indicated by three key factors: median income, the proportion of people living below the poverty line, and the percentage of individuals with health insurance.

REFERENCES

[1] Bishop, Christopher M., Pattern Recognition and Machine Learning, 2006.
[2] Bevans, R. (2020, October 26). Simple Linear Regression: An Easy Introduction and Examples.
[3] Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining., Introduction to Linear Regression Analysis
[4] [Online sources] https://www.analyticsvidhya.com/blog/2021/04/gradient-descent-in-linear-regression/
[5] https://statisticsbyjim.com/regression/interpret-r-squared-regression/