DAL 2023 Assignment 1: A Mathematical Essay on Linear Regression

Arun Palaniappan
Department of Mechanical Engineering
Indian Institute of Technology Madras
Chennai, India
me20b036@smail.iit.ac.in

***Abstract*—This essay provides an overview of the mathematical formulations and applications of Linear Regression. First, the fundamentals of Linear Regression are explored, along with the specifics required to comprehend the study. The problem statement and data-set are introduced in a detailed way. The data-set under consideration is one of cancer incidence and mortality among various socioeconomic classes of the population of the United States of America. To accomplish this efficiently, the data has been thoroughly evaluated, and both visual and quantitative insights have been provided in this essay. Finally, the results of the linear regression model application are reported to better understand which features are essential when associating socioeconomic status with the Incidence and Mortality Rate in different counties. The "Linear Regression" section has been improved by introducing the algorithm to obtain the normal equation of regularized and unregularized Linear Regression.**

***Index Terms*—linear regression, cancer, socioeconomic groups, visualization**

I. Introduction

Linear regression is a type of statistical analysis used to predict the relationship between variables. It is a supervised learning algorithm which assumes a linear relationship between the independent variable and the dependent variables, and aims to find the best-fitting line that describes the relationship. Using this best-fit line we can predict the independent variable for unforeseen data-sets.

This is accomplished by an optimisation algorithm. We use a Loss function, by Loss function we mean a way to quantify how far off we are from the underlying ground truth model. Hence, we try to minimize this Loss function to obtain the best fit line. To see how our model will perform on future unforeseen data-sets, we create a train-validation split, then train the model(minimize the Loss function) on the train split and then check the accuracy of the prediction on the validation split.

The challenge we're attempting to solve is to use Linear Regression techniques to better comprehend the association between different socioeconomic groups in the United States of America and cancer incidence and mortality rates in those groups. Cancer is the second largest cause of death in the US, trailing only heart disease. Cancer is responsible for one out of every four deaths in the United States. Despite advances in cancer detection and therapy, not everyone benefits equally from them.

In this paper, a comprehensive understanding of Linear Regression fundamentals is provided. The paper also undertakes a detailed exploration of the targeted problem and elucidates how the resulting insights are underpinned by a combination of visual representations and quantitative data.

II. Linear Regression

In this section, we will describe the mathematical details of linear regression. Linear regression is a supervised learning algorithm used for modeling the relationship between two types of variables, the target(dependent) and features(independent). The following is an overview of the jargons and mathematics utilised in the Linear Regression approach.

- *Features/Independent variables:* The features or independent variables are used as predictor features to predict the target variable. They can be classified into three two categories i.e., continuous and categorical features. Continuous features are those that take on a range of numerical values, allowing for fine-grained variations(for example, the average salary of a person in a state). Categorical features encompass distinct categories or labels. These features can be nominal, where categories hold no inherent order(for instance, it might be the state variable in a data-set).
- *Target/Dependent variables:* Dependent variables, known as target variables, exhibit either continuous or discrete traits. When implementing Linear Regression for a specific problem, the target variable takes on a continuous distribution.

A. Assumptions in Linear Regression

Linear regression is a powerful statistical technique, but its validity and reliability rests upon certain assumptions that must be met for accurate and meaningful results. Here, we will delve into the key assumptions underlying linear regression:

- *Linearity:* The response variable is assumed to be a linear combination of the predictor variables.
- *Independence:* The data points used in linear regression should be independent of each other.
- *No or Little Multicollinearity:* Multicollinearity refers to a high correlation between independent variables in the regression model. This can make it challenging to isolate

the individual effects of each variable, leading to unstable coefficient estimates.
- *Homoscedasticity:* This implies that the variance of the errors (residuals) are constant across all levels of the independent variables.
- *Normality:* The dependent variable is normally distributed for any fixed value of the independent variable.

B. Types of Linear Regression

- *Simple Linear Regression:* When the target variable is predicted by only one predictor variable, simple linear regression is employed. It is mathematically formulated as follows:

$$y = \beta_0 + \beta_1 X \tag{1}$$

Where β_0 and β_1 are the parameters of the model.

- *Multilinear Regression:* When the target variable is predicted by more than one predictor variable, Multilinear regression is employed. The input variables and weights are defined as:

$$\boldsymbol{X} = [x_1, x_2, x_3, ..., x_{n-1}, x_n] \tag{2}$$

$$\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, ..., \beta_{n-1}, \beta_n] \tag{3}$$

It is mathematically formulated as follows:

$$y = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X} \tag{4}$$

Where β_0 is the bias of the model.

- *Regularised Linear Regression:* While training any model the most frequent problem we run into is over-fitting. It is the case where the model has learnt the noise from the training data hence creating a lot of variance while testing. This problem can be tackled by modifying the loss function to give penalty for large magnitudes of coefficients. The two types of regularisation are Lasso(L1) regularisation and Ridge(L2) regularisation.

C. Loss function (or objective function)

In linear regression, the loss function is like a guide that tells the model how well it's doing. It checks how much the model's guesses are different from the actual answers. The goal is to make these differences as tiny as possible. One common loss function is called the *Mean Squared Error*. It looks at how far off the model's guesses are from the real answers and adds up those distances. The model works to reduce these distances by adjusting its guesses.

The general loss function is formulated as:

$$loss = \frac{1}{N}\left[\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right] \tag{5}$$

Where N is the total number of data points, y_i is the ground truth and $\hat{y}_i$ is the predicted value.

The loss function with *L1 regularisation* is formulated as:

$$loss = \frac{1}{N}\left[\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right] + \frac{\lambda}{N}\left[\sum_{j=1}^{p}|\beta_j|\right] \tag{6}$$

Where p is the total number of parameters and β_j's are the parameters.

The loss function with *L2 regularisation* is formulated as:

$$loss = \frac{1}{N}\left[\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right] + \frac{\lambda}{N}\left[\sum_{j=1}^{p}(\beta_j)^2\right] \tag{7}$$

Where p is the total number of parameters and β_j's are the parameters.

D. Normal Equation Algorithm

Eqn. 1 can be transformed to Eqn. 8 by including a unit scalar to each data row of X, including β_0 in the β vector itself and also taking transpose of these matrices and vectors.

$$\hat{y} = X\hat{\beta} \tag{8}$$

1) Vanilla Linear Regression:

Let y be the ground truth values. Then loss function can also be written as,

$$\begin{aligned}\text{Loss} &= (y-\hat{y})^T(y-\hat{y}) \\ &= (y - X\hat{\beta})^T(y - X\hat{\beta}) \\ &= (y^T - \hat{\beta}^T X^T)(y - X\hat{\beta}) \\ \text{Loss} &= y^T y - y^T X\hat{\beta} - \hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta}\end{aligned}$$

Now, to minimize loss, we take the derivative of it with respect to $\hat{\beta}$ and equate it to zero.

$$\begin{aligned}\frac{\partial\ (\text{Loss})}{\partial\hat{\beta}} &= \frac{\partial y^T y - y^T X\hat{\beta} - \hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta}}{\partial\hat{\beta}} \\ &= \frac{\partial y^T y}{\partial\hat{\beta}} - \frac{\partial y^T X\hat{\beta}}{\partial\hat{\beta}} - \frac{\partial\hat{\beta}^T X^T y}{\partial\hat{\beta}} + \frac{\partial\hat{\beta}^T X^T X\hat{\beta}}{\partial\hat{\beta}} \\ &= 0 - y^T X - (X^T y)^T + 2\hat{\beta}^T X^T X \\ &= 2\hat{\beta}^T X^T X - 2y^T X\end{aligned}$$

Equating this derivative expression to zero, we get,

$$\begin{aligned}Xy^T &= \hat{\beta}^T X^T X \\ \implies \hat{\beta}^T &= Xy^T(X^T X)^{-1} \\ \implies \hat{\beta} &= (X^T X)^{-1} X^T y\end{aligned}$$

Hence we have derived the normal equation for unregularised linear regression.

2) L2-Regularised Linear Regression:

The regularised loss function in Eqn. 7 becomes,

$$\text{Loss} = (y-\hat{y})^T(y-\hat{y}) + \lambda\beta^T\beta$$

Hence the normal equation which give the estimates of the parameters $\hat{\beta}$ becomes,

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

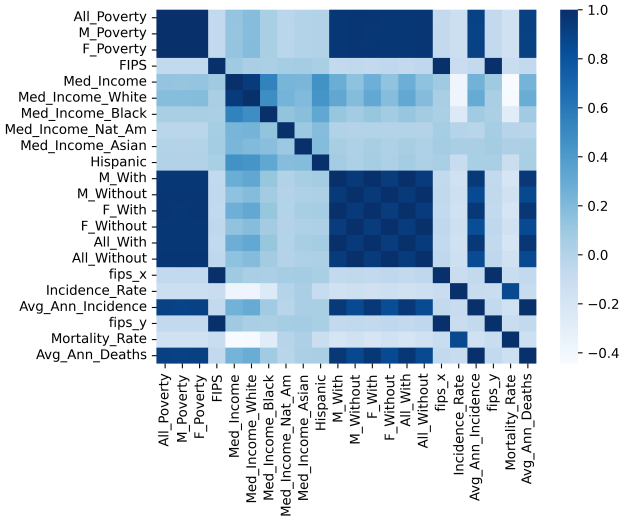

Fig. 1. Correlation heat-map within the variables

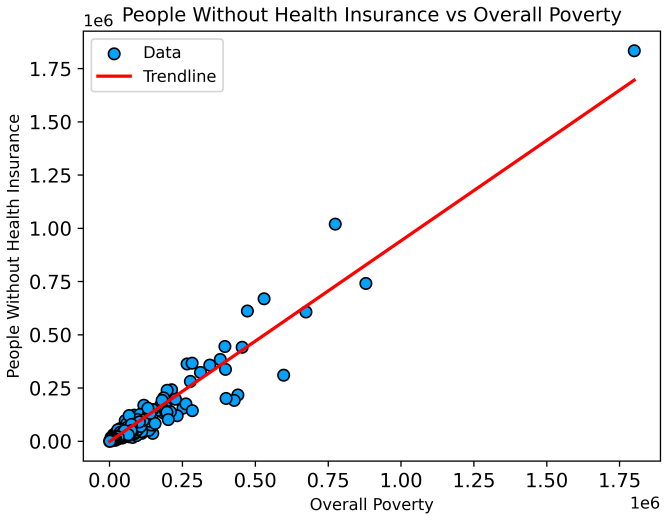

Fig. 2. People Without Health Insurance vs Overall Poverty

E. Pearson correlation coefficient

The Pearson correlation coefficient is a correlation coefficient in statistics that assesses the linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; consequently, the result always falls between -1 and 1.

$$\rho_{x,y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y} \tag{9}$$

It can also be used to measure the performance of a linear regression model.

III. The Problem

A data-set with information on cancer incidence and mortality rates in various American counties is made available to us. The ultimate aim of the study is to determine whether or not there is a relationship between a person's socioeconomic status and their risk of developing cancer or suffering death from it. The significance of various features as determined by the linear regression model is also explored in this study.

A. Outline

We begin by analysing, cleaning and visualizing the data, as it has a lot of missing values and unclean data points. Then we decide which variables must be used in our linear regression model according to its importance and multicollinearity. Then we impute the missing data using imputation techniques. A linear regression model is then fit on the data and inferences are made from the model coefficients.

B. Data description

In this section, we describe the structure of the data-set. The income, poverty and health insurance data were obtained from the US Census Bureau and the cancer data was obtained from the National Cancer Institute. Some of the important columns from the data are:

- The state that the county belongs to in the US
- The number of people below the poverty line by gender
- The number of people with and without health insurance by gender
- The median income for five racial groups (White, Black, Native American, Asian and Hispanic)
- Recent trends have also been given for each county.
- Cancer incidence and mortality rates along with the average annual incidences and deaths.

C. Data cleaning and imputation

The data-set contained many missing values in both the predictor variables and the target variables (incidence rate and mortality rate). They were dealt with as follows:

- *Predictor variables:* The missing values in these columns (for example, median income overall or median income for a certain ethnic group) has been replaced by the state-wise mean. This will not tend to over-fitting because the median values are normally distributed, so replacing it with the mean will have no effect on the model's robustness.
- *Target variables:* The missing values in the response variables were dealt with by removing the respective rows since imputing with mean or median could lead to erroneous results.

D. Visualizing the data

Here is a heat map of the correlation values between variables in the data-set(variables like state, area name and FIPS were dropped since they did not provide any useful information):

We can conclude that:

- Number of males and females in poverty and with/without life insurance were highly correlated with the overall numbers for the same quantity.

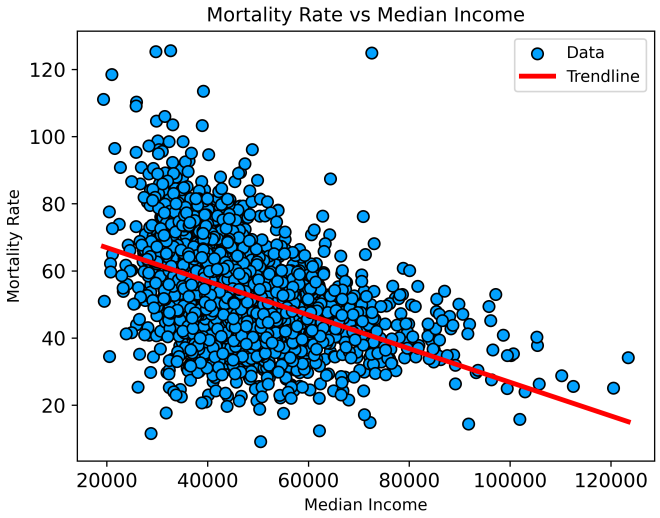

Fig. 3. Mortality Rate vs Median Income

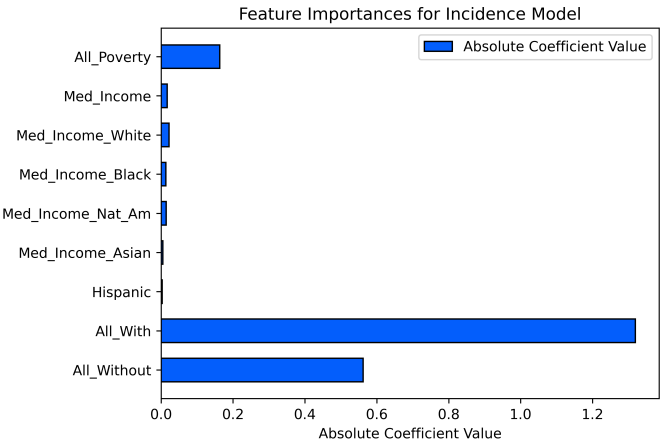

Fig. 4. Feature Importances for Incidence Model

- Average annual incidence and deaths were highly correlated with the number of people below poverty line and the average annual income of people.
- From Fig. 2 we can infer that Number of people without health insurance is highly and positively correlated with overall poverty.
- From Fig. 3 and the trend-line, we can observe that the Mortality rate decreases with increasing Median income

E. Linear regression modelling

We accomplish this by using *scikit-learn* package in python. The predictor variables that we use are *'All Poverty', 'Median Income', 'Median Income White', 'Median Income Black', 'Median Income Nat Am', 'Median Income Asian', 'Median Income Hispanic', 'All With Health Insurance', 'All Without With Health Insurance'*. We now instantiate two different linear models, for *'Average Annual Incidence'* and *'Average Annual Deaths'* respectively. We use *root mean square error(RMSE)*

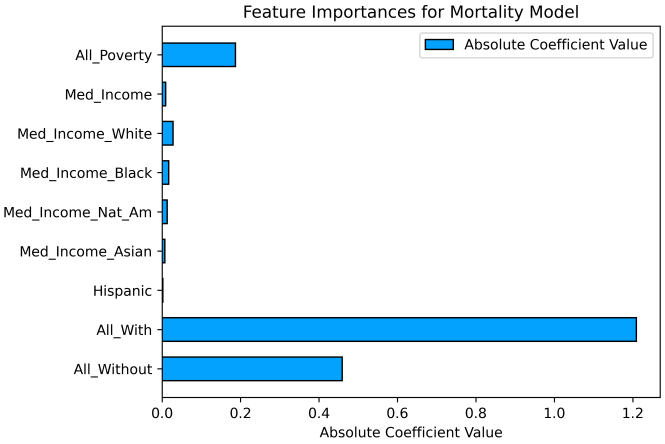

Fig. 5. Feature Importances for Mortality Model

and R^2 score as our evaluation criteria on the validation dataset. After testing on the Validation Split, the linear model for *'Average Annual Incidence'* obtains an RMSE of 43.54 and R^2 score of 0.91 where the *'Average Annual Incidence'* ranges from 2 to 3701. The linear model for *'Average Annual Deaths'* obtains an RMSE of 25.72 and R^2 score of 0.94 where the *'Average Annual Deaths'* vary from 3 to 2876.

IV. Conclusion

The conclusions that can be drawn from all the above analysis are:

- Data analysis concluded that two overall aspects of socioeconomic status(Overall Poverty and Median Income) were significantly correlated and each of them was correlated with the target variables (Incidence and Mortality counts).
- The most important characteristics according to Feature Importances are the number of people with insurance in a county, followed by number of people in poverty.
- Health insurance was highly correlated with the number of people in poverty which was pivotal in the linear models.
- The plots shown, support all the above quantitative evidence in a visual way.

Future studies could explore on using regularised regression techniques(L1 and L2). We can also apply Principle Component Analysis(PCA) to find important combination of features to explain the variance in the data.

References

[1] American Cancer Society, "Cancer Information and Resources," Cancer. *https://www.cancer.org* (accessed Aug. 28, 2023)
[2] "Linear Regression." Wikipedia, Wikimedia Foundation, 19 Aug. 2023, *en.wikipedia.org/wiki/Linear_regression* (accessed Aug. 28, 2023)