A Mathematical Essay on Linear Regression

Aryan Pandey
Department of Ocean Engineering
Indian Institute of Technology, Madras
Chennai, India
Email ID: na19b030@smail.iitm.ac.in

***Abstract*—This essay is an overview of the mathematical formulations and applications of Linear Regression. First, the basic aspects of Linear Regression is explained with all the necessary details needed to better understand the study that has been performed. Next the problem statement and the data-set are introduced. The data-set under study is one of Cancer incidence and mortality among different socioeconomic groups in the United States of America population. In order to effectively do this, the data is thoroughly analysed and any insights (both visual and quantitative) have been published in this essay. Finally, the results from the application of the linear regression model is described. The model is applied to the data-set after transforming some of the features in order to better understand which features are important when it comes to correlating the socioeconomic status with the Incidence and Mortality Rate in different counties.**

I. Introduction

Linear Regression is a form of Supervised Machine Learning where one tries to map the feature space to a target (or dependent) variable using a linear equation. Based of off this learned map, one can try to then make predictions about the target variable. In this algorithm, one strives to find the best possible linear fit to the target variable given the feature space.

This is done via an optimisation algorithm. We choose an objective function which measures the distance between the underlying true distribution and the current fit that we have. Via the optimisation algorithm, one aims to minimise the objective function in order to obtain minimum separation between the underlying true distribution and the fit that we have proposed via our linear model. To test the robustness of the model, we create a split in the data, train the model on the first split (the training data) and evaluate it on the second split(the test data).

The problem that we're trying to tackle is to apply the concepts of Linear Regression to better understand the relationship between the different socioeconomic groups in the United States of America and the cancer incidence and mortality rates in them. The data available has details on poverty status, income, health insurance, gender, incidence rates and mortality rates for each county. The study aims to examine whether low income groups or certain sections of society are at a greater risk of being diagnosed and dying from cancer.

Through this essay, one can understand the basics of Linear Regression, dive into the problem that is being examined and understand how the insights produced are backed with the help of visual or quantitative facts.

II. Linear Regression

Linear regression is a statistical analysis which depends on modeling a relationship between two kinds of variables, target (or dependent) and features (independent). The main purpose of regression is to examine if the features are successful in predicting the target and which features are significant predictors of the target. Given below is a discussion on the terminology and the mathematics used in a Linear Regression Approach.

A. Features/Independent Variable

The features or Independent Variables are those set of variables that are being used to predict the target variable. An ideal set of such variables would be a basis set, since this would allow us to map any point in an n dimensional space onto the target variable. The Features or Independent variables can either be discrete (for example, number of sales calls made in a day) or continuous (for example, the time taken by a ball to fall from the top of a building). The discrete features can also be in a textual format (for example, gender of a person). In such cases, care should be taken to get these features into a format which is understandable by a machine (more details in Section III). These set of variables are collectively represented in a matrix often referred to as X.

B. Target/Dependent Variable

Target or Dependent variables can be either continuous or discrete. When we apply Linear Regression to a certain problem statement, the target variable assumes a continuous distribution. This variable is often represented in a vector and if often referred to as Y.

C. Assumptions made in Linear Regression

Following are the assumptions that any Linear Regression framework makes:

- There is a linear relationship present in the underlying true distribution between X and Y
- For any value of X, the variance of the residual is the same (Homoscedasticity)
- Observations are distinct from one another (Independence)
- Y is regularly distributed for any fixed value of X (Normality)

D. Types of Linear Regression

- Uni-Variate Linear Regression: In this kind of Linear Regression we have one predictor and one target variable. The model then finds a linear relationship between this as:

$$Y = w_0 + w_1 X \tag{1}$$

w_0, w_1 are the parameters/weights

- Multi-Variate Linear Regression: In this kind of Linear Regression, we have multiple predictors which combine to form the same target variable. The input is a vector of individual features

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

n is the number of predictors that we have

The linear relationship between the predictors and the targets is then given by:

$$Y = w^T X^* \tag{3}$$

Where X^* represents the X vector and the bias term combined and w is the parameter/weight vector

E. Objective Function (or Loss Function)

The disparity between the underlying true distribution and the predicted values we get from our fit is quantified by an Objective or a Loss function. Most frequently, we use a Root Mean Square Error function which measures the mean of the square of the errors. Such a loss function keeps in mind that the X vector is normally distributed. In certain situations when we would like to assume the prior for X to be a Laplacian Distribution, we would use the Mean Absolute Error in which we take the mean of the Absolute errors.

III. The Problem

We are provided with a data-set that has details about different cancer incidence and mortality in different counties in the United States of America. Our end goal of the study is to establish whether or not there exists a correlation between the socioeconomic status of a person and the cancer incidence or mortality rate. This study also explores the importance given to different features by the linear regression model.

A. Data Description

The data-set consists of county-wise information (represented by each row) on:

- The number of people living in poverty (segregated by gender)
- The median income (segregated by ethnic race)
- Health Insurance (segregated by gender)
- Annual Incidence and Death rates

Information on the State the county is in along with its FIPS code is also present to identify the county with much more ease. For each county, we are also given the recent trend that the country has seen (as a categorical variable, for example, rising, falling etc.). In our study we observe that including the name of the State as a feature was causing the model to create a good fit to the data since the state was one of the major distinguishing factor. One possible analysis that could further be done is to identify if there are some states where there is a higher number of socioeconomic groups that have high cancer incidence and mortality (if at all there is a correlation). For this study, we remove the State feature from our analysis and focus purely on the columns that directly involve a relation to a socioeconomic group. The target variables for our analysis are the incidence rate and mortality rate, since these are numbers that are population normalised.

B. Data Handling

If we take a quick view of the data, we see that missing values exist in the target variables as well as income columns. The way in which the rows with missing values in these columns are handled is as mentioned:

- Income Related 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 mean of the values of all other counties in the same state. This is possible because the median values are normally distributed, so replacing it with the mean will have no effect on the model's robustness.
- Target Variables: Some states' incidence and mortality rates were not provided either due to confidentiality reasons or because there simply weren't enough cases of incidence or mortality. All counties with such entries were removed from the data-set since any attempt to replace these values with the mean or median could hamper the model's effectiveness. In doing so, we also take care of the missing values in the recent trend feature.

The spacial characters in Incident Rate, Mortality Rate and Average Annual Values of Incidence and Mortality were removed and the remainder, converted to float values. Following this data cleaning we are left with 2618 data points.

C. Exploration

The Pearson Correlation coefficients between the data are examined first. They are a measure of how linearly related any two given variables are. The coefficient of Pearson Correlation between two variables X and Y is given by:

$$\rho_{x,y} = \frac{E[(x - \mu_x)(y - \mu_y)}{\sigma_x \sigma_y} \tag{4}$$

If two variables are the same then the Pearson Correlation Coefficient has a value of 1 and when they are the exact opposite of each other, it has a value of -1. Given figure shows the correlation heatmap between the features of our given data-set

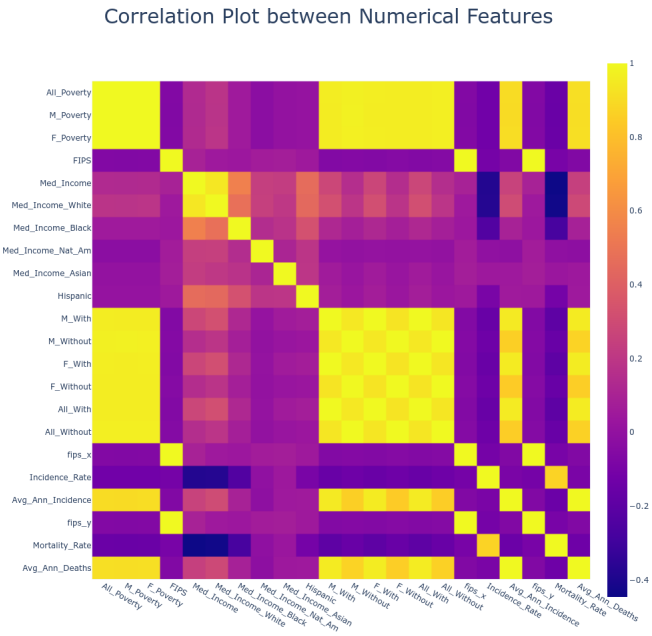


Fig. 1: Correlation Plot between Numerical Features

From this we can find a few relations to Average Annual Deaths and Average Incidence Rate:

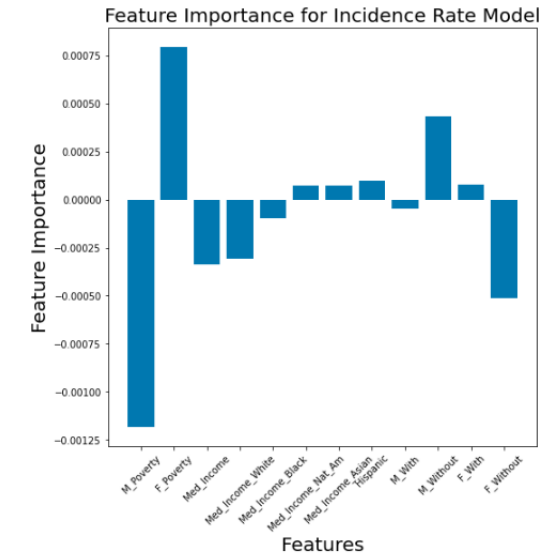
- Not having Health Insurance (which is also highly correlated to poverty) is a big factor in deciding Incidence and death rate
- Poverty also plays a huge role in this regard

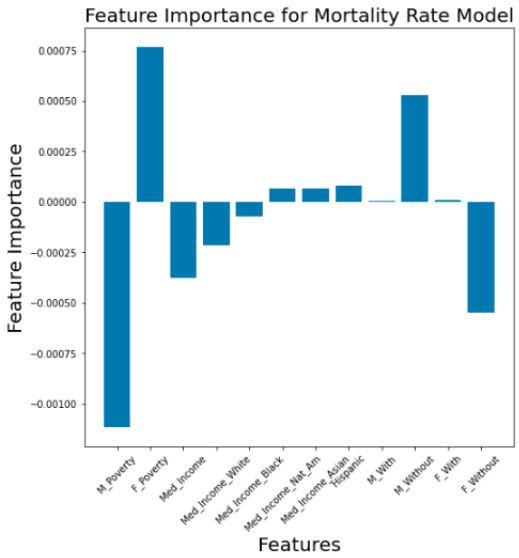
Moreover we can now say that we can remove the overall columns of the Gender and Ethnicity data since we want to get an idea of which groups are impacted the most and these columns are highly correlated with the individual groups.
Given on the right is a scatterplot of the income of different ethnic groups plotted against the Average Annual Incidence.

D. Applying Linear Regression

In order to do this we use the sklearn package in Python. The variable set that we use finally as our features X are M Poverty, F Poverty, Med Income, Med Income White, Med Income Black, Med Income Nat Am, Med Income Asian, Hispanic, M With, M Without, F With and F Without. We fit two linear models to the data, one for predicting Incidence Rates and the other for predicting mortality rates. We use the root mean square error as our evaluation metric after we have split the data-set into 2 splits: the train split and the validation split. On the Validation Split, the linear model for predicting Incidence Rate achieves an RMSE of 16.32 where the Incidence Rates themselves range from 13.5 to 203.7. The linear model for predicting Mortality Rate achieves an RMSE of 13.33 where the mortality rates themselves vary from 9.2 to 125.6. One thing to note here is that these rates are defined as the number of cases per 100000 people in the county.

Given below is a plot of the feature importance brought out by the two models





IV. Conclusion

- Because the correlations with the related overall statistics were quite high, the poverty and insurance data for overall population were redundant.
- Exploratory data analysis found that two overall aspects of socioeconomic status(Poverty Ratio and Median Income) were significantly associated (-0.79) and each of them was correlated with the target variables (Incidence and Mortality Rate).
- The most important characteristics according to Feature Importances are the number of Males in Poverty in a county, followed by number of Females in a county and then cases when each gender does not have health insurance.
- The Insurance Ratio and the incidence/mortality rate were both inversely associated (-0.55)
- While the Insurance Ratio didn't have a very strong relationship with the incidence rate, it had a strong correlation with the mortality rate.
- Socioeconomic Characteristics had a stronger relationship with the Death Rate when compared to the incidence rate.
- The Plots shown visually support all the above quantitative evidence.
- The percentage error in the death rate model was lower when compared to the incidence rate model since the features showed a higher correlation to the death rate than the incidence rate