Assignment [1]

A Mathematical Essay on Linear Regression.

Mustefa Abrahim
Deprt. Data Science
Indian Institute of Technology Madras (IITM)
ge22m014@smail.iitm.ac.in

***Abstract*—In this task, we estimate the average annual mortality based on changes in a number of parameters, including how these parameters, such as whether or not individuals have medical insurance, how the people's poverty condition also affects the average death, and other parameters.Here we used Linear Regression fits a linear model with coefficients w [2] to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. we also done such as cleaning, processing, exploratory analysis, and modelling the data.**

keyword: Explore, Modeling , Analysis, Regression.

I. Introduction

Linear Regression is a supervised learning-based machine learning algorithm. Linear regression involves two variables: a dependent variable that responds to change and an independent variable. Based on independent variables, regression models a target prediction value. A linear regression analysis is a technique for determining the relationship between two variables. It is important to note that we are only calculating the association, not the dependency of the dependent variable on the independent variable. [3]

In this problem we solved the given problem by using the linear regression and in this method first we import all the required python library sklearn, numpy, pandas and etc. From the sklearn we import sub-library such as the linear model which we used to model our results. Predicting the average death is not a novel issue; numerous studies [3] have been published in an effort to address and comprehend the variables that affect people's lives. But understanding the data and exploring its various features to see how they relate to one another and how they might affect the average death target variable during the modelling stage is a crucial step that should help us understand the steps we took to get to the machine learning model running step in our process. [1]

We have discovered that characteristics like medical insurance have a greater impact on the average death than other features due to the fact that Python and Microsoft Power BI assist us to understand our data very well from the features that we should handle first before studying our data. More so than the median income, we discovered that persons without health insurance die on average less frequently than those who do. This difference is not very large, [1]

We will examine the data appearance and the effectiveness of feature visualisation in the second section. In the third section, we will attempt to explain the mathematics underlying the linear regression model and how it was applied to our data. Finally, we will draw a conclusion about our research. [3]

II. DATA SECTION

The amount of data we have today is enormous, and there are many different ways that it was created. For example, data may have been created by machines, by your posts on social media, by your purchases from online retailers, or by a variety of other methods. Another approach is to just design your data using surveys and other methods [2].

A. GATHER, CLEAN AND PREPARE DATA

One of the key goals of our research is to examine the data, but before that, the data needs to be understood and cleaned in a way that would enable exploration. After loading the data and displaying it as a table, we discovered that some of the categorical and numerical features, also known as columns or attributes, contained the punctuation marks [*, -], while other features also contained empty values (Null). We therefore replaced these empty values and other non-representative values with the appropriate values based on the feature itself. Since mean will be impacted by higher values, we first had to display the feature to determine what we should replace the values with. We discovered that some features were skewed to the right, so we had to replace the missing values with median. Some missing values in categorical features were also replaced with the most frequent category, while others were replaced based on our knowledge from the data dictionary. As a result of the loading and cleaning processes, the data is now ready for exploration.

B. Exploratory, Analysis and Visualizations

Following the preparation of the data, we utilised some visualisation to examine the distribution of the features as well as the impact of various features on the average death rate. As you can see from the graphs below, our analysis of the data has shown us that medical insurance and poverty have a greater impact than median income. Additionally, we discovered that poverty has no effect on mortality but has a strong link with either the incidence of cancer or the average death rate. Additionally, the majority of the features have distinct distributions with a range of numbers, which can have an impact on the model because it can only be trained using

data with high or low values. And this data-scaling phase gets our features into the normal distribution that most things in real life have. It can also help avoid outliers in the data rather than deleting the instances (rows) that are connected with these outliers.

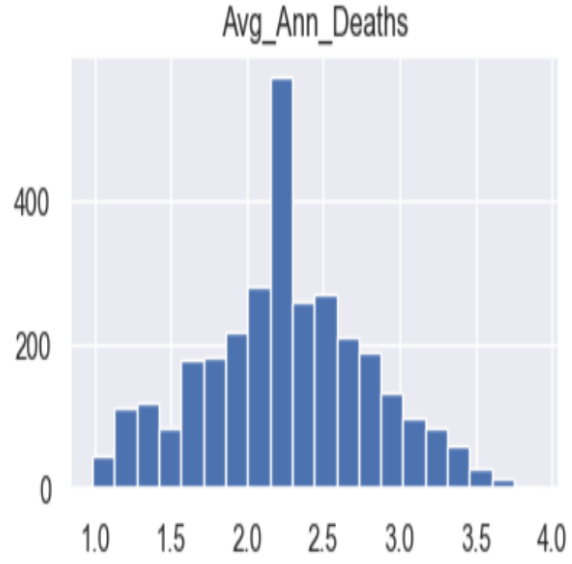

Fig 1: Average Annual Death

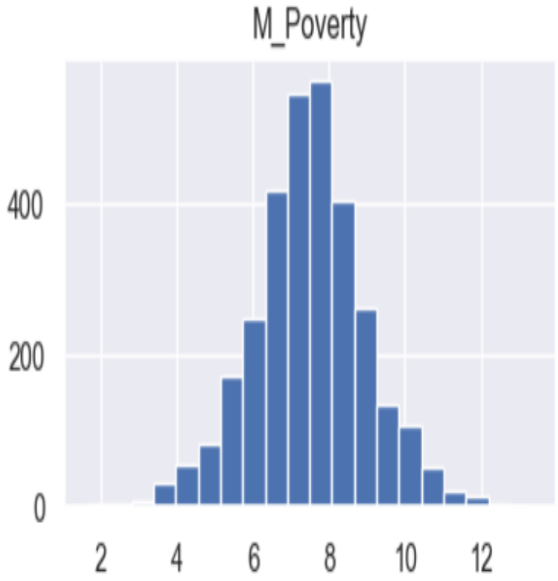

Fig 2: M Poverty After Scaling

C. MODELING

The linear regression model is the method that we use to deal with continuous output while attempting to predict it, but first we must train the model on some data, evaluate it, and determine the optimum parameters that minimise the cost function. The normal equation, which helps us obtain the optimum values of these weights directly, may also be used to train a linear regression model. However, this method only works when there are a limited amount of features. [2] In this study, we update these weights using gradient descent optimization in order to reduce the final cost function. Gradient Descent is a fairly general optimization approach that can locate the best answers to a variety of issues. Gradient Descent's general goal is to iterative adjust parameters in order to minimise a cost function. [3] Theta (weights) are updated using the equation below for each iteration in order to reduce the cost function. So, in linear regression, we look for the weights that minimise the cost function, or "mean square error." Gradient descent is used in a number of iterations to optimise these weights to minimise the cost function. The rate at which these weights are decreased varies from iteration to iteration and is referred to as the learning rate. This symbol is used to multiply the cost gradient of the cost function. [1]

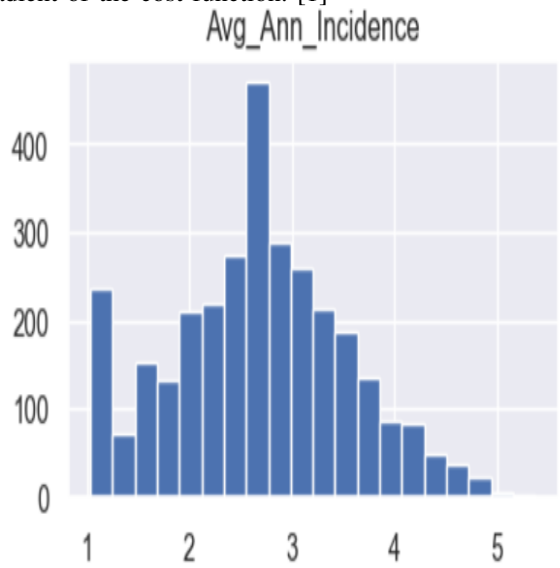

Fig 3: Average Annual Incidence

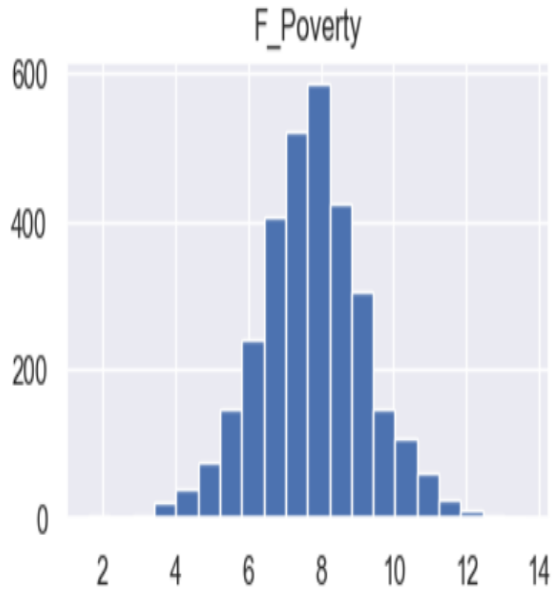

Fig 4: F Poverty After Scaling

D. Conclusion

In this experiment, we discovered that poverty and cancer incidence are major causes of death for people, but poverty and cancer are not the only causes of death for people.

REFERENCES

[1] Hastie T.J. Springer; New York: 2008. The elements of statistical learning: data mining, inference, and prediction. [Google Scholar]

[2] Drucker H. Advances in neural information processing systems. MIT Press; 1997. Support vector regression machines; pp. 155–161. [Google Scholar]

[3] Boccaletti S. Modeling and forecasting of epidemic spreading: the case of COVID-19 and beyond. Chaos Solitons Fractals. 2020 [PMC free article] [PubMed] [Google Scholar]