A mathematical essay of linear regression

Devansh Sanghvi
Department Of Biotechnology
Indian Institute of Technology, Madras
Chennai, India
be19b002@smail.iitm.ac.in

***Abstract*—This paper aims to develop an understanding of the mathematical properties of linear regression and showcase its use in estimating the relationships between the various factors that have an effect on the health data obtained from the United States.**

I. Introduction

Linear Regression is the most basic and commonly used predictive analysis. Once we have acquired data with multiple variables, one important task is to understand how the variables are related. 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. Despite its simplicity, linear regression has proven itself a powerful tool for analyzing data and revealing trends in the data. Its simplicity is evident in its representation:

$$y = \beta_0 + \beta_1 X$$

In the equation above, y denotes the independent variable (to be predicted) and X denotes the dependent variables. Multiple dependent variables can be expressed as a vector. The weights/ parameters of the model are represented by β_0 and β_1. In the most trivial case, as expressed above, β_0 represents the y-intercept and β_1 represents the slope of the line. Once we form the equation of the line, we can predict the y value for the given value of x. The objective of linear regression is to identify the parameter values for which the sum squared total of errors for all data points is minimum. These values provide us a quantitative description of the relationship between the variables.

To demonstrate the effectiveness of linear regression, we use standard socioeconomic variables like income and health insurance access in a population to identify putative correlations with the incidence of cancer and mortality caused by cancer in the same population. The appropriate data has been collected from [1]. It is important to preprocess the data before it is being used for data analysis.

This paper will serve as a part of the requirements to be fulfilled for receiving credits in the course **EE4708: Data Analytics Laboratory**.

II. Linear Regression

It was stated in the introduction section that despite its simplicity, linear regression has proven itself a powerful tool in data analysis. Inside its simplicity lies certain assumptions that must be mentioned before we delve into the mathematical details of linear regression.

A. Assumptions

- The first assumption is that the relationship between Y and X is linear. Violations of this assumption are especially serious - and probably the most common in the literature. It should be checked for initially using a scatter plot of Y against X.
- The second assumption is that the errors are independent. It is assumed that successive residuals are not correlated over time. This can be checked by plotting the serial correlation coefficient against time lag, and/or by using the Durbin-Watson test. The error terms also have a constant variance, which is called homoskedasticity. If the error terms have varying variance, it is called as heteroskedasticity. In this case, the model will be accurate in some parts of the data only.
- The last assumption is that X has to be measured without error or values of X are fixed by the experimenter. This assumption must be met if the parameter values are of interest; if the regression is purely descriptive, or is being used for prediction, then this assumption can be disregarded.

B. Errors

The goal of linear regression is to fit a linear plot across the input variables and it does so by checking the error terms of multiple lines and trying to improve the model by updating the weights. Hence, correctly measuring the error term is one of the most crucial steps of Linear Regression. The most used error terms are mentioned first in the list of error terms given below:

1) Root Mean Squared Error (RMSE) = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{pred} - y_{actual})^2}$
2) Mean Squared Error (MSE) = $\frac{1}{n}\sum_{i=1}^{n}(y_{pred} - y_{actual})^2$
3) Mean Absolute Error (MAE) = $\frac{1}{n}\sum_{i=1}^{n}|y_{pred} - y_{actual}|$
4) Mean Average Percent Error (MAPE) = $\frac{100\%}{n}\sum_{i=1}^{n}(\frac{y_{actual} - y_{pred}}{y_{actual}})^2$

5) Mean Percent Error (MPE) = $\frac{100\%}{n}\sum_{i=1}^{n}(\frac{y_{actual}-y_{pred}}{y_{actual}})$

Another metric used to evaluate the model's performance is coefficient of determination R^2 which is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{actual} - y_{pred})^2}{\sum_{i=1}^{n}(y_{actual} - y_{mean})^2}$$

The values of R^2 lie between -1 and 1. If the R^2 of the linear regression model is closer to 1, then it is able to explain a higher proportion of the variance in y and consequently performs better in estimating y.

C. Model Parameters

There are two broad types of linear regression: simple and multiple regression.

- In simple regression, there is one input variable X and two parameters β_0 and β_1 that estimate the value of y.
- In multiple regression, there are multiple input variables X_i, each of which have a β_i scaling factor. The scaling factors and the constant β_0 along with the X_is estimate the value of y.

The coefficients/scaling factors represent the slope of the line for a particular variable in the hyperplane. Each coefficient for the corresponding input variable indicates the correlation between the input and the output variables. If the coefficient β_i is positive, the input variable X_i is positively correlated with the output variable, i.e. with an increase in X_i, the value of the predicted value Y increases as well. If the coefficient β_i is zero, it means that the particular input variable is not correlated with the output.

III. Does the socioeconomic status determine cancer risk?

In this section, we will visualize and analyze the Cancer risk data provided in the assignment to determine the impact of poverty/ average income on the incidence and mortality because of cancer. This will be done using the principles of linear regression mentioned in the section above. Some important terms that will be used in the analysis:

1) Avg_Ann_Incidence: defined as the number of cancer cases recorded in a year per 100,000 people at risk. The population is taken to be the denominator assuming that cancer is not related to age.
2) Avg_Ann_Deaths: defined as the average number of death because of cancer in a particular county.
3) MedIncome: The median of the incomes in a particular county.
4) All_With: defined as the number of people who have a health insurance in a particular county.

A. Preparing the data for analysis

Data collection is rarely perfect and since this data has been collected from thousands of counties across the United States, some pre-processing has to be done before visualising the data. In the given dataset, there were 24 columns out of which 20 columns were redundant and therefore are not required for the analysis. For example, the columns with information for males and females separately are not required because they do not affect the socioeconomic status directly. The AreaName and the county code, given by the columns FIPS, fips_x and fips_y, the income for different races are not helpful for our analysis too. We choose to the average annual values for incidence and deaths instead of the mortality and the incidence rates because they are a better measure to be considered. After the data processing, we will have 4 columns left to be visualized and analysed using linear regression.

B. Visualisation

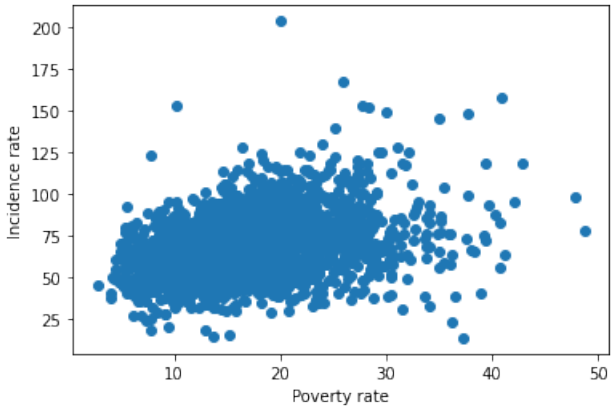

Fig. 1. Plot of average annual incidence rate against poverty rate.

First, we obtain the population of each county by adding numbers of people with and without health insurance. Poverty rate is calculated by dividing number of people in poverty by population size. The picture itself shows a positive correlation between these two variables which the model will confirm. Second, we plot the cancer mortality rate against poverty rate. Again, there is a positive correlation between these two variables which the model will confirm.

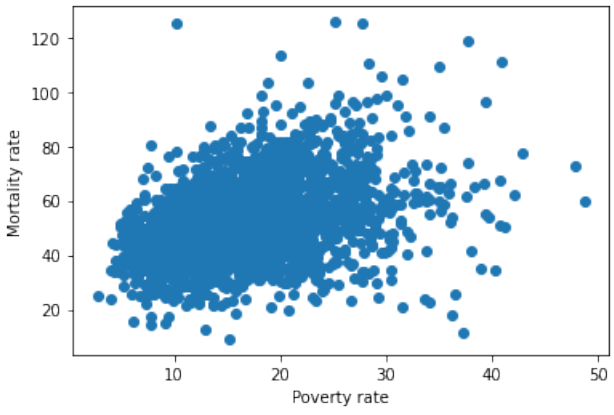

Fig. 2. Plot of average annual cancer mortality rate against poverty rate.

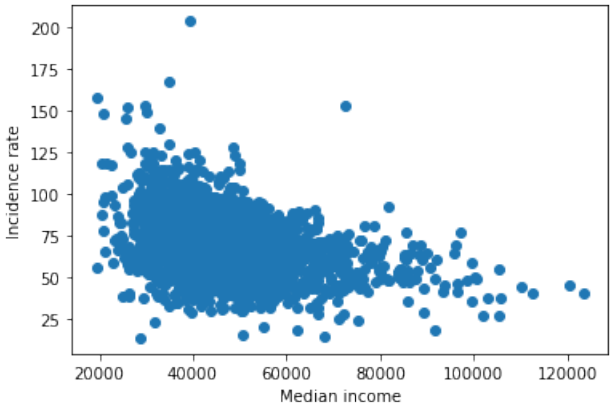

Fig. 3. Plot of average annual cancer incidence rate against median income.

The plot of average annual cancer incidence rate against median income shows clear negative correlation and so does the plot between average annual cancer mortality rate and median income (not shown here).

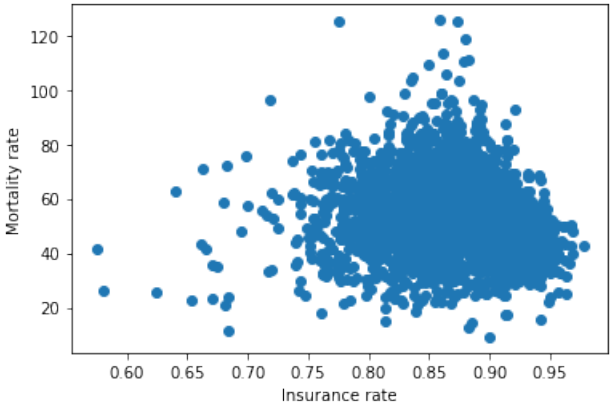

Fig. 4. Plot of health insurance rate against cancer mortality rate.

The plots of cancer mortality rate against health insurance rate (Fig. 4) and cancer incidence rate against health insurance rate (not shown) do not show a clear correlation between these variables.

The visual data shows that the poor people are likely to get cancer or die because of it. This can be seen through the positive correlation between poverty rate and incidence rates and the negative correlation between median incomes and the incidence rates. Now we have to show it with statistical proof using a **linear regression model**.

C. The linear regression model

A simple linear regression model was used with median income as X and average annual mortality rate as y. The downward sloping lines shows a negative coefficient for X with b1 = -0.00049949. Another simple linear regression model was used with median income as X and average annual incidence rate as y. The downward sloping lines shows a negative coefficient for X with b1 = -0.00053473. From the quantitative results obtained and the previous visualisation plots, we can see that the distribution is not really homoscedastic. As discussed, this can really hinder the performance of a linear regression model.

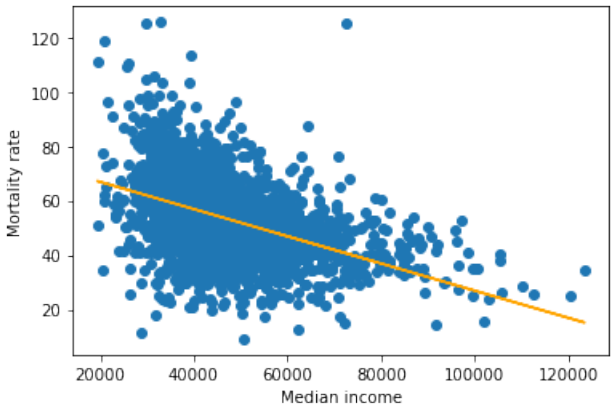

Fig. 5. Plot of median income against cancer mortality rate.

IV. Conclusion

While the model's performance was successfully demonstrated, both visually and experimentally, that there exists a negative correlation between incidence rate and median income, and between mortality rate and median income. The results of the linear regression model are not satisfactory so the relationships in question are not, strictly speaking, linear. This can be attributed to flaws in the data collection or in the model.

Plots may also suggest the need to take into account additional features that we have avoided in our analysis.
For the purposes of fundraising for the NGO, it can be argued that visual proof is more compelling than statistical evidence. Hence, while we can discuss the nature of the relationship between income and rates of incidence or mortality, it is clear that the mortality and incidence rates decrease with an increase in income. The figures make a compelling case for the health authorities to prioritise improving the health standard of the poorer section of the American society.

References

[1] "State Cancer Profiles", Statecancerprofiles.cancer.gov, 2021. [Online]. Available: **Link**.