Assignment 1: a mathematical essay on linear regression

Krishna Sah Teli
Cyber-Physical Systems (I2MP)
Indian Institute of Technology
Madras - 600036, India
ge22m018@smail.iitm.ac.in

***Abstract*—In this paper, linear regression has been implemented to depict the correlation among the different variables and to extract the insight from the data. It shows how the death of people is related to their socioeconomic status. Interestingly, it has been found out that about 67% people died of cancer are below the poverty line and 72% do not have health insurance. Health status of population from the different states of America have been compared to analyse how low-income groups are at greater risk for being diagnosed and dying from cancer. It also illustrates how income varies from black group to white group of people and American to Asian population. In this revised paper, I have worked to improve abstract and have more focused on data analysis and visualization under Data Set section. I have also added one more section modeling.**

***Index Terms*—Linear regression, Socioeconomic, Cancer, Health Insurance, Poverty, Analyse**

I. Introduction

In our daily life, knowingly or unknowingly we do predict or estimate something based on the probable factors which correlate the different data stored in our mind. From the data science point of view, everything can be represented in data and can be processed on it. Based on the data and its processing methodology, profound insight can be extracted including estimation, prediction, recognition, etc. from it. Now a day, data science has been incorporated in various fields from computing, biology, to astrology [10]. Similarly, this paper shows how the data is helpful to a non-profit consulting to advocate for the better health outcomes for low-income population of United States. To analyse and visualize the data, [1] linear regression has been used which is a statistical technique used for estimating the correlation between the variables.

The main focus in this paper is to demonstrate how the population of United States who is suffering and dying from cancer and to find out the reason and way to improve their health status. It correlates the health status of population with their socioeconomic status. With the help of data analysis and visualization techniques, health insurance with respect to poverty,and income, cancer diagnosed population has been studied and analysed.

To build an estimation model, a linear regression supervised machine learning has been implemented. Linear regression has been an easy statistical technique to illustrate the correlation among the variables. It builds an equation which relates the dependent and independent variables. It has been used widely in various fields such as e-commerce, IT-fields [9], banking area, scientific studies, forecasting, biology [10], etc.

With help of linear regression, it has been found out that there is strong correlation between the population under the poverty line and their health status. The mortality of population of low-income is significantly higher. About 72% of the population don't have health insurance and about 23% of the population diagnosed and dying from cancer have health insurance.

The rest of the paper has been organized as follows: section II outlines the data set related to our problem, section III explains the mathematics behind the linear regression and tools used for developing our model. Section IV is Findings and section V is Modeling which explains about the model and conclusion concludes the paper with key contributions made and directions for future works.

II. Data set

The data set has been taken from the American Community Survey(ACS)'s survey data on poverty, income, health insurance and cancer.(Citation) By preprocessing and feature engineering techniques, raw data can be converted into meaningful data and later used for training a linear regression model. The dataset contains a set of socioeconomic features that can be divided into different groups of columns. The first group is associated to poverty, containing the count of male, female and total population below poverty line. Similarly, another group of columns is associated to income containing median income of all ethnic groups, white people, black people, native Americans, Hispanics and Asians. The third group of column is associated to health insurance, containing the total count of male, female and sum total that have/don't have health insurance. Finally, the last group of columns consists of target variable related to count of cancer incidence cases and deaths resulting to it. Here, only the relevant columns from these first three groups are selected as the independent variables and from the last group, the column that gives average count is the dependent/target variable is selected for predicting model.

With the data analysis and data visualization tools, the correlation among the data features has been studied and correlated to the each other and target class. From the correlation chat, it has been found that there are some attributes such as Incidence_rate, mortality_rate, Hispanic, Med_income_assian and Med_income_nat_ann don't have significant correlation with average annual incidence or poverty. However, the features such as poverty, insurance and annual incidence have great effect on average annual death. From fig.1, we can see that how the average annual death differ from one state to another. Here it can be seen that CA state of America has highest average annual death while DC has least. It may be because DC is the capital city and CA is backward region.

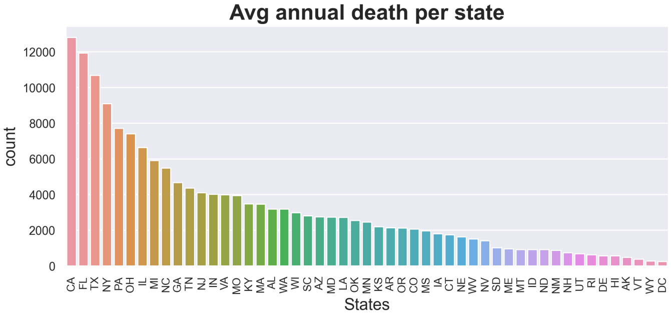

Fig. 1. Average annual death in different states

Similarly, proportion of female poverty is higher than that of male poverty. Thus, the average annual death count also can be higher. The degree of poverty among the female and male population can be seen in the fig.2. It also depicts that the poverty has significant impact on average annual death count.

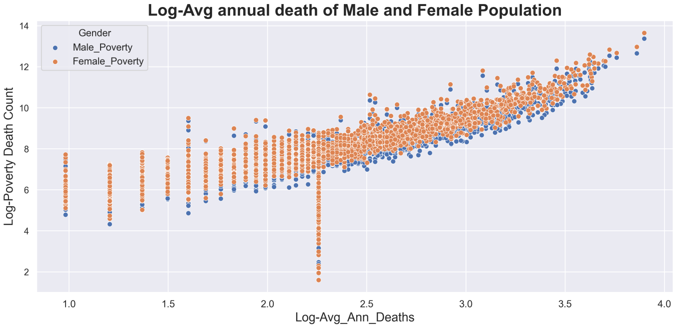

Fig. 2. Visualization of poverty

From income point of view, Asian population, white American have good average income where as the black people and Hispanic have lower average income. However, Hispanic group have not significant correlation with annual death which signifies that they are good in health status despite of low income. Similarly, data visualization shows that the number of diagnosed population among black race is higher despite of good income as shown in fig.3. It may be because, they may not be giving care to their health and life style. However, the white people who is diagnosed with cancer having low income. It shows they are suffering because of poor economic status.

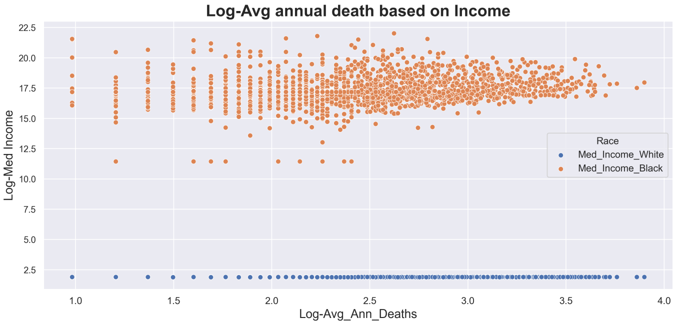

Fig. 3. Income Vs Death based on races

Analysis as shown in fig.4 illustrates male population hold more insurance than female population. However, the large number of diagnosed population is population not holding insurance in both cases i.e male and female group. It may be because large number of female population don't work in professional area. Hence they do miss insurance where as male do have as health insurance is mandatory in various companies. Interestingly, the fig.4 shows, regardless of insurance, average annual death is increasing with increase in total diagnosed population.

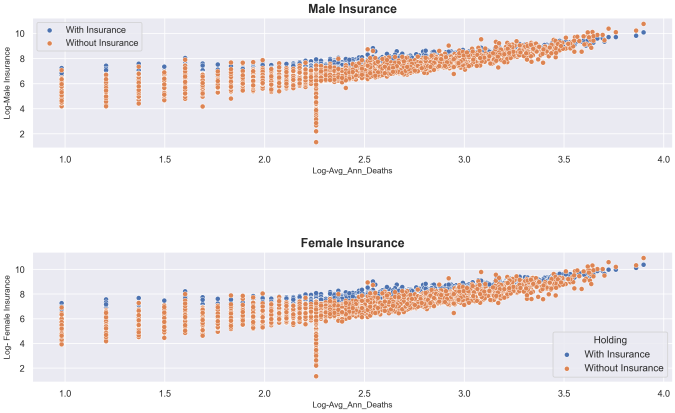

Fig. 4. Insurance Visualization

As we can extract insight from fig. 5 the impact of poverty on annual incidence and annual death. It shows that with increase in total poverty population, the death rate as well as cancer diagnosis increase. Hence, just by controlling the poverty, average annual death can be controlled significantly.

III. Linear Regression

Regression Analysis is a statistical technique of constructing mathematical models to estimate relationships existing between the response variable Y and the predictor variables X. Linear Regression is one of the most common regression analysis which assumes a linear relationship between one or more independent variables and a dependent variable. Based on the number of independent variables, linear regression is categorized into Simple Linear Regression and Multiple Linear Regression.

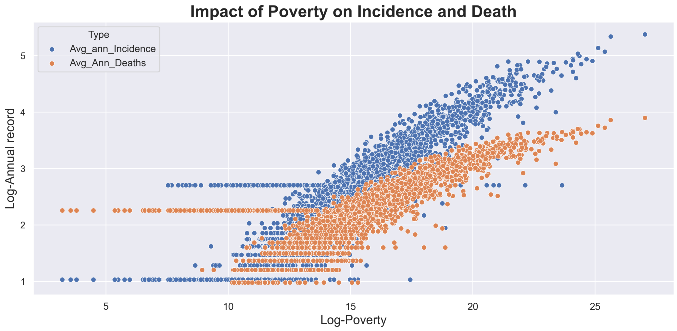

Fig. 5. Analysis of impact of poverty on diagnosis and death

Simple Linear Regression consists of single independent variable x and a dependent variable y in the form $y = \beta 0 + \beta 1 x$, where $\beta 0$ is the y-intercept or bias i.e. value of y when x=0 and $\beta 1$ is the coefficient of x or the slope of regression line. Likewise, Multiple Linear Regression generalizes simple linear regression by allowing more than one independent variable with the dependent variable in the form $y = \beta 0 + \beta 1 x 1 + \beta 2 x 2 + ... + \beta \mathrm{n} x \mathrm{m}$. Here $\beta 1 ... \beta \mathrm{m}$ are the estimated regression coefficients for independent variables $x 1 ... x \mathrm{m}$.

In real life, while modelling a problem with the linear regression approach, the linear regression line which we approximate does not always pass through all the data points, but tries to incorporate as much data points as possible. As the independent variables are estimator of the dependent variable and never the perfect predictors, there is always an error associated with the estimate of the output made by the linear regression. This error is called random error ϵ, which is the difference between the actual response value(y) and the predicted response value ($\hat{y}$). Mathematically, $\hat{y} = \beta 0 + \beta 1 x 1 + \beta 2 x 2 + ... + \beta \mathrm{m} x \mathrm{m}$ and y = $\hat{y} + \epsilon$. We can calculate the error ϵ as, $\epsilon = y - \hat{y}$. Our goal is to reduce the summation of the squared errors(ϵ) in each iteration, which is also considered as one of the loss function of linear regression, called sum of squared of the errors(SSE). SSE calculates the minimum amount of variance that cannot be described by the linear model. Our task is to estimate the regression coefficients βs. For n set of observations, we can write the equations as:

$$y_1 = \beta 0 + \beta 1 x 11 + \beta 2 x 12 + \cdots + \beta \mathrm{m} x 1\mathrm{m} + \epsilon 1$$

$$y_2 = \beta 0 + \beta 1 x 21 + \beta 2 x 22 + \cdots + \beta \mathrm{m} x 2\mathrm{m} + \epsilon 2$$

...

...

$$y_n = \beta 0 + \beta 1 x \mathrm{n} 1 + \beta 2 x \mathrm{n} 2 + \cdots + \beta \mathrm{m} x \mathrm{nm} + \epsilon \mathrm{n}$$

x_{mn} is the m_{th} observation for nth feature or independent variable. These m set of equations can be written in matrix form as:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & . & . & x_{1m} \\ 1 & x_{21} & . & . & x_{2m} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{n1} & . & . & x_{nm} \end{bmatrix} \tag{1}$$

where y is a $n \times 1$ column matrix with n output variables. X is a $n \times (m + 1)$ matrix. So, there are $m + 1$ unknown parameters to be determined. Our objective is to find a column matrix or a column vector, β that minimizes the loss function SSE. This leads to the optimization problem:

$$minSSE = \sum_{n=1}^{\infty} (y_i - \hat{y}_i)^2$$

$$minSSE = \sum_{n=1}^{\infty} (y_i - X_i \beta)^2$$

$$minSSE = \sum_{n=1}^{\infty} (\epsilon_i)^2$$

Since,

$$\epsilon_i^T \epsilon = \sum_{n=1}^{\infty} (\epsilon_i)^2$$

by property. Using equation (1),

$$= \epsilon_i^T \epsilon$$

$$= (y - X\beta)^T (y - X\beta)$$

This positive quadratic error function or objective function is always a convex surface facing upwards. The value of parameters at the minimum point is obtained by setting the first derivative of the objective function, with respect to the parameters, equal to 0. The partial derivative of the objective function, with respect to β, results the value for the column matrix, $\beta = (X^T X)^{-1} X^T y$.

This normal equation is the solution to the unknown parameters in linear regression. Since the parameters are estimates, the normal equation to determine estimated parameters is:

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

Our task is to find the value of these parameters which will find us the best fit linear regression line.

IV. Findings

After pre-processing followed by augmentation, the data was cleaned and necessary columns are selected and visualized using seaborn, matplotlib and pandas in order to extract meaningful and insight. It has been found out poverty plays crucial role in causing the people suffering from cancer or other diseases. Not only the population below the poverty line is suffering, but also mediocre and rich people who have good income do suffer from cancer. Out of the population diagnosed or dying from cancer, about 72% of them do not have health insurance. The female population number is found comparatively more than male population overall as well as in case of below poverty line and not holding insurance. The female population is diagnosed more, may be because breast cancer is more common and frequent in America among the female.

V. Modeling

Regression Analysis is a statistical technique of constructing mathematical models to estimate relationships existing between the response variable Y and the predictor variables X. Linear Regression is one of the most common regression analysis which assumes a linear relationship between one or more independent variables and a dependent variable. Here average annual death has been taken as target variable and other numerical variables have been taken as independent variables. I have already removed some dependent variables such as total_poverty, total_with, total_without because they can be derived by summation of other some variables. By plotting the correlation matrix, still some variables are not strongly correlated with target variables such as Incidence_rate, mortality_rate, Hispanic, Med_income_assian and Med_income_nat_ann. SO, two different linear regression models were trained and their least square errors were compared. Contradictorily, it has been found the LSE of the model trained with all numerical variables is 9.27% with training dataset and 9.17% with testing dataset and LSE of the model trained with selected variables is 10.65% with training dataset and 10.26% with testing dataset.

Acknowledgement

It is my pleasure to express with deep sense of gratitude to Professor Gaurav Raina, for his guidance, continual encouragement, understanding.

I would like to extend my gratitude to IIT Madras, for providing with an environment to work in and for his inspiration during the tenure of the course.

Conclusion

In this paper, linear regression has been used to get insight from the given data set. It has been applied in order to find out the correlation among the variables and cause-effect. It has been depicted from the dataset is that more female population is diagnosed and dying from cancer than male population. It is more frequent in the population under the poverty line. So in order to boost the health status of American population, government has to focus on the life style of poor people. Data also shows that health insurance does not play crucial role as low-income population cannot afford for health insurance. This paper also guides the new enthusiastic data science researcher to find insight in linear regression.

References

[1] Uyanık, Gülden Kaya and Güler, Neşe, ”A study on multiple linear regression analysis,” Procedia-Social and Behavioral Sciences, Elsevier, vol. 106, pp. 234–240, 2013.

[2] Yn, Xin and Su, Xiao Gang, ” Theory and Computing, World Scientific, 2003.

[3] Dhar, Vasant, ”Data science and prediction,” Communications of the ACM, vol. 56, pp. 64–73, nn. 12, 2013.

[4] Seber, George AF and Lee, Alan J, ”Linear regression analysis,” John Wiley & Sons, 2012.

[5] Montgomery, Douglas C and Peck, Elizabeth A and Vining, G Geoffrey,”Introduction to linear regression analysis,” John Wiley & Sons, 2021.

[6] Aiken, Leona S and West, Stephen G and Pitts, Steven C, ”Multiple linear regression,” Wiley Online Library, Handbook of psychology, pp. 481–507, 2003.

[7] Andrews, David F,”A robust method for multiple linear regression,” Technometrics, Taylor & Francis, vol. 16, pp. 523–531, nn. 4, 1974.

[8] Hayes, Andrew F and Montoya, Amanda K, ”A tutorial on testing, visualizing, and probing an interaction involving a multicategorical variable in linear regression analysis,” Communication Methods and Measures, Taylor & Francis, vol. 11, pp. 1–30, nn. 1, 2017.

[9] Abuella, Mohamed and Chowdhury, Badrul, ”Solar power probabilistic forecasting by using multiple linear regression analysis,” SoutheastCon 2015, IEEE, pp. 1–5, 2015.

[10] Liu, Yingxia and Heuvelink, Gerard BM and Bai, Zhanguo and He, Ping and Xu, Xinpeng and Ding, Wencheng and Huang, Shaohui, ”Analysis of spatio-temporal variation of crop yield in China using stepwise multiple linear regression,” Field Crops Research, Elsevier, vol. 264, pp. 108098, 2021.

Source code link: https://drive.google.com/drive/folders/1GFvOuibiN4djuAhC6dC3GpfWENTnF2EI?usp=share_link