International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

Heart Disease Analysis using Multiple Linear Regression

Adithya Mohanavel
Department of Computer Science,
B.S.A. Crescent Institute of Science and Technology
Chennai, Tamil Nadu

Joy Mathew
Department of Electrical and Instrumentation,
B.S.A. Crescent Institute of Science and Technology
Chennai, Tamil Nadu

***Abstract*— Smoking has been identified as the major cause of cardiovascular disease, ever since this fact was scientifically proven numerous activities have been pointed out to be life-saving as to avoid cardiovascular disease for men and women smokers alike. In this paper, we have explained why cycling is a better way to avoid cardiovascular disease. To prove our hypothesis, we surveyed 500 smokers and recorded their healthcare data. The paper continues by constructing a mathematical relationship using multiple linear regression and statistically proves our hypothesis thus effectively making it a theory.**

Keywords—Linear regression; Multiple Linear Regression; Machine Learning; Heart Disease; Smoking Data; Cycling Data;

I. INTRODUCTION

Smoking is one of the most addictive activities to act ever since humans discovered the relaxing effects of smoking tobacco, though the act provides relaxation it is never recommended to smoke, with the increased chances of cancer and heart-related diseases the loss clearly outweighs the gain. Ever since the ill effects of smoking were scientifically proven, government and non-profit organizations alike have been rallying to create awareness about the negative effects of smoking. More research has been poured into finding a way to stop this addiction but has not been very successful to date. Smokers find it extremely difficult to overcome the grip of nicotine partly because of biology, in extreme cases the fight to quit smoking becomes almost impossible as once the human body starts to intake too much nicotine and becomes comfortable to work with nicotine in the system quitting smoking is now not the matter of will power but biology. Going cold turkey will lead to withdrawal symptoms which in return might harm the person furthermore. Until there is a scientific breakthrough to stop nicotine addiction, we should focus on reducing the ill effects of smoking in hopes of prolonging the life of smokers. In this paper, we will be focusing on reducing heart disease more specifically cardiovascular disease (cardiovascular disease) that is associated with smoking. According to the CDC (centers for disease control and prevention), one person dies every 36 seconds in the United States due to cardiovascular disease that is about 655000 deaths in a year. To sum it up 1 in 4 deaths is related to heart disease, the number becomes even more haunting when all the countries are considered. Surely all is not lost yet as it's important to know that cardiovascular disease can be prevented if we lead a healthy lifestyle. In this paper, we have identified cycling as one of the most common and easy exercises to significantly reduce the chances of getting a cardiovascular disease. Before we move further it's important to note that cycling does not prevent smokers from the possibility of obtaining cancer and in no way do, we support or endorse the act of smoking.

Currently, with mass media intervening in this matter to spread awareness there are tons of information that suggest various types of activities that can reduce this tragedy as cardiovascular disease is identified as the leading of all deaths in the United States.

Though mass media does a good job in creating awareness about a cause it also does a far better job in spreading misinformation, so we have composed this research article to show what activity does help in reducing the risk of cardiovascular disease. Through our research, we have identified cycling as the simplest and effective way of reducing the risk of contracting cardiovascular disease. We were able to arrive at this conclusion by obtaining raw health data from 500 participants. We processed the raw data by creating a mathematical relationship using linear regression and plotted out the data where activity is considered independent and heart disease as dependent data.

II. RELATED WORKS

Dinesh Bhuriya and his team [1] have used linear regression to predict stock prices values and have gone further by proving its accuracy is better than that of polynomial and RBF regression.

Imran Naseem and his team [2] have used linear regression to improve face recognition technology, they've taken samples from a specific object class that lies on linear subspace. Using this concept they developed class-specific models of users by using downsampled images, thus successfully defining face recognition as a problem of linear regression.

Andrew Grøntved and his Swedish research team [3] have concluded that bicycling can be adopted as an effective strategy to avoid cardiovascular disease among middle-aged men and women alike by surveying 23732 people twice over 10 years.

Solveig Nordengen and his team [4] Concluded that any form of cycling is associated with lowering the risk of CARDIOVASCULAR DISEASE

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

Derek C.Monroe and his team [5] have proved that cycling after smoking significantly reduces mood disturbances.

P. Oja and his team [6] conducted a systematic review on 16 cycling-specific studies and reaffirmed the fact that cycling improves cardiovascular health thus effectively reducing the risk of contracting CARDIOVASCULAR DISEASE.

III. PROPOSED WORKING METHODOLOGY

The proposed working methodology makes use of an advanced statistics principle called linear regression. In layman's terms, linear regression can be simplified as a way of plotting and representing a set of data to predict future values. In the Multiple Linear regression model, 2 data must be independent of each other and both the data must be dependent on the third data to be efficient.

This method was chosen because of its ability to model the relationship between scalar response and numerous explanatory variables. The data set that is being used in this analysis has 3 variables. Thus, multiple linear regression can be applied to those data to get an insight and analyze it further.
Here the activity under observation (cycling) is considered as independent data and heart disease as dependent data.

A. Dataset

The dataset contains observations on the percentage of people biking to work each day, the percentage of people smoking, and the percentage of people with heart disease in an imaginary sample of 500 towns.

In this data, the percentage of people who smoke and the percentage of people who ride a bicycle are independent of each other. The percentage of people who smoke is dependent on both percentages of people who smoke and the percentage of people who do bicycling.

To apply multiple linear regression into this data, initially, it should be made sure that the data is normally distributed and the percentage of people who smoke and take bicycle should be linear concerning the data of the percentage of people who has heart disease.

B. Applying regression to the data

Once the above conditions of normality and linearity are verified then the data is ready to apply the linear regression.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

Fig 1: Multiple Linear Equation formula.

The data is then applied to multiple linear regression using the formula given in Fig 1. The percentage of the population who has heart disease will be the independent variable which is given as "X".

Once the formula is applied and the data is plotted then an insightful conclusion can be made between the relationship between the data of the population of people who have heart disease, smoke, and take a bicycle.

IV. IMPLEMENTATION ALGORITHM

A. Step 1:

Initially started by installing the required packages and libraries.

B. Step 2

Loading the sample dataset into R studio for prediction.

C. Step 3: Independence of observations

Test the relationship between independent variables biking and smoking to make sure they aren't too highly correlated.

D. Step 4 Normality

The next step is to check for normality to test whether the dependent variable follows a normal distribution.

E. Step 5 Linearity

Plotting the data in a graph to check the linearity between them. Later, linear relationships between biking to work, smoking, and heart disease are calculated in the dataset.

F. Step 7:

Before proceeding with data visualization, it should be made sure, that models fit the homoscedasticity assumption of the linear model.

G. Step 8

Creating a new data frame with the information needed to plot the model. It will create a sequence from the lowest to the highest value of our observed biking data.

H. Step 9

Predict the values of heart disease based on a linear model and save. our 'predicted y' values as a new column in the dataset we just created.

I. Step 11

Changing the 'smoking' variable into a factor allows for plotting the interaction between biking and heart disease at each of the three levels of smoking.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

J. Step 12

Plotting the original data as a graph and adding regression line to the plotted data. Annotating and adding other data to make it understandable.

V. OUTPUT AND OBSERVATIONS

A. Independence of observations

In order to plot data in a multiple linear regression algorithm, the data must be independent of each other.

```
> cor(heart$biking, heart$smoking)
[1] 0.01513618
>
```

Fig 2: Value of independence between biking and smoking

Through the observation of the results from Fig 2, the correlation between biking and smoking is small (0.015 is only a 1.5% correlation), so we can include both parameters in our model.

B. Normality of data

The dependent variable in Linear regression should follow a normal distribution.

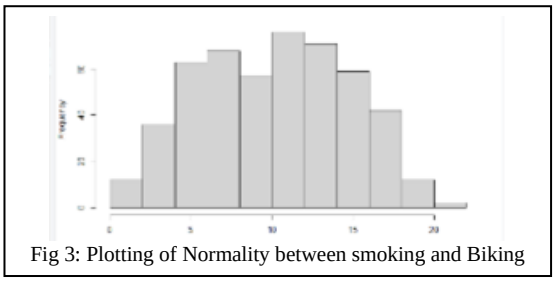

Fig 3: Plotting of Normality between smoking and Biking

The distribution of observations is roughly bell-shaped, so the data can proceed with the linear regression.

C. Linearity

Linearity can be checked here using two scatterplots: one for biking and heart disease, and one for smoking and heart disease.

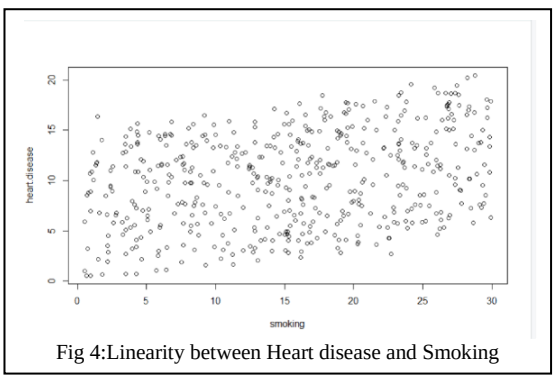

Fig 4:Linearity between Heart disease and Smoking

Although the relationship between smoking and heart disease in fig 4 is a bit less clear, it still appears linear. We can proceed with linear regression.

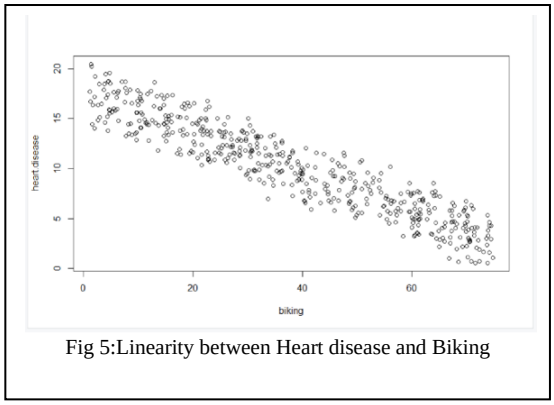

Fig 5:Linearity between Heart disease and Biking

From the further observation of both the graph fig 4and fig 5, the estimated effect of biking on heart disease is -0.2 and the estimated effect of smoking on heart disease is 0.178.

D. Homoscedasticity

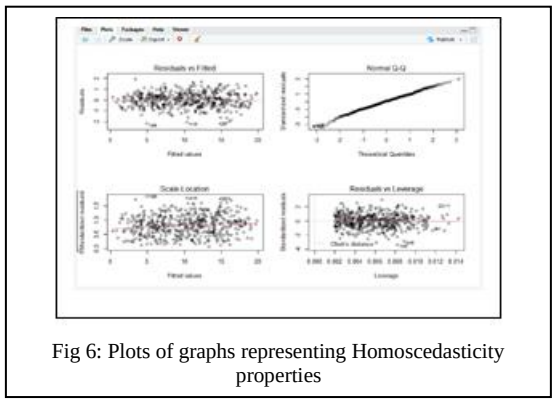

Fig 6: Plots of graphs representing Homoscedasticity properties

The residuals show no bias, so it is clear that the model fits the assumption of homoscedasticity.

E. Final graph of Multiple Linear regression

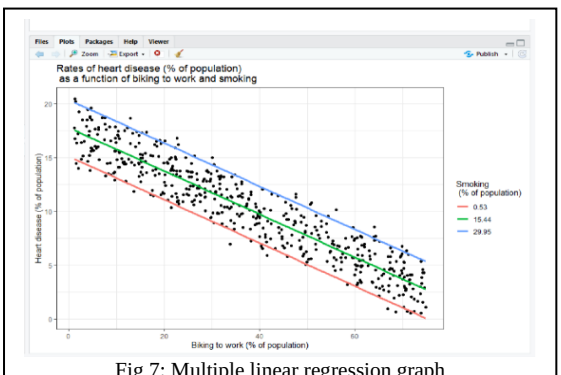

Fig 7: Multiple linear regression graph

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

Thus, with the final graph of the regression model from fig 7, it can be seen that the data is plotted in 3 ways. That is heart disease and Smoking data as Y1 and Y2 respectively. Biking to work data is plotted as X-axis.

From this, it is found a 0.2% decrease (± 0.0014) in the frequency of heart disease for every 1% increase in biking.
It also found a 0.178% increase (± 0.0035) in the frequency of heart disease for every 1% increase in smoking.

VI. CONCLUSION

Thus, from the sample data and applied Multiple linear regression it is clear that the risk of heart disease increases with smoking and decreases with any form of physical exercise.

Our modern world operates on information, all our policies and statements are backed up by precise statistics, unlike the older days when values were arbitrary. Looking at the devastating effects that cardiovascular disease has been having on the smoking population both men and women alike, although it is widely regarded by doctors to be easily avoidable. We believe it is due to the ocean of misinformation that is available online people are having a hard time avoiding it.

To avoid the spread of misinformation we have carefully studied and concluded what activity is more effective in reducing the chances of cardiovascular disease using statistical methods, linear regression in particular. By collecting a wide variety of data and testing it out using linear regression we were able to conclude that cycling is the most effective way to reduce the risks. It is important to note that by concluding the positive effects cycling has had on smokers we in no way endorse or support smoking. This paper doesn't account for the possibility of smokers contracting cancer.

REFERENCES

[1] Bhuriya, Dinesh & Kaushal, Girish & Sharma, Ashish & Singh, Upendra. (2017). “Stock market prediction using a linear regression”,510-513.10.1109/ICECA.2017. 8212716.
[2] Imran Naseem, Roberto Togneri, Mohammed Bennamoun. “Linear regression for face recognition”, IEEE transactions on pattern analysis and machine intelligence. Volume 32,issue 11.
[3] Ried-Larsen, Mathias & Rasmussen, Martin & Blond, Kim & Overvad, Thure & Overvad, Kim & Steindorf, Karen & Katzke, Verena & Andersen, Julie & Petersen, Kristina & Aune, Dagfinn & Tsilidis, Kostas & Heath, Alicia & Papier, Keren & Panico, Salvatore & Masala, Giovanna & Pala, Valeria & Weiderpass, Elisabete & Freisling, Heinz & Bergmann, Manuela & Grøntved, Anders. (2021). Association of Cycling With All-Cause and Cardiovascular Disease Mortality Among Persons With Diabetes: The European Prospective Investigation Into Cancer and Nutrition (EPIC) Study. JAMA Internal Medicine. 181. 10.1001/jamainternmed.2021.3836.
[4] P. Oja, S. Titze, A. Bauman, B. de Geus, P. Krenn, B. Reger-Nash, T. Kohlberger,” Health benefits of cycling: a systematic review”, Scandinavian Journal of Medicine & Science in Sports, Volume 41, Issue 4.
[5] Nordengen S, Andersen LB, Solbraa AK, et alCycling is associated with a lower incidence of cardiovascular diseases and death: Part 1 – systematic review of cohort studies with meta-analysis British Journal of Sports Medicine 2019;53:870-878.
[6] Derek C. Monroe, Neil R. Patel, Kevin K. McCully, Rodney K. Dishman, “The effects of exercise on mood and prefrontal brain responses to emotional scenes in smokers”, Physiology & Behavior, Volume 213,2020,112721, ISSN 0031-9384.