Advances and Applications in Statistics

http://www.pphmj.com
http://dx.doi.org/10.17654/AS070020219
Volume 70, Number 2, 2021, Pages 219-227 **ISSN: 0972-3617**

FUZZY LINEAR REGRESSION: AN APPLICATION TO HEART DISEASE

A. M. C. H. Attanayake

Department of Statistics and Computer Science
University of Kelaniya
Sri Lanka

Abstract

Disorders in heart condition refer to heart disease. Several risk factors are associated with causing the heart disease. Physical inactivity and smoking are leading risk factors among other risk factors. The aim of this study is to investigate the relationship of heart disease with physical activity and smoking. Regression analysis is one of the key areas that can be utilized in finding the relationship of variables. By considering heart disease as the output variable (dependent variable) and correlated other factors as input variables, one can model the relationship through multiple linear regression. Fuzzy regression is an application of fuzzy platform for conventional regression analysis. Fuzzy regression analysis gives a fuzzy relationship between dependent and independent variables which represents vagueness in the data. The input data may be crisp values or fuzzy numbers whereas the conventional ordinary least squares regression can handle only crisp measures. The model output is in the form of fuzzy representative which has lower and upper approximation models to represent the fuzziness of the output. Fuzzy models are especially suitable in modelling and predicting heart disease as the disease

Received: August 23, 2021; Accepted: September 3, 2021
2020 Mathematics Subject Classification: 62J86.
Keywords and phrases: heart disease, fuzzy regression, possibilistic linear regression with least squares method.

associated with various unknown and uncontrollable factors. One of the fuzzy regressions, namely, possibilistic linear regression with least squares (PLRLS) method was applied in the study as the modelling procedure. The predicted values from the fuzzy regression model and the actual values of the validation data set were within the upper and lower approximation models, which indicated the possibility of the prediction of heart disease through PLRLS method.

1. Introduction

Heart disease is one of the leading diseases in the world which causes a large number of deaths annually. Disorders in heart condition refer to heart disease. Heart attack, coronary heart disease, congestive and congenital heart disease are the common types of heart disease [1]. Quitting smoking, reducing cholesterol levels, maintaining blood pressure, having a healthy diet, and engaging in exercises are considered as preventive measures for the disease. The chest pain is the most common symptom of the disease. It is important to reduce the risk factors in order to reduce the prevalence for the heart disease. The aim of this study is to investigate the relationship of heart disease with physical activity and smoking. The physical activity refers to movement in the body that expends energy. Common physical activities include walking, cycling (biking), doing sports, etc. One in five deaths in U.S. related to smoking [2]. Smoking is a bad habit that causes heart disease. Reducing physical inactivity and smoking will add years for lives. Regression analysis is one of the key areas in statistics that can be utilized in finding the relationship of heart disease with associated risk factors. The multiple linear regression develops the relationship using more than one independent variable. The coefficients of the regression model are usually estimated by using the ordinary least squares method which minimizes the error represented by the actual and fitted values of the output variable. Fuzzy linear models deal with vague and imprecise information in order to represent better models [3]. Fuzzy regression is an application of fuzzy platform for conventional regression analysis. Fuzzy regression analysis gives a fuzzy relationship between dependent and independent variables

which represents vagueness in the data. The input data may be crisp values or fuzzy numbers whereas the conventional ordinary least squares regression can handle only crisp measures. The model output is in the form of fuzzy representative which has lower and upper approximation models to represent the fuzziness of the output. Fuzzy models are especially suitable in modelling and predicting heart disease as the disease associated with various unknown and uncontrollable factors. Numerous fuzzy regression procedures are available in theory which would address input/output variables as fuzzy numbers or crisp measures. In this study, one of the fuzzy regression approaches: possibilistic linear regression with least squares (PLRLS) method was applied as the modelling procedure. Pavel and Jaroslav in [4] stated that fuzzy regression is an alternative efficient approach for conventional statistical regression and summarized theories behind various fuzzy regression approaches. Romero et al. in 2019 illustrated the advantages of fuzzy logic-based approaches in modelling vector-borne diseases [5]. This study analyzes the relationship of heart disease with physical activity (biking to work) and smoking using PLRLS method.

2. Materials and Methods

Data for the study consists of heart disease (% of population), biking to work (% of population) and smoking (% of population) in 498 cities. The data set is freely available in R software under name ‘heart’. Analysis is done using R software package [6]. Last 5% of the data (25 observations) were used to validate the PLRLS method and rest of the data for model development.

2.1. Fuzzy number

Fuzzy numbers can be defined on uncertainty situations and applicable in any scenario where imprecise information is involved. Suppose a real value number x belongs to the fuzzy set B, with a degree of membership which ranges from 0 to 1. The degrees of membership of x are defined by the membership function:

$$\mu_B(x) : x \to [0, 1], \text{ where } \mu_B(x^*) = 0 \quad (1)$$

means that the value of x^* is not included in the fuzzy number B while

$$\mu_B(x^*) = 1 \quad (2)$$

means that x^* is included in B.

2.2. Triangular fuzzy number (TFN)

TFN is a special class of fuzzy numbers which has three parameters. One of the parameters is for the central value where the degree of membership is equal to 1. The others are for left spread and right spread of the data. If the left spread is equal to the right spread, then it is defined as a symmetric triangular fuzzy number. A triangular fuzzy number is depicted in the following (Figure 1).

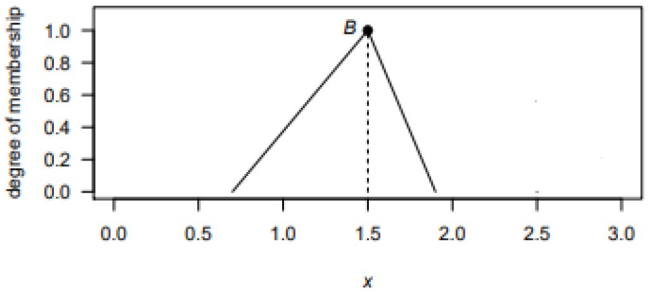

Figure 1. Triangular fuzzy number.

2.3. Fuzzy regression

Fuzzy regression is an application of fuzzy platform for conventional regression analysis. Fuzzy regression analysis gives a fuzzy relationship between dependent and independent variables where vagueness is present in the data [7]. The input data may be crisp values or fuzzy numbers whereas the conventional ordinary least squares regression can handle only crisp measures. When the output is in the form of fuzzy representative, then there is a chance to obtain lower and upper approximation models which represent the fuzziness of the output.

2.4. Possibilistic linear regression with least squares (PLRLS) method

This is one of the methods available under the fuzzy regression techniques. This method was proposed by Tanaka and Lee in 1999 [9] to deal with crisp inputs and fuzzy output. This method fits the model which compromises spreads and the central tendency by using the possibility and the least squares approach [8]. The input data represents crisp measures and the model output is in the form of non-symmetrical triangular fuzzy number.

The basic idea of the method is to minimize the fuzziness of the model by minimizing the total spread of the fuzzy coefficients subject to including all the given data. The least squares application minimizes the distance between the output and the actual observed output. The general fuzzy regression model is given as:

$$Y = A_0 + A_1 x_1 + \cdots + A_n x_n,$$

where Y is the fuzzy output, $A_i; i = 1, 2, \ldots, n$ are the fuzzy coefficients, and x_1 to x_n are non-fuzzy input variables. Fuzzy coefficients are assumed as triangular fuzzy numbers. The 'PLRLS' function available in 'fuzzyreg' package of R can be used to fit the method.

3. Results and Discussion

The relationship of heart disease (% of population) and biking to work (% of population) is depicted in Figure 2 whereas that is for smoking (% of population) is depicted in Figure 3. The graphical representations of variables show that there is a linear relationship between the variables.

Pearson's product-moment correlation of heart disease and biking to work is –0.935 (p value < 2.2e-16) and heart disease and smoking is 0.309 (p value < 1.72e-12). Therefore, there is a significant correlation between variables at 5% significance level and hence worth to model the variables. Density curve of the dependent variable (heart disease) is shown in Figure 4 and it is apparently normally distributed.

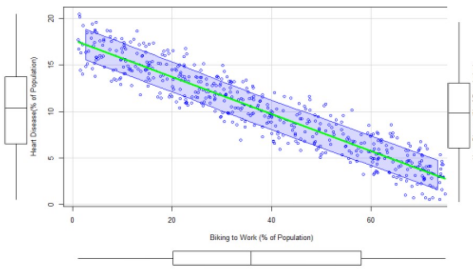

Figure 2. The scatter plot of heart disease and biking to work.

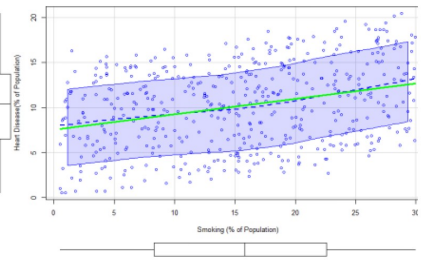

Figure 3. The scatter plot of heart disease and smoking.

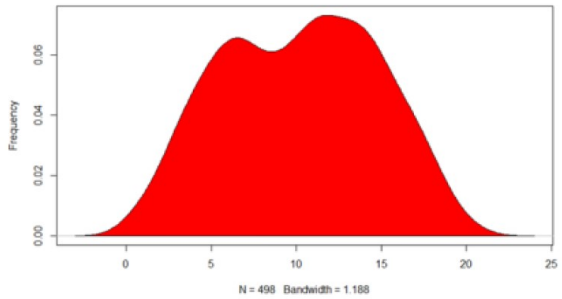

Figure 4. Density curve of the dependent variable.

```
Fuzzy linear model using the PLRLS method

Call:
fuzzylm(formula = heart_disease$heart.disease ~ heart_disease$biking +
    heart_disease$smoking)

Coefficients in form of non-symmetric triangular fuzzy numbers:

                          center   left.spread right.spread
(Intercept)           15.0431110  1.567829e+00  1.571012711
heart_disease$biking  -0.1973466 -1.192093e-07  0.003616328
heart_disease$smoking  0.1677251  2.754403e-02  0.000000000
```

Figure 5. Output of the PLRLS method.

The VIF values of biking to work and smoking are 1.0001 and 1.0001, respectively. Hence, there is no multicollinearity. Last 5% of the data (25 observations) were used to validate the PLRLS method and rest of the data for model development. The PLRLS model is provided in Figure 5. The central tendency with left and right spreads determines the support interval of predictions. Upper approximation model from the right spread and lower

approximation model from left spread can be obtained to reflect the fuzziness of the heart disease in cities.

Residual analyses of the center model of PLRLS model are shown in Figure 6 to Figure 8. Figure 6 shows that there is no heteroscedasticity in residuals. Further, Shapiro-Wilk normality test confirmed the normality of residuals at 5% significance level (p value = 0.2735).

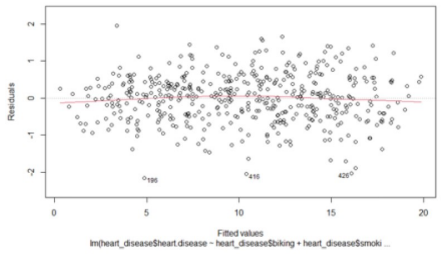

Figure 6. Residuals vs fitted values of the center model.

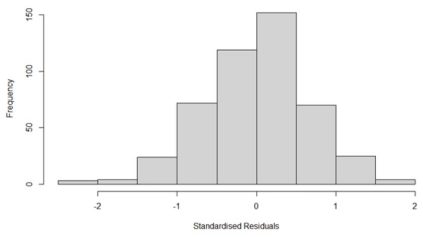

Figure 7. Histogram of residuals of the center model.

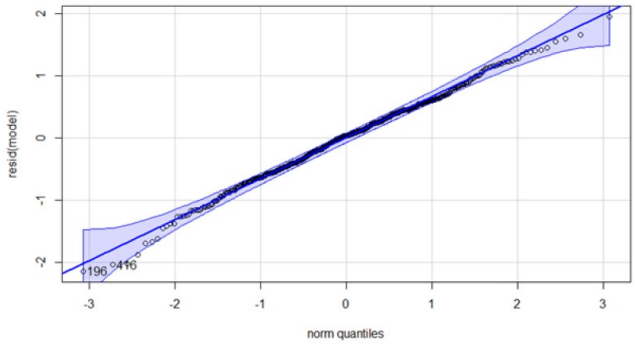

Figure 8. Probability plot of residuals of the center model.

Figure 9 shows the results of the validation set. The predicted values of the fuzzy regression model and actual reported frequency of heart disease are within the lower and upper approximation models indicating the possibility of prediction through the fitted fuzzy regression model. Therefore, fitted fuzzy regression model can be used to predict frequency of heart disease in cities.

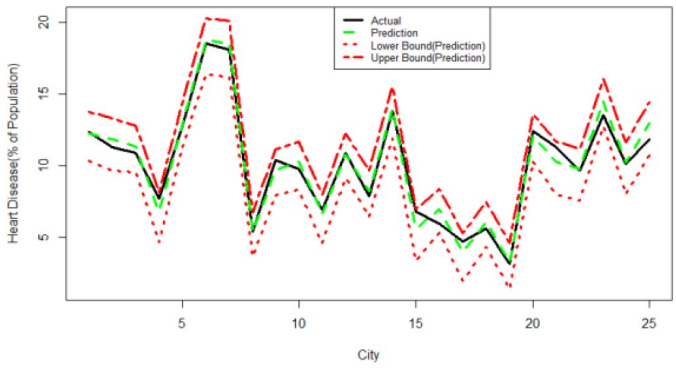

Figure 9. The validation plot.

4. Conclusion

The aim of this study was successfully achieved which was to analyze the relationship of heart disease with physical activity and smoking. One of the fuzzy regressions, namely, possibilistic linear regression with least squares (PLRLS) method was applied in the study as the modelling procedure. Last 5% of the data were used to validate the model. The predicted values from the fuzzy regression model and the actual values of the validation data set were within the upper and lower approximation models, indicating the possibility of the prediction of heart disease through PLRLS method. The upper and lower approximation models show the fuzziness of the center model which is useful in deciding possible deviations of the output (frequency of heart disease). Fuzzy regressions provide useful insights for the conventional regression procedures and applicable in any area where exact predictions are not essential.

References

[1] U.S. National Library of Medicine, Mediline Plus [Accessed: 2021 August 03]. https://medlineplus.gov/heartdiseases.html.

[2] Centers for Disease Control and Prevention, Smoking and Tobacco Use CDC [Accessed: 2021 April 20].

[3] N. Vilém, P. Irina and D. Antonín, Insight into Fuzzy Modeling, John Wiley & Sons, Inc., 2016, doi:10.1002/9781119193210.

[4] S. Pavel and M. Jaroslav, Models used in fuzzy linear regression, 17th Conference on Applied Mathematics, Slovak University of Technology, 2018, pp. 955-964.

[5] D. Romero, J. Olivero, R. Real and J. C. Guerrero, Applying fuzzy logic to assess the biogeographical risk of dengue in South America, Parasites and Vectors 12 (2019), Article number: 428, 13 pages.

[6] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2021. https://www.R-project.org/.

[7] C. Kahraman, A. Beşkese and F. T. Bozbura, Fuzzy regression approaches and applications, C. Kahraman, ed., Fuzzy Applications in Industrial Engineering, Studies in Fuzziness and Soft Computing, Vol. 201, Springer, Heidelberg, 2006.

[8] H. Tanaka and J. Watada, Possibilistic linear systems and their application to the linear regression model, Fuzzy Sets and Systems 27 (1988), 275-289.

[9] H. Tanaka and H. Lee, Interval regression analysis with polynomials and its similarity to rough sets concept, Fundamenta Informaticae 37 (1999), 71-87.