Analysis of heart disease using statistical techniques

E.Priyadarshini1, G.Raj Gayathri2, Samuel E Chakravarthy3, M.Vidhya4, and W.Abitha Memala5

1,2,4 Department of Mathematics, Sathyabama Institute of Science and Technology, Chennai, Tamilnadu, India
5 Department of EEE, Sathyabama Institute of Science and Technology, Chennai, Tamilnadu, India
3 Research Scholar, Icfai University, Ranchi, India

Email: priyadarshini.maths@sathyabama.ac.in

Abstract. Binary Logistic Regression analysis is one of the popular method used for prediction and forecasting of medical related data. The aim of this paper is to build a logistic model by identifying the various factors that trigger the heart disease of a person. The interaction effects of the various factors influencing the heart disease of a person was evaluated and the performance of the method were evaluated based on standard statistical measures such as MAPE, RMSE, etc. The results obtained shows that BLR model can be used to monitor the heart disease conditions satisfactorily.

1. Introduction
The term “heart disorder” is typically associated with cardiovascular disorder (CVD). The situations for CVD is caused by blocked or narrowed blood vessels which leads to chest ache, stroke or coronary heart attack. Other conditions including affection of heart’s muscle, rhythm, valves are also distinctive types of coronary heart disease. The signs and symptoms for coronary heart sickness are shortness of breath, pain within the throat, neck, jaw, upper abdomen, chest ache and tightness, angina and unusual heart beats. The seven major risk elements diagnosed for coronary heart disorder are age, gender, obesity, ex-people who smoke, BMI, dyspnea, regular chest ache, st-elevation and st-depression.

2. Literature Survey
Amir W et al[1] had identified the chance factors consisting of age, body mass index (BMI) and systolic to be related to high blood pressure by the use of logistic regression. Abedin [2] concluded that the logistic regression consequences about the hypothetical facts that elements such as age and smoking repute is associated with his/her CVD status. However, given the identical age, smokers are more likely to have CVD than non-people who smoke. Mythili T et al.[3] recommended a standard basically based model to assess the correctnesses of applying rules to the character results of help vector machine and logistic regression in Heart Disease Database at Cleveland to get an exact model for analysing coronary illness. Rahman et.Al (four) applied logistic regression to take a look at the chance elements and sociodemographic elements of stroke or myocardial infraction amongst hypertensive sufferers in Bangladesh.
3. Research Methodology

Binary Logistic Regression might be considered as a technique that is like Multiple Regression, with a difference that the categorical variable should be chosen as the dependent variable. Logistic Regression was created by Truett et al. (5) when he utilized multivariate investigation in the Framingham coronary examination. From that point forward, the Logistic Regression model has got a favored paired or dichotomous data examination in different territories of drugs. Data were arranged, cross examined and statistically investigated using PASW 18. Binary logistic regression analysis was performed with detailing of chances odds ratio proportion to set up the opportunity for heart disorder. The numerical thought of strategic relapse is to communicate the connection between conclusive yield variable (dependent variable) and indicator factors (independent factors) of logit function. Y is considered as a dependent variable taking values “1” and “0” and X can be considered as a continuous predictor variable. Logistic regression allows this situation with the aid of logit transformation at the final Y variable.

Binary logistic regression can be represented as

\[ \text{Logit}(y) = \ln \left( \frac{\pi}{1-\pi} \right) = \beta_0 + \beta_1 x \]  

(1)

Here \( \pi \) is the likelihood of event of the outcome Y and \( \pi/(1-\pi) \) is the chances of success; the proportion of the likelihood of occurring in the outcome Y and the possibility of not occurring in the outcome Y. By taking antilog on the two sides of condition (1), the occurrence of the chances of outcome Y for a given predictor X can be estimated.

\[ \pi = P \left( \frac{y}{x=x} \right) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \]  

(2)

The logistic model for more number of predictors is represented as,

\[ \text{Logit}(y) = \ln \left( \frac{\pi}{(\pi - 1)} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \]  

(3)

4. Results and Discussions

4.1. Logistic Regression Models

In this evaluation, logistic regression is used to construct a model by the usage of the interplay effects of the different factors influencing the coronary heart disease of a person and the overall performance of the method had been evaluated based on general statistical error measures consisting of MAPE, RMSE, and SSE. The major risk factors diagnosed for heart ailment are age, gender, obesity, excessive blood stress, ex-smokers, BMI, dyspnea, regular chest pain, st-elevation and st-melancholy. The outcome variable here is CVD status, whether affected person has normal or in stenosis (1=yes, 0=no)(Table 4.1)

| Step 1 | RCA | Predicted Percentage Correct |
|--------|-----|-----------------------------|
|        | RCA | normal | stenosis |               |
|        | normal | 152     | 37       | 80.4          |
|        | stenosis | 66     | 48       | 42.1          |
| Overall |     |         |          | 66.0          |

SPSS also provides two pseudo R² degree specifically Cox and Snell R² and Nagelkerke R² (Table 4.2). Both are variant of R² idea in simple linear regression but it is not the same. It is advised by researchers to document those information as an additive to the other goodness of fit measures.
Table 4.2 - Model Summary

| Step | -2 Log likelihood | Cox & Snell $R^2$ | Nagelkerke $R^2$ |
|------|-------------------|-------------------|-----------------|
| 1    | 369.365*          | 0.67              | 0.86            |

4.2 Assessing Goodness of Fit

Hosmer and Lemeshow Test is led by gathering the cases as risk factors of the deciles. (Table 4.3). The observed probability is contrasted along with the expected inside the deciles. Since $p$ is $0.112 > 0.05$, that there is no significant difference in the observed and expected values which implies that model fits the data well (6).

Table 4.3: Hosmer and Lemeshow Test

| Chi-square | df | Sig. |
|------------|----|------|
| 12.603     | 9  | 0.112|

4.3 Model determination for Cardiovascular Disease

Based on the significant values of the predictors of CVD, the binary logistic model was fitted as follows:

\[
\text{Predicted logit of CVD} = (-1.016) + (1.223 \times \text{chest pain}) + (0.177 \times \text{st.depression}) + (0.151 \times \text{Ex-smoker}) + (0.096 \times \text{obesity})
\]

The predictors chest pain, st.depression, Ex-smoker and obesity are associated with increased risk of CVD while sex, Dyspnea, St.Elevation are associated with decreased risk of CVD. Statistical significance of the regression coefficient can generally be tested using Wald’s test and overall model significance can be tested by likelihood ratio test or pseudo $R^2$ test. The parameter estimates of the model is given in the Table 4-4.

Table 4-4 Estimates of parameters of logistic regression model for cardiovascular diseases (CVD)

| Predictors       | B     | S.E.  | Wald   | df | Sig. | Exp(B) |
|------------------|-------|-------|--------|----|------|--------|
| Sex              | -0.545| 0.272 | 4.017  | 1  | 0.045| 0.580  |
| Obesity          | 0.096 | 0.281 | 0.115  | 1  | 0.034| 1.100  |
| Ex-Smoker        | 0.151 | 0.680 | 0.049  | 1  | 0.025| 1.163  |
| Dyspnea          | -0.230| 0.260 | 0.779  | 1  | 0.007| 0.795  |
| St Elevation     | -0.001| 0.587 | 0.000  | 1  | 0.008| 0.999  |
| StDepression     | 0.177 | 0.299 | 0.350  | 1  | 0.004| 1.194  |
| chestpain        | 1.223 | 0.264 | 21.440 | 1  | 0.000| 3.398  |
| Constant         | -1.016| 0.325 | 9.767  | 1  | 0.002| 0.362  |
5. Conclusion

The logistic regression model has turned out to be a general technique of analysis in many circumstances over the past decade and are used significantly within the scientific research in medical related problems and also helpful to make decisions in critical medical diagnosis. The risk factors identified in this paper using Binary Logistic Regression accentuated the requirement for identification of the these factors through clinical examination in the particular area for preventing and controlling cardio vascular malady and its outcomes.

References

[1] Amir W, Mamat M, & Ali Z.2014, Association of hypertension with risk factors using logistic regression, Applied Mathematical Sciences, 8, pp 2563 – 2572
[2] Abedin T, Chowdhury Z, Afzal AR, Yaseen F, Turin TC 2016, Application of Binary Logistic Regression in Clinical Research, JNFB, 5 , pp 8-11
[3] Mythili T., Dev Mukherji M, Padalia N, & Naidu A 2013, A heart disease prediction model using SVM-decision trees-logistic regression (SDL) , International Journal of Computer Applications, 68, pp 11-14
[4] Rahman M, Nakamura K, Seino K, Kizuki M 2015, Sociodemographic Factors and the Risk of Developing Cardiovascular Disease in Bangladesh, Am J Prev Med, 48(4):pp 451-461.
[5] Truett J, Cornfield J, Kannel W 1967, A multivariate analysis of the risk of coronary heart disease in Framingham , J Chron Dis ,20: pp 511–524.
[6] D. W. Hosmer and S. Lemeshow 2000, Applied logistic regression, John Wiley and Sons