Binary Logistic Regression in Ascertaining Risk Factors for Atherosclerosis Heart Disease

K. Srividhya, A. Radhika

Abstract - The main intention of this study is to scrutinize factors that contribute significantly to enhancing the risk of Atherosclerosis Heart Disease (AHD). The dependent variable of the study is diagnosis of heart disease - whether the patient has the Atherosclerosis Heart Disease or not and the independent variables are cp (Chest Pain Type) and thalach (Maximum Heart Rate Achieved) were discussed. Logistic Regression is a predictive analysis. It is used to describe data and also to explain the relationship between one dependent binary variable and one or more independent variables (nominal, ordinal, interval or ratio-level variables). The Logistic Regression usage is increased in health care data. In this study we conclude that the type of Chest Pain and Maximum Heart Rate achieved are the imperative attributes of the Atherosclerosis Heart Disease (AHD) by using Binary Logistic Regression.

Index Terms – Atherosclerosis Heart Disease (AHD), Heart disease Attributes, Logistic Regression, Analysis of Variance.

_ _

1 Introduction

Atherosclerosis Heart Disease (AHD) is the most common type of heart disease in recent years. AHD causes impaired blood flow in the arteries that supply to the blood to the heart. It is also the leading cause of death for both men and women in the United States. The data for this study were taken from UCI Machine Learning Repository of Atherosclerosis Heart Disease (AHD). When the heart does not get enough arterial blood a few of the following symptoms are experienced. Angina (Chest Discomfort) is the most common symptoms of AHD. People may describe this discomfort as Chest Pain. Consequently in this study the type of Chest Pain and Maximum Heart Rate achieved are chiefly considered. The risk for AHD is mainly increasing by type of Chest Pain, and Maximum Heart Rate achieved. For reducing the prevalence of AHD, there is a need of exploring the factors that are responsible to enhancing the risk of this disease.

Logistic Regression model is discussed briefly to investigate the factors that to enhancing the risk of ischemic heart disease data. Logistic Regression analysis is applied for exploring the factors affecting the disease [1]. The desired patterns for the application of Logistic Regression methods testing a research hypothesis were illustrated. Recommended for appropriate reporting formats of Logistic Regression results and the minimum observation-to-predictor ratio [3]. Multiple Logistic Regression is used to explicate the prognostic factors for the development in elderly patients and to study the possible involvement of QTc interval prolongation [4]. The Multiple Logistic Regression Model was discussed to evaluate the risk factors associated with anemia and iron deficiency [7]. Binary Logistic Regression Analysis was conferred and identified factors that influence the use of family planning [8].

Research Scholar, Department of Statistics, Periyar University, Salem-11, E-mail: srividhyastat@gmail.com,
Assistant Professor, Department of Statistics, Periyar University, Salem-11, E-mail: radhisaran2004@gmail.com

2 Data Description

This data were taken from UCI Machine Learning Repository of the Atherosclerosis Heart Disease (AHD) patients. The AHD database consists of 303 patients with 10 attributes. In this paper only the 10 important attributes are used. These 10 attributes are main cause of the AHD patients. They are 1. Age (Continuous Variable) (age in years), 2. Sex (1 = male; 0 = female), 3. cp (Chest Pain Type) (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain and 3 = asymptomatic), 4. trestbps (Continuous Variable) (Resting Systolic Blood Pressure (in mm Hg on admission to the Hospital), 5. chol (Continuous Variable) (Serum Cholesterol in mg/dl), 6. thalach (Continuous Variable) (Maximum Heart Rate Achieved), 7. exang (Exercise Induced Angina) (1 = yes, 0 = no) , 8. oldpeak (ST depression induced by exercise relative to rest), 9. ca (Number of Major Vessels Colored by Fluoroscopy) (Value 0-4) and 10. target (0 = AHD no; 1 = AHD yes).

3 Materials And Methods

3.1 Logistic Regression Model

The Logistic Regression model is to express the relationship between outcome variable (dependent variable) and predictor variable (independent variable). The outcome variable is categorical (dichotomized) and the predictor variable can be continuous or categorical. The Logistic Regression model can be written as:

$$Logit(Q) = \ln\frac{\pi}{1-\pi} = \gamma_0 + \gamma_1 P \cdots\cdots\cdots(*)$$

Here π is the probability of occurring the outcome Q and $\frac{\pi}{(1-\pi)}$ is the odds of success; γ_0 is called intercept and γ_1 is called slope (regression coefficient). Taking the Antilog on both sides of equation (*) we can estimate the probability of the occurrence of outcome Q given predictor P (P can be either continuous or categorical):

$$\pi = \Pr\left(Q/P = p\right) = \frac{Exp(\gamma_0 + \gamma_1 P)}{1 + Exp(\gamma_0 + \gamma_1 P)} \cdots\cdots\cdots(**)$$

The logistic model for more than one predictor, we get

$$Logit(Q) = \ln\frac{\pi}{1-\pi} = \gamma_0 + \gamma_1 P_1 + \cdots + \gamma_s P_s \cdots\cdots\cdots(***)$$

The above equation is the general form of logistic regression model for s number of predictors.

3.2 Interpretation of Coefficients Using Odds

The Logistic Regression model in terms of the odds of an event is defined as the ratio of the probability of success to the probability of failure. The Logistic regression model in terms of log of the odds can be defined as:

$$Logit(Q) = \ln\frac{\pi}{1-\pi} = \gamma_0 + \gamma_1 P_1 + \cdots + \gamma_s P_s$$

The Logistic Regression equation can be written in terms of odds as:

$$\frac{\pi}{1-\pi} = Exp(\gamma_0 + \gamma_1 P_1 + \cdots + \gamma_s P_s)$$

$$Odds\ Ratio = Exp(\gamma)$$

3.3 Reviewing the Goodness of Fit of the Model

The main objective of goodness of fit is to know how well the model fits not only the sample of data from which it is derived, but also the population from which the sample data were selected. Define the Log-Likelihood function as:

$$Log - Likelihood = \sum_{i=1}^{m}\left[X_i \ln\left(\hat{X}_i\right) + (1 - X_i)\ln\left(1 - \hat{X}_i\right)\right]$$

where X_i' s are actual outcome and $\hat{X}_i$' s are the predicted probabilities of event occurring. Some of the goodness of fit statistics used for the model are Cox and Snell R^2 and $\tilde{R}^2$. The Cox and Snell R^2 can be defined as,

$$R^2 = 1 - \left[\frac{L(0)}{L(\gamma)}\right]^{2/M}$$

where L(0) is the Likelihood for the model with only a constant, L(γ) is the likelihood for the model in consideration and sample size is M. The problem with this measure for Logistic Regression is that it cannot achieve a highest value of 1. The Cox and Snell R^2 so that the value of 1 could be achieved. The Nagelkerke $\tilde{R}^2$ can be defined as,

$$\tilde{R}^2 = \frac{R^2}{R^2_{MAX}}$$

where $R^2_{MAX} = 1 - \left[L(0)\right]^{2/M}$. Nagelkerke $\tilde{R}^2$ reveals about the variation in the outcome variable which is explained by the Logistic Regression model. An additional approach for testing goodness of fit is Chi-Square test. Define Chi-Square statistic as:

$$\chi^2 = 2\left[\begin{matrix}\left(\log - likelihood of\ larger\ model\right) \\ -\left(\log - likelihood of\ smaller model\right)\end{matrix}\right]$$

Degrees of freedom (df) are the difference between the larger model and smaller model.

4 Model Results

The odds ratio can be calculate by the following formula,

Table 4. 1: Results of Logistic Regression of Atherosclerosis Heart Disease at Chest Pain Type

Attributes	B	S.E.	Wald	Df	Sig.	Exp(B)	95.0 % C.I. for EXP(B)	
							Lower	Upper
Cp			48.014	3	.000			
cp(1)	-1.503	.515	8.513	1	.004	.222	.081	.611
cp(2)	.510	.610	.698	1	.403	1.665	.503	5.509
cp(3)	.554	.550	1.016	1	.313	1.740	.593	5.111
Thalach	.033	.007	22.206	1	.000	1.034	1.019	1.048
Constant	-4.246	1.163	13.333	1	.000	.014		

From the results (Table 4. 2) we can write down the fitted logistic regression model as:
Logit (AHD status) = -4.246 + (-1.503) * cp(1) + 0.510 * cp(2) + 0.554 * cp(3) + 0.33 * thalach.

Both attributes "cp" (cp(2) & cp(3)) and "thalach" are positively related to the log of odds having AHD (Atherosclerosis Heart Disease), While the other attributes are negatively related. So the interpretations of results from Logistic Regression for the positively related attributes are given in the following manner:

The Chest Pain type attribute provides the fascinating results. As shown in Table 1, four separate Chest Pain contrasts were scrutinized. Each of these contrasts is evaluated against the reference category of Chest Pain Type (cp ie., Typical Angina). Two components of this attribute, Chest Pain Type 2 (cp(2) ie., Non-Anginal Pain) and Chest Pain Type 3 (cp(3) ie., Asymptomatic) are insignificant (p value > 0.05) at the 0.05 level. This indicates that, relative to being of these alternatives increases the chances of Atherosclerosis Heart Disease. The odds of patients having AHD for Chest Pain type 1 (cp(1) ie., Atypical Angina) are increased by a factor of 0.222. The odds of a patient having AHD for thalach are increased by a factor of 1.034.

Table 4. 2: Goodness of Fit Statistics of Logistic Regression on Hypothetical Data

Goodness of fit test	Test Statistics	P Value
Hosmer and Lemeshow Test	3.416	0.906
Cox & Snell R Square	0.306	-
Nagelkerke R Square	0.409	-
-2 Log likelihood	306.945	

Hosmer-Lemeshow (H-L) test is commonly used to measure the goodness of fit. Under Null hypothesis the data fits the model well. Our hypothetical data resulted a χ^2 of 3.416 with p value is 0.906 which is greater than the level of significance $\alpha = 0.05$ (insignificant, p value > 0.05), suggesting the model was fit to the data well. Hence our fitted Logistic Regression model is good fit for the AHD.

SPSS also afford two pseudo R^2 measures. Namely, Cox and Snell R^2 and Nagelkerke R^2. The Cox and Snell R^2 indicates that 30.6% of the variation in the dependent variable was explained by the explanatory variable. Nagelkerke R^2 specifies that 40.9% of the variability in the dependent variable was clarified by the explanatory variable. The Logistic Regression results support our hypothetical data that subject's chest pain type and maximum heart rate achieved is positively related to the Atherosclerosis Heart Disease (AHD).

Table 4. 3: Descriptive Statistics

	trestbps	thalach	Oldpeak
N	303	303	303
Mean	131.62	149.65	1.040
Std. Deviation	17.538	22.905	1.1611

Table 4.4 : ANALYSIS OF VARIANCE (ANOVA)

		Sum of Squares	df	Mean Square	F	Sig.
Trestbps	Between Groups	2643.080	3	881.027	2.919	.034
	Within Groups	90248.029	299	301.833		
	Total	92891.109	302			
Thalach	Between Groups	24029.827	3	8009.942	17.818	.000
	Within Groups	134413.388	299	449.543		
	Total	158443.215	302			
Oldpeak	Between Groups	51.000	3	17.000	14.273	.000
	Within Groups	356.125	299	1.191		
	Total	407.125	302			

From this study we conclude that the Chest Pain type has association with the rest of other factors such as trestbps (Resting Systolic Blood Pressure), thalach (Maximum Heart Rate Achieved) and oldpeak (ST depression induced by exercise relative to rest) as the significant value is (0.034, 0.000 and 0.000 respectively) less than 0.05.

5 Conclusion and Discussion

The aim of this study is to analyze the attributes that contribute significantly to enhancing the risk of AHD by using Logistic Regression. There are so many risk factors for AHD such as Age, Sex, Maximum Heart Rate, Chest Pain Type, Family history, High Blood Pressure, High Blood Cholesterol Level, Diabetes and Obesity. Chest Pain Type and increased Heart Rate both are the important risk factors of AHD. In this study the Chest Pain Type (cp) and Maximum

Heart Rate Achieved (thalach) are considered effectively. Both attributes "cp" and "thalach" are positively related to the log of odds having AHD whereas the other attributes are negatively related by using Logistic

Regression model. At the same time ANOVA also provides Chest Pain Type has association with the rest of other attributes such as trestbps (Resting Systolic Blood Pressure), thalach (Maximum Heart Rate Achieved) and oldpeak (ST depression induced by exercise relative to rest). A heart-healthy life style is important for all, not just for people with existing health problems. We keep our heart and blood vessels healthy by taking step toward a healthier lifestyle.

REFERENCES

[1] I. P. Bhatti, H. D. Lohano, Z. A. Pirzado and I. A. Jafri, "A Logistic Regression Analysis of the Ischemic Heart Disease Risk", *Journal of Applied Sciences*, vol. 6, pp. 785-788, 2006.

[2] R. Krishnan, R. Thandavarayan and S. Dipika, "Multinomial Logistic Regression Model for the Inferential Risk Age Groups for Infection Caused by Vibrio cholerae in Kolkata, India", *Journal of Modern Applied Statistical Methods*, vol. 6, pp. 324-330, 2007.

[3] C. J. Peng, K. L. Lee, G. M. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting", *The Journal of Educational Research*, No. 96, 2002.

[4] T. Yasuyuki, T. Gen, O. Miki, S. Akihisa and I. Takeshi , "Multiple logistic regression analysis of risk factors in elderly pneumonia patients: QTc interval prolongation as a prognostic factor", *Multidisciplinary Repository Medicine*, 2014.

[5] P.L. Michael, "Logistic Regression", *Circulation Journal of the American Heart Association*, 2008, 2395-2399.

[6] S. W. Pyke & P. M. Sheridan, "Logistic Regression Analysis of Graduate Student Retention", *The Canadian Journal of Higher Education*, No. 23, 1993.

[7] J.M. Schneider, M. L. Fujii, C. L. Lamp, B. Lonnerdal, K.G. Dewey and S. Zidenberg - Cherr , "The use of multiple logistic to identify risk factors associated with anemia and iron deficiency in a convenience sample of 12–36-mo-old children from low-income families1–3", *American Journal of Clinical Nutrition*, pp. 614-620, 2008.

[8] O.Chandra Sekhra Reddy, H. T. Likassa, L. Asefa, "Binary Logistic Regression Analysis in Assessing and Identifying Factors that Influence the Use of Family Planning: The Case of Ambo Town, Ethiopia", *International Journal of Modern Chemistry and Applied Science*, vol. 2, pp. 108-120, 2015.

[9] G.K. David, K. Mitchel, "Logistic Regression: A Self - Learning Text", Third Edition, *Springer,* New York, 2010.