https://doi.org/10.33472/AFJBS.6.13.2024.3884-3891

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

ADVANCED INSIGHTS INTO ATHEROSCLEROSIS: UTILIZING MULTINOMIAL LOGISTIC REGRESSION TO IDENTIFY KEY RISK FACTORS FOR HEART DISEASE

Dr. K. Srividhya[1*], Dr. B. Sendilkumar[2], Mrs. R. Tamilchudar[3] and Dr. A. Radhika[4]

[1]Associate Professor, Department of Biostatistics, School of Allied Health Sciences Vinayaka Mission's Research Foundation-DU, Salem, Tamil Nadu, India. Orcid ID: 0000-0002-1915-5948

[2]Dean, School of Allied Health Sciences Vinayaka Mission's Research Foundation-DU, Salem, Tamil Nadu, India,

[3]Professor, Department of Public Health, School of Allied Health Sciences Vinayaka Mission's Research Foundation -DU, Salem, Tamil Nadu, India

[4]Department of Statistics, Periyar University Salem, Tamil Nadu, India.

Corresponding Author E-mail: srividhyastat@gmail.com

Article Info

Volume 6, Issue 13, July 2024

Received: 04 June 2024

Accepted: 05 July 2024

Published: 31 July 2024

doi: 10.33472/AFJBS.6.13.2024.3884-3891

ABSTRACT:

The primary objective of this research is to meticulously examine the factors that significantly elevate the risk of Atherosclerosis Heart Disease (AHD). By employing a Multinomial Logistic Regression model, we analyzed the relationships between the non-binary dependent variable and independent variables such as age, sex, and exercise-induced angina (exang). This predictive analysis not only characterizes the data but also elucidates the connections between the dependent variable and various independent variables, whether they are nominal, ordinal, interval, or ratio-level. The application of logistic regression has become increasingly prevalent in healthcare data analysis. Our study concludes that the type of chest pain and the maximum heart rate achieved are crucial determinants of Atherosclerosis Heart Disease. By comparing chest pain type with independent attributes through the Multinomial Logistic Regression model, we have identified these key attributes, providing valuable insights for better understanding and managing AHD risks.

Keywords: Atherosclerosis heart disease (AHD), Risk factors, Multinomial logistic regression, Pearson Chi-square test, Parameter estimation.

1. INTRODUCTION

In the ever-evolving realm of data analysis, understanding and predicting complex outcomes is essential. Multinomial logistic regression stands out as a powerful tool for this purpose, extending beyond binary outcomes to handle dependent variables with three or more categories. By modeling the probabilities of multiple outcomes based on a diverse set of predictors, this method offers a nuanced and flexible approach to data analysis. Ideal for applications in healthcare, marketing, social sciences, and more, multinomial logistic regression provides deep insights and robust predictions, making it indispensable for modern statisticians and analysts seeking to unravel the complexities of their data.

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. 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.

LITERATURE REVIEW: EVALUATION ON LOGISTIC REGRESSION MODELS

Daryl Pregibon (1981) established diagnostic methods to assist analysts in recognizing and evaluating such observations impact on various parts of the maximum likelihood fit. The diagnostics are practically "free for the taking" a correctly constructed computational package for fitting the conventional maximum-likelihood model. Data analysis for logistic regression models, in particular, does not have to be costly or time-consuming. Multiple logistic regressions demonstrated by James Lee (1986), is a widely used statistical method for determining the relationship between an antecedent characteristic and a quantal outcome, while statistically controlling other covariates possible confounding effects. The stated goal of this paper was to highlight some of the issues that can arise when using logistic regression analysis.

Sandar W. Pyke and Peter M. Sheridan (1993) employed logistic regression analysis to forecast the retaining of 477 masters and 124 PhD applicants at a huge Canadian institution. An independent factor, selected demographic, academic, and financial assistance variables were employed. The dependent variable is dichotomous and verifies whether the student completed the degree satisfactorily or not. Hosmer et al. (1997) had discussed an association of goodness-of-fit for the logistic regression model. Recent research has demonstrated that using chi-square like goodness-of-fit model for the logistic regression established by Hosmer and Lemeshow (1980) that uses static sets of calculated chances may have disadvantages.

Peng et al. (2001) demonstrated the usage of logistic regression in health care study. Stepwise logistic regression techniques, both forward and backward, were implemented systematically to the cancer patients and a collection of descriptive factors were used in this real-world data set.

Irfana Bhatti et al. (2006) investigated the factors that contribute considerably to increasing the risk of ischemic heart disease. The dependent variable of the study is diagnosis - regardless of whether the patient is ill or not. Logistic regression analyses are used to examine the illness factors. The results of the study demonstrate factors that considerably increase the risk of ischemic cardiovascular diseases. Researcher Rajendran et al. (2007) discussed the Multinomial Logistic Regression (MLR) modeling, which is an efficient method for categorical results, as contrasted in discriminant function analysis and log-linear

simulations for profiling categories of the dependent variable. Multiple logistic regressions to identify and assess risk variables related to anemia and iron deficiency (ID) in a sample of children who were enrolled in or applied for the exceptional additional nourishment platform for women, infants, and children (WIC) were anticipated by Julie M Schneider et al. (2008). Yasuyuki Taooka et al. (2014) demonstrated a multivariate logistic regression analysis of risk variables. The predictive criteria for pneumonia development in older individuals were clarified and the probable participation of QTc interval prolongation was examined.

DATA DESCRIPTION

The data for this study were obtained from the UCI Machine Learning Repository, specifically from the Atherosclerosis Heart Disease (AHD) dataset, which includes records from 303 patients. Only the 10 most significant attributes related to AHD are analyzed in this research. These attributes are:

1. Age (continuous variable) - age in years
2. Sex (1 = male, 0 = female)
3. CP (Chest Pain Type) - categorized as 0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic
4. Trestbps (continuous variable) - resting systolic blood pressure (in mm Hg on admission)
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) - values range from 0 to 4
10. Target - 0 = no AHD, 1 = AHD present

These attributes are considered pivotal in understanding the factors contributing to Atherosclerosis Heart Disease among the patient population studied.

2. MATERIALS AND METHODS

LOGISTIC REGRESSION MODEL

The logistic regression model serves as a powerful tool for elucidating the intricate relationship between a categorical outcome variable (dependent variable) and predictor variables (independent variables), which can encompass both continuous and categorical data types. This statistical framework is adept at quantifying how these predictors influence the probability of specific outcomes, providing valuable insights into complex phenomena across various domains of research and practice.

The logistic regression model can be written as:

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

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 above equation we can estimate the probability of the occurrence of outcome Q given predictor P (P can be either continuous or categorical):

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

The logistic model for more than one predictor, we get

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

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

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:

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

The odds ratio can be calculate by the following formula,

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

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 obtained, but also the population from which the sample data were designated. 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 anticipated probabilities of event occurring. Some of the goodness of fit statistics utilized for the model are Cox and Snell R^2 *and* $\tilde{R}^2$. The Cox and Snell R^2 can be defined as,

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

Here, $L(0)$ represents the likelihood for the model with only a constant $L(\gamma)$represents 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 $R^2_{MAX} = 1 - [L(0)]^{2/M}$. 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:

$$\chi^2 = 2\left[(Log-likelihood\ of\ larger\ model) - (Log-likelihood\ of\ smaller\ model) \right]$$

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

MODEL RESULTS: **MULTINOMIAL LOGISTIC REGRESSION**

Binary logistic regression is employed to establish predictive relationships between one or more independent variables and a binary dependent variable, making it the standard choice in many analytical scenarios. However, beyond binary outcomes, multinomial logistic regression emerges as a versatile alternative. This advanced model proves invaluable when dealing with non-binary dependent variables or when categories within the variable are

unordered or ordered. Multinomial logistic regression expands upon the logistic regression framework, effectively tackling multi-class classification challenges by leveraging one or more independent variables to predict diverse outcome possibilities across various domains of analysis and decision-making.

Here in this data set the dependent variable cp ((chest pain type), 0=typical angina, 1=atypical angina, 2=non-anginal pain and 3=asymptomatic) is non-binary, and the independent variable age, sex, exang had considered to verify the most important influencing attributes using multinomial logistic regression.

Table 1: Descriptive Statistics

	trestbps	**thalach**	**oldpeak**
N	303	303	303
Mean	131.62	149.65	1.040
Standard deviation	17.538	22.905	1.1611

Table 2: Model Fitting Information using Multinomial Logistic Regression

Model	**Model Fitting Criteria**			**Likelihood Ratio Tests**		
	AIC	**BIC**	**-2 Log Likelihood**	**Chi-Square**	**df**	**Sig.**
Intercept Only	519.453	530.594	513.453			
Final	453.046	497.611	429.046	84.407	9	.000

The entire model is statistically significant when compared to the null model in terms of fit.

Table 3: Goodness of Fit

	Chi-Square	**df**	**Sig.**
Pearson	408.541	339	.006
Deviance	301.811	339	.928

The Pearson Chi-square test suggests that the model does not match the data well, but the deviance Chi-square indicates that the model fits the data well. Therefore the non-significant test indicates that the model fits for the data well.

Table 4: Pseudo R-Square

Cox and Snell	.243
Nagelkerke	.267
McFadden	.116

These are pseudo R-square values that are used as rough analogues for R-square values in ordinary least square regression.

Table 5: Likelihood Ratio Tests

Effect	**Model Fitting Criteria**			**Likelihood Ratio Tests**		
	AIC of Reduced Model	**BIC of Reduced Model**	**-2 Log Likelihood of Reduced Model**	**Chi-Square**	**df**	**Sig.**
Intercept	459.556	492.980	441.556	12.510	3	.006
age	455.390	488.813	437.390	8.343	3	.039
exang	512.860	546.283	494.860	65.814	3	.000

sex	452.721	486.144	434.721	5.675	3	.129

The discrepancy in -2 log-likelihoods among the final model and a reduced model is the chi-square statistic. The reduced model is created by omitting one of the final model's effects. The null hypothesis states that all of the effect's parameters are zero. These findings include likelihood ratio tests of each independent variable's overall contribution to the model. Using the conventional $\alpha = 0.05$ threshold, we see that age and exang both was significant predictor in the model, although sex was non-significant.

Table 6: Parameter Estimates

	cp[a]	B	Std. Error	Wald	df	Sig.	Exp (B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
1	**intercept**	2.534	1.138	4.962	1	.026			
	age	-.050	.020	6.319	1	.012	.951	.915	.989
	exang	-2.607	.552	22.274	1	.000	.074	.025	.218
	sex	-.267	.385	.481	1	.488	.766	.361	1.627
2	**intercept**	1.888	.982	3.699	1	.054			
	age	-.027	.017	2.509	1	.113	.974	.942	1.006
	exang	-2.103	.367	32.814	1	.000	.122	.059	.251
	sex	-.406	.319	1.619	1	.203	.666	.356	1.245
3	**intercept**	-2.453	1.651	2.209	1	.137			
	age	.011	.027	.175	1	.676	1.011	.959	1.066
	exang	-1.895	.581	10.633	1	.001	.150	.048	.470
	sex	.854	.602	2.013	1	.156	2.348	.722	7.634

a: The reference category is: 0

These results provide information comparing each chest pain type (1=atypical angina, 2=non-anginal pain and 3=asymptomatic) against the reference category “cp” (0=typical angina).

The first set of coefficients shows the comparisons among typical angina (coded 0) and atypical angina (coded 1). The age (b=-.050, s.e.=.020, p<0.05) and exang (b=-2.607, s.e.=.552, p<0.05) both was a significant predictors in this model. The odds ratio of 0.951 (for age) and 0.074 (for exang) indicates that it contributes significantly to enhancing the risk on chest pain, the odds of a patients having chest pain (atypical angina) for age are increased by a factor of 0.951 and for exang are increased by a factor by 0.074. The factor sex is insignificant (p value > 0.05) at the 0.05 level. This indicates that, relative to being of these alternatives increases the chances of chest pain (in other words, the odds were decreasing).

The second set of coefficients shows the comparisons among typical angina (coded 0) and non-anginal pain (coded 1). Only exang (b=-2.103, s.e.=.367, p<0.05) was a significant predictor in this model. The odds ratio of 0.122 indicates that an increase on exang (exercise induced angina), the odds of a patients having chest pain for non-anginal pain increased by a factor of 0.122. The factors age and sex is insignificant (p value > 0.05) at the 0.05 level. This indicates that, relative to being of these alternatives increases the chances of chest pain (in other words, the odds were decreasing)

The final set of coefficients shows the comparisons among typical angina (coded 0) and asymptomatic (coded 1). Only exang (b=-1.895, s.e.=.581, p<0.05) was a significant predictor in this model. The odds ratio of 0.150 indicates that an increase on exang (exercise induced angina), the odds of a patients having chest pain for asymptomatic increased by a factor of 0.150. The predictor’s age and sex were not significant in this model (p value >

0.05) at the 0.05 level. This indicates that, relative to being of these alternatives increases the chances. of chest pain (in other words, the odds were decreasing).

Table 7: Model Classification

Observed	Predicted				
	0	**1**	**2**	**3**	**Percent Correct**
0	103	0	40	0	72.0%
1	15	1	34	0	2.0%
2	27	0	60	0	69.0%
3	14	0	9	0	0.0%
Overall Percentage	52.5%	0.3%	47.2%	0.0%	54.1%

These are classification statistics used to determine which types of chest pain were predicted by the model. Typical angina was correctly predicted by the model 72.0%. Non-anginal pain was correctly predicted by the model by 69.0%. Atypical angina and asymptomatic both were predicted by the model by poor ratios.

3. CONCLUSION AND DISCUSSION

The purpose of this study is to analyze the attributes that contribute significantly to enhancing the risk of AHD by using Multinomial logistic regression model. There are so many risk factors for AHD such as age, sex, maximum heart rate, chest pain type, exercise induced angina, 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 Multinomial regression model the three chest pain types were compared with the independent attribute age, sex, and exang, by this comparison we observed the results that the third chest pain type (asymptomatic pain) influencing high risk on the attributes age and sex because it is positively related to the log of odds having chest pain. Adopting a heart-healthy lifestyle is crucial for everyone, not just those with existing health conditions. By embracing healthier habits, we actively nurture our heart and blood vessels, paving the way for long-term well-being.

4. REFERENCES

1. Abbott, R. D. (1985). Logistic regression in survival analysis. *American journal of epidemiology*, *121*(3), 465-471.
2. Abedin, T., Chowdhury, Z., Afzal, A. R., Yeasmin, F., & Turin, T. C. (2016). Application of binary logistic regression in clinical research. *JNHFB*, *5*, 8-11.
3. Bhatti, I. P., Lohano, H. D., Pirzado, Z. A., & Jafri, I. A. (2006). A logistic regression analysis of the ischemic heart disease risk. *Journal of Applied Sciences*, *6*(4), 785-788.
4. David, K., & Mitchel, K. (1994). Logistic regression: A self learning text.Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*, *16*(9), 965-980.
5. Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*, *16*(9), 965-980.
6. Lee, J. (1986). An insight on the use of multiple logistic regression analysis to estimate association between risk factor and disease occurrence. *International journal of epidemiology*, *15*(1), 22-29.

7. Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, *96*(1), 3-14.
8. Pyke, S. W., & Sheridan, P. M. (1993). Logistic regression analysis of graduate student retention. *Canadian Journal of Higher Education*, *23*(2), 44-64.
9. Taooka, Y., Takezawa, G., Ohe, M., Sutani, A., & Isobe, T. (2014). Multiple logistic regression analysis of risk factors in elderly pneumonia patients: QTc interval prolongation as a prognostic factor. *Multidisciplinary respiratory medicine*, *9*, 1-6.
10. Pregibon, D. (1981). Logistic regression diagnostics. *The annals of statistics*, *9*(4), 705-724.
11. Rajendran, K., Ramamurthy, T., & Sur, D. (2007). 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*, *6*, 324-330.
12. Reddy, O. S., Likassa, H. T., & Asefa, L. (2015). 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*, *2*(2), 108-120.
13. Schneider, J. M., Fujii, M. L., Lamp, C. L., Lönnerdal, B. O., Dewey, K. G., & Zidenberg-Cherr, S. (2008). The use of multiple logistic regression to identify risk factors associated with anemia and iron deficiency in a convenience sample of 12–36-mo-old children from low-income families. *The American journal of clinical nutrition*, *87*(3), 614-620.