BINOMIAL LOGISTIC REGRESSION OF INCIDENCE OF CARDIOVASCULAR DISEASE AMONG STAFF IN HIGHER INSTITUTION

Olorunlogbon, O. R. , Adetunji, A.A., and Olorunlogbon, A.
Department of Statistics, Federal Polytechnic, Ile-Oluji, Nigeria

Corresponding Author: Olorunlogbon, O. R.

E-mail: *oluolorunlogbon@fedpolel.edu.ng, +234 8062607364*

Manuscript received 25/09/22 and accepted on the 25/10/22

ABSTRACT

Cardiovascular disease is a cluster of diseases and injuries that affect the heart and blood vessels of people across the globe. This study evaluates the incidence of the disease among 200 randomly sampled staff of the Federal Polytechnic, Ile-Oluji, Nigeria from existing records of Health Services Unit of the Institution. Variables include Age, Gender, Weight, Pulse Rate, Heart Disease Condition and Maximal Aerobic Capacity ($V0_2$ Max). The logit transformation of continuous dependent variables (Age, Weight and $V0_2$ Max) reveals that they are linearly related to the dependent variable (Heart Disease Status). Evaluating the observed and predicted classification with respect to staff heart disease condition, the logit transformation: correctly classifies 84.0% of the overall cases with the addition of the independent variables; correctly predicted 97.6% of staff not having the disease; correctly predicted 66.7% cases with the observed characteristic compared to the total number of cases predicted as having the disease; and correctly predicted 85.1% cases without the observed characteristic compared to the total number of cases predicted as not having the disease. Chi-square test of independence reveals that age and gender are dependent on the heart disease status. Wald test reveals that Age (P-value = 0.001) and Weight (P-value =0.030) added significantly to the model, but Gender (P-value = 0.069) and VO_2 Max (P-value = 0.235) are not significant. Hence, prediction is made such that the probability that a staff will have heart disease on a one-unit change is 2.602 times greater for males compared to females.

Keywords: Binomial Logistic Regression, Heart Disease, Chi-Square, Wald Test, Logit Transformation

1. INTRODUCTION

Cardiovascular disease (CVD), in all its forms, is the largest single cause of death on the planet (WHO, 2004) and therefore, it is a very important matter of life and death. CVD is a cluster of diseases and injuries that affect the heart and blood vessels (the cardiovascular system). They are mostly diseases of the heart and of the blood vessels of the heart and brain which generally affect people in later life with incidence rising sharply after the age group of (30-44); although, according to a leading cardiologist, most people with a form of CVD already have the beginnings of the disease by around 35 years old (Edward, 2006).

It was estimated in 2005 that 30% of the 58 million people that died globally were due to cardiovascular disease (CVD) of which infectious diseases, nutritional deficiencies, maternal and prenatal conditions combined were responsible for this proportion (WHO, 2005). It is also essential to acknowledge that a substantial proportion (46%) of these deaths were of people under 70 years of age, in the more productive period of life and 79% of the disease burden attributed to cardiovascular disease is in this age group (WHO, 2002).

In developing countries, it is estimated that cardiovascular diseases (CVD) are growing radically with 12 million people dying each

year (WHO, 2013). Consequently, coronary heart disease (CHD) is becoming the main cause of cardiac surgery globally (Go et al., 2013).

Several researches have been carried out and it has been found out that heart disease is the leading cause of death in many countries of the world. For example, more than 600,000 Americans die of heart disease every year and that accounts for one in every four deaths (CDC, 2009). In Canada, men and women die majorly as a result of cardiovascular disease (Canadian Institute for Health Information, 2002). Generally, the onset of CVD is approximately 10 years later in women than in men; women present with a wider range of symptoms (Milner et al., 1999). In CVD risk factors, sex differences have also been described which include cigarette smoking, depression, low income, elevated serum lipids, hypertension, obesity and lack of physical activity (Abbey & Stewart, 2000; Lonn, 2000). In the middle-income countries, Ischaemic Heart Disease and stroke are the leading cause of death and Disability-Adjusted Life Years (DALYs) in most countries (Naghavi, 2017). In Indonesia both morbidity and mortality are as a result of CVDs and it is responsible for a third of all deaths in the country (WHO, 2015). It is estimated that 3,299 and 2,555 are years of life lost per 100,000 due to the premature mortality from coronary heart disease (CHD) and cerebrovascular diseases respectively (Institute of Health Metric Evaluation, 2016).

Federal Polytechnic, Ile-Oluji (FEDPOLEL) is a federal government-owned and operated Nigerian polytechnic established in 2014. The polytechnic is located in Ile-Oluji/Oke-Igbo, Ondo State, South-West Nigeria. Federal Polytechnic Ile-Oluji offer National Diploma and Higher National Diploma courses at undergraduate levels. The polytechnic aims at becoming a prime manpower provider in the field of technology, innovation and management for national development. Federal Polytechnic, Ile-Oluji has been officially accredited and/or recognized by the National Board for Technical Education (Federal Polytechnic Ile-Oluji, 2021).

This study is aimed at evaluating the incidence of heart disease among Polytechnic staff by using binomial logistic regression, and it is divided into different sessions: Session 1 consists of the Introduction, Session 2 Materials and Methods, Results and Discussion in session 3, and we have the Conclusion and Recommendation in session 4 of the paper.

2. METHODOLOGY

In this research, the team randomly selected 200 staffs (teaching and non- teaching) from the Federal Polytechnic, Ile-Oluji, Ondo State, Nigeria. Data on Age (years), Gender (male or female), Weight (kg), Pulse Rate (per 20 seconds) and Heart Disease Condition or Status ("Yes" or "No") were collected from existing records of the Health Services Unit of the Institution. Data on Maximal Aerobic Capacity - $V0_2$ Max (mL/kg/min) for each of the 200 staffs were derived from their age and pulse rate (number of heartbeat in 20 seconds). SPSS for Windows (version 23) is used to perform the statistical analysis. Box-Tidwell Test is used to determine the linear relationship between the continuous independent variables (Age, Weight and $V0_2$ Max) and the logit transformation of the dependent variable (Heart Disease Status). Also, Chi-square test of independence is used to determine whether age and gender of the staff are related to their health status.

The most popular technique for analyzing data with binary responses is binomial logistic regression (Adejumo & Adetunji, 2013). When the response, Y, may only take one of two potential values, it is used to model interactions between the response variable and numerous explanatory variables, which may be discrete or continuous. Yes/No, Alive/Dead, Present/Absent, etc.,

are possible results. It is simpler to represent the two levels as 0 and 1 ("failure" or "success") for analysis purposes. The logistic curve's data sufficiency and the likelihood of an event occurring can both be predicted using the model. Based on the values of the independent variables, it may also be used to predict the likelihood of getting involved in a case (Kleinbaum & Klein, 2010; Park, 2013)

The main goal is to find out the relationship between the response probability $\pi = \pi(x)$ and the explanatory variables $x_1, x_2, \dots, x_n$. Thus, a binomial random variable can be defined as:

$$Y = \begin{cases} 1 \; if \; the \; result \; is \; a \; success \\ 0 \; if \; the \; result \; is \; a \; failure \end{cases}$$

with probabilities P(Y = 1) = π and P(Y = 0) = 1 – π.
Let X_i be a random variable, $Y_i = 1$ if the event occurs and $Y_i = 0$ if otherwise. Then,

$$P_i = E(Y = 1/X_i) = \left(\frac{1}{1+e^{-z}} \right)$$

Where $Z = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{1k}$

The (cumulative) logistic distribution function is represented by the equation. It is simple to demonstrate that P_i is non-linearly connected to Z through X_i and that, as Z varies between -∞ and ∞, P_i varies between 0 and 1. Since P_i is non-linear in both X_i and β_i's, as seen in the equation below, an estimation problem will be devised in an effort to meet these conditions. This assumption violation renders the OLS approach ineffective for the estimation of β_i's. A set of regression coefficients that predicts the likelihood of the desired outcome are estimated to fit a binary logistic regression model. Given a person's likelihood of having heart disease as P_i and their likelihood of not having heart disease as 1-P_i, the linear combination of the probabilities becomes:

$$ln\left(\frac{probability \; (event)}{1 - probabibility \; (event)} \right) = ln\left(\frac{Pi}{1 - Pi} \right)$$

$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

3. RESULTS AND DISCUSSION

The linear relationship between the continuous independent variables (*Table 1*) in the study and the logit transformation of the dependent variable was carried out by Box-Tidwell Test and it reveals that the original continuous independent variables (Age, Weight and $V0_2Max$) is linearly related to the logit transformation of the dependent variable (Heart Disease Status) as the result is not significant at 0.065, 0.053 and 0.387 respectively.

***Table 1**: Box-Tidwell Linearity Test*

	B	S.E.	Wald	Df	p-value	Exp(B)
Age	1.883	0.970	3.765	1	0.052	6.571
Weight	-0.847	0.450	3.545	1	0.060	0.429
VO2_Max	1.538	1.795	0.734	1	0.391	4.657

trAge	-0.367	0.199	3.402	1	0.065	0.692
trWeight	0.162	0.084	3.728	1	0.053	1.176
trVO2_Max	-0.340	0.394	0.748	1	0.387	0.711
Constant	-23.437	15.108	2.406	1	0.121	0.000

Examining the significance of the overall model through the Omnibus Tests of Model Coefficients, it can be deduced from *Table 2*that it is statistically significant at $\chi^2_{(4)} = 30.872$, p = 0.000 (< 0 .05).

Table 2: *Omnibus Tests of Model Coefficients*

	Chi-square	Df	p-value
Step	30.872	4	0.000
Block	30.872	4	0.000
Model	30.872	4	0.000

Table 3 contains the Cox & Snell R^2 and Nagelkerke R^2 values, which are both methods of calculating the explained variation. These values are sometimes referred to as pseudo R^2 values (and will have lower values than in multiple regression). However, they are interpreted in the same manner, but with more caution. Therefore, the explained variation in the dependent variable based on our model ranges from 14.3% to 23.4%, depending on whether you reference the Cox & Snell R^2 or Nagelkerke R^2 methods, respectively.

Table 3: *Cox & Snell R^2 and Nagelkerke R^2 Values*

-2 Log likelihood	Cox & Snell R^2	Nagelkerke R^2
157.685	0.143	0.234

To ascertain the null hypothesis that predictions made by the model fit perfectly with observed group memberships, the Hosmer-Lemeshow Test (*Table 4*)is carried out. A chi-square statistic is computed comparing the observed frequencies with those expected under the linear model. The non-significant chi-square on the table indicates that the data fit the model well.

Table 4: *Hosmer-Lemeshow Test*

Chi-square	Df	p-value
14.399	8	0.072

Table 5 explains or evaluates the observed and predicted classification with respect to the heart disease status or condition of the staff of the Polytechnic. This evaluation is done in light of the following measuring parameters:

A***ccuracy***: The model now correctly classifies 84.0% of cases overall with the addition of the independent variables.

Sensitivity: 22.2% of participants who had heart disease were also predicted by the model to have heart disease. This means that we have $\frac{8}{36} \times 100\% = 22.2\%$

Specificity: 97.6% of participants who did not have heart disease were correctly predicted by the model not to have heart disease. In this case, we have $\frac{160}{164} \times 100\% = 97.6\%$

Positive Predictive Value is the percentage of correctly predicted cases with the observed characteristic compared to the total number of cases predicted as having the characteristic. Here, we have $\frac{8}{12} \times 100\% = 66.7\%$

Negative Predictive Value is the percentage of correctly predicted cases without the observed characteristic compared to the total number of cases predicted as not having the characteristic. Finally, we have: $\frac{160}{188} \times 100\% = 85.1\%$. That is, of all cases predicted as not having heart disease, 85.1% were correctly predicted.

Table5: *Observed and Predicted Classification*

		Predicted			
		Heart Disease Status			Percentage
	Observed	No	Yes	∑	Correct (%)
Heart Disease Status	No	160	4	164	97.6
	Yes	28	8	36	22.2
	∑	188	12	200	
Percentage Correct (%)		66.7	85.1		
Overall Percentage (%)					84.0

Determining the significance of each of the independent variables, the Wald Test is employed (Table 6). It shows, from the result, that age (p = 0.001) and weight (p = 0.030) added significantly to the model/prediction, but gender (p = 0.069) and VO_2Max (p = 0.235) did not add significantly to the model. Hence, prediction can be made such that the probability that a staff will be having heart disease on a one-unit change in one of the independent variable (say "Age") when all other independent variables are kept constant. For example, the table shows that the odd of having heart disease ("Yes" category) is 2.602 times greater for males as opposed to females.

Table 6: *Variables in the Equation*

							95% C.I.for EXP(B)	
	B	S.E.	Wald	df	p-value	Exp(B)	Lower	Upper
Age	0.085	0.025	11.714	1	0.001	1.088	1.037	1.142
Gender(1)	0.956	0.526	3.312	1	0.069	2.602	0.929	7.290
Weight	0.027	0.012	4.695	1	0.030	1.027	1.003	1.052
VO2_Max	-0.047	0.039	1.409	1	0.235	0.954	0.884	1.031
Constant	-6.154	2.199	7.831	1	0.005	0.002		

Table 7 reveals that out of the 200 staff examined in this research, 36 (18.0%) of them have heart disease condition while 164 (82.0%) do not have heart disease issue. Also, considering the age groups, the result shows that age (40-49) with 18 (9.0%) has the

highest cases of staff with heart disease condition while groups (20-29) and (70-79) both with 0 (0.0%) have no cases of staff with heart disease. The Chi-square test of independence reveals that age and gender are related with heart disease status for p-values 0.000 and 0.030 respectively. This supports the finding by (Levy, 1981) that age and gender (male sex) and family history are non-modifiable risk factors responsible for coronary heart disease. Thus, we can conclude that both age and gender have influence on the staff heart disease condition.

Table 7: *Staff characteristics and heart disease status*

Characteristics	**Heart Disease Status or Condition**			**Chi-Square p-value (α = 0.05)**
	No n = 164 (82.0%)	**Yes n = 36 (18.0%)**		
Age (Grouped)			Total	0.000
20-29	6 (3.0%)	0 (0.0%)	6 (3.0%)	
30-39	89 (44.5%)	6 (3.0%)	95 (47.5%)	
40-49	54 (27.0%)	18 (9.0%)	72 (36.0%)	
50-59	13 (6.5%)	9 (4.5%)	22 (11.0%)	
60-69	1 (0.5%)	3 (1.5%)	4 (2.0%)	
70-79	1 (0.5%)	0 (0.0%)	1 (0.5%)	
Gender			Total	0.030
Female	59 (29.5%)	6 (3.0%)	65 (32.5%)	
Male	105 (52.5%)	30 (15.0%)	135 (67.5%)	

4. CONCLUSION AND RECOMMENDATION

A logistic regression was performed to ascertain the effects of age, weight, gender and VO_2Max on the likelihood that participants have heart disease. The logistic regression model was statistically significant, $\chi^2_{(4)} = 30.872$, $p < 0.05$. The model explained 23.4% (Nagelkerke R^2) of the variance in heart disease and correctly classified 84.0% of cases. Males were 2.602 times more likely to exhibit heart disease than females. Increasing age was associated with an increased likelihood of exhibiting heart disease. However, increasingVO_2Max was associated with a reduction in the likelihood of exhibiting heart disease. Furthermore, it was found out that age (grouped) and gender are significant in relation with heart disease status for p-values 0.000 and 0.030 respectively. Thus, we can conclude that both age and gender have influence on the staff heart disease condition.

From the result of this research (Table 7), it is shown that age group (40-49) with 18(9.0%) has the highest number of staff having one form of heart disease or the other. This is followed by age groups (50-59) and (30-39) with 9(4.5%) and 6(3.0%) respectively. This result upholds (Edward, 2006) and thus, staffs of the Polytechnic, especially those within the age bracket 30 to 60 years, are strongly encouraged to always embark on consistent medical check-ups that will provide them with necessary and sufficient information about the state of their health and heart such that appropriate measures can be taken to combat and overcome any medical experience or challenge they may be facing. The Management is encouraged to incorporate regular and consistent health orientation and awareness into the Institution calendar such that proper information will be given to the staffs on their health as the heart is the key to

the health of anyone. This would be achieved by employing the skill and expertise of health professionals on a regular basis. These professionals will, among others, educate the staff on proper dietary plans, health essentials, heart matters and general well-being of the staff as noted by (Kelley, 2014; Briggs et al., 2017; WHO, 2007).

REFERENCES

Abbey S, Stewart DE: Gender and psychosomatic aspects of ischemic heart disease. *J Psychosom Res* 2000, 48(5):417-423.

Adejumo A. O. and Adetunji A. A. (2013): Application of Ordinal Logistic Regression in the Study of Students' Performance, *Journal of Mathematical Theory and Modelling, 3(11), pp 10 – 19.*

Briggs, M. A.; Petersen, K. S.; Kris-Etherton, P. M. Saturated Fatty Acids and Cardiovascular Disease: Replacements for Saturated Fat to Reduce Cardiovascular Risk. Healthcare 2017, 5, 29. https://doi.org/10.3390/healthcare5020029

Canadian Institute for Health Information: *Hospital Mortality Database.Ottawa: CIHI* 2002.

CDC: Deaths: Final Data for 2009. *www.cdc.gov/nchs/data/nvsr/nvsr60/nvsr6003.pdf*

Edward Nason's interview with a British Heart Foundation (BHF) professor of cardiovascular sciences, on 15th May 2006.

FEDERAL POLYTECHNIC, ILE-OLUJI. www.fedpolel.edu.ng. Retrieved 2021-09-04

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Borden, W. B., & Turner, M. B. (2013). on behalf of the American Heart Association StatisticsCommittee and Stroke Statistics Subcommittee. Heart disease and stroke statistics-2013 update: a report from the American Heart Association. *Circulation*, 127(1), e1-e240.

Institute of Health Metric Evaluation, Global Burden of Diseases Compare: Indonesia 2016. https://vizhub.healthdata.org/gbd-compare/

Kelley, Deeanna (2014) "Heart Disease: Causes, Prevention, and Current Research,"*JCCC Honors Journal*: Vol. 5: Iss. 2, Article 1. Available at: http://scholarspace.jccc.edu/honors_journal/vol5/iss2/1

Kleinbaum, D. G. and Klein, M. (2010) Modeling Strategy Guidelines. In: Logistic Regression, Part of the Series Statistics for Biology and Health, Springer, Berlin, 165-202. http://dx.doi.org/10.1007/978-1-4419-1742-3_6.

Levy, R. I. Declining mortality in coronary heart disease. Arteriosclerosis. 1981 Sep-Oct;1(5):312-25. doi:10.1161/01.atv.1.5.312.PMID: 7052033.

Lonn E: *Epidemiology of ischemic heart disease in women: women and ischemic heart disease. Canadian Cardiovascular Society, Consensus Conference.* 2000.

Milner, K. A., Funk, M., Richards, S., Wilmes, R. M., Vaccarino, V., Krumholz, H. M.: Gender differences in symptom presentation associated with coronary heart disease. *Am J Cardiol*1999, 84:396-399.

Naghavi M, Abajobir AA, Abbafati C, Abbas KM, Abd-Allah F, Abera SF, et al. Global, regional, andnational age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet. 2017; 390(10100): 1151–1210. https://doi.org/10.1016/S0140-6736(17)32152-9

Park H. (2013). An Introduction to Logistic Regression: From Basic Concepts to

Interpretation with Particular Attention to Nursing Domain, J Korean AcadNurs 43(2), pp. 154-164.

Preventing chronic disease: a vital investment. Geneva, World Health Organization, 2005.

The World Health Report 2002: reducing risks, promoting healthy life. Geneva, World Health Organization, 2002.

World Health Organization (2004) "Atlas of Heart Disease and Stroke", available at http://www.who.int/cardiovascular_diseases/resources/atlas/en/index.html

World Health Organization (2013). Cardiovascular disease.Retrieved March 20th, 2013 from http://www.who.int/ncd/cvd/.

World Health Organization. Non-communicable Diseases (NCD) Country Profiles Indonesia. 2015. http://www.who.int/nmh/countries/idn_en.pdf?ua=1

World Health Organization. (2007). Prevention of cardiovascular disease: guidelines for assessment and management of total cardiovascular risk.World Health Organization. https://apps.who.int/iris/handle/10665/43685