Table of Contents

[**CHAPTER ONE** 2](#_Toc127090453)

[**1.0** **INTRODUCTION** 2](#_Toc127090454)

[**1.1** **BACKGROUND OF STUDY** 2](#_Toc127090455)

[**1.2** **SIGNIFICANCE OF STUDY** 2](#_Toc127090456)

[**1.3** **AIMS AND OBJECTIVES** 3](#_Toc127090457)

[**1.4** **SOURCE OF DATA** 3](#_Toc127090458)

[**CHAPTER TWO** 5](#_Toc127090459)

[**2.1 LITERATURE REVIEW** 5](#_Toc127090460)

[**CHAPTER THREE** 7](#_Toc127090461)

[**3.0** **METHODOLOGY** 7](#_Toc127090462)

[**3.1** **LOGISTIC REGRESSION** 7](#_Toc127090463)

[**3.2** **ASSUMPTIONS OF LOGISTIC REGRESSION** 7](#_Toc127090464)

[**3.3** **DICHOTOMOUS DEPENDENT VARIABLE** 8](#_Toc127090465)

[**3.4** **INDEPENDENT OBSERVATIONS** 8](#_Toc127090466)

[**3.4.1** **Testing for Independence** 8](#_Toc127090467)

[**3.5** **MULTICOLLINEARITY** 9](#_Toc127090468)

[**3.5.1** **Detecting multicollinearity** 10](#_Toc127090469)

# **CHAPTER ONE**

## **1.0 INTRODUCTION**

## **1.1 BACKGROUND OF STUDY**

The heart is an organ in the circulatory system that serves as pump, to help pump oxygenated blood through the body, thereby transporting oxygen and nutrients to other parts of the body. The heart is located in the middle of the chest between the lungs, slightly to the left of the breastbone.

The heart beats at an average of 100,000 times daily, therefore pumping about 2,000 gallons (7,571 litres) of blood through the human body. The heart is one of the key organs that makes up the circulatory system. The human heart is divided into 4 chambers, which are; the left and right artricles, and, the left and right ventricles.

Cardiovascular diseases (heart related diseases) are conditions that involve diseased vessels, irregular functioning of the heart, structural problems and blood clots. There are different conditions of the cardiovascular diseases, such as heart attack, heart failure, angina, arrhythmia, congenital artery disease, etc.

## **1.2 SIGNIFICANCE OF STUDY**

According to W.H.O fact sheet, cardiovascular diseases [CVD] are the leading causes of death globally with an estimated death of 17.9 million in the year 2019, which is about 32% of all deaths globally, of which 85% of the heart related diseases were due to heart attack and stroke.

Over 33.33% of deaths from cardiovascular diseases occur in the low- and middle-income countries, such as Nigeria, Burundi, Afghanistan, etc.

About 90% of cardiovascular diseases can be prevented by addressing behavioural risk factors, such as Obesity, use of tobacco, unhealthy diet, harmful use of alcohol and physical inactivity.

## **1.3 AIMS AND OBJECTIVES**

Modelling the behaviour of heart diseases in relation to other medical vital statistics is essential and useful in the diagnosis of the presence of heart diseases in the health sector.

Due to the nature and importance of the heart in the cardiovascular system, its in-depth and proper diagnostic is required and beneficial.

The objective of this study is as follows:

1. Model the available data on heart diseases using the logistic regression model.
2. Estimate the model parameters for more efficient analysis and predictions.
3. Carry out diagnostic to check the adequacy of the model.

## **1.4 SOURCE OF DATA**

The dataset used in this project is from a secondary source, which was curated from 303 observations from Cleveland; 123 observations from Switzerland; 200 from V.A. medical centre, Long Beach; 270 observations from Stalog (heart) dataset; and 294 Hungarian observations. With a total of 1190 observations with 272 duplicate observations, which finalized to 918 observations.

----------------------------------------------------------------------------------------------------------------

The project is aimed at modelling heart diseases using the logistic regression model, considering some important health vitals, aimed at exploring in-depth, the knowledge encompassed in the data collected, for a better and data driven decision-making process in the health sector. This helps with:

1. Improvement of use of statistics in decision making process in the health sector, for a better data-driven insight.

# **CHAPTER TWO**

## **2.1 LITERATURE REVIEW**

Statistical analysis is useful for solving various set of problems. One of the aspects of application of statistics is in the area of predictive analysis, which is of great value in the healthcare sector, as it is great to curb some health irregularities at their early stages for proper treatment.

The cardiovascular diseases have been identified as one of the largest causes of death, even in the developed countries. C. Beulah et al (2019) noted that one of the reasons for fatality due to heart diseases, is due to the fact that the risks are either not identified, or are identified at the later stage. However, it is possible to forecast cardiovascular disorders, using predictive statistical techniques like Logistic regression, Naive Bayes, Ridge regression, etc.

Apurb Rajdha et al (2020) emphasis on the mandatory need to develop a system to predict cardiovascular diseases effectively and accurately, due to its increasing mortality rate. In their studies, they compared the accuracy score of Decision Tree, Logistic Regression, Random Forest and Naive Bayes for predicting heart diseases. A result of which the Random forest algorithm came top, with an accuracy score of 90.16%.

Ashok Kumar Dwivedi (2016) reported the highest classification accuracy of logistic regression at 85%, with a sensitivity and specificity of 89% and 81% respectively.

Peter C. Austin et al (2010) concluded that logistic regression predicted in-hospital mortality in patients hospitalized with heart failure more accurately than the regression tree algorithm.

The current research has employed ensemble approaches to increase classification precision in heart disease prediction. The prediction success of a model is not equally contributed to by all the independent variables. Imamovic D. et al (2020) noted that multiple attributes lead to increased complexity and decreased performance. As a result, cautious feature extractions are required while maintaining the model’s performance. A number of methods, including entropy and information gain, can be used to accomplish this goal.

# **CHAPTER THREE**

## **3.0 METHODOLOGY**

## **3.1 LOGISTIC REGRESSION**

Logistic regression is the appropriate regression analysis to conduct, when the categorical variable is independent. Logistic regression like other regression analysis is a predictive analysis with a dichotomous dependent variable and one or more nominal, ordinal, interval or ratio level independent variables.

At the core of the logistic regression analysis is a task of estimating the log odds of an event. mathematically, logistic regression estimates a multiple linear regression function defined as;

Where, is the probability of occurrence of the event of interest being true. is the slope, and are the slopes corresponding to the m independent variables respectively.

## **3.2 ASSUMPTIONS OF LOGISTIC REGRESSION**

The logistic regression as compared to other linear regression and general linear models that are based on the ordinary least square, does not necessarily require a linear relationship between the dependent and independent variables, secondly, the residuals do not need to follow the assumptions of normality. Thirdly, the error terms do not necessarily need to be homoscedastic. Also, the dependent variable is not measured in an interval or ratio scale.

However, some other assumptions still apply for logistic regression models, such as:

1. Dependent variable is dichotomous.
2. Observations are independent.
3. No severe multicollinearity.
4. Independent variables and log odds are linear.
5. Large sample size.

## **3.3 DICHOTOMOUS DEPENDENT VARIABLE**

A binary logistic regression requires the dependent variable to be of binary (dichotomous) possible outcome, such as positive or negative, true or false, yes or no, male or female, etc. While the ordinal logistic regression requires the dependent variable to be of ordinal data type.

## **3.4 INDEPENDENT OBSERVATIONS**

Logistic regression assumes that the observations of independent variables in the dataset are independent of each other. This implies that the observations should not come from repeated measurement of the same individual, or be related to each other in any way.

The logistic regression assumes that there is no correlation from observations within group (i.e. there is no autocorrelation).

When this assumption is violated, the standard error of the coefficients in a regression model are more likely to be under-estimated, which means that the predictor variables are more likely to be declared statistically significant, when they are actually not, leading to an erroneous conclusion.

### **3.4.1 Testing for Independence**

There are several ways to statistically test for independence within a set of a group of

observations, one of which is the Durbin-Watson test, which is used to detect autocorrelation

In the residuals of a regression.

Durbin-Watson test, test the following hypothesis;

With a test statistic d, given mathematically as;

Where; T is the total number of observations, and is the tth residual from the regression

model.

The test statistic always ranges from 0 to 4, where;

* d = 2 indicates no serial correlation,
* d < 2 indicates positive serial correlation,
* d > 2 indicates negative serial correlation.

In general, if d is lower than 1.5 or greater than 2.5, then there is potentially a serious serial correlation problem. Otherwise, if d is between 1.5 and 2.5, then the serial correlation is likely not a cause for concern.

## **3.5 MULTICOLLINEARITY**

Multicollinearity is the existence of a high inter-relationship between two or more independent variables in a multiple regression model. Multicollinearity can cause a model to yield screwed and misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a model.

Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model, making it hard to interpret models, and also creating a problem of overfitting. This means that an independent variable can be predicted from another independent variable.

In general, multicollinearity can lead to wider confidence intervals, producing less reliable probabilities in terms of the effect of the independent variables on a model.

### **3.5.1 Detecting multicollinearity**

Detecting multicollinearity can be done using different statistical methodologies, such as;

1. **Correlation matrix and scatter plot**

The correlation matrix and scatter plots can be used to detect multi collinearity by showing the bivariate relationship between independent variables.

1. **Variable inflation factor**

The variable inflation factors determine the strength of correlation between the independent variables. It is predicted by taking an independent variable and regressing it against other independent variables.

VIF is the reciprocal of tolerance, which is given mathematically as;

Where, is the coefficient of determination of the explanatory variable on all remaining variables.

Generalised variance inflation factor (GVIF)

1. **Eigen values, condition index and variance proportions**

Eigen values for the scaled uncentered cross-product matrix, condition indices and variance proportion for each explanatory variable are used to identify multicollinearity amongst regression variables. The regression parameters can be greatly affected by small changes in the explanatory variables or outcome, if any eigen value is larger than others. if eigen values are fairly similar, then the fitted model is likely to be unchanged by small changes in the measured variables.

The condition indices are computed as the square root of the ratio of the maximum eigen value and the eigen value of interest. It is defined as;

Where, and are the maximum and kth eigen values respectively. When there is no collinearity at all, the eigen values and condition indices will be equal 1. As collinearity increases, eigen values will be both greater and smaller than 1. There is an indication of multicollinearity, if one or more of the eigen value is close to zero, and the condition index is large. There is no fast and hard rule about the range of conditional index indicating multicollinearity, but an informal rule of thumb is that, when the condition index is within the range of 15 to 30, multicollinearity is a concern, and when its greater than 30, multicollinearity is a serious concern.

The variance of each regression coefficients can be broken down across the eigen values. The variance proportion explains the proportion of the variance of each regression coefficients that is attributed to each eigen values.