COMPARATIVE STUDY ON LOGIT AND PROBIT MODELS IN THE PREDICTION OF BRONCHO-PULMONARY DYSPLASIA STATUS OF INFANTS

BY

Ismail Adekunle KOLAWOLE

DEPARTMENT OF STATISTICS
AHMADU BELLO UNIVERSITY,

ZARIA, NIGERIA

JUNE, 2018

COMPARATIVE STUDY ON LOGIT AND PROBIT MODELS IN THE PREDICTION OF BRONCHO-PULMONARY DYSPLASIA STATUS OF INFANTS

BY

Ismail Adekunle KOLAWOLE
B.Sc. (A.B.U. 2014)

P14SCMT8024

A THESIS SUBMITTED TO THE SCHOOL OF POSTGRADUATE STUDIES,

AHMADU BELLO UNIVERSITY, ZARIA,

NIGERIA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE (M.Sc.) IN STATISTICS

DEPARTMENT OF STATISTICS,

FACULTY OF PHYSICAL SCIENCE

AHMADU BELLO UNIVERSITY,

ZARIA, NIGERIA

JUNE, 2018

DECLARATION

I declare that the work in this dissertation entitled "Comparative Study on Logit And Probit Models in the Prediction of Broncho-Pulmonary Dysplasia Status of Infants" has been performed by me in the Department of Statistics under the supervision of Dr. A. Yahaya and Dr. S.I.S Doguwa.The information derived from literature has been duly acknowledged in the text and a list of references provided. No part of this dissertation was previously presented for another degree or diploma at any University or Institution.

Kolawole Ismail Adekunle	_____	_____
Name of Student	Signature	Date

CERTIFICATION

The dissertation entitled "Comparative Study on Logit And Probit Models in the Prediction of Broncho-Pulmonary Dysplasia Status of Infants" by Ismail Adekunle KOLAWOLE (P14SCMT8024) meets the regulations governing the award of the degree of Master of Science of the Ahmadu Bello University, Zaria and is approved for its contribution to knowledge and literary presentation.

Dr. A. Yahaya Chairman, Supervisory Committee	_____ Signature	_____ Date
Dr. S. I. S. Doguwa Member, Supervisory Committee	_____ Signature	_____ Date
Dr. H. G. Dikko Head of Department	_____ Signature	_____ Date
Dr. Y. Musa External Examiner	_____ Signature	_____ Date
Prof. S. Z. Abubakar Dean, School of Postgraduate Studies	_____ Signature	_____ Date

DEDICATION

This dissertation is dedicated to Almighty Allah for His immeasurable Rahma and Sekinah, then to my beloved children.

ACKNOWLEDGEMENTS

My profound gratitude goes to Almighty Allah; the Nourished and Cherisher of the whole universe, who by His Mercy make this dream becomes a reality.

My sincere appreciation goes to my supervisors Dr. A. Yahaya and Dr. S.I.S. Doguwa for their relentless effort to make this work a success and their suggestions and strictness has really brought out the best in me.

I want to use this medium to appreciate my Lecturers especially Prof Asiribo, Dr. H.G. Dikko,Mallam Tasiu Musa and all other lecturers in the department for their guidance, constructive criticisms and direction all through this work.

My special thanks go toDr. S.A Abdulazeez and Dr. Aliyu Usman of Department of Mathematics and Statistic of Kaduna Polytechnic, Nigeria for their constant guidance and direction.

Unique thanks to my mother Hajia Muslimat Kolawole, my father Alhaji Abdulrasheed Kolawole and to my entire family. My appreciations go to my friends; Alh.Abideen Oladigbolu, Alh. Hakeem Adeniji and to my colleagues.

Finally to my wonderful wife, Hajia Nafisat Bello and children, I appreciate you all.

ABSTRACT

Broncho Pulmonary Dysplasia (BPD) is a form of chronic lung disease that develops in preterm neonates treated with oxygen and positive-pressure ventilation. The disease affects premature babies and contributes to their morbidity and mortality. This research seeks to fit and compare the predictive powers of Logistic Regression (Logit) Modeland Probability Regression (Probit) Model in tracking infants' BPD status using gender and weights at two different time intervals. The data used for the analysis were samples of 50 infants drawn from an underlying population of children with low birth weight (g) from Ahmadu Bello University Teaching Hospital Zaria. The children were confined to a neonatal intensive care unit, where they require intubation during the first 12 hours of life, and they survive for at least 28 days and their weights measured four weeks later. The results obtained found explanatory variables (weight at birth, weight after Four weeks of life and gender) used to be significantly associated with the occurrence of BPD in infants and suggested that, Probit fits BPD data more than Logit. It is therefore recommended that Clinics should adopt the use of the probit model fitted by this research to detect prevalence of BPD among infants so that adequate measures for prevention and control can be put in place early enough to signal the danger of the full manifestation of the disease.

Table of Contents

TITLE PAGE i
DECLARATION ii
CERTIFICATION iii
DEDICATION iv
ACKNOWLEDGEMENT v
ABSTRACT vi
TABLE OF CONTENTS vii
LIST OF TABLES ix
SYMBOLS AND ABBREVIATION x

CHAPTER ONE: INTRODUCTION 1
1.1 Background of the Study 1
1.2 Statement of The Problem 6
1.3 Aim and Objectives of The Study 7
1.4Significance of The Study 7
1.5 Scope and Limitation 7
1.6 Brief Methodology 8
1.6.1 Logistic Regression (Logit) Analysis 8
1.6.2 Probability Regression model (Probit) 8
1.7 Meaning and Definition of Terms 9

CHAPTER TWO: LITERATURE REVIEW 10
2.1 Introduction 10
2.2 Review of related Literature 10

CHAPTER THREE: MATERIALS AND METHODS 16
3.1 Introduction 16
3.2 Research Design 16
3.3 Population 17
3.4 Sample 17
3.5 Method of Data Analysis 17
3.6 Logit and Probit Regression 18
3.6.1 Logistic Regression (Logit) 18
3.6.2 Probability Regression (Probit) 26
3.6.3. Hypothesis and Confidence Interval for Logit and Probit 30
3.6.4 Deviance (G^2): A Measure of Goodness - of - Fit 31
3.6.5 Log-likelihood Ratio Test 32
3.6.6. Akaike Information Criterion (AIC) 33
3.6.7 Model's Performance Evaluation 33
CHAPTER FOUR: ANALYSIS, RESULT AND DISCUSSION 35

4.1 Introduction 35
4.2 Results of Model fit 35
4.3The results of the logistic regression of the BPD data 35
4.3.1 Accuracy of Logit Model 36
4.4 The results of the Probit Regression of the BPD data 38
4.4.1 Accuracy of Probit Model 38
4.5 Discussion 40
CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATION……..……..41
5.1 Summary 41
5.2 Conclusion 41
5.3 Recommendation 42
5.4 Contribution to Knowledge 42
References ..... …….43
Appendix ..... …….47

LIST OF TABLES

Table 1: Result of BPD data for Logit link.…….…….....36

Table 2: Logistic Model Accuracy……,…..……37

Table 3: Result of Logistic Regression AIC..……...……...37

Table 4: Estimation of sample logistic.....…….37

Table 5: Result of BPD data for Probit link …….…...38

Table 6: Probit Model Accuracy ……..……..……..……39

Table 7: Result of Probability Regression AIC ..……39

Table 8: Estimation of sample Probit ……...39

SYMBOLS AND ABBREVIATION

1. **BPD**: Broncho–Pulmonary Dysplasia.
2. **RDS**: Respiratory Distress Syndrome
3. **IUGR**: Intra-Uterine Growth Retardation.
4. **CPAP**: Continuous Positive Airway Pressure
5. **ELBW**: Extremely Low Birth Weight
6. **INSURE**: Intubation Surfactant Extubation
7. **NCPAP**: Nasal Continuous Positive Airway Pressure
8. **NICHD**: National Institute of Child Health and Human Development
9. **PAH**: Pulmonary Arterial Hypertension
10. **PDA**: Patent Ductus Arteriosus
11. **VLBW**: Very Low Birth Weight

CHAPTER ONE: INTRODUCTION

1.1 Background of the Study

The understanding of purpose of statistical science will play important roles to start a research of this kind. Usman (2016), in Bivariate and Multivariate Statistical Analysis, refers Multivariate statistical analysis as multiple advanced techniques for checking relationships among multiple variables at the same time. Researchers use multivariate techniques in a study that involve more than one response variable (phenomenon of interest) and more than one explanatory variable (also known as a predictor) or both. The statistical methods comes into play either when we have a medical theory to test or when we have a relationship in mind that has some importance in medical decision or policy analysis in public health.

According to Northway (1967), Broncho-Pulmonary Dysplasia (BPD)is a chronic lung disorder of infants and children and was first described in 1967. It is more common in infants with low birth weight and those who receive prolonged mechanical ventilation to treat respiratory distress syndrome (RDS).According to Namasiavayam (2014), BPD is a form of chronic lung disease that develops in preterm neonates treated with oxygen and positive-pressure ventilation. BPD is one of the most common chronic lung diseases in children. According to the National Heart, Lung, and Blood Institute (NHLBI), there are between 5,000 and 10,000 cases of BPD every year in the United States. Babies with highly low birth weight (less than 2.2 kilogram) are most at risk for developing BPD. In BPD, the lung and the airways (bronchi) are damaged in the neonatal period, causing destruction (dysplasia) of the tiny air sacs of the lung (alveoli). The pathogenesis of this condition remains complex and poorly understood; however various factors can not only injure small airways but also interfere with alveolarization (alveolar septation), leading to alveolar simplification with a reduction in the overall surface area for gas exchange.

The developing pulmonary microvasculature can also be injured. Many infants born with BPD exhibit signs and symptoms of respiratory distress syndrome, including the following: tachypnea, tachycardia, increased respiratory effort (with retractions, nasal flaring, and grunting), frequent desaturations.Namasiavayam (2014) found that, prematurely born infants, especially those born before 28 weeks of gestation, has few alveoli at the point of birth. The alveoli that are present are not matured enough to functioning well, and the infant requires respiratory support (a respirator) for breathing. Babies who are victim of premature or who have respiratory challenge shortly after birth are at risk of developing broncho pulmonary dysplasia, sometimes called chronic lung disease. Although life-saving, these treatments can also cause lung damage, referred to as "broncho [airway] pulmonary [lung] dysplasia [destruction]", or BPD. Broncho pulmonary dysplasia is a chronic lung disease that affects premature babies and contributes to their morbidity and mortality,Sahni (2005).

How BPD Affects Body

BPD directly affects both the lungs and the rest of the body. In the lungs, a significant number of alveoli that become fibrotic (scarred) and stop working. This damage affects not only the existing alveoli, but also those that continuously develop after birth. The low number of working alveoli means that the affected infant will need to remain on a breathing machine (ventilator) and/or receive oxygen for an extended period of time. This oxygen can cause further damage.

The damage to the alveoli also causes damage to the blood vessels around them, making the passage of blood through the lungs more difficult. In the long run, this leads to increases in the pressure inside blood vessels in the lungs and between the heart and lungs (pulmonary hypertension) and puts significant strain on the heart, which in severe cases may lead to heart

failure. Because of the low number of working alveoli, the affected infant although needs to breathe much faster and harder than healthy infants. This work may slow early growth because the infants neither have the energy nor the time to feed properly, thus taking fewer calories in than they should, and burning most of the calories just to breathe. This leaves fewer calories for them to grow, with poor growth or "failure to thrive" that in turn may cause problems to other organs of the body.

How Serious Is Broncho Pulmonary Dysplasia?

An estimated 10,000 newborns could develop BPD in the U.S. every year. Its severity varies from infant to infant. In mild cases, the infant may only have a faster than usual respiratory rate. In cases of moderate severity, the infant may require oxygen for several months. In uncommon but severe cases, the infant may have respiratory failure that requires not only oxygen but also prolonged need for mechanical ventilation.

BPD Symptoms, Causes and Risk Factors

Symptoms

The symptoms of BPD vary depending on its severity. Several risk factors make the development of BPD more likely but do not automatically lead to BPD. The most common symptoms of BPD are:

- Rapid breathing
- Laboured breathing (drawing in of the lower chest while breathing in).
- Wheezing (a soft whistling sound as the baby breathes out).
- Bluish discoloration of the skin around the lips and nails due to low oxygen in the blood.
- Poor growth.

- Repeated lung infections that may require hospitalization.

Causes

The cause of BPD is related to life saving oxygen and mechanical ventilation. While a relatively high amount of inhaled oxygen over several days may be necessary to support life, it may also cause damage to the alveoli. This is sometimes made worse when the ventilator blows air into the lung, overstretching the alveoli. Less well understood, inflammation can damage the inside lining of the airways, the alveoli and even the blood vessels around them. These effects are particularly damaging on the premature lung, and BPD is considered to be primarily a complication of prematurity.

Risk Factors

There are several conditions that do not cause but make the development of BPD more likely (risk factors) such as the following:

- **Degree of prematurity:** The less developed the lungs, the more they are likely to be damaged and result in BPD. BPD is rare in infants born after 32 weeks of pregnancy.
- **Prolonged mechanical ventilation:** Mechanical ventilation stretches the alveoli. When overstretched, and for longer periods of time they may be damaged.
- **High concentrations of oxygen:** The higher the concentration of oxygen and longer duration it is given, the higher the possibility of developing BPD. In general, concentrations of less than 60% oxygen are considered to be relatively safe.
- **Other risk factors.** These include:
    - **Patent ductus arteriosus**: The ductus arteriosus is a blood vessel that connects the right and left side of the heart that closes shortly after birth. This vessel is

more likely to remain open in premature infants causing lung damage when too much blood flows into the lungs.

- I**ntrauterine growth retardation (IUGR):** Different conditions may affect the growth of the fetus during the pregnancy and may also lead to premature labour. Relatively undeveloped lungs are more likely to develop BPD.

Logit and Probit regression models are members of Generalized Linear Model (GLM) that are widely used to estimate the functional relationship between binary response variable and predictors. The binary logit and probit models can be used to model functional relationship between a dichotomous response outcome and one or more predictors, (Krzanowki, 1998). When the outcome variable is dichotomous such as the case of BPD being considered in this study, both models are suitable for estimating the functional relationship between response variable and the predictors. Both models can therefore be used to analyze same data set for the same purpose, (Alison, 1999). Since the two models can be used for the same purpose, it is necessary to determine which model performs / predicts better.

Logit model is a technique for fitting a set of data when the response variable consists of proportion or binary coded data. Probit model is a type of binary classification model which is also appropriate in fitting regression curve when the response variable is a dichotomous variable and the predictors are either numerical or categorical, (Dobson 1990). Model fit can be improved by the selection of appropriate link for dichotomous data. The main focus of the research is to make comparison of the link function selection and model fits of logit and probit regression models in the fitting of BPD data.

Usman (2016), predictive modeling is one of the techniques used in statistical methodology by using historical information on a certain attribute to identify patterns which will help in predicting / determining a future value with a certain probability attached to it. Its application is valuable in the field of pharmacy and public health, particularly in medical settings. Some questions in medical research may contain dichotomous factor; in form of a person is a male or a female; a person does or does not have a disease in question, to mention but a few. In this study, we shall particularly fit and compare the two symmetrical dichotomous model; logit and probit, with prior knowledge for predicting BPD status of infants using gender and weight at two different survival time intervals of the infant as predictor variables and to establish the difference between the two stated models. In most cases, the model is used to make predictions in either the testing of a medical theory or the study of a policy's impact in pharmacy and public health.

This kind of research demands a careful control, so we have decided to use a record of BPD in ABU Teaching hospital Shika, Zaria.

1.2 Statement of the Problem

There have been reported cases of late discovery of BPD in infants which has been causing serious permanent health challenge for many people due to inability to discover it at an infancy as a result of clinical diagnosis of the disease which is rather expensive; however, a predictive model that will predict the disease could be a rare opportunity to detect the disease without passing through the rigour and expenses of the clinical diagnosis. Moreover, the model could be used to determine the prevalent rate of the diseasebased on the prior information available such as; weight at birth, weight after four weeks of birth and gender. Zysman *et al* (2013), discovered that gestational age and birth weight were correlated with the occurrence of BPD with each additional week of gestation. However, researches on statistical models that can appropriately

predict the disease using some other information are quite limited. Therefore, this research seeks to address the problem of predicting BPD in infants using weights and gender with the aid of Logit and Probit Models.

1.3 Aim and Objectives of the Study

The aim of this research is to fit and compare the symmetric dichotomous models that predict infants' BPD status using gender and weights. The following are specific objectives through which the stated aim would be achieved by;

i. fitting a Logistic Regression (Logit) model capable of tracking infants' BPD status.
ii. fitting a Probability Regression (Probit) model that can be used to predict infants' Broncho-pulmonary dysplasia (BPD) status based on gender and weight at two different times.
iii. comparing the two symmetric binary models fitted in (i) and (ii) above in order to assess the one that predict better.

1.4 Significance of the Study

This research will be of help to the research community especially, the medical practitioners to help them detect in time with the help of the fitted models, the prevalence of BPD among infants based on the weights and gender so as to take proper and adequate measures for controlling BPD.

1.5 Scope and Limitation

This study used samples of 50 infants drawn from an underlying population of children with low birth weight (g)from Ahmadu Bello University Teaching Hospital Zaria. These children were confined to a neonatal intensive care unit, they require intubation during the first 12 hours of life, and they survive for at least 28 days and their weights measured four weeks later. Infected

infants are denoted by (1) while normal infants by (0). Two statistical models were fitted which are; Logistic Regression (Logit) and Probability Regression Model (Probit) using BPD status, gender and weights at two different time interval.

1.6 Brief Methodology

1.6.1 Logistic Regression (Logit) Analysis

The goal of logit is to find the best fitting and most parsimonious model to describe the relationship between the outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables. The method is relatively robust, flexible and easily used, and it lends itself to a meaningful interpretation. In logit model the link function is the logit transform, $\ln\left(\mu/1-\mu\right)=\eta$. This research focuses on the case of a dichotomous outcome variable (Y). The logit model that will be fitted can be expressed as

$$P_i = \frac{e^{\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}}}{1+e^{\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}}} \quad \ldots \quad 1.1$$

Where P_i (the ps are independent Bernoulli random variables)is the probability that the ith infant has BPD,*fori= 1……n*

The coefficients of this model are estimated using the maximum likelihood method. Logistic regression model is discussed further by Hosmer and Lemeshow (1989).

1.6.2 Probability Regression model (Probit)

A Probit model (also called *Probit regression*), is a way to perform regression for binary outcome variables. Binary outcome variables are dependent variables with two possibilities like yes/no, positive testresult/negative test result or single/not single. The word “probit” is a

combination of the words probability and unit; the probit model estimates the probability that a value will fall into one of the two possible binary (i.e. unit) outcomes.

The Probit transformation or Probit link, is given by the inverse of the standard cumulative normal distribution function which gives;

$$P_i = \Phi\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}\right) \quad \ldots 1.2$$

Where P_i is the probability that the ith infant has BPD,*for* $i = 1 \ldots\ldots n$

1.7 Meaning and Definition of Terms

1. **Broncho** : Airways
2. **Dysplasia** : Destruction
3. **Alveoli** : Lung
4. **Ventilator** : Breathing Machine
5. **Ductus Arteriozus** : Blood vessel that connects the right and left side of the heart that closes shortly after birth

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

In every research, the review of related literature is necessary to achieving the desired research objectives. It is pertinent to begin every research work, particularly in this kind of statistical fitting, by outlining how other relevant literatures were consulted.

2.2 Review of Related Literature

A review of only those items relevant to the dissertation work has been made in this section, which has an immediate bearing to this research at hand. Broncho Pulmonary Dysplasia (BPD) continues to be a major cause of chronic morbidity among this population. Danan (2002) observed that there are large variations in the incidence and severity of this disease.

Some Key Facts about BPD

- BPD is associated with inflammation and scarring in the lungs
- BPD is much more common among low birth weight and premature infants.
- Most infants recover from BPD, but some may have long-term breathing difficulty.
- Infants are not born with BPD; the condition results from damage to the lungs which are caused by mechanical ventilation (respirator) and long-term use of oxygen.
- The severity of BPD is defined by the amount of oxygen an infant requires at time of birth and the length of use of supplemental oxygen or mechanical ventilation.

According to the National Institutes of Child Health and Human Development of USA (NICHD) consensus (2001),it reported that the incidence of BPD in Latin Americacomes from the neonatal group study of a very-low birth- weight (VLBW).

Tapia e*t al* (2006) examined that the BPD is a chronic pulmonary disease which affects premature infants and contributes to their morbidity and mortality. Despite substantial changes in incidence, risk factors and severity after the introduction of new therapies and mechanical ventilation (MV) techniques, BPD remains common, for more details refer to Tapia *et al* (2006).

Kumar *et al* (2011) conducted a study to determine the prevalence risk factors of nephropathy in type-2 diabetic patients. Here, it is remarked that Kumar *et al* (2011) discovered that as the duration of type-2 diabetes increases, the incidence of Nephropathy also increases significantly. Hence, all the type-2 diabetic patients, especially those with increased duration should be screened for Nephropathy and be made aware of the complications.

Carlos *et al* (2007) conducted a research which involves the building of model for the prediction of Broncho-pulmonary dysplasia model for seven-day old infants and their aim was to develop a predictive model capable of identifying which premature infants have the greatest probability of presenting (BPD) based on assessment at the end of the first week of life. Carlos *et al* (2007) concluded that at the end of the first week of life, the predictive model they developed was capable of identifying newborn infants at increased risk of developing BPD with high degree of sensitivity.

Boule *et al* (2001) proposed that adaptive control effects of exercise on glycemic control and body mass in type 2 diabetes mellitus is generally access by clinical trials. Maja *et al* (2004) worked on comparison of Logistic Regression and Linear Discriminant Analysis; a simulation study and their aim was the problem of choosing between the two methods and to set some guideline for proper choice. Maja *et al* (2004) found out that, LDA is a more appropriate method when the explanatory variables are normally distributed. In the case of categorized variables, LDA remains preferable and fails only when the number of categories is really small (2 or 3).The

results of LR, however, are in all these cases constantly close and a little worse than those of LDA. But whenever the assumptions of LDA are not met, the usage of LDA is not justified, while LR gives good results regardless of the distribution. As the estimates for LR are obtained by the maximum likelihood method, they have a number of nice asymptotic properties as well. Shah *et al* (2007), Inhaled corticosteroids have long been used as a therapy for patients who have developing or established BPD, but the evidence supporting their use is mixed.

A Cochrane systematic review of randomized controlled trials revealed that there is no evidence that early-inhaled corticosteroid therapy (at <2 weeks after birth) to ventilated preterm infants decreases the incidence of BPD. According to HIFI (High Frequency Ventilation in Premature) study in 1980, high frequency positive pressure ventilation, high frequency jet ventilation, and high frequency oscillation have been developed to provide artificial ventilation and reduce barotrauma. It is unclear whether any of these techniques offer any advantages over conventional mechanical ventilation in the routine treatment of respiratory failure of preterm infants. Their use does not seems to decrease the incidence of broncho pulmonary dysplasia, and may be associated with undesirable side effects such as increased incidence of grade III or IV intracranial haemorrhage.

Sauve *et al* (1985), observed that only one study looked at children as old as 8 years of age, and no significant effect of bronchopulmonary dysplasia on neurodevelopmental outcome was found.Kugelman and Durand (2001), observed that nasal ventilation techniques have been adopted as well, with the terms nasal intermittent mandatory ventilation and nasal intermittent positive pressure ventilation often used interchangeably.

Kugelman *et al* (2007), Nasal intermittent mandatory ventilation has been shown to decrease the rate of BPD when compared with NCPAP (Nasal Continuous Positive Airway Pressure) for

treatment of RDS (Respiratory Distress Syndrome), although the authors caution that the number of infants who have birth weights <1,500g in the study was small, and the results need to be validated in a powered study in the VLBW and ELBW populations. Zysman *et al* (2013), carried out a research to describe the characteristics of broncho-pulmonary dysplasia (BPD) and respiratory distress syndrome subjects, along with the trends in severity and mortality associated with BPD over the past three decades.

Zysman *et al* (2013), discovered that gestational age and birth weight were correlated with the occurrence of BPD with each additional week of gestation and 100g in birth weight being associated with an OR of developing BPD of 0.77 and 0.89, respectively.BPD severity was associated with male sex, and the occurrence of neonatal pneumonia.Significant trends were observed for lower mortality despite lower gestational age and birth weight, greater maternal age and multiple gestations.

Using retrospective study of BPD and respiratory distress syndrome, subjects born between 1980 and 2008, and admitted to Montreal Children's Hospital (Montreal, Quebec).Data were abstracted from hospital records, it was found out that, the mortality rate of infants with BPD has improved over the past three decades despite a significant trend toward more pronounced prematurity, lower birth weights, more advanced maternal age and multiple gestation.

Despite these factors and the introduction of routine use of pulmonary surfactant, the proportion of the various levels of BPD severity, from mild, moderate to severe, has remained unchanged.

Gerd *et al* (2012), wrote on development of lung function in very low birth weight infants with or without broncho pulmonary dysplasia. Longitudinal assessment during the first 15 months of corrected age to purposely compared functional lung development after discharge from hospital between VLBW infants with and without BPD.Gerd*et al* (2012) concluded that, the extent of

somatic growth, and the evolution of some lung function parameters, of very preterm infants with former BPD lag behind those characteristic of preterm infants without BPD for the first 15 months of life. The differences between the groups in most lung function parameters disappear after the somatic growth retardation of former BPD infants is taken into account.

Longitudinal LFT of preterm infants after discharge from hospital may help to identify at risk of incomplete recovery of respiratory function, which can lead to development of respiratory problems in childhood and adolescence.

Brunnella *et al* (2012), in their study of evaluating the prevalence and factors associated with bronch pulmonary dysplasia at a neonatal intensive care unit found out that the prevalence of broncho pulmonary dysplasia was high; the high prevalence was related to extreme prematurity, patent ductus arteriosus, a longer period under mechanical ventilation and prolonged hospitalization. The increased survival of infants with low gestational age makes this disorder a public health issue.

Eugene and Refik (2005), worked on Probit and Logit Models: Differences in the Multivariate Realm, and discovered that model fit can be improved by the selection of the appropriate link even in small data sets. Eugene and Refik (2005) observed that in multivariate link function models, the logit link provides better fit in the presence of extreme independent variable levels. Conversely, model fit in the random effects models with moderate size data sets is improved generally by selecting the probit link.

Gill (2001), puts it in discussing link functions including the cloglog, he indicated that they "provide identical substantive conclusions.Vishwa *et al*(2015), in application of Discriminant Analysis on Brocho-Pulmonary Dysplasia among infants: A case study of UMTH and UDUTH Hospitals, in Maiduguri and Sokoto respectively, Nigeria discovered that the prediction of BPD

is better done with discriminant model in UMTH, Maiduguri, while it has misclassified one of five new cases in UDUTH, Sokoto.

Chambers and Cox (1967), argued in their paper that probit and logit will yield different results in the multivariate context.Sadam *et al* (2017), worked on the comparison of some link functions of binary response analysis under symmetric and asymmetric assumptions using simulated sample size of n=50 and showed that, the link functions can be distinguished even with small value of (n<1000) under the two assumptions.

Gunduzand Fokouse(2013) said, it is not surprising that one would empirically notice virtually no difference when the two are compared in the same binary regression task. Despite this apparent in distinguishability due to many of their similarities, it is fair to recognize that the two functions different at least by definition and by their very algebra.

Gunduz and Fokouse(2013), persistently reveal the performance indistinguishability of the links in univariate settings, but some sharp difference begin to appear as the dimension of the input space (number of variables measured) increased.

Despite these developments, the properties of the link function for binary response models in the multivariate remain largely unexplored. This research seeks to fit symmetric Binary Models to predict BPD status of infant and to compare between the probit and logit models with real data.

CHAPTER THREE: MATERIALS AND METHODS

3.1 Introduction

In this case, we have to critically examine our methods and procedures as a precondition to achieving the desired objectives. Hence, this study will critically fit and compare the prediction of dichotomous outcome models in the classification of Infants' Broncho-Pulmonary Dysplasia as regards to the proper applications of biomedical modeling. The methodology must be carefully outlined; as the research would be conducted in the following format.

3.2 Research Design

In order to achieve the research objectives, this work employed cross – sectional classification as a research design. In the classification design, the researcher is not interested in a mere collection of haphazard facts but models would be used to classify the BPD status of an infant whose BPD is not known earlier.

Consider the three selected predictor variables which are capable of characterizing a BPD infant. From experience and records of medical practice, these variables are also believed to vary significantly between normal infants (π_1) and BPD infants (π_2). These variables are;

$X_1 =$ Weight at birth (g)

$X_2 =$ Weight after four week of birth (g)

$X_3 =$ Gender

The dichotomous variables are used to represent normal infants (π_1) and BPD infants (π_2) as the case may be. As used in the model, the dichotomous variable for normal infants (π_1) is 0 and that of BPD infants (π_2) is 1.

3.3 Population

Population is a collection of a known *N* number of identifiable units. In this case, *N* is called the population size. The population of this study is somewhat infinite as infants are given birth on daily basis. Any baby at birth is a potential BPD infant; hence, the population size cannot be specified at any point in time.

3.4 Sample

A sample is a part, a fraction, or a subset of the population. Samples are usually drawn with the aim of estimating the population quantities. Sampling is the act of drawing samples from the population; which saves time and cost. Usually *n* units are selected from the entire *N* units of the population. In this case, *n* is called the sample size. In this research work, samples will be drawn from the target population based on a statistically determined, efficient sample size so as to estimate some parameters of the population.

3.5 Method of Data Analysis

We shall fit Logit and Probit in analyzing the data. The data used in this study were collected through a well - designed clinical survey. In the process of analysis, data were duly classified into logical categories. The possible categories were considered when plans were made for collecting the data to facilitate analysis. Therefore, the process of analysis was partially concurrent with collection and presentation. Hence, there is the need to, first and foremost, present the data in their original form before extracting the desired analytical tables for the actual data analysis.it will be attached at the appendix part of this work. The birth weight (g), weight four weeks after birth (g), and gender, and the BPD status of infants will be collected and tabulated.

3.6 Logit and Probit Regression

Logistic and Probability regression models (sometime called Binomial regression) are adequate for those situations where the dependent variable of the regression problem is binary. That is, the dependent variable has only two possible outcomes, e.g., "success/failure" or "normal/abnormal". We assume that these binary outcomes are dummy coded as 1 and 0. Among the link functions for Binomial regression are logit and probit transformation, (Agresti 2002, Krzanowki 1998, and Dobson 1990). The choice of the link function can be critical to the accuracy of the result of binary modeling of a data set.

3.6.1 Logistic Regression (Logit)

Logistic regression models are adequate for those situations where the dependent variable of the regression problem is binary. That is, the dependent variable has only two possible outcomes, e.g., "success/failure" or "normal/abnormal". We assume that these binary outcomes are coded as 1 and 0. The application of linear regression models to such problems would not be satisfactory since the fitted predicted response would ignore the restriction of binary taking on values for the observed data.

According to Efron and Hastie (2016), defined logit model of a binary response variable as follows;

$$\ln(odds) = \text{logit}(p) = \ln\left(\frac{p_i}{1-p_i}\right) = \sum_{k=0}^{k=n} \beta_k X_{ik} \quad \ldots 3.1$$

Where;

P_i is the probability that the ithinfant has BPD. *For i=1,2 ...n*

$$X_i = \begin{pmatrix} x_{10} & x_{11} & x_{12}\ldots & x_{1k} \\ x_{20} & x_{21} & x_{22}\ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & x_{n2}\ldots & x_{nk} \end{pmatrix} \quad \ldots 3.2$$

In most regression – type models, the first column is a vector of ones to allow for a constant term, so that $x_{i0} = 1$

$$X_i = \begin{pmatrix} 1 & x_{11} & x_{12}\ldots & x_{1k} \\ 1 & x_{21} & x_{22}\ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2}\ldots & x_{nk} \end{pmatrix} \quad \ldots 3.3$$

and

$$\beta = \begin{pmatrix} \\ \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \ldots 3.4$$

The ratio $\left(\frac{p_i}{1-p_i}\right)$ is the odd of having BPD against a normal infant at $P = \frac{1}{2}$ and $Y_i = 1$

The inverse transformation of odd ratio is therefore given by

$$\Lambda^{-1}(p) = \ln\left(\frac{p_i}{1-p_i}\right) \quad \ldots 3.5$$

It is the log of odds that Y_i is 1 rather than 0. Any value of P in the range (0,1) is transformed into a value of the $\text{logit}(p)$ in $(-\infty, +\infty)$ so that as $p \to 0$, $\text{logit}(p) \to -\infty$ (Lawal 2003).

Adekanmbi (2017), logit model can be generalized to k explanatory variables which require a linear predictor which is a function of several predictors

$$P_i = \Lambda(\eta_i) = \Lambda(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}) \quad \ldots 3.6$$

$$P_i = \Lambda(X_i^{'}\beta) \quad \ldots 3.7$$

$$P_i = \frac{1}{1+\exp\left[-(X_i^{'}\beta)\right]} \quad \ldots 3.8$$

Also

$$\text{logit}(P_i) = \ln\left(\frac{p_i}{1-p_i}\right) \quad \ldots 3.9$$

$$\text{logit}(P_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} \quad \ldots 3.10$$

Using equation 3.3 and 3.4, we have

$$\text{logit}(P_i) = X_i^{'}\beta \quad \ldots 3.11$$

The odd can very on a scale of $(-\infty, +\infty)$, so that the log – odds can vary on the scale of $(-\infty, +\infty)$

There is no error term in logit regression model, unlike in classical linear regression, Adekanmbi (2017).

Efron and Hastlie (2016), exponentiated equation 3.9 as;

$$\text{logit}(P_i) = \exp\left[\ln\left(\frac{p_i}{1-p_i}\right)\right] \quad \ldots 3.12$$

$$\text{logit}(P_i) = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}) \quad \ldots 3.13$$

$$\text{logit}(P_i) = (e^{\beta_0})(e^{\beta_1})^{X_{i1}} (e^{\beta_2})^{X_{i2}} ... (e^{\beta_k})^{X_{ik}} \quad \ldots 3.14$$

The (e^{β_j}) is the multiplication effect on the odds of increasing X_j by 1 while holding other $X's$ constant.

In this work, we shall fit the logit model as follows; when studying linear regression, we attempted to estimate a population regression equation;

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon \quad \ldots 3.15$$

By fitting the model of the form;

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_k X_{ik} \quad \ldots 3.16$$

The response *Y* is continuous, and is assumed to follow a normal distribution. We were concerned with predicting or estimating the mean value of the response corresponding to a given set of values for the explanatory variable.

There are many situations, however, in which the response of interest is dichotomous rather than continuous. Examples of variables that assume only two possible values are disease status (the disease is either present or absent) and survival following surgery (a patient is either alive or dead). In general, the value 1 is used to represent a "success" or the outcome we are not interested in, and 0 represents a "failure". The mean of the dichotomous random variable *Y*, designated by *p*, is the proportion of times that it takes the value 1. Equivalently;

$$p = P(Y = 1) = P(success) \quad \ldots 3.17$$

Just as we estimated the mean value of the response when Y was continuous, we would like to estimate the probability p associated with a dichotomous response (which, of course, is also its mean) for various values of an explanatory variable. To do this, we use the technique of *logistic regression*. A simple regression model for this situation is:

$$Y_i = g(x_i) + \varepsilon_i \qquad \ldots 3.18$$

With

$$y_i \in \{0,1\}$$

In this case we shall only consider multiple logistic regression– that is, logistic regression models with three explanatory variables. Our first strategy might be to fit a model of the form:

$$p = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} \qquad \ldots 3.19$$

This is simply the standard linear regression model in which x_s represents the explanatory variables and y – the outcome of a continuous, normally distributed random variable which has been replaced by p. As before, α is the intercept and β_i is its slope. On inspection, however, this model is not feasible. Since p is a probability, it is restricted to taking values between 0 and 1. The term $\alpha + \beta_1 x_{i1} + \beta_{i2} x_{i2} + \beta_3 x_{i3}$, in contrast, could easily yield a value that lies outside this range. Instead we might try to solve this problem by fitting the model;

$$p = e^{\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}} \qquad \ldots 3.20$$

This equation guarantees that the estimate of p is positive. We would soon realize, however, that this model is also unsuitable. Although the term $e^{\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}$ cannot produce a negative

estimate of *p*, it can result in a value that is greater than 1. To accommodate this final constraint, we fit a model of the form;

$$P_i = \frac{e^{\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}}}{1+e^{\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}}} \qquad \ldots 3.21$$

This expression on the right, called a *logistic function* (*logit model*), cannot yield a value that is either negative or greater than 1; consequently, it restricts the estimated value of *p* to the required range.

Therefore, in order to handle the binary valued response, we apply a mapping from the predictor domain onto the $[0,1]$interval. This mapping is an example of S-shaped functions, also called the *sigmoidal* functions; which is an example of non-linear regression. The logistic response enjoys the interesting property of simple linearization. As a matter of fact, denoting as before the mean response of the probability *p*, and if we apply the *logit* transformation.

The term "***Logit***" as a contraction of the phrase "**log**arithmic un**it**" was introduced by Berkson (1944). Recall that if an event occurs with probability *p*, odds in favour of the event are;

$$\frac{p}{1-p} \quad to \quad 1$$

Thus, if a success occurs with probability

$$P_i = \frac{e^{\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}}}{1+e^{\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}}} \qquad \ldots 3.22$$

the odds in favour of success are

$$\frac{p}{1-p} = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3} / (1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3})}{1-(e^{\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3} / 1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3})} \quad \ldots 3.23$$

$$= \frac{e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}} / (1+e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}})}{(1+e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}} - e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}}) / (1+e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}})} \quad \ldots 3.24$$

$$= \frac{e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}} \times (1+e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}})}{(1+e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}})} \quad \ldots 3.25$$

$$= e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}} \quad \ldots 3.26$$

Taking the natural logarithms of each side of this equation;

$$\ln\left(\frac{p}{1-p}\right) = \ln(e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}}) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} \quad \ldots 3.27$$

Where; the link function is

$$\ln\left(\frac{p}{1-p}\right)$$

Thus, modeling the probability *p* with a logistic function is equivalent to fitting a linear regression model in which the continuous response *y* has been replaced by the logarithms of the odds of success for a dichotomous random variable. Instead of assuming that the relationship between *p* and *x* is linear, we assume that the relationship between;

$\ln\left(\frac{p}{1-p}\right)$ and *x* is linear.

The technique of fitting a model of this form is known as logistic regression; the estimated relationship between the explanatory variable and the odd in favour of success is given as:

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\alpha} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} \qquad \ldots 3.28$$

However, we cannot apply the method of least squares, which assumes that the response is continuous and normally distributed, to fit a logistic regression model; instead we use the method of maximum likelihood. The method of least squares cannot be applied because the random error component is not normally distributed. The method of maximum likelihood uses the information in a sample to find the parameter estimates that are most likely to have produced the observed data. Let us see how this method is applied for the proposed logit model. We start by assuming a Bernoulli random variable associated to each observation y_i; therefore, the joint distribution of the n observations is given as;

$$P(y_1,\ldots,y_n) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \qquad \ldots 3.29$$

Taking the natural logarithms of this likelihood function, we obtain;

$$\ln P(y_1,\ldots,y_n) = \sum_i y_i \ln\left(\frac{p_i}{1-p_i}\right) + (1-y_i)\sum_i \ln(1-p_i) \qquad \ldots 3.30$$

Using the fact that;

$$P_i = \frac{e^{\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}}}{1+e^{\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}}} \qquad \ldots 3.31$$

And

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} \quad \ldots 3.32$$

The logarithms of this likelihood function (*log-likelihood*), which is a function of the coefficients, $L(\beta)$, can be expressed as follows:

$$L(\beta) = \sum_i y_i(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + (1-y_i)\sum_i \ln\left[1+\exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})\right] \quad \ldots 3.33$$

The maximization of $L(\beta)$ function can be carried out using one of the many numerical optimization methods, such as the Quasi-Newton method, which iteratively improves current estimates of function maxima using estimates of its first and second order derivatives. The numerical optimization methods cannot be discussed in text; rather, the use of computer packages will be employed to obtain the parameter estimates of the models.

Assumptions of Logit

1. It has standard logistic distribution of errors
2. Independently observation.
3. The dependent variable Y_i does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family.
4. Logit assumes linear relationship between the transformed response in terms of the link function and the explanatory variables;
5. The homogeneity of variance does NOT need to be satisfied.

3.6.2 Probability Regression (probit)

The term "***Probit***" as a contraction of the phrase "**Pro**bability un**it**" was introduced by Bliss (1935). Probit is also referred to as inverse normal function (Efron and Hastlie 2016, and Lawal

2003). In order to ensure that P is between 0 and 1, a positive monotone function that maps the linear prediction, $(\eta = \alpha + \beta X_i)$ into the unit interval.

$$P_i = P(\eta_i) = P(\alpha + \beta X_i) \quad \ldots 3.34$$

Where $P(.)$: cumulative distribution function, α and β are parameters to be estimated. A reasonable a priori should be both smooth and symmetric and should approach $P = 0$ and $P = 1$ as asymptotes, (Adekanmbi 2017).

The probit model has not been widely applied to social and biological science problem. The model is especially useful in epidemiological and demographic research in the assessment of the effect of explanatory factors on the relative risk of outcomes such as; fertility, mortality, and on the onset of disease or illness. Probit models are adequate for those situations where the dependent variable of the regression problem is binary. That is, the dependent variable has only two possible outcomes, e.g., "success/failure" or "normal/abnormal".

The early origins of the Probit models can be traced to psycho-physics (Thurstone,1927). Modern developments of the Probit models, however, were developed in the field of bioassay or dose-response methodology (Cox 1970; Finney 1971). Binomial response models can be motivated by considering the situation where the outcome of an event is dichotomous given number of independent variables. The Probit model is a class of multivariate symmetric function. Again, a nonlinear model in p is transformed so that a monotonic function of p is linear with respect to explanatory variables. The probability in the ith cell or the ith observation, *pi* is given by the standard cumulative normal distribution function:

$$P_i = \int_{-\infty}^{\eta_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) du \quad U \sim N(0,1) \quad \ldots 3.35$$

The equation in 3.35 can be more conveniently written as

$$P_i = \Phi(\eta_i) \quad \ldots 3.36$$

$$P_i = \Phi(\alpha + \beta X_i) \quad \ldots 3.37$$

$$\Phi^{-1}(P_i) = \eta_i = \sum_{k=0}^{k=n} \beta_k X_{ik} \quad \ldots 3.38$$

The probit model can also be generalized to k explanatory variables such that

$$P_i = \Phi\left(\sum_{k=0}^{k=n} \beta_k X_{ik}\right) \quad \ldots 3.39$$

$$P_i = \Phi(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}) \quad \ldots 3.40$$

For this research work, we shall fit

$$P_i = \Phi(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}) \quad \ldots 3.41$$

Where P_i is the probability that the ith infant has BPD. *for i= 1……n*

Where $\Phi(.)$ denotes the cumulative distribution function of the standard normal distribution.

In probit regression, the errors are assumed to have a standard normal distribution if Λ is defined in terms of the inverse normal probability integral as $\Lambda = \Phi^{-1}(P)$ then Λ is referred to as the probit of p , Adekanmbi (2017).

It is known that, we cannot apply the method of least squares, to fit a probability regression model; instead we use the method of maximum likelihood. We start by assuming a Bernoulli random variable associated to each observation y_i; therefore, the joint distribution of the n observations is given as;

$$f(P) = \Pr\left(y/np\right) = \binom{n}{y} P^y (1-P)^{n-y} \quad \ldots 3.42$$

Assuming independence of observations, the joint density or likelihood is the product of the individual densities:

$$L = \prod_i f\left(P_i\right) = \prod_i \binom{n_i}{y_i} P_i^{y_i} \left(1-P\right)^{n_i - y_i} \quad \ldots 3.43$$

Letting the individual probability depends on a smaller set of covariates (x_i) and unknown parameters (β),the likelihood can be expressed as:

$$L = \prod_i \binom{n_i}{y_i} F\left(x_i^1 \beta\right)^{y_i} \left[1 - F\left(x_i^1 \beta\right)\right]^{n_i - y_i} \quad \ldots 3.44$$

And

$$\log L = \sum \left\{ \log \binom{n_i}{y_i} + y_i \log F\left(x_i^1 \beta\right) + \left(n_i - y_i\right) \log \left[1 - F\left(x_i^1 \beta\right)\right] \right\} \quad \ldots 3.45$$

Where $x_i^1 \beta = \sum_{k=0}^{k=n} \beta_k x_{ik}$ and $F(.)$ is a cumulative probability distribution function for the logistic, normal, or other suitable distributions. The binomial coefficient $\binom{n_i}{y_i}$ appearing in equation 3.44 and 3.45 is simply a constant multiplier, which does not involve unknown parameters. Therefore, in practice we maximize the log-likelihood that is proportional to equation 3.45

$$\log L = \sum \left\{ y_i \log F\left(x_i^1 \beta\right) + \left(n_i - y_i\right) \log \left[1 - F\left(x_i^1 \beta\right)\right] \right\} \quad \ldots 3.46$$

This expression is maximized to yield optimal values of β. Different statistical packages rely on different methods to find maximum likelihood estimator of probit models.

Adekanbi (2017), the difference between logit and probit models lies in the assumption about the distribution of errors, logit has standard logistic distribution of errors while probit has normal

distribution of errors. The probit and logit models usually produce almost identical marginal effects. McCullah and Nelder (1992), and Long (1999), for logit and probit models, the predicted probabilities are limited between 0 and 1. Cakmakyapan and Goktas (2013), when the variance of logit and probit transformations are equated, the two transformations become so similar that it becomes difficult to establish the different. There are practical advantages of logit transformation over the probit transformation. Efron and Hastlie (2016), and Long (1999), observed that equation of logistic cumulative distribution function (CDF) is simply compared to CDF of probit especially for binary outcome data. Efron and Hastlie (2016), the inverse linearizing transformation for the logit model $\Lambda^{-1}(p_i)$ is directly interpreted as log-odds as opposed to the inverse transformation $\Phi^{-1}(p_i)$ in probit that does not have a direct interpretation.

Assumptions of Probit

1. It has normal distribution of errors
2. Independent observations.
3. It assumes linear relationship between the transformed response in terms of the link function and the explanatory variables;
4. The homogeneity of variance does NOT need to be satisfied.

3.6.3. Hypothesis and Confidence Interval for Logit and Probit

Hypothesis tests for logit and probit models are based on Wald statistic, Alison (1999). For an individual coefficient to test the hypothesis: $H_0 : \beta_j^0$, the Wald statistic should be calculated as;

$$Z_0 = \frac{\beta_j - \beta_j^{(0)}}{S.E\left(\beta_j\right)} \quad \ldots 3.47$$

Where

$S.E\left(\beta_j\right)$: The asymptotic standard error of β_j.

Z_0, Follows an asymptotic uni-normal distribution under the null hypothesis.

The $100\left(1-\alpha\right)\%$ confidence interval for β_j is

$$\beta_j = \beta_j \pm Z_{\alpha/2} S.E\left(\beta_j\right) \quad \ldots 3.48$$

Where

$Z_{\alpha/2}$: The value from $Z \sim N\left(0,1\right)$ with a probability of $\alpha/2$ to the right.

3.6.4Deviance (G^2): A Measure of Goodness - of - Fit

A commonly reported indicator of mode fit is given by the likelihood – ratio statistic (or deviance/G^2), Daniel and Yu (1999). It is an extent to which a generalized linear model adequately represents a set of binary data, Krzanowski (1998). The goodness-of-fit of GLM can be assessed by computing the scaled deviance and comparing the result with a χ^2 distribution with the relevant degree of freedom (McCullagh 1992)

$$G^2 = 2\left[l\left(\hat{\theta}_{\max}; y\right) - l\left(\theta; y\right)\right] \quad \ldots 3.49$$

Where

$\hat{\theta}_{\max}$: is the maximal model likelihood estimate.

Generally, a generalized linear model that is based on estimate of p parameters from a data set with n observations will have its test statistic distributed as χ^2_{n-p} (Krzanowski 1998).

3.6.5 Log-likelihood Ratio Test

The omnibus Chi-square test is a log-likelihood ratio test for investigating the model coefficients in logistic and probability regression. The test procedures are as follows:

Hypothesis for Omnibus Chi-square Test:

H_0: The model coefficients are not statistically significant

H_1: The model coefficients are statistically significant

Test statistic:

$$\chi^2 = 2\left[\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \ln\left(\frac{O_{ij}}{e_{ij}}\right)\right] \quad \ldots 3.50$$

Or

$$\chi^2 = 2\left[\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \ln O_{ij} - \sum_{i=1}^{r} R_i \ln R_i - \sum_{j=1}^{c} C_j \ln C_j - n \ln n\right] \quad \ldots 3.51$$

Where R = number of rows

C = Number of columns

n = Number of observations

Decision Rule:

Reject H_0if $p<0.05$ otherwise accept H_0 at the 5% level of significance. The computations are obtained using Stata as employed for data analysis in this study.

Justification for the Use of Omnibus Chi-square Test

It is justifiable and even necessary to investigate the significance of the model coefficient in the logistic and probability regression model. Hence, the Omnibus test is applied.

3.6.6. Akaike Information Criterion (AIC)

There are several criteria for selecting the best model in GLM fitting as advocated by several authors such as (Lawal 2003 and McCullagh 1992). One of criteria method that will be used for this research is Aikaike Information Criterion (AIC). Clayton etal (1986), AIC is one of the model selection criteria and it is defined as

$$AIC = -2\ln(L) + 2p \qquad \ldots 3.52$$

Where

L: maximized value of the likelihood function for the estimated model.

P: number of parameters in the model.

Agresti (2002), when comparing competing models fitted by maximum likelihood to the same data, the smaller the AIC the better the fit.

3.6.7 Model's Performance Evaluation

The evaluation of the performance of the logit and probit models fitted is based on some statistical criteria or performance metrics. The performance metrics employed in this work include accuracy (AC), sensitivity (SE) and specificity (SP).

Sensitivity (SE): Measures its ability to correctly identify subjects with the disease (true positive rate, e.g., the percentage of BPD patients who are correctly identified as having BPD)

$$SE = \frac{TP}{(TP + FN)} \times 100\% \quad \ldots 3.53$$

Specificity (SP): is the ability to correctly identify those without the disease (true negative rate, e.g., the percentage of healthy patients who are correctly identified as not having the condition).

$$SP = \frac{TN}{(TN + FP)} \times 100\% \quad \ldots 3.54$$

Accuracy (AC): The accuracy of diagnostic tests is often assessed with the two conditional probability ;sensitivity and specitivity

$$AC = \frac{TP + TN}{(TN + TP + FN + FP)} \times 100\% \quad \ldots 3.55$$

Where;

TP refers to the number of true positives

TN refers to the number of true negative

FN refers to the number of false negatives, and

FP is the number of false positives

CHAPTER FOUR: ANALYSIS, RESULTS AND DISCUSSION

4.1 Introduction

This chapter contains the analyses and results obtained using the prescribed methods in the previous chapter. We have used the sample of infants drawn from an underlying population of children with low birth weight (g) from Ahmadu Bello University Teaching Hospital Shika Zaria, Nigeria. These children were confined to a neonatal intensive care unit, they require intubation during the first 12 hours of life, and they survive for at least 28 days and their weights measured four weeks later. Infected infants are denoted by (1) while normal infants by (0). The explanatory variables used for the fitted models are; weight at birth (X_1), weight after four weeks of life (X_2) and gender (X_3). The Logistic regression (logit) and Probability regression (probit) models were fitted, the models fit and performances were carried out on both logit and probit models to check their accuracy and see the one that performs better.

4.2 Results of Model fit

Fitting and evaluating the performance of the logit and probit model to the BPD data as well as comparing the models and select the one that can best fit the data is carried out with STATA package.

4.3 The results of the logistic regression of the data

The fitted logit model for the BPD of infant has the following equation;

$$\text{logit}(p) = 13.8227 + 0.0401x_1 - 0.0452x_2 + 1.9138x_3 \quad \ldots 4.1$$

The results of logistic regression of the BPD data are shown in the Table 1. The logit model contains three predictor variables and could be concluded that the model fit the BPD data and can be used to predict the occurrence of the disease. The p-valueslessthan of 0.05 and could be concluded that the weight at birth, weight after four weeks of life and gender were significantly associated with the BPD status of an infants as it can be seen in the Table1.

Logistic regression Number of obs = 50
LR chi2(3) = 34.47
Prob > chi2 = 0.0000

Log likelihood = -16.414689 Pseudo R2 = 0.5122

Table 1. *Result of BPD data for Logit link*

BPDStatus1	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
WeightAtBirth	.0400614	.018124	2.21	0.027	.004539	.0755837
WeightAfter4weeks	-.0452054	.0193445	-2.34	0.019	-.0831199	-.007291
Gender1	1.913838	.9693105	1.97	0.048	.014024	3.813651
_cons	13.82268	5.426074	2.55	0.011	3.18777	24.45759

4.3.1 Accuracy of Logit Model

Table 2 is on how accurate the logit model is in fitting BPD status of infants using the stated explanatory variables (*weight at birth, weight after four weeks of life and gender*). The specificity which measures the percentage of healthy patients who are correctly identified as not having the condition showed 80.00% and the sensitivity which is the percentage of BPD patients who are correctly identified as having BPD showed 83.33%. The total accuracy of logit model using the conditional probabilities; specificity and sensitivity in fitting the BPD data is 82%.

Table 2. *Logistic Model Accuracy*

Classified	True D	True ~D	Total
+	16	5	21
-	4	25	29
Total	20	30	50

Classified + if predicted Pr(D) >= .5
True D defined as BPDStatus1 != 0

Sensitivity	Pr(+| D)	80.00%
Specificity	Pr(-|~D)	83.33%
Positive predictive value	Pr(D| +)	76.19%
Negative predictive value	Pr(~D| -)	86.21%
False + rate for true ~D	Pr(+|~D)	16.67%
False - rate for true D	Pr(-| D)	20.00%
False + rate for classified +	Pr(~D| +)	23.81%
False - rate for classified -	Pr(D| -)	13.79%
Correctly classified		82.00%

Logistic model for BPDStatus1, goodness-of-fit test

number of observations	=	50
number of covariate patterns	=	45
Pearson chi2(41)	=	24.55
Prob > chi2	=	0.9804

Table 3. *Result of Logistic regression AIC*

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	50	-33.65058	-16.41469	4	40.82938	48.47747

Table 4. *Estimation sample logit*

Variable	Mean	Std. Dev.	Min	Max
BPDStatus1	.4	.4948717	0	1
WeightAtBi~h	1387.08	349.6371667	1892	
WeightAfte~s	1168.2	333.5688550172	0	
Gender1	.42	.4985694	0	1

4.4 The results of the Probit Regression of the BPD data

The fitted probit model for the BPD of infant has the following equation;

$$P_i = \Phi(\eta_i) = \Phi(8.0341 + 0.0234x_1 - 0.0264x_2 + 1.1115x_3) \quad \ldots 4.2$$

The Table 5 depicts the results of probability regression of the BPD data. The probit model contains three predictors' variables and could be concluded that the model fit the BPD data and can be used to predict the occurrence of the disease. The p-values less than 0.05 and could be concluded that the weight at birth, weight after four weeks of life and gender were significantly associated with the BPD status of an infants as can be seen in the table5.

Probit regression Number of obs = 50
LR chi2(3) = 34.63
Prob > chi2 = 0.0000
Log likelihood = -16.337331 Pseudo R2 = 0.5145

Table 5. *Result of BPD data for Probit link*

BPDStatus1	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
WeightAtBirth	.0234096	.0105571	2.22	0.027	.002718 .0441012
WeightAfter4weeks	-.0268842	.0112469	-2.34	0.019	-.048411 -.0043241
Gender1	1.111506	.5572554	1.99	0.046	.0193055 2.203707
_cons	8.034131	3.14733	2.55	0.011	1.865478 14.20278

4.4.1 Accuracy of Probit Model

Table 6 depicts how accurate the logit model is in fitting BPD status of infants using the stated explanatory variables (*weight at birth, weight after four weeks of life and gender*). The specificity which measures the percentage of healthy patients who are correctly identified as not having the condition showed 80.00% and the sensitivity which is the percentage of BPD patients who are correctly identified as having BPD showed 83.33%. The total accuracy of probit model using the conditional probabilities; specificity and sensitivity in fitting the BPD data is 82%.

Table 6. *Probit Model Accuracy*

Classified	True D	True ~D	Total
+	16	5	21
-	4	25	29
Total	20	30	50

Classified + if predicted Pr(D) >= .5
True D defined as BPDStatus1 != 0

Sensitivity	Pr(+| D)	80.00%
Specificity	Pr(-|~D)	83.33%
Positive predictive value	Pr(D| +)	76.19%
Negative predictive value	Pr(~D| -)	86.21%
False + rate for true ~D	Pr(+|~D)	16.67%
False - rate for true D	Pr(-| D)	20.00%
False + rate for classified +	Pr(~D| +)	23.81%
False - rate for classified -	Pr(D| -)	13.79%
Correctly classified		82.00%

Probit model for BPDStatus1, goodness-of-fit test

number of observations	=	50
number of covariate patterns	=	45
Pearson chi2(41)	=	23.71
Prob > chi2	=	0.9859

Table 7. *Result of Probability regression AIC*

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	50	-33.65058	-16.33733	4	40.67466	48.32275

Table 8.*Estimation of sample probit*

Variable	Mean	Std. Dev.	Min	Max
BPDStatus1	.4	.4948717	0	1
WeightAtBi~h	1168.2	333.5688	550	1720
WeightAfte~s	1387.08	349.6371	677	1892
Gender1	.42	.4985694	0	1

4.5 Discussion

One of the criterion-based model selection used in this research work to select the best model on the response variable *BPD status* and the three predictors; *weight at birth, weight after four weeks of life and gender* for the logit and probit regression models is AIC. The models were fitted differently for logit and probit, and the model that produced the lowest values of AIC is selected as best model to fit BPD data. Considering the results in Table 3 and Table 7,logit model has AIC value of (40.82938) which is the maximum value as compare to probit model which has (40.67466). According to the procedure in this work, the probit model is found to fit BPD data better than logit model.The performance evaluation for both logit and probit modelscan be seen from Table 2 and Table 6 outputs and showed that both have 82%accuracy to identified subject. Checking the Table 1 and Table 5 of the logit and probit, it could be seen that, the p-values are less than 0.05 and be concluded that the explanatory variables; *weight at birth, weight after four weeks of life and gender* were significantly associated with the BPD status of an infants.

CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATION

5.1 Summary

Our goal in this work as initially stated in our objectivesto fit symmetric dichotomous models; logit and probit models that capable of predicting infants' BPD status using gender and weights at two different times and compare the results. The study used the data of BPD status obtained from ABU teaching hospital Shika, Zaria in Nigeria to see the performance of the two stated models. The performance of the logit and probit models were checked with AIC and the accuracy of predicting the actual patient into the normal infants or BPD infants was carried out using category evaluation metrics: accuracy, sensitivity and specificity. Stata package was used for the analysis and it was discovered that probit model is more accurate in fitting BPD data as a result of its least AIC as compare to logit model. Furthermore, the model performance which measures the performance of the classification models using the stated evaluation metrics, probit and logit models were found to have 82% accurately categorize the BPD infants and normal infants to their respective categories. It was also observed that the p-values are less than 0.05 which indicated that the explanatory variables used were significantly associated with the occurrence of BPD in infants.

5.2 Conclusion

The results of our analysis indicated that both the logit and probit model fit BPD data and the explanatory variables were found to be significantly associated with the occurrence of BPD in infants. The probit model was found to fit the BPD data more than the logit model. As far as accuracy of the logit and probit link is concerned to categorize the normal infant and BPD infants to their respective classes, probit and logit models were found to have 82% accuracy to classify

infants intotheir respective categories. We have been able to make an improvement in the field of medical diagnosis using statistical methods to detect the occurrence of the disease using the stated explanatory variables without passing through the rigour and expenses of the clinical diagnosis.

5.3 Recommendation

From the analysis and evaluation of results in the previous discussions in this study so far, the following recommendations are proffered.

1. In the light of the above it is recommended that Clinics should adopt the use of the probit model fitted by this research to detect prevalence of BPD among infants so that adequate measures for prevention and control can be put in place early enough to signal the danger of the full manifestation of the disease.
2. It is also recommended that the model should be adopted to reduce the cost of clinical diagnosis.

5.4 Contribution to Knowledge

We have been able to make an improvement in the field of medical diagnosis using statistical methods to detect the occurrence of the disease without going through the expense of medical diagnosis through;

a) Logistic Regression (Logit) model and
b) Probability Regression (Logit) model fitted by this research work.

References

Adekanmbi, D. B. (2017). Comparison of Probit and Logit Models for Binary Response Variable with Applications to Birth data in South-Western, Nigeria. *American Journal of Mathematics and Statistics* 7(5):199-208.

Agrestic, A. (2002). *Categorical Data Analysis.* 2ed. New-York: John Willey.

Alison, P.D. (1999). Comparing logit and probit coefficients across groups. Sociology Methods and Research. 28:186-208.

Berkson, J. (1944). Application of the logistic function to Bio-Assay:*Journal of the American Statistical Association, 39: 194-227.*

Bliss C. I. (1935). The Comparison of Dosage-Mortality Data: *Annal of Applied Biology*

Brunnella, A., Chagas F., Mirene P. (2012)., Guilherme Lobo da Silveira3, *Giana Zarbato Longo4 Rev Bras Ter Intensiva. 24(2):179-183*

Boule, N. G., Haddad, E., Kenny, G. P., Wells, G.A. and Sigal, R. J. (2001). Effects of exercise on glycemic control and body mass in type 2 diabetes mellitus: a meta-analysis of controlled clinical trials.*The Journal of the American Medical Association,286(10), 1218-1227,*

Carlos A., Bhering Christieny C., Mochdece M., Moreira E.L., Jose R. Rocco, and Guilherme M. (2007). Broncho Pulmonary dysplasia prediction model for 7-day-old infants. *Journal of Pediatria. (Rio J.), 83(2):111-121,*

Cakmakyapan, S., and Goktas A. (2013). A comparison of binary and probit models with a simulation study. *Journal of Social and Economic Statistics.2(1): 1-17.*

Chambers, E.A. and Cox D.R.(1967). Discrimination between alternative binary response models. *Biometrika 54, 573-578*

Clayton, M.K, Geisser, S., and Jenning, D.E. (1986). A Comparison of several model selection procedures. Bayesian Inference and Decision Technoques, Amsterdam: North Holland. 196.

Cox, D.R. (1970).*The Analysis of Binary Data*, London: Methuen.

Danan C. (2002), Gelatinase activities in the airways of premature infants and development of Bronchopulmonary dysplasia. *American Journal of Physiol. Lung Cell Mol Physiol., 283: L1086- L1093,*

Daniel A. P. and YU X. (1999). *Statistical Methods for Categorical Data Analysis*. Academic Press, Inc.

Dobson, A.J. (1990).*An introduction to generalized linear models*. London: Chapman and Hall,

Efron, B. and Hastie, T. (2016).*Computer age statistical inference: algorithms, evidence and data science*. Cambridge: Cambridge University Press.

Eugene, D. A andRefik, S. (2005). Probit and Logit Models: Difference in the Multivariate Realm. USA

Finney, D. J. (1971). *Probit Analysis*,3rd ed., Cambridge: Cambridge University Press

Gerd S., Silke W., Charles C. R., Hans, P. and Christoph B. (2012). *http://www.biomedcentral..com/1471-2431/12/37*

Gill, J. (2001). General Linear Models: A Unified Approved. Thousand Oaks, CA:Sage.

Gunduz, N. and Fokouse, E. (2013). On the predictive analytics of the probit and logit link function. Accessed from *http:scholarworks.rit.edu/article/1235*

Hosmer, D. W. and Lemeshow, S. (1989): *Applied Logistic Regression*. New York: Wiley.

Krzanowski W.J. (1998).*An Introduction to Statistical Modeling*. New-York: Oxford University Press.

Kugelman A., and Durand M.(2001). A comprehensive approach to the prevention of Broncho Pulmonary Dysplasia. *Pediatr Pulmonol.doi 10.1002/ ppul.21508:112-130*

Kugelman A., Feferkorn I., Riskin A., Chistyakov I., Kaufman B. and Bader D. (2007). Nasal intermittent mandatory ventilation versus nasal continuous positive airway pressure for respiratory distress syndroame: a randomized, controlled, prospective study. *J Pediatr. 150(5): 521–526.*

Kumar, V., Moses, A. and Padmanaba N.(2011), Prevalence and risk factors of nephropathy in type 2 Diabetic patients:*Journal of Collaborative Research on Internal Medicine and Public Health.101(5): 221–256*

Lawal, B. (2003). *Categorical data analysis with SAS and SPSS Application*. New-Jersey: Lawrence Erlbaum Associates.

Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables.* CA: sage Press

Maja, P. Mateja, B. and Sandra, T. (2004). Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. *Metodološki zvezki, 1, 143-161*

McCullagh, P. and Nelder, J. (1992). *Generalized linear models (2nd)*. London: Chapman and Hall.

Namasiavayam A. (2014). *Medscape Education Cardiology*,

National Institutes of Child Health and Human Development of USA (NICHD) Consensus (2001). National Center for Biotechnology Information, U.S. National Library of Medicine8600 Rockville Pike, BethesdaMD, 20894USA

Northway, W.H Jr, Rosan R. C., Porter, D.Y,(1967). Pulmonary disease following respirator therapy of hyaline-membrane disease. Bronchopulmonary dysplasia. The New England *JournalofMedicine.* ***276*** *(7): 357–68.*doi:10.1056/NEJM196702162760701. PMID 5334613.

Sadam, A. D., Musa T., Yusuf B. Musa M. N., Nurudeen A.A., and Samson A. (2017).On the Comparison of some link functions of Binary Response Analysis Under Symmetric and Asymmetric Assumptions. *Biomedical Statistics and information. 2(5),245-149. doi: 10.11648j.bsi.20170205.15*

Sahni, R. (2005). "Is the new definition of bronchopulmonary dysplasia more useful?". Journal of perinatology : *Official journal of the California Perinatal Association.* ***25*** *(1): 41–6. doi*:*10.1038/sj.jp.7211210*. *PMID 15538399.*

Sauve, R. S. and Singhal N. (1985). Long term morbidity of infants with Broncho Pulmonary Dysplasia. *Journal of Pediatrics,76:725-33.*

Shah, V., Ohlsson A., Halliday, H. L., Dunn, M. S.(2007). Early administration of inhaled corticosteroids for preventing chronic lung disease in ventilated very low birth weight preterm neonates. Cochrane Database Syst Rev.(4):CD001969

Tapia, J. L., Agost D., and Alegria A. (2006). Bronchopulmonary dysplasia: incidence, risk factors and resource utilization in a population of South American very low birth weight infants: *Journal de Pediatria, 82(1), 15–20.*

Thurstone, L. L. (1927). "A law of Comparative Judgement", *Psychological Review*,34:273-286.

Usman, A. (2016). *Bivariate and Multivariate Statistical Analysis (1st ed).* Millennium Printing and Publishing Company Limited, Kaduna, Nigeria.

Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of Statistical Society Series B, 58:267-2888.*

Vishwa, N., Madaki, U., Vijay, V. and Babagana, M. (2015). Application of Discriminant Analysis on Broncho-Pulmonary Dysplasia among Infants: A Case Study of UMTH and UDUS Hospitals in Maidguri, Nigeria. *American Journal of Theoretical and Applied Statistics*. 4, 2-1, 44-51.

Zysman Z. C., Tremblay, G. M., Bandeali, S. and Landry, J. S.(2013).Broncho Pulmonary Dysplasia – trends over three decades. Paediatr Child Health;18(2):86-90.