UNIVERSITY FOR DEVELOPMENT STUDIES, TAMALE

INVESTIGATIVE MODELLING OF ANAEMIA IN PREGNANCY IN GHANA: A CASE STUDY IN BOLGATANGA

BY

AHMED RABIU ZANGWIO (B.SC. ACTUARIAL SCIENCE)

(UDS/MAS/0021/12)

THESIS SUBMITTED TO THE DEPARTMENT OF STATISTICS, FACULTY OF MATHEMATICAL SCIENCES, UNIVERSITY FOR DEVELOPMENT STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF MASTER OF SCIENCE DEGREE IN APPLIED STATISTICS

JANUARY, 2015

DECLARATION

Student

I hereby declare that this thesis is the result of my own original work and that no part of it has been presented for another degree in this university or elsewhere:

Candidate's Signature:.. Date: 09-02-15.............................

Name: AHMED RABIU DANGWID..

Supervisor

I hereby declare that the preparation and presentation of the thesis was supervised in accordance with the guidelines on supervision of thesis laid down by the University for Development Studies:

Supervisor's Signature:.. Date: 09/02/2015..........................

Name: DR. ALBERT LUGUTERAH..

ABSTRACT

Anaemia is the condition of low levels of haemoglobin in the blood, which results in a reduced amount of oxygen being transported in the body. In this study data on 580 pregnant women were obtained from the ante-natal care unit of the Bolgatanga regional hospital was used in investigating the presence of anaemia in pregnancy in the municipality and some of the factors which contribute significantly to anaemia in pregnancy using the Artificial Neural Network model, the Logit model and the Probit model to investigate and get best appropriate model for determining some of the factors and prevalence of anaemia in pregnant women and in the municipality respectively. The results revealed that the Artificial Neural Network model was the best model for investigating and determining some of the factors and prevalence of anaemia in pregnancy. This Artificial Neural Network model had the least AIC and BIC of 147.487 and 173.665 respectively across the Logit and Probit models, with a better predictive potential of 64.8%. The training and testing of the data for the Artificial Neural Network also showed the model being adequate. The Artificial Neural Network was proposed to be the best model for modelling anaemia in pregnancy. However, each model had it's limitation as there is no standard method for finding the ideal number of hidden nodes in a feed-forward network or for determining the best activation function for model equations when using the Artificial Neural Network. The Probit and Logit model equations limit the researcher by requiring a large amount of sample data to estimate the parameters of the equation and to find the data trend. Therefore more research needs to be done to get a better understanding of how both models are related.

ACKNOWLEDGEMENTS

It is a tradition in academic circles to acknowledge the assistance one received from colleagues in the writing of an academic document. It is in this foregoing context that all those who contributed in diverse ways to the success of this work were duly acknowledged.

Many thanks to my supervisor, Dr. Albert Luguterah for his professional guide, endless effort, and urge to ensuring that this work emerges as one of the standard essays in the university.

I would also like to thank Madam Monica Achalwe (In-Charge of ANC, BRH), Madam Elizabeth Abodan (In-charge of RCH, BRH), Mrs Leticia Denteh (midwife, BRH) and Evelyn Tretu (midwife, BRH), for all the support and help you gave me to make this work a successful one, am extremely grateful.

My gratitude is also due the staff of the Faculty of Mathematical Sciences. I have learned so much from you, and I thank you for showing me such kindness. Words cannot express how grateful I am to each of you especially Mr. Nasiru Suleman.

Lastly to my friends and family; to my parents for their support and love throughout this two years, my siblings Samira Zangwio, Rekiya Sackey, Nadia Sackey and Forzia Sackey, and now to my friends Ida Anuwoje Lugubayom, Joyceline Annor, Nana Ama Nyarko Appiah, Ahmed Sadiq and Priscilla Nana Ama Nkrumah.

I would also like to extend my gratitude to my friends, especially, those I have met in UDS for being so nice and supportive.

DEDICATION

To my parents Mr. Harrison Zangwio and Mrs Jemilatu Cobblah whose courage inspires me to dream big.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

ADALINE	Adaptive Linear Elements
AIC	Akaike Information Criteria
AIDS	Acquired Immune Deficiency Syndrome
ALSPAC	Avon Longitudinal Study of Parents and Children
ANN	Artificial Neural Network
ANV	Ante-Natal Visit
BIC	Bayesian Information Criteria
BMA	Bolgatanga Municipal Assembly
BMI	Body Mass Index
BRH	Bolgatanga Regional Hospital
CDF	Cumulative Distribution Function
DPB	Diastolic Blood Pressure
FNN	Feed-Forward Neural Network
GDHS	Ghana Demographic Health Survey
GHS	Ghana Health Service
GNA	Ghana News Agency

HIV	Human Immune Virus
LBW	Low Birth Weight
MADALINE	Multiple Adaptive Linear Elements
MCV	Mean Cell Volume
MLP	Multi-Layer Perceptron
MPR	Multiple Polynomial Regression
OR	Odds Ratio
PTB	Preterm Birth
SBP	Systolic Blood Pressure
SGA	Small Gestational Age
UNDP	United Nations Development Programme
WHO	World Health Organization

CHAPTER ONE

INTRODUCTION

1.0 Background of study

For years now anaemia in pregnant women has been a global problem in developing countries. Africa which is one of the developing continents is far hit with this canker. Worldwide, about 41.8% of pregnant women are anaemic as compared with 30.2% non-pregnant women; the most severely affected areas are South-East Asia (48.2%) and Africa (57.1%).A large proportion of the 17.2 million anaemic women in Africa live in the west-African sub-region (WHO, 2005).

The prevalence rate in some of the countries range from 50.2% in Togo, 66.7% in Nigeria, 68.3% in Burkina Faso, 72.7% in Benin,75.1% in Gambia and 75% in Ghana (WHO, 2005).

Many measures have been taken to reduce or eradicate this canker but Africa still remains to be the continent with the highest rating of this problem (Broek *et al*., 1999). Anaemia in pregnancy is high in the west- Africa sub-region, being one of the major causes of maternal mortality in the sub-region. In most developing countries anaemia in pregnancy makes an important contribution to maternal mortality and morbidity (Broek *et al*., 1999). Anaemia is the condition of low levels of haemoglobin in the blood. This results in a reduced amount of oxygen being transported in the body. Iron is a main component of haemoglobin and iron deficiency is estimated to be responsible for over half of all anaemia globally. Other causes of anaemia include malaria, hookworm and other helminthes; other nutritional deficiencies such as vitamin A, vitamin B12 or folic acid deficiency, chronic infections and genetic conditions which vary by region (such as sickle cell and thalassemia); HIV/AIDS; and high fertility. Anaemia is a serious concern for pregnant women and children. It increases the risk of maternal mortality and poor birth outcomes. In children anaemia impairs cognitive development, stunts growth and increases morbidity from infectious diseases. It is estimated that more than two billion people globally, mostly women and children, are anaemic. Anaemia is considered a public health problem when more than 5% of the intended population is anaemic, a significant public health problem in need of immediate action when prevalence exceeds 20%, and a severe public health problem when prevalence exceeds 40% (WHO/UNICEF/UNU 2001).

The etiological factors for anaemia in pregnancy are multiple and their relative contributions vary by geographical area and by season. In West Africa, the most common cause is iron deficiency. Other causes include parasitic infestations such as malaria and hookworm; infections like HIV and haemoglobinopathies. The predisposing factors are grand-multiparity, young age, low socioeconomic status, and illiteracy (WHO, 1999).

Iron-deficiency anaemia accounts for 85% of all cases of anaemia that are identified and is characterized by low mean cell volume (MCV). It is usually caused by nutritional deficiency or low iron stores resulting from previous pregnancy or previous heavy menstrual blood loss (Blackwell, 2008). Haemoglobin is a commonly used, well validated, and widely accepted indicator for anaemia. Mean haemoglobin is one useful way to present this indicator. However, anaemia is also commonly presented based on cutoffs. Pregnant women with haemoglobin level less than 11g/dl are said to be non-anaemic, pregnant women with haemoglobin level between 10g/dl-10.9g/dl are said to have mild anaemia, pregnant women with haemoglobin level between 7g/dl-9.9g/dl have moderate anaemia and pregnant women with haemoglobin level less than 7g/dl have severe anaemia (WHO, 2001).

Even though anaemia has long been recognized in Africa as a major health problem, it is highly prevalent in Ghana and the fourth leading reason for admissions in many hospitals. It has been described as the second leading cause of death (GHS, 2004). The normal physiological change of an increase in plasma volume causes hemodilution in a pregnant woman. Although the red cell mass increases, plasma volume increases disproportionately, resulting in a lowering of the haemoglobin level to approximately 11g/dl which leads to anaemia (Willacy, 2012).

The main causes of anaemia can be attributed to poor diet that has low iron bio-availability (extent to which a nutrient is absorbed into the blood stream), low absorption enhancers, inhibitors, excessive blood loss, hook worm infestation, breakdown of the red blood cells (GNA, 2004). Fruits, meat, fish, peanuts, grains, green vegetables and others are in abundance but the sad aspect is that our mothers and sisters in the villages during the harvest sell the best part of their produce with the hope to get a good price and consume the 'bad' ones forgetting that they also deserve the best of the farm produce (GNA, 2004).

Chronic infections, parasitic worm infestations and genetic disorders have also been identified as other sources of anaemia. Statistics from the Nutrition Unit of the Ministry of Health's national study conducted in 1995, using the World Health Organization (WHO) standards indicated that 64.5 per cent of pregnant women were anaemic whilst 59 per cent were lactating women (GNA, 2004).

According to GNA (2004), anaemia has been identified as the major cause of maternal death associated with delivery, it has also been identified as directly related to increased risk of premature delivery, low birth weight, impaired physical growth, still birth and infant deaths. Anaemia in pregnancy could also lead to anaemia in the fetus, damaging the brain of the fetus, which could cause a possible permanent impairment in mental ability of school children and adolescents. It is sad to note that anaemia; something that could be prevented and even controlled has taken a bigger percentage of the causes of death in our hospitals, having all the sources of food and fruits that can enhance our iron level abundantly (GNA, 2004). Preliminary findings of the 2008 Ghana Demographic Health Survey (GDHS) has revealed a high prevalence rate of anaemia among women and children (GDHS, 2009).

Pregnant women account for 65 per cent, 41 per cent for women of child bearing age and 76 per cent for pre-school children account for the high prevalence rate of anemia among women and children (GDHS, 2009).

Anaemia, also accounted for 20 per cent of maternal death with the current micro-nutrient deficiencies situation having an immense impact on child mortality (GHS, 2012).

Ghana is losing lots of workers' productivity due to effects of anemia among men and women, especially for those engaged in manual labour, indicating that based on the profile's analysis.

It is projected that between 2011 and 2020, GH¢1.9 billion (US$1.3 BN) will be lost in manual labour, including agricultural productivity, as a consequence of iron deficiency anemia. The refusal of pregnant women to take their drugs, poor eating habits and the refusal to use insecticide treated bed nets have contributed to anaemia among pregnant women and children less than five years.

Although, many researches have been done on anaemia, researchers have mostly used the Logistic regression model to find out some of the significant factors which contribute significantly to anaemia in pregnancy in the world and in Ghana. Research has rarely compared the logistic regression model, the Probit Regression model and the Artificial Neural Network model (ANN). The models used for the purpose of this research were employed due to the fact that, the Logit model and the Probit model are non-linear statistical models with the Artificial Neural Network model being an algorithm that can perform non-linear statistical modelling. The three models were chosen because their methods are the most commonly used methods for developing predictive models for dichotomous outcomes in medicine.

1.1 Problem Statement

Anaemia currently affects 40–80% of women in Africa of which Ghana is not an exception. It is projected that about 9,000 pregnant Ghanaian women would die by the year 2020, if the high anaemia levels amongst them persist (WHO, 2012). It has been estimated that 20 per cent of the country's disturbing maternal mortality ratio of 451 per 100,000 live births is caused by this health condition (GNA 2012).

Life expectancy in the Bolgatanga municipality is only 50 years; compared to the national figure of 55 years. The high level of illiteracy and poverty as well as limited access to safe drinking water and the existence of poor sanitation and unhygienic practices have exposed many people to health hazards which contribute to the lowering of the living standards of the people (BMA, 2006).The prevalence of diseases like anaemia, acute respiratory infections and gynecological disorders as well as the outbreak of epidemics such as cholera, anthrax and CSM can be traced to the above factors. Other conditions such as malnutrition and micro nutrient deficiency (mostly among children) are also prevalent in the municipality (BMA, 2006).

Anaemia continues to be a disease of major public health problem. It is the number one cause of morbidity and accounts for 46% of OPD attendance and 11% of all institutional deaths in the municipality. A total of 32,128 anaemic cases representing 43% occurred in children 5 years of age were recorded during the year 2004. About 47 deaths occurred due to anaemia, 36 of which occurred in children 5 years of age representing 0.1% under 5 case fatality rate (BRH, 2003). Children in the Upper East and Upper West regions are the most likely to be anaemic (88-89 percent). The prevalence of anaemia in children has also increased slightly over the past five years, from 76 percent in 2003 to 78 percent in 2008 (GDHS, 2008). While models for investigating anaemia have been intensively studied to determine some predictors of anaemia, research has rarely used the neural network model for the modelling of independent predictors of anaemia. Research has also rarely compared models between the Logistic regression model, the Probit regression model and the Artificial Neural Network model to try finding out the appropriate model to determine some of the factors which contribute to this problem.

1.2 Significance of Study

The main aim of this research is to investigate between the models Logistic Regression, Probit Regression and Artificial Neural Network with regards to model the probability of a pregnant woman being anaemic or non-anaemic to select which model is the most appropriate for use when it comes to determining some of the significant factors which contribute to anaemia. The study will unearth an appropriate model that determines independent factors which contribute to anaemia in pregnancy.

It would be useful for policy makers and health workers for planning interventions and effective strategies for addressing the problem of anemia in pregnancy on a long-term treatment.

This study would help management and control of anaemia in pregnancy, which is enhanced by the availability of local prevalence statistics, which is however not adequately provided in the Bolgatanga municipality.

1.2 Objectives of the study

i. To unearth an appropriate model for predicting the incidence or severity of anaemia in pregnancy.

ii. To investigate the factors that contributes to anaemia in pregnancy.

iii. To determine the prevalence of anaemia in pregnancy in the municipality.

1.4 Structure of the Thesis

The thesis is organised into five chapters. Chapter one contains the introduction of the research work. Chapter two contains literature review. Chapter three outlines the methodology employed in this research while chapter four presents the analysis and discussion of results. Chapter five is devoted to conclusion and recommendations.

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

This chapter reviews empirical works done on anaemia in pregnancy and some relevant models that have been used in modelling anaemia in pregnancy. The chapter is divided into eight main headings namely; Introduction, Prevalence of anaemia in Ghana, causes of anaemia in pregnancy, reduced work productivity, Increased maternal mortality, adverse birth outcomes, empirical researches on anaemia in pregnancy and review of some, logistic, probit and neural network methods.

2.1 Prevalence of Anaemia in Ghana

Anaemia, defined as blood haemoglobin level below established cut-off points, is a pervasive global public health problem. An estimated 2 billion people are affected, or more than one third of the world's population. Anaemia prevalence is highest in developing countries. Although both males and females of all ages are affected, the most vulnerable groups are pregnant women and young children. Worldwide, more than 50% of pregnant women and over 30% of all women suffer from anaemia (WHO, 2001).

An estimated 58.27 million women worldwide are anaemic during pregnancy, 95.7% of whom live in developing countries. In Africa, the prevalence of anaemia in pregnancy has been estimated to be 35-75%. It continues to be a topical issue in many developing countries because of its association with adverse pregnancy outcome such as increased rates of maternal and perinatal mortality, premature delivery, low birth weight among others (WHO, 1992).

In Ghana, seventy-seven percent of children age 6-59 months and 47 percent of mothers are anaemic (GDHS, 2003).

Anaemia rates for children are highest in the Upper East and Northern regions (84 percent) and lowest in Greater Accra region (62 percent) (GDHS, 2003). Anaemia rates for mothers are highest in the Upper East region (55 percent) and lowest in Brong-Ahafo region (32 percent) (GDHS, 2003).

A review of the health, disease profiles and pathology reports in selected hospitals in Ghana ranks anaemia as the fourth leading reason for hospital admissions and the second cause of maternal death (GHS, 2004). About 83.5 per cent of pre-school children, 71.3 per cent of school age children, 64.5 per cent of pregnant women and 59 per cent of lactating women are anaemic. An analysis of the economic impact of anaemia estimated that from 2001 to 2005, Ghana's female workforce would lose over ¢141.6 billion in economic output due to anaemia (GHS, 2004).

In Ghana, the maternal mortality rate is estimated to be 214 per 100 000 live births. There are considerable differences between the regions, with the deprived northern regions showing MMR (Maternal Mortality Ratio) of over 800 deaths per 100,000 live births (GHS, 2004).

The major causes of maternal mortality in Ghana include hemorrhage, sepsis/infections, pregnancy induced hypertension, unsafe abortions, and obstructed labor, anemia. Other broader factors that have been identified as contributing to maternal deaths include poverty, high fertility, malnutrition, and traditional beliefs and practices, with anemia having a percentage of 12 (GHS, 2004).

2.2 Causes of Anaemia in Pregnancy

In developing countries like Nigeria and Ghana the cause of anaemia is multi-factorial and varies greatly by geographical location, season and dietary intake. The most common causes of anaemia in Nigeria and Ghana include nutritional deficiencies of iron and folate, parasitic diseases such as malaria and hookworm, haemoglobinopathies such as sickle cell disease and recently human immunodeficiency virus infection. Most of these causes of anaemia in pregnancy are preventable. However, despite the use of iron and folate supplementation and anti-malarial prophylaxis, which are prescribed for pregnant women in ante-natal clinics for the prevention of anaemia, the prevalence of anaemia is still high in these countries. This shows that there are other likely underlying factors that are contributing to the high prevalence of anaemia recorded in Nigeria and Ghana (Vanderjagt *et al.*, 2007).

Apart from permanent organ damage, worm infections cause anaemia, poor physical growth, poor intellectual development and impaired cognitive function all cause anaemia (Crompton *et al.*, 2002). Poor nutrition in general and anaemia in particular are the main underlying causes of poor pregnancy outcomes in the developing world including low birth weights, result in the spontaneous abortion, neonatal and maternal deaths (Jilly, 1969; McGregor *et al.*, 1983; Brabin, 1991; Egwunyenga *et al.*,1997). In areas of Africa with stable malaria transmission, malaria infection during pregnancy is estimated to cause 400,000 cases of severe maternal anaemia and from 75,000 - 200,000 infant deaths each year (Steketee, 2001). Local prevalence of risk factors for iron deficiency and anaemia may vary broadly between populations. While malaria infection or the acquired immunodeficiency syndrome are common in the African continent and are important contributors to anaemia in women of reproductive age (Bouvier *et al.*, 1997; Verhoeff *et al.*, 1997), hookworm infections whose prevalence and intensity vary by geographic region, may also serve as an important cause of anaemia in women of reproduction age (Dreyfuss *et al.*, 2000).

2.3 Reduced Work Productivity

Haemoglobin transports oxygen in the blood for delivery to the body's tissues. Therefore one of the first signs of low haemoglobin, or anaemia, is fatigue, due to lack of oxygen for physical activity. For the world's many anaemic women, this causes work productivity and incomes to suffer, as well as the ability to carry out daily tasks and to nurture and care for children. The relationship between anaemia and reduced productivity has been well documented. A literature review examining the association between iron deficiency and work capacity identified a strong causal effect of severe and moderate iron deficiency anaemia on aerobic capacity, which translates into reduced physical activity and productivity. Studies in a variety of countries have shown an improvement in work capacity for laborers in various occupations with the provision of iron supplementation. A review of studies comparing work output in relation to changes in haemoglobin found consistent results across countries and contexts. A 10 percent increase in haemoglobin levels was associated with a 10 to 20 percent increase in work output (WHO, 2001).

2.4 Increased maternal mortality

Severe anaemia (Hb<7g/dL) in pregnancy is associated with increased risk of maternal mortality. It has been estimated that the increased risk of death for pregnant women with Hb<7g/dL is 1.35, and that those with Hb<5g/dL have a 3.5 times greater risk of dying from obstetric complications compared with non-anaemic women (WHO, 2001).

2.5 Adverse birth outcomes

The results of several studies have shown an association between maternal iron deficiency anaemia in early pregnancy and a greater risk of preterm delivery and consequent low birth weight. Low birth weight greatly increases the risk of neonatal mortality and morbidity, and is also associated with a variety of deficits in health, development and cognitive growth for the surviving infant. In addition, infants of anaemic mothers have reduced iron stores continuing into the first year of life, increasing their vulnerability to iron deficiency and anaemia (WHO, 2001).

2.6 Researches on Anaemia in pregnancy

Many researches have been made on anaemia in pregnancy, with most researchers using the main cause of anaemia in pregnancy due to lack of iron deficiency and haemoglobin concentration as a proxy for iron deficiency anaemia accepted by the World Health Organization due to its low cost and relative ease of determination. Most researchers used logistic regression models and probit regression to model the prevalence and predictors of anaemia in pregnancy.

Pala and Dundar (2007) modelled prevalence and risk factors of anaemia among women of reproductive age in Bursa, Turkey by using multivariate logistic regression analysis. A sample of 530 women using stratified random sampling among 6,506 women in 15-49 age group and 488 women (92.1%) participated in the study. Pregnant women or women who were not sure of their pregnancy were not included in the study. They then studied the usage of sanitary pads used to determine the risk factors of anaemia among women of reproductive age. Presence of anaemia was accepted as the dependent variable in the model; the independent variables were age, education, marital status, job, parity, body mass index, and menstruation characteristics (regularity of cycle, length of cycle, length of flow, sanitary pad usage). They found out that the prevalence of anaemia was significant using the standard acceptable level of haemoglobin (<12g/dl) with multivariable logistic regression analysis, consumption of more than 2 sanitary pads during menstruation and more than five days of menstrual bleeding were found to be risk factors for anaemia. No independent relation was observed between anaemia and age, education,

marital status, job, parity, BMI, regularity of cycle and length of cycle.

Next, Karaoglu *et al.,* (2010) modelled prevalence of nutritional anaemia in pregnancy in the east Anatolian province in Turkey by using Chi-square test to detect any association between anaemia prevalence and independent variables. Mann-Whitney U and Kruskall Wallis tests were used in making comparisons between the red blood cell indices. A p value of < 0.05 was considered statistically significant. Backward logistic regression analysis was performed to evaluate the independent association existing between the potential risk factors and anaemia. Independent variables that were significant at the $p = 0.05$ level in univariate analysis were included in multivariate analysis to control for confounding in regression models. Trimester, family income, number of live births, number of living children, family structure, PICA, receiving social support were included in regression model as dichotomous variables. The results were presented in odds ratios (OR) and 95% confidence intervals. Pregnant women with haemoglobin level (Hb< 11.0 g/dl) were considered anaemic for the general sample while doing the statistical analysis.

Also, Nwizu *et al.*, (2011) used multivariate logistic regression to compute adjusted odds ratio (OR) and to determining the independent predictors of anaemia in pregnancy in kano state, Nigeria. They sampled 300 pregnant women using systematic sampling and tested them for anaemia based on the WHO criteria for mean minimum acceptable haemoglobin level during pregnancy which is taken to be 11g/dl (PCV of 33%) in the first half of pregnancy and 10.5 g/dl in the second half of pregnancy, and used predictors of anaemia, such as educational status, marital status, socio-economic class, parity, inter-pregnancy interval and gestational age at booking against the response variable anaemia and found out that there were 51 pregnant women with anaemia out of the 300 giving a prevalence of 17% and anaemia in pregnancy was significantly associated with all the predictors (educational status, marital status, socio-economic class, parity, inter-pregnancy interval and gestational age at booking), by using the multivariate logistic regression at a significance level of 0.05.

Next, Sukrat *et al.*,(2013) did a research on articles about haemoglobin concentration and pregnancy outcomes from January 1, 1990 to April 10, 2011. Based on the fact that adverse pregnancy outcomes thought to be affected by anaemia include maternal mortality, perinatal mortality, preterm birth (PTB), low birth weight (LBW), and small for gestational age (SGA). They used a mixed logit model with a random intercept (i.e., to account for between-study variation) was applied to assess haemoglobin effects on pregnancy outcome. The estimated pooled odds ratio (OR) along with 95% confidence interval (CI) was estimated by exponential logit coefficients, with a 5% level of significance for all analysis. It was found out that most of the eligible studies included pregnant women from Asian populations (12 studies), followed by European (4 studies), Africans (2 studies), and North Americans (2 studies).Their results suggested that lower haemoglobin concentration is associated with a higher risk of poor pregnancy outcomes.

The risk of PTB, LBW, and small gestational age was approximately higher in pregnant women who had haemoglobin concentration below 10 and 11 g/dl. With a final result showing that low haemoglobin concentration in early pregnancy (<20 weeks gestation) was associated with PTB.

Also, Brion *et al.*,(2008) used the Avon Longitudinal Study of Parents and Children (AL-SPAC) which is a prospective cohort study investigating the health and development of children, to assess the association between maternal iron status during pregnancy and offspring blood pressure (BP) in Bristol, United Kingdom.

They sampled 14,541 women from three health districts in Bristol, with an expected date of delivery between 1 April 1991 and 31 December 1992, of these 13,678 had a singleton live born child with BP measurements taken for 7638 singleton children at 7 years of age. The WHO criterion for haemoglobin levels was used to check for anaemia in pregnant women and Systolic BP (SBP) and diastolic BP (DBP) were measured with a Dinamap 9301 Vital Signs Monitor.

It was found that maternal anaemia during pregnancy was associated with lower offspring SBP at 7 years of age.

Also, Gaviria and Hoyos, (2011) used the ordered probit regression to model the effects of anaemia on child education in Colombia, by estimating the impact of nutrition on school attendance and on the probability of being overage.

Two types of measures of nutrition were used: anthropometric measures (z-scores of height-for-age, z-scores of weight-for-age and Body Mass Index (bmi)) and micronutrients measures (haemoglobin) using the WHO criterion for haemoglobin. It was found out that anaemia appears to increase the probability of being over-age but do not have a discernable impact on school attendance. Malnutrition, defined by anthropometric measures, does not have an impact either on school attendance or on the probability of being over-age.

Furthermore, Melku *et al.*, (2013) used bivariate and multivariate logistic regression to determine the predictors of maternal anaemia during pregnancy in Gondar, Northwest Ethiopia using the WHO criterion for acceptable haemoglobin levels due to anaemia, to determine the pregnant women who were anaemic. The levels of anaemia were divided into Mild, Moderate and Severe anaemia with the main predictors being Low family income, high family size, hookworm infection, and living with HIV/AIDS. The overall prevalence of anaemia was found to be small with a percentage of 17%, with mild anaemia being the highest among the pregnant women. All the main predictors of anaemia were found to be associated with the prevalence of anaemia.

Also, Oue´draogo *et al.*, (2013) used the multilinear regression model to investigate the effectiveness of routine preventive measures for anaemia in Beninese pregnant women. Pregnant women were tested for anaemia upon their first or second antenatal visit (ANV1 or ANV2) and those found to be anaemic (haemoglobin<11g/dl), were given the preventive treatment of malaria and intestinal helminths with sulfadoxine-pyrimethamine intermittent preventive treatment (SP-IPTp) and mebendazole (or albendazole), all administered at antenatal visits (ANVs). The results showed that the prevalence of anaemia had reduced upon the second antenatal visit as compared to the number who were anaemic at the first antenatal visit.

Next, Fuseini *et al., (*2010)examined the association between anaemia and Plasmodium and or intestinal helminth infections during pregnancy in the kassena-Nankana district in

Northern Ghana, by using the stool and blood specimens of pregnant women. The result showed that there was an association between plasmodium and anaemia and also indicated that worm infections have greater impact on the haemoglobin levels of the mothers in the district especially hookworm infections.

Also, Ndukwu and Dienye (2012) researched to see if there was an association between prevalence of anaemia in pregnancy and socio-demographic factors in River State, Nigeria. The results showed an association between anaemia in pregnancy and socio-demographic factors using chi-square test with a p-value of less than or equal to 5%.

Next, Dey *et al.*, (2010) used the logistic regression model, to assess the contribution of the predictors on anaemia in Meghalaya, India. The study was carried out between anaemic and non-anaemic pregnant women using haemoglobin level of (<11g/dl) with data on predictors taken on age (grouped), type of place of residence, highest educational level, wealth index, pregnancy status, nutritional status, working status, occupation, total children ever born, habit of cigarette. It was found that anaemia prevalence was high in pregnant women and the logistic regression showed that all the predictors except children ever born had an effect on the prevalence of anaemia in pregnant women.

2.7 Review of some Artificial Neural Network, Logistic model and Probit model methods

2.7.1 Artificial Neural Network methods

In 1943, neurophysiologist Warren McCulloch and mathematician Walter Pitts wrote a paper on how neurons might work. In order to describe how neurons in the brain might work, they modeled a simple neural network using electrical circuits.

Next, Hebb (1949) wrote The Organization of Behavior, a work which pointed out the fact that neural pathways are strengthened each time they are used, a concept fundamentally essential to the ways in which humans learn. If two nerves fire at the same time, he argued, the connection between them is enhanced.

Also, Rochester (1950) from the IBM research laboratories made it finally possible to simulate a hypothetical neural network using computers as computers became more advanced.

Next, Widrow and Hoff (1959) developed models called "ADALINE" and "MADALINE." which comes from (Multiple Adaptive Linear Elements). ADALINE was developed to recognize binary patterns so that if it was reading streaming bits from a phone line, it could predict the next bit. MADALINE was the first neural network applied to a real world problem, using an adaptive filter that eliminates echoes on phone lines. Widrow and Hoff developed a learning procedure that examines the value before the weight adjusts it (i.e. 0 or 1) according to the rule: Weight Change = (Pre-Weight line value) * (Error / (Number of Inputs)). It is based on the idea that while one active perceptron may have a big error, one can adjust the weight values to distribute it across the network, or at least to adjacent perceptrons. Applying this rule still results in an error if the line before the weight is 0, although this will eventually correct itself. If the error is conserved so that all of it is distributed to all of the weights than the error is eliminated.

Next, Theil (1969) developed a similar network independently of one another. They both used matrix mathematics to describe their ideas but did not realize that what they were doing was creating an array of analog ADALINE circuits. The neurons are supposed to activate a set of outputs instead of just one. The first multilayered network was developed in 1975, an unsupervised network.

Also, Sarle (1994) showed that Multi-Layer Perceptrons (MLP) could be viewed as nonlinear regression models. He showed that simple linear multivariate regression can be represented with a single linear perceptron. Sarle also made simple comparisons between nonlinear regression and feed-forward networks (with nonlinear activation functions) and showed how neural networks can be designed to represent polynomial regression (using different polynomial activation functions). While the sigmoid activation function was compared to nonlinear regression, it was not compared to Multiple Polynomial regression (MPR). Sarle suggested that one may potentially be able to design an artificial neural network to represent the structure of any regression model, and vice versa. However, his research was limited by providing only a basic, theoretical comparison between the models. No consideration was made to the number of hidden nodes in the network. Another limitation is that Sarle did not determine any formal equations to relate the parameters of ANNs to those of statistical regression equations.

Next, Warner and Misra (1996) continued this train of thought by performing empirical tests with synthetic data to show how FNNs and statistical regression models perform similarly. Both models were fit to linear and nonlinear data and results showed that the neural network was able to produce a best-fit line comparable to linear and power regression. They demonstrated that a feed-forward neural network with a sigmoid activation function can act as a function approximator and that this is an advantage over traditional regression when the underlying function of the system is unknown. They further suggested that if the physical relationship is known between the input and output of a given system, then a specific regression equation would be more desirable.

Next, Tokar and Johnson (1999) investigated the modeling of daily runoff of the Little Patuxent River watershed, measured in the form of streamflow. They compared the prediction abilities of feed-forward networks and regression equations. The models were tested using different combinations of input parameters. The feed-forward networks had sigmoid activation functions and the number of hidden nodes ranged from one to four hundred. Selection of the number of hidden nodes was made in part in relation to the data size of the training set (either one, two or three year's worth of data). The structure of the regression equations were a combination of linear and power models. Using error values to measure performance, the best-fit ANN model was better than all of the regression models by a considerable degree (validation error for the best ANN was 0.42 and for the best regression model it was 0.64). Tokar and Johnson noted that the number of parameters in the best-fit neural network was fifty-one, while the number of parameters in the regression equations was never greater than nine. They believed that this difference in parameters allowed the ANN to reach a higher level of flexibility and complexity. Another reason for the advantage of the network models was the fact that they used a nonlinear activation function, while most of the regression equations were linear.

Cobourn *et al*., (2000) compared a nonlinear regression model with a feed-forward neural network for their ability to predict ozone concentrations. The regression model was a combination polynomial and power model. No specific mention was made of the activation function or the number of hidden nodes in the FNN. Both models produced practically equivalent predictions for daily ozone concentration in Louisville, Kentucky.

However, it was found that both models used a different set of optimal input variables, with the regression model including more inputs than the FNN. In particular, the FNN did not include one parameter, air-mass trajectory, which was highly significant in the nonlinear regression equation.

Schnabel and Maneta (2005) investigated the comparison between FNNs and multiple quadratic regression. Both models were applied to the issue of estimating sediment transport in rivers. Sensitivity analysis was used to determine the most effective input variables for each model. The feed-forward network was defined to have ten hidden nodes, but it was not explained why this number was used. The activation function for the nodes was not specified. For the regression equation, the linear and quadratic forms of each variable were used, but any cross-terms were ignored. The results found that both neural networks and statistical regression could effectively predict sediment transport, and that the performance of the two models were similar.

However, the best-fit model determined by the regression equation used a different set of input parameters than that used by the feed-forward network. The authors did not offer an explanation for this difference and did not elaborate on the structures of the FNNs and regression equations.

Many experiments have compared the results of using both artificial neural networks and statistical regression equations for modeling and predicting biological systems. Most papers concluded that the use of artificial neural networks produced data predictions more accurate or at least comparable to regression models. However, many of these papers found inconclusive or conflicting results, such as Schnabel and Maneta (2005). There is enough evidence to support the connection between neural networks and statistical regression, but it is obvious that more research needs to be done to get a better understanding of how both models are related. Few articles explored the theoretical and mathematical connections between the structure of the two models, such as the articles written by Sarle (1994) and Chon (1997).

2.7.2 Logistic Regression Model Methods

The logistic function was invented in the 19th century for the description of the growth of populations and the course of autocatalytic chemical reactions, or chain reactions Malthus (1789). Alphonse Quetelet (1795-1874), the Belgian astronomer turned statistician, was well aware that the indiscriminate extrapolation of exponential growth must lead to impossible values. He experimented with several adjustments to the model and also asked his pupil, the mathematician Pierre-Francois Verhulst (1804-1849), to look into the problem. Like Quetelet, Verhulst approached the problem by adding an extra term to the model to represent the increasing resistance to further growth, which Verhulst named the logistic function. As a model of population growth the logistic function was discovered anew in 1920 by Pearl and Reed. They were unaware of Verhulst's work and they arrived independently at the logistic curve. When this was fitted to census figures of the U.S., again by making the curve pass through three points, it gave a good fit for the period from 1790 to 1910. Pearl and his collaborators in the next twenty years went on to apply the logistic growth curve to almost any living population from fruit flies to the human population of the French colonies in North Africa as well as to the growth of cantaloupes. Next, Reed and Berkson (1929) hinted there is another early root of the logistic function in chemistry, where it was employed (again with some variations) to describe the course of autocatalytic or chain reactions, where the product itself acts as a catalyst for the process while the supply of raw material is fixed. This leads naturally to a differential equation and hence to the logistic function for the time path of the amount of the reaction product. The basic idea of logistic growth is simple and effective, and it is used to this day to model population growth and market penetration of new products and technologies. The introduction of mobile telephones is an autocatalytic process, and so is the spread of many new products and techniques in industry.

Also, Cox (1969) was among the first to explore (and exploit) the possibilities of the logit model to logistic regression where binary outcomes are related to a number of determinants, without a specific theoretical background, and this statistical model proved as fertile as linear regression in an earlier era. Later, the link of the logistic model with discriminant analysis was recognized, and its ready association with log-linear models in general. In epidemiology, case-control studies began even earlier, and since these are directly concerned with odds, and odds ratios, the log-odds or logit transformation arises naturally.

Next, McFadden (1973) made an important contribution on the logit model by linking the multinomial logit to the theory of discrete choice from mathematical psychology. This provided a theoretical foundation of the logit model that is much more profound than any theory.

2.7.2 The Probit Model Methods

The invention of the probit model is usually credited to Gaddum (1933) andBliss (1934),but the roots of the method and in particular the transformation of frequencies to equivalent normal deviates can be traced to Fechner(1860). Who recognized that human response to an identical stimulus is not uniform, and he was the first to transform observed differences to equivalent normal deviates.

Next, Gaddum and Bliss (1934) published their contributions. Both authors regard the assumption of a normal distribution as commonplace, and attach more importance to the logarithmic transformation of the stimulus, where the stimulus is determinate and responses are random because of the variability of individual tolerance levels. Bliss introduced the term probit (short for 'probability unit') originally as a convenient scale for normal deviates, but abandoned this within a year in favour of a different definition which has since been generally accepted. For any (relative) frequency f there is an equivalent normal deviate $\check{Z}$such that the cumulative normal distribution at $\check{Z}$equals f.

Next, Berkson (1951) showed that, the probit could be related to an underlying (normal) distribution of tolerance levels. Aitchison and Brown (1957) proved the probit has a well recognised and manageable frequency distribution of tolerances in a natural way.

Also, Hoetker (2007)researched on how probit and logit models differ from ordinary least squares and how this can lead researchers to misunderstand their statistical results and draw incorrect conclusions regarding the theory they are testing. He identified four critical issues in their use: interpreting coefficients, modeling interactions between variables, comparing coefficients between groups (e.g., foreign and domestic firms), and measures of model fit.

Furthermore, Stigler (1986) researched to show how the probit model was a solution to the deficiencies of the linear probability model as estimated by the ordinary least squares by constraining the estimated probabilities to be between 0 and 1, and relaxes the constraints that the effect of the independent variables is constant across different predicted values of the independent variable.

In addition, Farrell (1954) used a probit model for the ownership of cars of different vintage as a function of household income, and Adam (1958) fits lognormal demand curves to survey data of the willingness to buy cigarette lighters and the like at various prices. The classic monograph on the lognormal distribution of Aitchison and Brown (1957) brought probit analysis to the notice of a wider audience of economists.

2.8 Theoretical Discussion of Models

Resop (2006), researched to identify potential equivalences between Artificial Neural Networks and statistical regression and to verify these equivalences when applied to modeling biological resources systems by investigating the relationship between feed-forward neural networks (FNN) and multiple polynomial regression (MPR) equations and the relationship between Recurrent Neural Networks (RNN) and Auto-Regressive Moving Average (ARMA) equations. The results The results from the research supported the theory that feed-forward neural networks and polynomial regression equations are mathematically equivalent models. Formal tests confirmed that FNNs with linear and polynomial activation functions can perfectly replicate target MPR equations, as long as there are enough hidden nodes in the network. It was also demonstrated that FNNs with a sigmoid activation function can model the output function of polynomial equations with accuracy similar to that of FNNs with a polynomial activation function.

Karlson *et al.,* (2010) compared regression coefficients between logit and probit models because common practice of comparing the coefficients of a given variable across differently specified models does not warrant the same interpretation in logits and probits as in linear regression. Unlike linear models, the change in the coefficient of the variable of interest cannot be straightforwardly attributed to the inclusion of confounding variables. The reason for this is that the variance of the underlying latent variable is not identified and will differ between models. They proposed to this as the problem of rescaling. By allowing researchers to assess the influence of confounding relative to the influence of rescaling, and they developed a test to assess the statistical significance of confounding.

Artificial Neural Networks are commonly thought to be used just for classification because of the relationship to logistic regression: neural networks typically use a logistic activation function and output values from 0 to 1 like logistic regression. However, the worth of neural networks to model complex, non-linear hypothesis is desirable for many real world problems including regression.

Fox (2010), explained that despite the similarity of the Logit and the Probit models there are two practical advantages of the Logit over the Probit model.

Simplicity: The equation of the logistic CDF (Cumulative Distribution Function) is very simple, while the normal CDF of the Probit model involves an unevaluated integral. The difference is trivial for polytomous data, where we will require the multivariate logistic or normal distribution, the disadvantage of the Probit model is more acute.

Interpretability: The inverse linearizing transformation for the Logit model, is directly interpretable as the log-odds, while the inverse transformation does not have a direct interpretation.

Also, the Probit model is used for designed experiments and the logit for observational studies, the Probit model reports effective values for various rates of response as the logit model reports estimates of odds ratio for the independent variables.

The main purpose of the Probit model is to estimate the probability that an observation with particular characteristics would fall into one of the specific categories, the main purpose of the Logit model is to generate an equation that would reliably classify an observation into one of two outcomes and the artificial Neural Network is for solving classification problems.

2.9 Summary

The chapter dealt with reviewing of literature that is relevant to the study. Reviewing of the literature has exposed us to the diverse techniques that researchers have employed in the comparison of models and modelling of anemia. However, among the diverse techniques reviewed the neural network model. The Artificial Neural Network model, Probit Regression model and the Logistic Regression model were employed in this study to model anemia in pregnancy because they were the techniques used frequently in literature.

CHAPTER THREE

RESEARCH METHODOLOGY

3.0 Introduction

This chapter deals with the data and statistical techniques that were employed in order to achieve the objectives of the study. The chapter is divided into five main headings namely; data and source, study area, statistical models, model diagnostics and model selection criteria.

3.1 Data and Source

In order to achieve the objectives of the study, data for this research was mainly a secondary data on five hundred and eighty (580) pregnant women obtained from the ante-natal unit of the Bolgatanga Regional Hospital (BRH). The data for the study was analysed using R, SAS, STATA and SPSS statistical softwares.

3.2 Study Area

The Bolgatanga Municipality was established in 2004 by Legislative Instrument (LI) 1797(2004). Located in the centre of the Upper East Region, approximately, between latitudes 10°30'and 10°50' North and longitudes 0°30' and 1°00' West (Map 2.1), it is also the regional capital. Bolgatanga Municipality is bordered to the north by the Bongo District, south and east by Talensi-Nabdam District, and to the West by Kassena-Nankana District. It covers a total land area of $729km^2$ (UNDP, 2010). It is one of two municipalities (the other being Bawku Municipal) which together with six other districts constitute the Upper East Region of Ghana. A major irrigation project, the Vea Project covering 850 hectares, is located in the municipality. The municipality falls within the Birimian-Tarkwaian and Voltaian rocks of Ghana; and there is ample evidence of the presence of minerals particularly gold in the area (UNDP, 2010). The main rivers are the White and Red Volta, and their tributaries. The landscape has gentle slopes ranging from 1 per cent to 5 per cent with some isolated rock outcrops and some uplands which have slopes of over 10 per cent (UNDP, 2010).

3.3 Statistical Models:

3.3.1 The Probit Regession Model

The Probit Regression model proposed by Chester Bliss(1935), is a type of regression where the dependent variable can only take two values in the simple linear case eg; married or not married and more than two dependent variables in the multinomial case. The purpose of the model is to estimate the probability that an observation with particular characteristics will fall into a specific one of the categories.

For the simple regression model

$$P_r(Y = 1|x) = \varphi(\boldsymbol{X'\beta}) \qquad (3.1)$$

where; $P_r$denotes probability, $Y$is the response or dependent variable

Φ is the Cumulative Distribution Function (CDF) of the standard normal distribution.

The parameters β are typically estimated by maximum likelihood, $\boldsymbol{X}$ is a vector of regressors

For the multiple Regression model

$$\pi_i = \varphi(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) \tag{3.2}$$

where; π_i is a conditional probability, φ is a cumulative distribution function.

$$y_i = \begin{cases} 1, & y_i * > 0 \\ 0, & y_i * \leq 0 \end{cases} \tag{3.3}$$

$$y_i *= \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i \tag{3.4}$$

Where, y^* is an unobservable variable which can take all values in $(-\infty, +\infty)$

$P(y_i = 1|x) = (y_i * > 0|x) = P(\boldsymbol{x_i}'\boldsymbol{\beta} + \varepsilon_i > 0|x) = P(\varepsilon_i > -\boldsymbol{x_i}'\boldsymbol{\beta}|x) =$

$$1 - F(-\boldsymbol{x_i}'\beta) \tag{3.5}$$

Assumption: Error terms are independent and normally distributed

$$P(y_i = 1|x) = 1 - \varphi\left(\frac{x_i'\beta}{\sigma}\right), \sigma \equiv 1 \tag{3.6}$$

$$P(y_i = 1|x) = \varphi(\boldsymbol{x_i}'\boldsymbol{\beta}) \text{ because of symmetry.} \tag{3.7}$$

Giving a joint density:

$$f(y|x, \beta) = \prod F(\boldsymbol{x_i}'\boldsymbol{\beta})^{y_i}[1 - F(\boldsymbol{x_i}'\boldsymbol{\beta})]^{(1-y_i)} = \prod F_i^{y_i}(1 - F_i)^{1-y_i} \tag{3.8}$$

With a Log Likelihood Function:

$$l_n L = \sum_i y_i l_n F_i + (1 - y_i) ln(1 - F_i) \tag{3.9}$$

Maximum Likelihood:

$$\frac{\partial l_n L}{\partial \beta} = \sum_i \left[\frac{y_i f_i}{f_i} + \frac{(1-y_i)(-f_i)}{1-F_i}\right] x_i = \sum_i \left[\frac{y_i - F_i}{F_i(1-F_i)} f_i\right] x_i = 0 \quad (3.10)$$

3.3.2 Artificial Neural Networks:

Artificial Neural Networks was first proposed by Warren McCollough and Walter Pitts (1943), when combining two or more artificial neurons we are getting an artificial neural network.

An Artificial Neural Network (ANN) is a mathematical model that tries to simulate thestructure and functionalities of biological neural networks. The ANN are algorithms that can perform non-linear statistical modelling since it has the ability to detect complex non-linear relationships between the dependent variable and independent variables and also has the ability to detect all possible interactions between predictor variables.For the purpose of this research the Multilayer-perceptron neural network was employed.

3.3.3 Multi-Layer Perceptron Neural Network (MLP)

A MLP differs from the simple perceptron in two ways. Firstly, it has additional layers of neurons between the input and output layers, known as hidden layers in order to improve the learning power of the network. Secondly, it uses a transfer function to modify the input to neuron. Multilayer perceptrons (MLP) are layered feed-forward networks typically trained with static back propagation. These networks are called multilayer and feed-forward since they have multiple layers and data flow in one direction in the network.

They have found their way into countless applications requiring static pattern classification. Their main advantage is that they are easy to use, and that they can approximate any input/output map. The key disadvantages are that they train slowly, and require lots of training data. Neural networks are non-linear learning machines built from many Processing Elements (PE).

Each PE receives connections from other PEs and/or itself. The signals flowing on the connections are scaled by adjustable parameters called weights(w_i). The PEs sum all these inputs and produce an output that is a non-linear function of the sum. This output becomes either a system output or is sent to the same or other PEs.

A processing element is simply a sum of products followed by a threshold nonlinearity (transfer function).

Its input-output equation can be expressed as:

$$y(k) = f(\sum_{i=0}^{m} w_i(k) \times x_i(k) + b) \qquad (3.11)$$

where:

$x_i(k)$is input value in discrete time k where i goes from0 to m,

$w_i(k)$ is weight value in discrete time k where i goes from 0 to m,

b is bias, f is a transfer function, $y_i(k)$is output value in discrete time k

The ANN model has three simple sets of rules: multiplication, summation and activation. At the entrance of artificial neuron the inputs are weighted, which means that every input value is multiplied with individual weight. In the middle section of the artificial neuron is the sum function that sums all weighted inputs and bias. At the exit of artificial neuron the sum of previously weighted inputs and bias is passing through activation function that is also called transfer function as shown in (Figure 3.1) below.

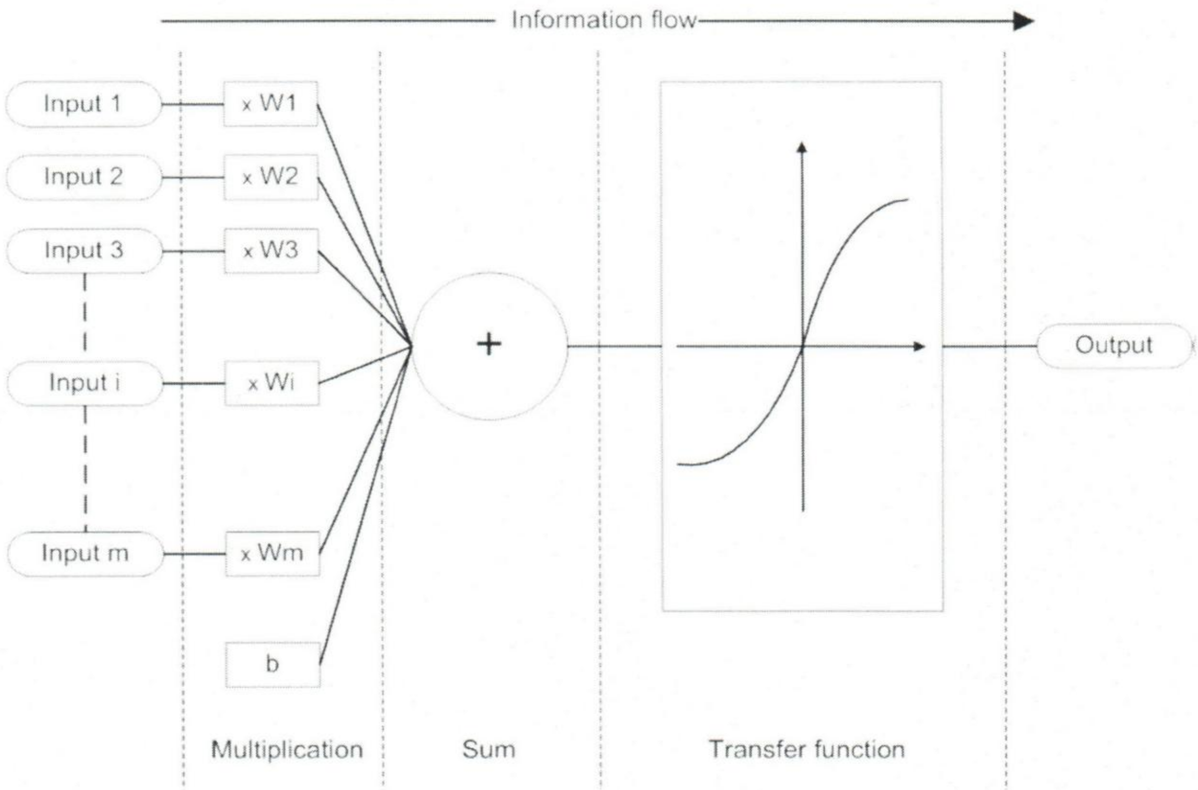

Figure 3.1: Working principle of an Artificial Neural Network

3.3.4 The Logistic RegressionModel

Logistic Regression was proposed by Pierre-Francois Velhurst (1838). It is well suited for studying the relationship between a categorical or qualitative outcome variable and one or more predictor variables. In the simplest case of one predictor X (say, IQ score) and one dichotomous outcome variable Y (say, diagnosed to be learning disabled), the logistic model predicts the logit of Y from X. the logit is the natural logarithm(ln) of odds of Y. the simple logistic model has the form:

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \log(odds) = logit = \beta_{\circ} + \beta x, \qquad (3.12)$$

$$p(x) = \text{probability}(\text{Y} = \text{outcome of interest } |\text{X} = \text{x}) = \frac{e^{\beta_\circ + \beta x}}{1+e^{\beta_\circ + \beta x}} \tag{3.13}$$

where p = probability of the outcome of interest, or the event, under variable Y, $\beta_\circ$ is the Y intercept, β is the slope of the parameter

X can be categorical, continous, or mixed whereas Y is always categorical.

Multiple logistic regression: The logistic regression with more than one predictor variable is given by:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_\circ + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K \tag{3.14}$$

$$\frac{\pi}{1-\pi} = e^{\beta_\circ + \beta x} \rightarrow \pi = (1-\pi)e^{\beta_\circ + \beta x} = e^{\beta_\circ + \beta x} - e^{\beta_\circ + \beta x}\pi \rightarrow \pi + e^{\beta_\circ + \beta x}\pi = e^{\beta_\circ + \beta x} \tag{3.15}$$

$$\pi\left(1 + e^{\beta_\circ + \beta x}\right) = e^{\beta_\circ + \beta x} \rightarrow \pi = \frac{e^{\beta_\circ + \beta x}}{1+e^{\beta_\circ + \beta x}} \tag{3.16}$$

Maximum likelihood estimation for the Logit model.

$$-\infty < \beta_\circ + \beta_1 x < +\infty, \qquad F(-\infty) = 0 \; and \; F(+\infty) = 1$$

Because logistic regression predicts probabilities, rather than classes, we can fit it using likelihood. For each training data-point, we have a vector of features, x_i, and an observed class, y_i. The probability of the class is either p, if $y_i = 1$, or $1 - p, if\ y_i = 0$. Then the likelihood is;

$$l(\beta_\circ, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i)^{1-y_i}) \tag{3.17}$$

Then the log-likelihood becomes;

$$l(\beta_\circ, \beta) = \sum_{i=1}^{n} y_i log p(x_i) + (1 - y_i) log 1 - p(x_i) \quad (3.18)$$

$$= \sum_{i=1}^{n} log 1 - p(x_i) + \sum_{i=1}^{n} y_i log \frac{p(x_i)}{1 - p(x_i)} \quad (3.19)$$

$$= \sum_{i=1}^{n} log 1 - p(x_i) + \sum_{i=1}^{n} y_i (\beta_\circ + \beta x_i) \quad (3.20)$$

$$= \sum_{i=1}^{n} -log 1 + e^{\beta_\circ + \beta x_i} + \sum_{i=1}^{n} y_i (\beta_\circ + \beta x_i) \quad (3.21)$$

where in the next to last step we use equation (3.12)

Typically to find the maximum likelihood estimates we would differentiate the log likelihood with respect to the parameters, set the derivates equal to zero, and solve. To start that we take the derivative with respect to one component of β say β_j.

$$\frac{\partial l}{\partial \beta_j} = -\sum_{i=1}^{n} \frac{1}{1 + e^{\beta_\circ + \beta x_i}} e^{\beta_\circ + x_i \beta x_{ij}} + \sum_{i=1}^{n} y_i x_{ij} \quad (3.22)$$

$$= \sum_{i=1}^{n} (y_i - p(x_i; \beta_\circ \beta)) x_{ij} \quad (3.23)$$

We are not going to be able to set this to zero and solve exactly (that is a transcendental equation, and has no closed-form solution), we can however approximately solve it numerically.

Multinomial Logistic Regression

Multinomial logistic regression is used when the dependent variable in question is nominal equivalently categorical, meaning that it falls into any one of a set of categories which cannot be ordered in any meaningful way) and for which there are more than two categories.

Multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes. That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables (which may be real-valued, binary-valued, categorical-valued. Which is given by;

$f(k,i) = \beta_{0,k} + \beta_{1,k}X_{1,i} + \beta_{2,k}X_{2,i} + \cdots + \beta_{M,k}X_{M,i,}$(3.24)

Where; $f(k,i)$ is a linear function predictor,i is the observation,k is the outcome, $\beta_{M,k}$ *is* a regression coefficient associated with the m^{th} explanatory variable and the k^{th} outcome and ***X*** is a vector.

3.4 Model Diagnostics

3.4.1 Wald Test

This is a parametric statistical test which is used whenever a relationship within or between data items can be expressed as a statistical model with parameters to be estimated from a sample, the Wald test can be used to test the true value of the parameter based on the sample estimate. In the univariate case it's given by;

$\theta = \frac{(\widehat{\theta}-\theta_0)^2}{var(\widehat{\theta})}$, which can be compared against a chi-square distribution (3.25)

$\theta = \frac{\widehat{\theta}-\theta_0}{SE(\widehat{\theta})}$, which can be compared to a normal distribution (3.26)

where; θ_0 is the proposed value, θ the parameter of interest

$\widehat{\theta}$ is the maximum likelihood estimate of the parameter(s) of interest θ.

SE is the standard error of the maximum likelihood estimate, given by;

$SE = \frac{1}{\sqrt{I_n(MLE)}}$, (3.27) where, I_n is the fisher information of the parameter

3.4.2 Likelihood Ratio Test

This is a test statistic that is usually used to compare the fit of the model with and without the predictor(s). For a particular parameter the test compares the likelihood of obtaining the data when the parameter is zero (L_0) with the likelihood (L_1) of obtaining the data evaluated at the maximum likelihood estimator of the parameter.

$$-2 \times l_n(likelihood\ ratio) = -2 \times l_n\left(\frac{L_0}{L_1}\right) = -2 \times (l_n L_0 - l_n L_1) \tag{3.28}$$

3.4.3 Score Test

This is a statistical test of a simple null hypothesis that a parameter of interestθ is equal to some particular value$\theta_\circ$. It is the most powerful test when the true value of θ is close to $\theta_\circ$.the advantage of the score test is that it does not require an estimate of information under the alternative hypothesis or unconstrained maximum likelihood. This makes testing feasible when the unconstrained maximum likelihood estimate is a boundary point in the parameter space. The test statistic is given by;

$$U(\theta) = \left(\frac{\partial logL\left(\frac{\theta}{x}\right)}{\partial\theta}\right)_{\theta=\theta_\circ} \geq C \tag{3.29}$$

Where; L is the likelihood function, θ parameter of interest, $\theta_\circ$ value of parameter of interest under the null hypothesis, C constant set depending on the size of the test desired and x is the data.

3.4.4 Hosmer-Lemeshow Test

This is used for assessing the goodness of fit of a model and allows for any number of explanatory variables, which may be continuous or categorical. The test assesses whether or not the observed event rates match expected event rates in subgroups of the model population. The Hosmer–Lemeshow test specifically identifies subgroups as the deciles of fitted risk values. The test statistic is given by;

$$H = \sum_{g=1}^{G} \frac{\left(O_g - E_g\right)^2}{N_g \pi_g \left(1 - \pi_g\right)} \,. \tag{3.30}$$

Where O_g denote the observed events, E_g, expected events, N_g, observations,

π_g, , predicted risk for the g^{th} risk decile group, G is the number of groups

3.5 Model Selection Criteria

3.5.1 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a measure of the relative quality of a statistical model, for a given set of data. As such, AIC provides a means for model selection. AIC deals with the trade-off between the goodness of fit of the model and the complexity of the model (Akaike and Hirotugu., 1974).

Given a set of models for the data, the preferred model is the one with the minimum AIC value. Hence AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages over-fitting (increasing the number of free parameters in the model improves the goodness of the fit, regardless of the number of free parameters in the data-generating process).

$$AIC = 2k - 2l_n(L) \tag{3.31}$$

Where k is the number of parameters in the statistical model

L is the maximized value of the likelihood function for the estimated model

3.5.2 Bayesian Information Criterion

The Bayesian Information Criterion (BIC) is a criterion for model selection among a finite set of models. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC) (Schwarz, and Gideon E., 1978).

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in over-fitting. Both BIC and AIC resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC.

$$\text{BIC} = \log(\sigma_e^2) + \frac{k}{n}\log(n) \tag{3.32}$$

k is the number of parameters in the statistical model, n is the number of observations in the data, σ_e^2 is the error variance.

3.6 Justification of Models

The models employed for this research were used because all three models are non-linear statistical models whose methods are the most commonly used for developing predictive models for dichotomous outcomes in medicine. Although the Logit and Probit models are similar they differ in terms of Simplicity and Interpretability, the Artificial Neural Network has the ability to implicitly detect complex non-linear relationships between dependent and independent variables. The data for the research was modelled using the probability of a subject (pregnant woman) being anaemic or non-anaemic making it a dichotomous data for the response variable.

CHAPTER FOUR

ANALYSIS AND DISCUSSION OF RESULTS

4.0 Introduction

This chapter presents, analyses, interprets and discusses the results of the study. It consists of preliminary analysis, further analysis and discussion of results obtained.

4.1 Preliminary Analysis

This section presents and explains the frequencies of all variables, which are anaemicst (anaemic status), parity (number of children), gravida (gravidity), bp (blood pressure of mother), BMI (underweight, thin for weight, healthy weight, overweight and obesity), season(dry/wet), sickling (sicklecell status), educst (educational status), maritalst (marital status), religionst(religious status), trimesterst (stage of pregnancy), agegroup (age of mother), occupation(occupation of mother), HIVst (HIV status of mother), and the descriptive statistics of hbatreg (haemoglobin level at registration) and gestation(gestational age of mother in weeks).

Table 4.1 Descriptive Statistics of Data

VARIABLE	FREQUENCY	PERCENT
Severity of Anaemia		
Non- anaemic(Hb>11 g/dl)	247	42.6
Mild(Hb between 10.0 to 10.9g/dl)	175	30.2
Moderate(Hb between 7.0 to 9.9g/dl)	136	23.4
Severe(Hb less than 7g/dl)	22	3.8
Maternal age-group 15-19	38	6.6
20-24	159	27.4

	25-29	216	36.2
	20-34	109	18.8
	35-39	46	7.9
	40-44	12	2.1
Gravidity	1-3 pregnancies	475	81.9
	4-6 pregnancies	104	17.9
	< or equal to 7 pregnancies	1	0.2
Parity	no child	221	38.1
	1-3 children	322	55.5
	4-6 children	37	6.4
Blood Pressure	normal	349	60.2
	abnormal	231	39.8
Body Mass Index (kg/m^2)			
	Underweight(<18)	18	3.1
	Thin for height(<18.5)	22	3.8
	Healthy weight(18.6-24.9)	286	49.3
	Over weight(25-29.9)	159	27.4
	Obesity(>30)	95	16.4
Sickling status	Positive	544	93.8
	Negative	36	6.2
Season	dry	363	62.6
	wet	217	37.4
Occupation	unemployed	153	26.4

	farmer	104	17.6
	Trader	55	9.3
	student	22	3.6
	hairdresser	195	33.6
	others	54	9.3
Religion	christian	517	89.1
	muslim	63	10.1
Educational level	none	55	9.5
	primary	74	12.8
	jhs	200	34.5
	shs	145	25.0
	tertiary	106	18.3
Marital status	single	91	15.7
	married	489	84.3
Trimester	(1-12) early stage	214	36.9
	(13-24) mid stage	239	41.2
	(24-36) late stage	127	21.9
HIV status	negative	573	98.8
	positive	7	1.2

Table 4.1 results shows that the pregnant women who were anaemic (haemoglobin<11 g/dl) were 333 with a percentage of fifty seven (57.4%) as compared to non-anaemic (haemoglobin>11g/dl) pregnant women who were 247 representing 42.6 percent (42.6%).

The severity of anaemia was recorded to be 247(42.6%) for non-anaemic (haemoglobin>11g/dl) pregnant women, 175(30.2%) for mild anaemia (haemoglobin between 10.0 to 10.9g/dl), 136(23.4%) for moderate anaemia (haemoglobin between 7.0 to 9.9g/dl) and 22(3.8%) for severe anaemia (haemoglobin less than 7g/dl). The population of women between 25-29 age-group who were likely to be anaemic were 216representing (36.2%). Women who have had between 1-3 pregnancies and were more likely to be anaemic were 475 representing (81.9%) and 322(55.5%) women who have between 1-3 children are more likely to be anaemic. The number of women who had normal blood pressure were 349 representing (60.2%) and 289 representing (49.6%) women showing a healthy weight. The number of pregnant women who showed a positive sickling status were 544 representing (93.8%) and 363(62.6%) women were more likely to be anaemic in the dry season. The number of pregnant women who were Christians were 517 representing (89.1%) and were mostly hairdressers 195 (33.6%) were found to be anaemic. The number of women being married were 200 representing 35% of the total population and 489 representing (84.2%) women had their highest education at the JHS level. The number of women who were in their mid- stages (13-24 weeks) of pregnancy were 239 representing (49.2%) and also tested negative to HIV were 573 representing (98.8%).

The minimum, average and maximum haemoglobin levels were 4.50g/dl, 10.60g/dl and 16.30g/dl respectively. The minimum, average and maximum Gestational age in weeks of the pregnant women were 2 weeks, 17 weeks and 37 weeks respectively, as shown in Table 4.2 below.

Table 4.2: Descriptive statistics of Haemoglobin level and Gestational Age

VARIABLE	N	Mean	Std. Deviation	Minimum	Maximum
Haemoglobin Level	580	10.5928	1.57932	4.50	16.30
Gestational age (in weeks)	580	17.32	8.586	2	37

A chi-square test of association was done to show whether there was an association between the dependent variable (anaemic status) and the independent variables. It was evident from the results that BMI (body mass index), Season(dry/wet) and Occupation were significantly associated with the anemic status (anaemicst) of the pregnant women. However the variables that were found to be insignificant at the 5% significance level were age-group, gravidity, parity, occupation, trimester, educational level, religion, HIV status, marital status, blood pressure, and sickling) as shown in Table 4.3 below.

Table 4.3: Chi-square Test of Association

VARIABLE	TEST STATISTIC	DEGREE OF FREEDOM	P-VALUE
Age group	5.237	5	0.388
Gravidity	1.382	2	0.501
Parity	1.342	2	0.511
Blood pressure	0.004	1	0.949
BMI	22.142	4	0.000
Sickling	0.337	1	0.561
Season	3.922	1	0.048
Occupation	15.202	5	0.010
Religion	0.002	1	0.923
Education	2.530	4	0.639
Marital status	0.269	1	0.604
Trimester	2.251	2	0.325
HIV status	0.569	1	0.451

4.2 Further Analysis

To model the probability of a subject being anaemic, the haemoglobin level was considered as dichotomous (anaemic or non-anaemic) for all models. The hypothesis for both the Probit regression model and the Logistic regression model is given by:

$$H_0: All\ coefficients\ except\ the\ intercept = 0$$

$$H_1: All\ coefficients\ \neq 0$$

4.2.1 The Probit Regression results

Table 4.4 shows the inferential statistics of individual parameter estimates of the probit model to check if the model is significant at 5% significance level. Likelihood Ratio test the Score test and the Wald test gave test statistics that were significant at the 5% significance level, indicating that at least one of the independent variables is not equal to zero as shown in Table 4.4 below.

.Table 4.4: Testing Global Null Hypothesis: BETA=0

TEST	CHI-SQ	DF	Pr>ChiSq
Likelihood Ratio	32.7574	14	0.0031*
Score	32.1156	14	0.0039*
Wald	31.2379	14	0.0051*

***Means significant at 5% significance level**

The full model was then fitted to check the variables that were significant at 5% significance level as shown in Table 4.5 below, after which the significant variables were selected from the full model and the insignificant variables were dropped from the model as shown in Table 4.5 below.

Table 4.5: Analysis of Maximum Likelihood Estimates of the probit model

Parameter	DF	Estimate	Standard error	Wald Chi-Sq	Pr>ChiSq
intercept	1	1.1156	0.4532	6.0595	0.0138
Agegroup	1	-0.0199	0.0502	0.1572	0.6918
Gravidity	1	-0.00860	0.1584	0.0029	0.9567
Parity	1	-0.0679	0.1064	0.4071	0.5234
Blood pressure	1	-0.00803	0.1132	0.0050	0.9434
BMI	1	-0.2030	0.0609	11.1232	0.0009*
Sickling	1	-0.2275	0.2274	1.0008	0.3171
Season	1	0.2476	0.1145	4.6722	0.0307*
Occupation	1	-0.0696	0.0308	5.1100	0.0238*
Religion	1	0.0708	0.1760	0.1619	0.6874
Education	1	0.0384	0.0490	0.6144	0.4331
Marital status	1	0.0638	0.1486	0.1843	0.6677
Trimester	1	-0.5182	0.1999	6.7212	0.0095*
HIV status	1	0.2743	0.5169	0.2815	0.5957
Gestational age	1	0.0521	0.0175	8.8294	0.0030*

***Means significant at 5% significance level**

Table 4.5 shows the Maximum likelihood estimates of the parameters of the probit model. The results also showed an AIC, BIC and -2Loglikelihood of 788.495, 853.940 and 758.495 respectively. The significance level shows a more likely or less likely outcome of the predictor on the dependent variable. It is evident from table 4.6 that, the predictors which were significant at 5% significance level were BMI, season(dry), gestational age, occupation and trimester status. The insignificant variables were then dropped from the full model and the significant variables were fitted to give a reduced model using the stepwise algorithm. The analysis of the significant variables produced results as shown in Table 4.6below.

Table 4.6: Analysis of maximum likelihood estimates of significant variables

Parameter	DF	Estimate	Standard Error	Wald chi-square	Pr>ChiSq
intercept	1	1.1044	0.2743	16.2079	<.0001
BMI	1	-0.2111	0.0592	12.7012	0.0004
Gestational age	1	0.0501	0.0174	8.2503	0.0041
Season(dry)	1	0.2272	0.1109	4.1974	0.0405
Occupation	1	-0.0635	0.0291	4.7515	0.0293
Trimester	1	-0.4946	0.1983	6.2246	0.0126

From Table 4.6, the parameter estimates of the significant variables produced a reduced probit model of:

$$\pi_i = \varphi(1.1044 - 0.2111x_1 - 0.0501x_2 + 0.2272x_3 - 0.0635x_4 - 0.4946x_5) \qquad (4.1)$$

where: π_i is a conditional probability, φ is the cumulative distribution function,x_1 is bmi, x_2 is gestational age, x_3 is season(dry), x_4 is occupation, x_5 is trimester.

4.2.2 Model Diagnostics

The reduced model was diagnosed to check if the reduced model was significant, fits the data well and how well the model predicts its decisions. The inferential statistics to check if the reduced model was significant at 5% significance level was shown by the results in Table 4.7 below.

Table 4.7: Testing Global Null Hypothesis: BETA=0 for reduced Probit model

TEST	CHI-SQ	DF	Pr>ChiSq
Likelihood Ratio	29.4133	5	<.0001*
Score	28.9827	5	<.0001*
Wald	28.0292	5	<.0001*

***Means significant at 5% significance level**

From Table 4.7 it is evident the Likelihood Ratio test, the Score test and the Wald test were all significant at 5% significance level. Indicating that, at least one of the independent variables is not equal to zero. Implying the reduced model is significant.

A goodness of fit statistics was performed to assess the fit of the reduced probit model against the data using the Hosmer-Lemeshow test to check whether the model fits the data or does not fit the data at 5% significance level, as shown in Table 4.8 below.

Table 4.8: Hosmer-Lemeshow goodness of fit statistic for reduced model

number of observations	number of groups	Hosmer-Lemeshowchisq	Prob>chisq
580	10	10.18	0.2525*

***Means insignificant at 5% significance level**

An insignificance level of 0.2525 at 5% significance level indicates that the model fits the data well making the model adequate. A test of association to measure the relationship between the predictors and the dependent variable in the model was then performed, to show how well the model correctly predicts probabilities for event outcomes and non-event outcomes using the c statistic as shown in Table 4.9 below.

Table 4.9: Association of Predicted Probabilities and Observed Responses

Percent Concordant	63.4
Percent Discordant	36.1
Percent Tied	0.5
Pairs	82251
Somers' D	0.273
Gamma	0.274
Tau-a	0.134
c	0.637

The tau-a is Kendall's rank-order correlation coefficient without adjustment for ties. The Gamma statistic is based on kendall's coefficient but adjusts for ties. Somer's D is an extension of the Gamma statistic. The c statistic is the proportion of observation pairs with different observed outcomes for which the model correctly predicts a higher probability for observations with the event outcomes than the probability for non-event observations. A value of 1 means the model assigns higher probabilities to all observations with the event outcomes compared to non-event observations. The c statistic from the table gives us a value of 0.637, which shows that the model assigns high probabilities to more observations with event outcomes compared to non-event observations, After the association of predicted probabilities, a classification table was used to check the validity of the predicted probabilities of the model and how well the model predicts its decisions as shown in Table 4.10 below.

Table 4.10: Classification table for the reduced Probit model

	Correct		Incorrect						
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensitivity	Specificity	False POS	False NEG
0.500	266	82	165	67	60.0	79.9	33.2	38.3	45.0

Table 4.10 shows the validity of the predicted probabilities with the rows representing the two possible outcomes and columns are high and low probabilities based on a cut-off point of 0.05.

From the table the correct column shows the percentage of correct classification of 60.0, implying the validity of the predicted probability of the model is 60.0%. The sensitivity test indicated that about 79.9 percent of the events of interest (anaemic) were correctly classified by the model while the specificity test indicated that about 33.2% of the non-event (non-anaemic) was correctly classified.

A test was then performed to show the effects of each predictor on the dependent variable as shown in Table 4.11 below.

Table 4.11: Marginal Effects of Predictors

VARIABLE	MEAN	STANDARD DEVIATION
BMI	0.0794811	0.0058016
Gestational age	-0.0188500	0.0013759
Season (dry)	-0.0855729	0.0062462
Occupation	0.0239031	0.0017448
Trimester	0.1862745	0.0135968

From Table 4.11 it is evident that, the variables for pregnant women which would cause them to be more likely anaemic were BMI, Occupation and Trimester. The variables of the pregnant women which would cause them to be less likely anaemic were gestational age and season (dry).

A test of overall estimated probability of the model was then performed to see the predictions made by the reduced model as shown in Table 4.12 below.

Table 4.12: Estimated probability of reduced Probit model

Variable	N	Mean	Std Dev	Minimum	Maximum
anemicst (Anemic Status)	580	0.5741379	0.4948998	0	1.0000000
pprobit Estimated Probability	580	0.5738145	0.1094442	0.2719831	0.8696715

From the Table 4.12, the pprobit (predicted probit) showed on average that fifty-seven percent (57%) of the pregnant women were anaemic (haemoglobin<11g/dl), with a minimum of 0.27 and a maximum of 0.87.

The AIC, BIC and -2loglikelihood of the reduced model produced values of 773.839, 800.017 and 761.839 respectively, based on information criteria as shown in Table 4.13 below.

Table 4.13: Model fit based on Information Criteria

CRITERION	Intercept and Co-variates
AIC	773.839
BIC	800.017
-2 LOG L	761.839

From Table 4.13, the probit model in terms of the AIC and the BIC(Schwartz criterion, SC), which is used to compare different models derived from the same sample, different models from different samples and nested and non-nested models, produced results of 773.839 and 800.017 respectively. A smaller value indicates a better model fit. The -2Log L (-2 log likelihood) is used to show how poorly the model predicts the decisions, the lesser the statistic the better the model in terms of model comparison, which produced a result of 761.839.

4.2.3 The Logistic Regression results

Table 4.14 shows the inferential statistics of individual parameter estimates of the logit model to check whether the model is significant at 5% significance level. Likelihood Ratio test the Score test and the Wald test gave test statistics that were significant at the

5% significance level, indicating that at least one of the independent variables is not equal to zero implying the overall model was significant as shown in Table 4.14 below.

Table 4.14: Testing Global Null Hypothesis: BETA=0 (Logit)

TEST	CHI-SQ	DF	Pr>ChiSq
Likelihood Ratio	32.9984	14	0.0029*
Score	32.1156	14	0.0039*
Wald	30.5058	14	0.0065*

***Means significant at 5% significance level**

The full model was then fitted to check the variables that were significant at 5% significance level as shown in Table 4.15 below using the stepwise algorithm, after which the significant variables were selected from the full model and the insignificant variables were dropped from the full model.

Table 4.15: Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard error	Wald Chi-Sq	Pr>ChiSq
intercept	1	1.8290	0.7370	6.1583	0.0131
Age group	1	-0.0329	0.0814	0.1636	0.6858
Gravidity	1	-0.0158	0.2569	0.0038	0.9509
Parity	1	-0.1093	0.1727	0.4004	0.5269
Blood pressure	1	-0.00935	0.1837	0.0026	0.9594
BMI	1	-0.3370	0.0996	11.4547	0.0007*
Sickling	1	-0.3632	0.3686	0.9710	0.3244
Season(dry)	1	0.4074	0.1866	4.7648	0.0290*

Occupation	1	-0.1130	0.0500	5.1076	0.0238*
Religion	1	0.1208	0.2857	0.1788	0.6724
Education	1	0.0632	0.0794	0.6350	0.4255
Marital status	1	0.1027	0.2406	0.1821	0.6696
Trimester	1	-0.8284	0.3246	6.5128	0.0107*
HIV status	1	0.4292	0.8597	0.2492	0.6176
Gestational age	1	0.0837	0.0285	8.5958	0.0034*

***Means significant at 5% significance level.**

Table 4.15 shows the Maximum likelihood estimates of the parameters of the logistic model, which is used to show the predictors which have an effect on the dependent variable (anaemic status), by being significant at 5% significance level. The results produced an AIC, BIC and -2Loglikelihood values of 788.253, 853.699, 758.253 respectively.

It is evident from Table 4.15 that, the predictors that were significant were BMI, Season, Gestational age, Occupation and Trimester. The insignificant variables were dropped from the model and the significant variables were fitted using parameter estimates from the analysis of maximum likelihood for the significant variables using the stepwise algorithm as shown in Table 4.16 below.

Table 4.16: Analysis of maximum likelihood for the significant variables

Parameter	DF	Estimate	Standard error	Wald Chi-Sq	Pr>ChiSq
intercept	1	1.8129	0.4493	16.2801	<.0001
Gestational age	1	0.0807	0.0283	8.1339	0.0043
BMI	1	-0.3501	0.0969	13.0575	0.0003
Season(dry)	1	0.3742	0.1804	4.3045	0.0380
Occupation	1	-0.1032	0.0473	4.7659	0.0290
Trimester	1	-0.7952	0.3215	6.1186	0.0134

From Table 4.16, fitting the model using the parameter estimates, the reduced logit model becomes:

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = 1.8129 + 0.0807x_1 - 0.3501x_2 + 0.3742x_3 - 0.1032x_4 - 0.7952x_5 \qquad (4.2)$$

Where: $p(x)$ = probability(Y = outcome of interest |X = x), x_1 is Gestational age, x_2 is BMI, x_3 is Season(dry), x_4 is Occupation, x_5 is Trimester.

Table 4.17 below, shows the odds ratio estimates of each significant predictor in the reduced logit model. The odds ratio (OR) is a measure of association between an exposure and an outcome, the OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

OR=1 Exposure does not affect odds of outcome.

OR>1 Exposure associated with higher odds of outcome.

OR<1 Exposure associated with lower odds of outcome.

Table 4.17: Odds ratio of the reduced logit model

Odds Ratio Estimates

Effects	Point Estimate	95% Wald Confidence estimates	
Gestational age	1.084	1.026	1.146
BMI	0.705	0.583	0.852
Season(dry)	1.454	1.021	2.070
Occupation	0.902	0.822	0.990
Trimester	0.452	0.240	0.848

It can be seen from Table 4.17 that, variables which have exposure associated with the high odds of having anaemia is the Season(dry). Variable whose exposure is not associated with the odds of having anaemia is the Gestational age. Variables whose exposures were associated with less odds of having anaemia were BMI, Occupation and Trimester.

The 95% confidence interval (CI) is used to estimate the precision of the OR. A large CI indicates a low level of precision of the OR, whereas a small CI indicates a higher precision of the OR.

4.2.4 Model Diagnostics

The reduced model was diagnosed to check if the model is significant, fits the data and how well the model predicts its decisions. An inferential statistics tests was done to show if the model was significant at 5% significance level using the Likelihood Ratio test, the Score test and the Wald tests as shown by the results in Table 4.18 below.

Table 4.18: Testing Global Null Hypothesis: BETA=0 of reduced logit model

TEST	CHI-SQ	DF	Pr>ChiSq
Likelihood Ratio	29.7030	5	<.0001*
Score	28.9827	5	<.0001*
Wald	27.6414	5	<.0001*

***Means significant at 5% significance level**

From Table 4.18, it was evident that, the likelihood ratio test, the score test and the Wald test were all significant at 5% significance level, therefore the null hypothesis of BETA=0 was rejected for all three tests making the reduced model significant.

A goodness of fit statistics was then performed to assess the fit of the logit model against the data using the Hosmer-Lemeshow test whether the model is adequate (insignificant at 5% significance level) at 5% significance level or not adequate as shown in Table 4.19 below.

Table 4.19: Hosmer-Lemeshow goodness of fit statistic for reduced Logit model

number of observations	number of groups	Hosmer-Lemeshowchisq	Prob>chisq
580	10	8.62	0.3753*

***Means insignificant at 5% significance level**

An insignificance level of 0.3753 at 5% significance level indicates that the model fits the data well making the model adequate.

A test of association to measure the relationship between the predictors and the dependent variable in the model was performed to show how well the model correctly predicts probabilities for event outcomes and non-event outcomes using the c statistic as shown in Table 4.20

Table 4.20: Association of Predicted Probabilities and Observed Responses

Percent Concordant	63.4
Percent Discordant	36.1
Percent Tied	0.5
Pairs	82251
Somers' D	0.272
Gamma	0.274
Tau-a	0.133
c	0.636

The tau-a is Kendall's rank-order correlation coefficient without adjustment for ties. The Gamma statistic is based on kendall's coefficient but adjusts for ties. Somer's D is an extension of the Gamma statistic. The c statistic is the proportion of observation pairs with different observed outcomes for which the model correctly predicts a higher probability for observations with the event outcomes than the probability for non-event observations. A value of 1 means the model assigns higher probabilities to all observations with the event outcome compared to non-event observations. The c statistic from the Table 4.21 gives a value of 0.636, which shows that the model assigns high probabilities to more observations with event outcomes compared to non-event observations.

Table 4.21: Classification Table for the reduced Logit model

	Correct		Incorrect						
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensitivity	Specificity	False POS	False NEG
0.500	266	85	162	67	60.5	79.9	34.4	37.9	44.1

Table 4.21 shows the validity of the predicted probabilities with the rows representing the two possible outcomes and columns are high and low probabilities, based on a cut-off point of 0.05.

From the table, the correct column shows the percentage of correct classification of 60.5, implying the validity of the predicted probability of the model is approximately 61%. The sensitivity test indicated that about 79.9 percent of the events of interest (anaemic) were correctly classified by the model while the specificity test indicated that about 34.4% of the non-event (non-anaemic) was correctly classified.

A test was then performed to show the effects of each predictor on the dependent variable as shown in Table 4.22 below.

Table 4.22: Marginal Effects of Predictors

VARIABLE	MEAN	STANDARD DEVIATION
BMI	0.0813060	0.0073425
Gestational age	-0.0187412	0.0016925
Season(dry)	-0.0869085	0.0078484
Occupation	0.0239637	0.0021641
Trimester	0.1846686	0.0166768

From Table 4.22, it is evident that the variables for pregnant women that would cause them to be more likely to have anaemia were BMI, Occupation and Trimester. The variables of the pregnant women that would cause them to be less likely to have anaemia were Gestation and Season (dry).

A test of overall estimated probability of the model was performed to check the predictions made by the model using the variables and the data as shown in Table 4.23 below.

Table 4.23: Estimated probability of the reduced Logit model

Variable	N	Mean	Std Dev	Minimum	Maximum
Anaemicst AnaemicStatus	580	0.5741379	0.4948998	0	1.0000000
Estimated probability (plogit)	580	0.5741379	0.1108220	0.2699720	0.8641184

Table 4.23 shows the estimated probability of the reduced logit model. The anaemic status of the pregnant women showed an average of 57% with a minimum of 0 and a maximum of 1. The estimated probability plogit (predicted logit) showed on average that, fifty-seven percent (57%) of the pregnant women were anaemic (haemoglobin<11g/dl), with a minimum of 0.11 and a maximum of 0.86.

The model fit for the logit model in terms of the AIC and the BIC (Schwartz criterion, SC) which is used to compare different models derived from the same sample, different models from different samples and nested and non-nested models. A smaller value indicates a better model fit.

The -2Log L (-2 log likelihood) is used to show how poorly the model predicts the decisions. The lesser the statistic, the better the model. This produced a result of 761.549 as shown in Table 4.24 below.

Table 4.24: Model fit based on information criteria for the reduced Logit model

CRITERION	Intercept and Co-variates
AIC	773.549
BIC	799.727
-2 LOG L	761.549

4.2.5 The Artificial Neural Network results

The Multilayer Perceptron (MLP) procedure produces a predictive model for one or more dependent (target) variables based on the values of the predictor variables. In order to get an Artificial Neural Network model the data must first be partitioned into training and testing the data in order to get a predictive model to ensure whether the model is good for the data in terms of its predictions. The network training will generally be most efficient if the testing sample is smaller than the training sample as shown in Table 4.25 below.

Table 4.25: Classification table for the Artificial Neural Network model

SAMPLE OBSERVED		PREDICTED		
		Non anemic	anemic	Percent correct
Training	non anemic	112	68	62.2%
	Anemic	49	178	78.4%
	Overall percent	39.6%	60.4%	71.3%
Testing	non anemic	34	33	50.7%
	Anemic	32	74	69.8%
	Overall percent	38.2%	61.8%	62.4%

Table 4.25 shows that using 0.5 as the pseudo-probability cutoff for classification, the network does considerably better at predicting anemic pregnant women than non-anaemic pregnant women. It also shows the validity of the predicted probability of the overall model. Which explains the predictions made by the model would have a 62.4% correct prediction.

The architecture of the neural network model shows how the model is structured to obtain the estimated weights of each of the predictor variables. The hidden layer contains unobservable network nodes (units). Each hidden unit is a function of the weighted sum of the inputs, the multilayer perceptron can have only two hidden layers, but for this research only one hidden unit was used. The function is the activation function, which links the weighted sums of units in a layer to the values of units in the succeeding layer and the values of the weights are determined by the estimation algorithm as shown in Table 4.26.

Table 4.26: Weights for Independent Variables

VARIABLE	HIDDEN LAYER	WEIGHTS
intercept	1	-0.792
Age group	1	5.223
Gravidity	1	-6.891
Parity	1	-16.944
Blood Pressure	1	-6.460
BMI	1	0.7812
Sickling	1	8.577
Season(dry)	1	10.792
Occupation	1	15.320
Religion	1	-18.601
Education	1	-3.747
Marital status	1	0.294
Trimester	1	5.652
HIV status	1	-23.548
Gestational age	1	2.257

The model fit for the neural network model in terms of the AIC and the BIC (Schwartz criterion, SC) was 202.144 and 346.124 respectively. The weights which are the parameter estimates of the model were used to fit the overall model. The independent variables need to be assessed to check the importance of each variable and its effect on the dependent variable. The importance of the independent variables performs a sensitivity analysis of each independent variable in determining the neural network. The importance of an independent variable is a measure of how much the network's model-predicted value changes for different values of the independent variable. Normalized importance is simply the importance values divided by the largest importance values and expressed as percentages. It can be seen that the three most important variables are the BMI with 85.1%, Gestational age with 100% and the Occupation with 86.1% as shown in Table 4.27 below.

Table 4.27: Independent variable importance

Variable	Importance	Normalised Importance
Body Mass Index(BMI)	.154	85.1%*
Gestational age (in weeks)	.181	100.0%*
Occupation	.155	86.1%*
Maternal Age(Grouped)	.074	41.0%
Gravidity (Number of Pregnancies)	.044	24.1%
Parity	.053	29.2%
Blood Presure(BP)	.032	17.7%
Sicklen status(Positive/Negative)	.044	24.6%
Season(Dry/ Wet)	.079	43.9%
Religious status	.023	12.5%
Educational status	.055	30.7%
Marital status	.020	11.3%
Trimester status	.063	34.8%

Hiv status	.023	12.7%

***Means largest Importance values above 50%**

The Pareto chart is simply a bar chart of the values in the importance table, sorted in largest value of important variables as shown in figure 2 below.

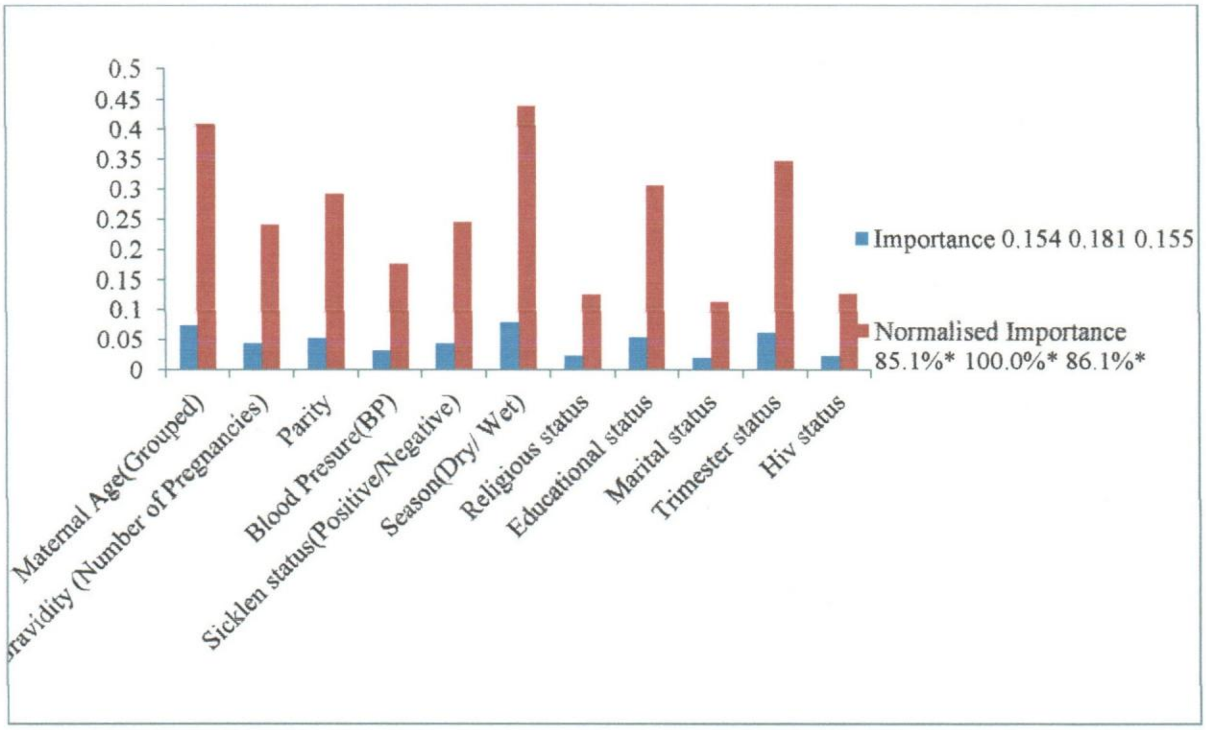

Figure 4.1: Pareto chart for Independent Variable Importance

It appears that variables related to the anaemic status of pregnant women above50% were BMI, Gestation and Occupation. These predictors have the greatest effect on how the network classifies anaemic status of pregnant women. The predictors which were least important in the model were then dropped from the overall model and the most important predictors were fitted using their weights to get a reduced predictive model. Figure 3 below shows the Multilayer perceptron architecture of the BMI, Gestational age and Occupation predictors, with Table 4.28 showing their fitted values respectively.

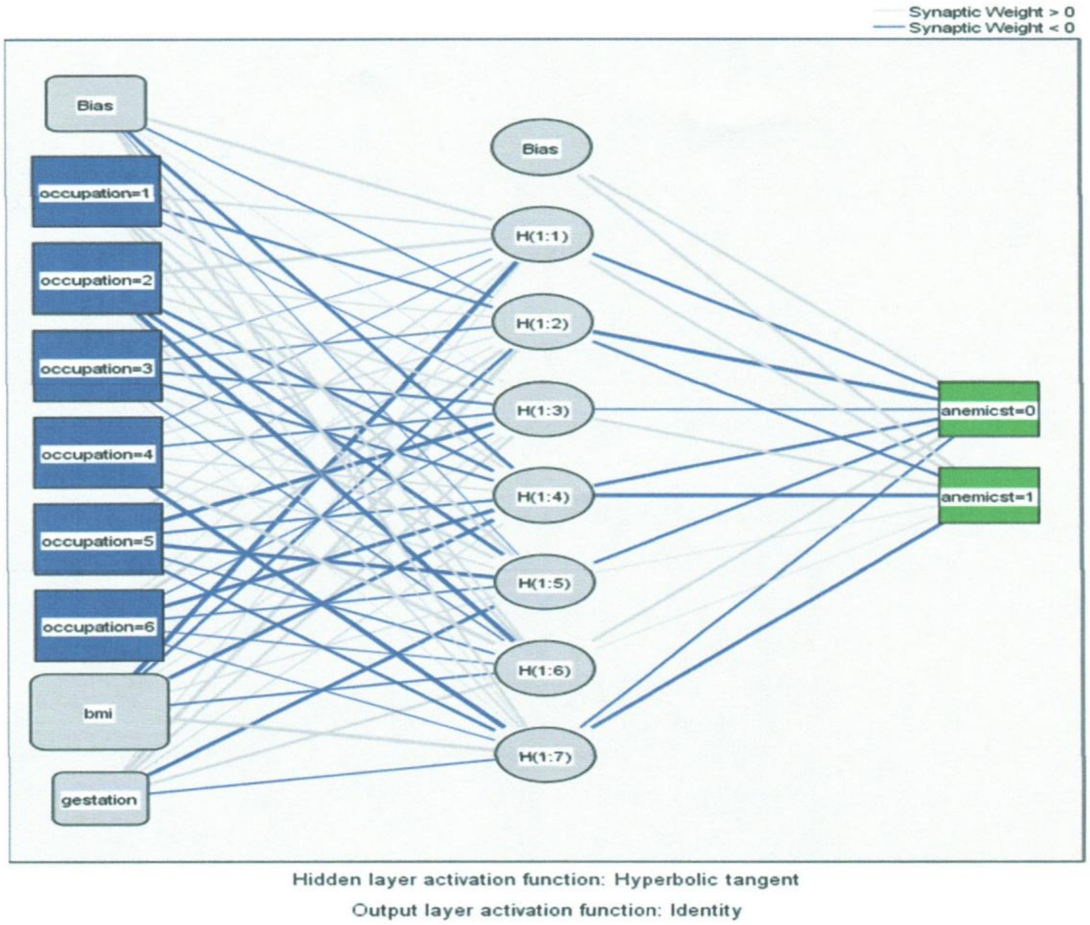

Figure 4.2: Architecture of the Reduced Neural Network model

Figure 4.2 shows the architecture of the Artificial Neural Network model. The first row containing the input for the largest important variables with a bias which is always set to 1. The middle row showing the hidden layer of 1 for each independent variable were all the processing is done through an activation function which leads to the last row giving an output of the anaemic status.

Table 4.28: Fitted values for the most important variables

Variable	Hidden layer	Weight
Intercept	1	-8.858
BMI	1	1.206
Gestational age	1	-0.036
Occupation	1	0.206

From Table 4.28, a reduced model based on the most important variables was developed from the overall fitted model by dropping the least important variables, giving a reduced model. The positive parameter estimates of the BMI and Occupation shows that, A unit increase in the BMI of a pregnant woman who is a trader would increase the probability of a pregnant woman being anaemic and the negative parameter estimate of the gestational age of the pregnant woman, indicates that, the lesser the Gestational age, the probability of the subject being anaemic is lower.

4.2.6 Model Diagnostics

The model was diagnosed to check if the model fits the data well, and how well the model predicts its decisions as shown in Table 4.29 below.

Table 4.29: Classification of reduced Artificial Neural Network model

SAMPLE OBSERVED		PREDICTED		
		Non anemic	anemic	Percent correct
Training	non anaemic	49	129	27.5%
	Anaemic	45	192	81.0%
	Overall percent	22.7%	77.3%	58.1%
Testing	non anaemic	24	45	34.8%
	Anaemic	13	83	86.5%
	Overall percent	22.4%	77.6%	64.8%

Table 4.29 shows that, the percentage of cases correctly classified by the Artificial neural network is 64.8%, implying the predictions made by the reduced Artificial Neural Network would have a64.8% correct prediction.

Figure 4.3 below shows the predicted-by-observed chart which displays clustered boxplots of predicted pseudo-probabilities for the combined training and testing samples. The x axis corresponds to the observed response categories, and the legend corresponds to predicted categories.

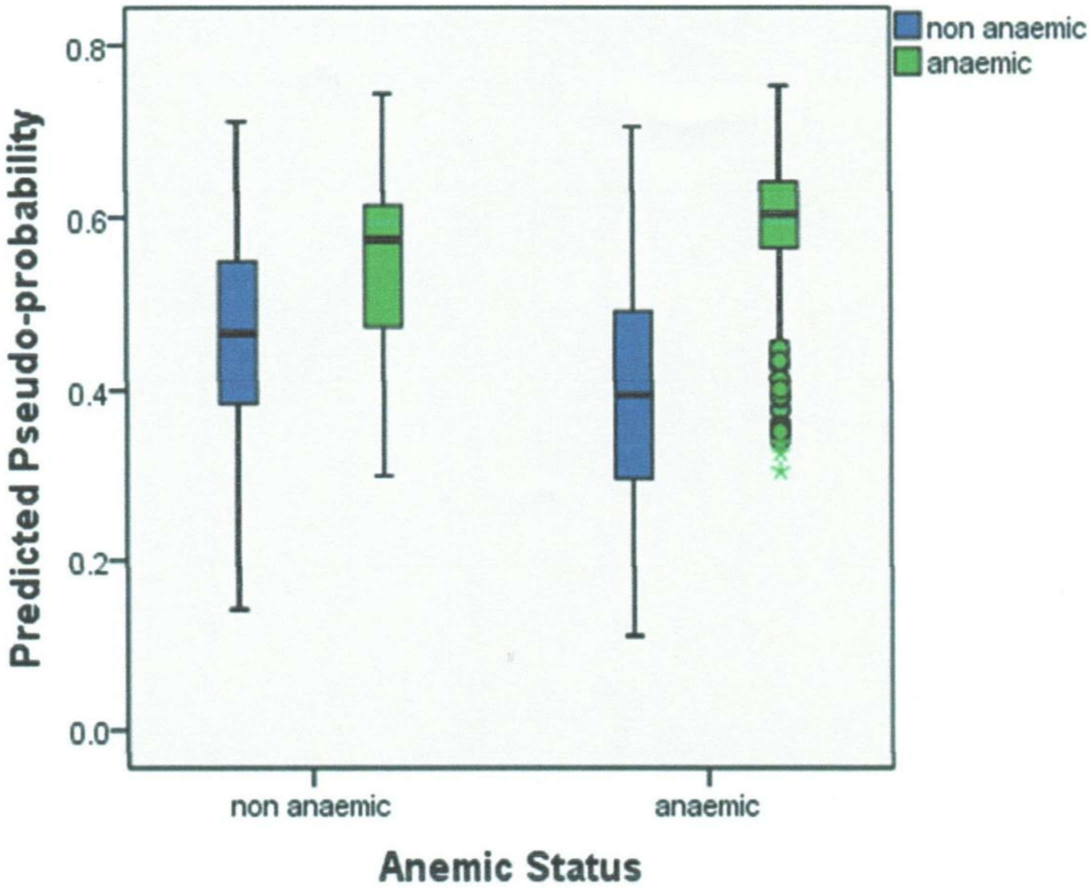

Figure 4.3: PREDICTED PSEUDO-PROBABILITY

- The leftmost boxplot shows, for cases that have observed category non anaemic, the predicted pseudo-probability of category non anaemic. The portion of the boxplot above the 0.5 mark on the *y* axis represents correct predictions shown in the classification table. The portion below the0.5 mark represents incorrect predictions.
- The next boxplot to the right shows, for cases that have observed category non anaemic, the predicted pseudo-probability of category anaemic. Since there are only two categories in the target variable, the first two boxplots are symmetrical about the horizontal line at 0.5.

- The third boxplot shows, for cases that have observed category anaemic, the predicted pseudo-probability of category non anaemic. It and the last boxplot are symmetrical about the horizontal line at 0.5.
- The last boxplot shows, for cases that have observed category anaemic, the predicted pseudo-probability of category anaemic. The portion of the boxplot above the 0.5 mark on the y axis represents correct predictions shown in the classification table. The portion below the 0.5 mark represents incorrect predictions.

The model fit for the neural network model in terms of the AIC and the BIC (Schwartz criterion, SC), produced results of 147.487 and 173.665 respectively. Threshold reached shows the value of the activation function (sigmoid) which is 0.010, as shown in Table 4.30 below.

Table 4.30: Model fit based on Information Criteria for Artificial Neural Network

ERROR	AIC	BIC	THRESHOLD REACHED	STEPS
67.743	147.487	173.665	0.010	5436.000

4.2.7 Severity of Anaemia

In addition, to model the probability of a subject suffering from various levels of anaemia (severe, moderate or mild), the Haemoglobin level of subjects was considered multichotomous (non-anaemic, severe anaemia, moderate anaemia and mild anaemia) and a multinomial logistic regression was fitted to the data using the stepwise algorithm.

The model was fitted without the intercept because the intercept was not significant at the 5% significance level. From Table 4.31, only the Occupation, Trimester, Gestational age and BMI are useful in modelling the probability of the various levels of anaemia. From Table 4.31 below, TheOccupation, Trimester, BMI and the Gestational age were useful in predicting the probability of a subject suffering from moderate and mild anaemia but not useful in predicting the probability of a subject suffering from severe anaemia; The BMI was useful in predicting the probability of the various levels of anaemia. TheOccupation, Trimester, and the BMI have negative coefficients indicating that an increase in the BMI of a pregnant woman who is eight weeks and is a hairdresser are likely to reduce the probability of a subject suffering from the various degree of anaemia. The Gestational age has a positive coefficient indicating that as the Gestational age increases, the probability of the subject suffering from the various degrees of anaemia increases.

Table 4.31: Parameter Estimates for the reduced Multinomial Logistic Regression

Parameter	Mild verses Non Anaemia	Moderate verses Non-Anaemia	Severe verses Non-Anaemia
BMI	-0.278 $(P-value = 0.016)$	-0.410 $(P-value = 0.001)$	-0.380 $(P-value = 0.128)$
Gestational Age	$.082$ $(P-value = 0.014)$	0.103 $(P-value = 0.005)$	-0.016 $(P-value = 0.829)$
Occupation	$-.137$ $(P-value = 0.019)$	-0.089 $(P-value = 0.161)$	-0.064 $(P-value = 0.622)$
Trimester	-0.830 $(P-value = 0.029)$	-1.031 $(P-value = .014)$	0.350 $(P-value = 0.686)$

4.2.8 Comparative analysis of Probit Regression, Logistic Regression and ANN models

The differences in the performance of the Probit model, Logit model and Artificial Neural Network model was assessed based on their predictive potential since the percent correct is the most commonly used goodness-of-fit measure for ANN models. These results are shown in Table 4.31 below, as shown in this table, the Artificial Neural Network has a significant edge over the Probit model and the Logit model in terms of the percentage of cases correctly classified. Overall, the ANN classifies 64.8% cases correctly while this rate is 60.0% for the Probit model and 60.5% for the Logit model.

Table 4.32: Classification table for models

MODEL	PERCENT CORRECT	PREDICTIVE POTENTIAL
PROBIT	60.0	GOOD
LOGIT	60.5	GOOD
NEURAL NETWORK	64.8	BETTER

The comparison of the model fit in terms of information criteria of the models is shown in Table 4.32 below, as shown in this table the ANN showed a better model fit over the probit model and the logit model in terms of the AIC and the BIC. Overall the ANN shows an AIC and BIC of 153.426 and 192.693 respectively whilst the AIC and BIC for the logit model is 770.106 and 800.647 respectively and AIC and BIC for the probit is 770.196 and 800.737 respectively, explaining that, the Artificial Neural Network model is better than the Logit and the Probit models because the lesser the statistic of the model the better the model.

Table 4.33: Model fit based on Information Criteria

MODEL	AIC	BIC
NEURAL NETWORK	147.487*	173.665*
LOGIT	773.549	799.727
PROBIT	773.839	800.017

*** Means the best model. The lesser the statistic the better the model**

The overall model comparison showed the Artificial Neural Network model surpassing the Logit and Probit models in terms of the AIC, BIC, and model classification, making the model the appropriate model as compared to the Logit and Probit models.

4.3 Discussion of Results

From the results of the descriptive statistics, it was evident that out of the five hundred and eighty pregnant women 333 (57%) were anaemic (Haemoglobin<11g/dl) and 247 (43%) were non-anaemic (Haemoglobin>11g/dl).

This prevalence rate of 57% is consistent with estimated prevalence rate of 40% to 60% in developing countries. However, the prevalence is comparably almost the same to 57.1% in the Sekyere West District of Ghana, (Glover-Amengor *et al.*, 2005), 76.9% in South Eastern Nigeria (Uneke *et al.*, 2007) and 62.6% in Eastern Sudan (Adam *et al.*, 2007), which may be attributable to geographical variations that influence certain anaemic influencing factors and improvement in health policies with time.

The minimum, average and maximum haemoglobin levels were 4.5g/dl, 10.6g/dl and 16.6g/dl respectively. However, the severity of anaemia was recorded to be 175 (30.2%) for mild anaemia (Haemoglobin between 10.0 to 10.9g/dl), 136 (23.4%) for moderate anaemia (Hemoglobin between 7.0 to 9.9g/dl) and 22 (3.8%) for severe anaemia

(Haemoglobin less than 7g/dl). The results revealed that, mild anaemia was common followed by moderate anaemia. This is consistent with findings in Africa and elsewhere in the world.

The chi-square test of association showed results of BMI(body mass index), Season(dry/wet) and Occupation being significantly associated with the anaemic status (anaemicst) of the pregnant women. Which means for BMI 286 (49.3%) of the pregnant women had a healthy weight and this could increase or decrease the probability of a pregnant woman being anaemic. This may be due to causes like, parasitic worm infestations and genetic disorders (WHO, 1995). In the dry season, pregnant women are more likely to be anaemic and pregnant women who are Hairdressers would increase the probability of being anaemic. However, the test showed no association for age group, Gravidity, Parity, Occupation, Trimester, Education, Religion, HIV status, Marital status, Blood Pressure, and Sickling at 5% significance level.

The results of the inferential statistics of the overall probit model showed that, the model was good for analyzing the data, since the inferential statistics produced results of significant tests for the Likelihood Ratio test, the Score test and the -2log likelihood at 5% significance level. From the analysis of maximum likelihood estimates for fitting the model the variables which were significant as compared to a 5% significance level were BMI, Gestational age, Season, Occupation and Trimester which resulted in a reduced probit model. This results contradicts results by researchers Pala and Dundar (2007) who found no independent relation between BMI and the anaemic status of pregnant women and Nwizu *et al.*, (2011) who found a significant relationship between anaemia in pregnancy and educational status, marital status and parity. This difference may be due to geographical factors, genetic disorders among others between the countries. The results also supports works by researchers karaoglu *et al.*, (2010) who also researched that, the Trimester of pregnancy was significantly associated with the anaemic status of pregnant women.

The parameter estimates of the BMI, Gestational age, Occupation and Trimester were negative indicating that the variables are likely to reduce the probability of a pregnant woman being anaemic, which explains that a woman whose BMI is between 18.6-24.9 (healthy weight) is less likely to be anaemic, A pregnant woman who is 8 weeks pregnant is less likely to be anaemic. A pregnant woman who is a hairdresser is less likely to be anaemic. A pregnant woman who is in her mid-stage (13-24) of pregnancy is less likely to be anaemic. The positive parameter estimate of the season (dry) indicates that in the dry season, the probability of a pregnant woman being anaemic is higher.The diagnostic checks of the reduced model showed that the model fitted the data well as the inferential statistics of the Wald test, the Score test and the Likelihood Ratio test were all significant compared to a 5% significance level making the reduced model significant. The Hosmer-Lemeshow test also gave an insignificant test result at 5% significance level which showed the model was adequate.The validity of the predicted probability results for the classification of the reduced model produced a correct classification result of 60.0% for the model implying predictions made by the model would have a validity of 60.0% correct.The estimated probability showed that, the reduced probit model predicted that the number of pregnant women said to be anaemic (haemoglobin<11g/dl) were 57%.

The results of the inferential statistics of the overall logistic model showed that the model was good for analyzing the data since the inferential statistics produced results of significant tests for the Likelihood Ratio test, the Score test and the Wald Test at 5% significance level. The maximum likelihood estimates of the overall logistic model showed that, the variables which were significant were BMI, Season (dry), Gestational age, Occupation and Trimester status at5% significance level, which resulted in a reduced logit model. The results contradicts results by researchers Melku *et al*., (2013) who found HIV to be significantly associated with anaemia in pregnancy, which may be due to geographical area, factors, illiteracy, season among others. The result also supports views by researchers karaoglu *et al*., (2010) who also researched that, the Trimester of pregnancy was significantly associated with the anaemic status of pregnant women. The parameter estimates of the BMI, Occupation and Trimester were negative indicating that the variables are likely to reduce the probability of a pregnant woman being anaemic, which explains that a woman whose BMI is between 18.6-24.9 (healthy weight) is less likely to be anaemic. A pregnant woman who is 8 weeks pregnant is less likely to be anaemic. A pregnant woman who is a hairdresser would decrease the probability of being anaemic. A pregnant woman who is in her mid-stage (13-24 weeks) of pregnancy is less likely to be anaemic. The positive parameter estimate of the season (dry) and gestational age indicates that there is a higher probability of a woman who is 8 weeks pregnant in the dry season being anaemic. The Odds Ratio (OR) of the logistic model for the significant variables which were BMI, Season, Occupation, Gestational age and Trimester, showed that for BMI the model predicts the odds that pregnant women with BMI between 18.6-24.9(healthy weight) are 0.705 times less likely to be anaemic. The Gestational age of women who are 8 weeks pregnant are 1.084 times not likely to be anaemic. In the dry season pregnant women are 1.454 times more likely to be anemic. A pregnant woman who is a Hairdresser is 0.902 times less likely to be anemic and a woman who is between 13-24 weeks (mid stage) pregnant is 0.452 times less likely to be anaemic. The diagnostics of the reduced model showed that, the model was significant as the inferential statistics of the Wald test, the Score test and the Likelihood ratio test were all significant at 5% significance level. The Hosmer-Lemeshow test also gave an insignificant test at 5% significance level which showed the model was adequate. The validity of the predicted probability results showed a result of 60.5% for the model, implying whatever predictions made by the model in terms of anemic pregnant women would have a validity of 60.5% correct.The estimated probability showed that for the predicted logit model, the number of pregnant women said to be anaemic (Haemoglobin level <11g/dl) were 57%.

The training data set for the overall model of the Neural Network showed a percentage of 71.3 whilst the overall percentage of the testing data set showed 69.8, which showed the training of the data is most efficient for the data with a hidden layer of 1. The neural network classification showed an overall percent correct of 64.8% which shows that the overall model is good for the data and also good in terms of predictions made by the model. The weights of the overall model showed the importance of each independent variable based on their percentages. The importance of an independent variable is a measure of how much the network's model-predicted value changes for different values of the independent variables. The variables which were most important at 50% were BMI, Gestational age and Occupation and the least important variables were dropped from the model. Fitting the model for the most important variables produced a reduced neural network model. From the reduced model the positive parameter estimates of BMI and Occupation variables would cause an increase in the probability of a pregnant woman being anaemic, which explains that an increase in the BMI of pregnant woman who is a Hairdresseris more likely to be anaemic. The negative parameter estimate of the Gestational age of a pregnant woman would cause a decrease in the probability of a pregnant woman being anaemic, which explains that a pregnant woman with low gestational age is less likely to be anaemic. The AIC and BIC of the neural network produced results which showed it to be a better model in terms of model comparison based on information criteria.

The results for all three models showed that, the data fitted the models well, the chi-square test of association which showed association between BMI, Season (dry) and Occupation, supports views made by researchers such as Ndukwu and Dienye (2012) and Nwizu *et al*., (2011) who showed an association between anaemia in pregnancy and socio-demographic factors such as occupation in terms of chi-square test. The results of the probit and logit models indicated that only BMI, Gestational age, Season, Occupation and Trimester are useful in determining the probability of a subject being anaemic. Studies in Burkina Faso (Meda *et al*., 1999), sudan (Adam *et al*., 2005) and Jimma (Belachew and Legesse, 2006) reported that Gravidity did not have a statistically significant contribution for difference in anaemia prevalence.

The neural network model was able to represent the structure of the logit and probit models and was able to relate the parameters of ANNs to those of the logit and probit equations, which supports views by Sarle (1994) who was not able to determine any formal equations to relate the parameters of ANNs to those of statistical regression equations. Although the neural network had a higher prediction value for its classification, the results proved that, neural networks, logistic regression and probit regression models could effectively predict the variables which affect anaemia in pregnant women, and that the performance of the three models were almost similar, as suggested by researchers Schnabel and Maneta (2005).

The probit and logit models both produced a reduced model with five variables with the neural network producing a reduced model with three variables, because the neural network model used a sigmoid function and a hidden layer which acts a universal function approximator and it's back propagation, thereby only paying attention to the best significant variables in the model, than the statistical regression models which paves way for all significant variables in the model in terms of fitting the model. The neural network model also had a limitation of the ANN training and application which involved the composition of nonlinear functions that makes it difficult to simplify and reduce to terms that can be understood physically. Another limitation is that there is no standard method for finding the ideal number of hidden nodes in a feed-forward network or for determining the best activation function. This is usually a time consuming, trial-and error process and can lead to inelegant ANN designs, just as suggested by researchers (Shigidi and Garcia, 2003). The limitations of the probit and logit models equations limit the user by requiring a large amount of sample data to estimate the parameters of the equation and to find the data trend. Also, there are difficulties that arise when manually determining the optimal structure of the statistical equation.

The contradiction of some results of this research to views of other researchers may be due to disparities in data on these variables in different countries due to differences in geographical area, geographical factors, genetic orders, nutrient availability among others.Also, many of these researches used data from most undeveloped African countries such as Somalia, Ghana, Sudan, Mali among others.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.0 Introduction

This chapter presents the conclusion and recommendations of the study. The chapter is further divided into conclusion and recommendations.

5.1 Conclusion

In this research the use of three modelling techniques, the Artificial Neural Network model, the Logistic regression model and the Probit regression model were investigated in terms of their applicability to anaemia in pregnancy. All methods generated strong results for the model fit, their predictive potential, and goodness of fit tests. The models were also adequate and appropriate for the data although the multi-layer perceptron Artificial Neural Network yielded a better predictive potential, predicting an average of 64.8% of cases correctly compared to 60.0% for the probit model and 60.5% for the logit model.

The Artificial Neural Network model also produced better results in terms of the AIC and BIC than the Probit model and the Logistic regression model. The Logistic regression model also yielded a better predictive potential of 60.5% cases correct than the Probit model with 60.0%. The Logistic regression model results for the AIC and BIC showed it was also a better model than the Probit model. The Independent predictors that were found to contribute significantly to anaemia in pregnancy were BMI, Gestational Age and Occupation.The results revealed that, Occupation, Gestational age and BMI were useful in determining the Haemoglobin level and the anaemic status of the pregnant women. The studies also revealed that as the Gestational age increases the pregnant women are likely to have low Haemoglobin levels which results in anaemia in pregnancy. The pregnant women who were anaemic (haemoglobin<11 g/dl) were 57.4% as compared to non-anaemic (haemoglobin>11g/dl) pregnant 42.6%, with mild anaemia being common followed by moderate anaemia. The prevalence of 57.4% of pregnant women being anaemic means anaemia is a severe public health problem that needs immediate attention in the municipality.

Although each model had its own advantage and disadvantage the Artificial Neural network model turned out to be the best model in terms of its applicability to anaemia in pregnancy. The independent predictors that were found to significantly contribute to anaemia in pregnancy were the BMI, Gestational age and Occupation. The comparison presented confirmed the robustness of Artificial Neural Network model. These types of models can generate better prediction results than the traditional discrete choice methods in much less time. However, it is likely that the “black box” nature of these procedures will continue to be a concern to practitioners, who generally are more comfortable with the greater transparency of statistically-based methods.

One of the major issues concerning the application of Artificial Neural Network models in an integrated modelling system is the difficulty of integrating these types of models into a larger modelling framework where other models have different modelling structures.

Additionally, while the logit model and the Probit model offers statistical interpretation (elasticity, goodness-of-fit, etc.), the ANN does not appear to offer explicit measure of sensitivity. This imposes some limitations on the applications of ANN for policy and sensitivity analysis.

5.2 Recommendations

Following the outcome of this research work, the following recommendations were made;

i. Medical experts should advise the pregnant women to take in foods and or supplements that are high in iron that will increase their Haemoglobin level as the Gestational age increases and also engage in activities such as exercises that will ensure that they have a healthy weight or normal Body Mass Index.

ii. Stakeholders should put better measures in place that will motivate people to know their anaemic status early so that treatment could be initiated as early as possible.

iii. Clinicians and health care workers as well as other stakeholders should put measures that will guide and monitor anaemic patients to continue to adhere to treatment, so that patients will continue to experience increase in their haemoglobin levels.

iv. While this study is useful as a guide for education to the Public, particularly Patients, and also to guide policy and management of treatment, further studies are recommended to expand the scope of study as well as include more predictors to determine more independent factors which contribute to anaemia.

v. More research needs to be done to get a better understanding of how all three models are related.

REFERENCES

Adam, D., (1958). *Les Reactions du ConsommateurDevant les Prix*.Number**15** in Observation Economique. Paris: Sedes.

Aitchison, J., Brown, C. A. J., (1957). *The Lognormal Distribution*.Number 5 in University of Cambridge, Department of Applied Economics Monographs. Cambridge: Cambridge University Press.

Belachew, T., and Legesse, Y., (2006). Risk Factors for Anemia Among Prgenant Women Attending Antenatal Clinic at Jimma University Hispital, South Wes Ethiopia. *Ethiopian Medical Journal*, **44**(3): 211-220.

Berkson, J., (1951). Why I prefer logits to probits. *Biometrics* 7, 327-339.

Blackwell S, Merck Manual 2008

Bliss, C. I., (1934). The method of probits. *Science* 79, 38-39.

Bouvier, P., Doumbo, O., Breslow, N., Robert, C., Mauris, A., Picquet, M., (1997). Seasonality, malaria and impact of prophylaxis in a west African village. Effect on anaemia in pregnancy.*Am. J. Trop. Med. Hyg*. **56**: 378-383.

Brabin, B. J., (1991). The risks and severity of malaria in pregnant women. Applied field Research in Malaria Reports 1: 21-25.

Brabin, B. J., Hakimi, M., Pelletier D., (1999) An analysis of anemia and pregnancy-related maternal mortality. *J Nutr*1999;**131**:604S-615S

Brion, A. M., Leary, D. S., Smith, D. G., McArdle,J. H., Ness R. A.,(2008). Maternal anemia, iron intake in pregnancy, and offspring blood pressure in the Avon Longitudinal Study of Parents and Children. *American Journal of Clinical Nutrition*2008; **88**:1126 –33.

Broek, V.D.N.R., Ntonya, C., E. Mhango, E., White S. A.,(1999) Diagnosing anaemia in pregnancy in rural clinics: assessing the potential of the Haemoglobin Colour Scale Bulletin of the World Health Organization, 1999, **77**(1)

Broek, V.D.N.R., (1999)Anaemia in pregnancy in developing countries.Review. *British journal of obstetrics and gynaecology*, 1998, **105**: 385–390.

Chon, K. H., Cohen, J. R., (1997) "Linear and Nonlinear ARMA Model Parameter Estimation Using anArtificial Neural Network."IEEE Transactions on Neural Networks.Vol. 44. **No. 3** (March 1997):168 - 174.

Cobourn, W. G., Dolcine, L., French, M., Hubbard, C. M., (2000)."A Comparison of Nonlinear Regression and Neural Network Models for Ground-Level Ozone Forecasting."Journal of the Air and Waste Management Association.**Vol. 50**. (November 2000): 1999 - 2009.

Cox, D. R., (1969).*Analysis of Binary Data*. London: Chapman and Hall.

Crompton, D. W. T., Nesheim, M. C., (2002). Nutritional impact of intestinal helminthiasis during the human life cycle, Ann. Rev. Nut.**22**: 35-59.

Dey, S., Goswami, S., Goswami, M., (2010). Prevalence of anaemia in women of reproductive age in Meghalaya: a logistic regression analysis Turkish Journal of Medicine and Science 2010; **40** (5):783-789

Dreyfuss, M. L., Stoltzfus, R. J., Shrestha, J. B., (2000) Hookworms, malaria and vitamin A deficiency contribute to anemia and iron deficiency among pregnant women in the plains of Nepal. *J Nutr* 2000; **130**:2527-2536.

Egwunyenga, O. A., Ajayi, J. A., Duhlinska-Popova, D. D. , Nmorsi, O. P. G.. (1997). Transplacental passage of Plasmodium falciparum and seroevaluation of newborns in Northern Nigeria.South Eastern, Asia. *J. Trop. Med. Public. Health* **28**: 741-745.

Farrell, M. J., (1954). The demand for motorcars in the United States.*Journal of the Royal Statistical Society, series A* **117**, 171-200.

Fechner, G. T., (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf und HÄartel.

Fuseini, G., Edoh, D., GumahKalifa, G. B., Abdul-Wahab H., Knight D., (2010). Parasitic infections and anaemia during pregnancy in the Kassena-Nankana district of Northern Ghana.*Journal of Public Health and Epidemiology* Vol. **2**(3), pp. 48-52, June 2010.

Gaddum, J. H., (1933). *Reports on Biological Standard III.Methods of Biological Assay Depending on a Quantal Response*. London: Medical Research Council. Special Report Series of the Medical ResearchCouncil, no. 183.

Gavira, A., Hoyos, A., (2011). Anaemia and Child Education: The case of Colombia, Desarrolloy sociedad, num. 68, diciembre, 2011, pp.47-77. ISSN 0120-3584 revistadesarrolloysociedad@uniandes.edu.co

Ghana Demographic and Health Survey (2003) Nutrition of young children and mothers in Ghana.

Ghana Demographic and Health Survey 2003

Ghana Health Service (2005).Upper East Regional Health Administration 2004 Annual Report, Upper East Region

Ghana News Agency (2004) *(The Ghanaian Times, Tuesday August 17, 2004)*

Ghana News Agency (2013), October, 17 Anaemia leading cause of maternal deaths in Ghana.

Ghana News agency NA, Accra, 27 Feb. 2004 Most children and pregnant women anaemic.

Ghana Statistical Service (GSS) and Macro International Inc. (MI). 1999. Ghana Demographic and Health Survey 1998. Calverton, Maryland: GSS and MI.

Ghana Statistical Service (GSS), Noguchi Memorial Institute for Medical Research (NMIMR), and ORC Macro. 2004. Ghana Demographic and Health Survey 2003. Calverton, Maryland: GSS, NMIMR, and ORC Macro.

Hebb, Donald (1949). *The Organization of Behavior*. New York: Wiley.

Hosmer, D. W., Lemeshow, S., (2000).*Applied Logistic Regression*. New York: Wiley.

Jilly, P., (1969). Anaemia in parturient women, with special reference to malaria infection of the placenta.Annals of Trop. Med. Parasitol.**63**:109-116

Karaoglu, L., Pehlivan, E., Egri, M., Deprem, C., Gunes, G., Genc F. M., Temel, I., (2010). The prevalence of nutritional anemia in pregnancy in an east Anatolian province, Turkey *BMC Public Health* 2010, **10**:329 http://www.biomedcentral.com/1471-2458/10/329

Malthus, T. E., (1798). *An Essay on the Principle of Population*.London:anon.

McCulloch, Warren; Walter Pitts (1943)."A Logical Calculus of Ideas Immanent in Nervous Activity". *Bulletin of Mathematical Biophysics* **5** (4): 115-133.doi:10.1007/BF02478259.

McFadden, D., (1973). Economic choices.*American Economic Review* 91, 352-370. Nobel prize acceptance speech.

McGregor IA, Wilson ME, billewicz WZ (1983). Malaria infection of the placenta in the Gambia, West Africa, its incidence and relationship to stillbirth, birth weights and placenta weights. Transaction of the Royal. Society. Trop. Med. 77: 232-244.

Melku, M., Addis, Z., Alem, M., Enawgaw B., (2013). Prevalence and Predictors of Maternal Anemia during Pregnancy in Gondar, Northwest Ethiopia: An Institutional Based Cross-Sectional Study. Hindawi Publishing Corporation Volume 2014, Article ID 108593, 9 pages http://dx.doi.org/10.1155/2014/108593

Ndukwu G. U., Dienye P. O., Prevalence and socio-demographic factors associated with anaemiain pregnancy in a primary health centre in Rivers State, Nigeria. *Afr J Prm Health Care* Fam Med. 2012;**4**(1), Art. #328,7 pages. http://dx.doi.org/10.4102/phcfm.v4i1.328

Nwizu, E., Iliyasu, Z., Ibrahim, S., Galadanci H., (2011)Socio-Demographic and Maternal Factors in Anaemia in Pregnancy at Booking in Kano, Northern Nigeria. *African Journal of Reproductive Health December 2011;* **15**(4): 33

Oue´draogo, S., K. Koura, K. G., Bodeau-Livinec, F., M. K. Accrombessi, M. K. M., Massougbodji, A., Cot, M., (2013) Maternal Anemia in Pregnancy: Assessing the Effect of Routine Preventive Measures in a Malaria-Endemic Area.*Am. J. Trop. Med. Hyg.*,**88**(2), 2013, pp. 292–300

Pala, K., Dundar N., (2007). Prevalence & risk factors of anaemia among women of reproductive age in Bursa, Turkey. *Indian journal of Medicinal Research***128**, September 2008, pp 282-286

Pearl, R., Reed, J. L., (1920). On the rate of growth of the population of the United States since 1870 and its mathematical representation. *Proceedings of the National*

Academy of Sciences 6, 275-288.Prevention and management of severe anaemia in pregnancy.Geneva,

World Health Organization,1993 (unpublished document) WHO/FHE/MSM/93.5; available upon request from Division of Reproductive Health, World Health Organization, 1211 Geneva 27, Switzerland).

Quetelet, A., (1848). *De systµeme social et des lois qui le regissent*. Paris :Guillaume.

Quetelet, A., (1850). Notice sur Pierre-Francois Verhulst. *Annuaire de l'Academie*

Reed, L. J., Berkson, J., (1929). The application of the logistic function to experimental data. *Journal of Physical Chemistry33*, 760-779.

WHO, (1999) Report at the UNICEF & WHO. *Regional consultation.Prevention and control of iron deficiency anaemia in women and children.* Geneva: World Heath Organization; 1999.

Rochester, N., Holland J.H., Habit L.H., and Duda W.L., (1956). "Tests on a cell assembly theory of the action of the brain, using a large digital computer".*IRE Transactions on Information Theory***2** (3): 80–93

Sarle, W. S., (1994) "Neural Networks and Statistical Methods."Proceedings of the Nineteenth Annual SAS Users Group International Conference. (April 1994).

Schnabel, S., Maneta M., (2004)."Comparison of a Neural Network and a Regression Model to Estimate Suspended Sediment in a Semiarid Basin."Proceedings of the International Conference, Catalonia, Spain. (May 2004): 91 - 100.

Steketee, R. W., (2001). The burden of malaria in pregnancy in malaria endemic areas. *Am. J. Trop. Med. Hyg.***64**(1, 2 S): 28-35

Stigler, S. M., (1986). *The History of Statistics*. Cambridge, Mass.:

Harvard University Press.

Sukrat, B., Wilasrusmee, C., Siribumrungwong, B., McEvoy, M., Okascharoen, C., Attia, J., and Thakkinstian A., (2013) Hemoglobin Concentration and Pregnancy Outcomes: A Systematic Review and Meta-Analysis BioMed Research International Volume 2013, Article ID 769057, 9 pages http://dx.doi.org/10.1155/2013/769057

Theil, H., (1969). A multinomial extension of the linear logit model.*Inter- national Economic Review* 10, 251-259.

Tokar, A. S., Johnson,A. P., (1999) "Rainfall-Runoff Modeling Using Artificial Neural Networks." *Journal of Hydrologic Engineering.* (July 1999): 232 - 239.

Uneke, C. J., Duhlinska, D. D., and Igbinedion, E. B., (2007). Prevalence and Public Health Significance of HIV Infection and AnaemiaAmong Pregnant Women Attending Antenatal Clinics in South Eastern Nigeria. *Journal of Health, Population and Nutrition*, **25**(3): 328-335

UNDP (2010).Bolgatanga Municipality

Vanderjagt, D. J., Brock, H. S., Melah, G. S., El-Nafaty, A. U., Crossey, M. J., Glew, R. H., (2007) Nutritional Factors Associated with Anaemia in Pregnant Women in Northern Nigeria. *J Health PopulNutr*. 2007; **25** (1):75–81.

Verhoeff, F. H., Brabin, B. J., Chimsuku, L., (1999) Malaria in pregnancy and its consequences for the infant in rural Malawi. Ann Trop Med Parasitol. 1999 Dec;93Suppl 1:S25-33.

Verhulst, P. F. (1847). Deuxiµeme memoire sur la loid'accroissement de la population. *Nouveaux Memoires de l'Academie Royale des Sciences, des Lettreset des Beaux-Arts de Belgique*20, 1-32.

Verhulst, P. F., (1838). Notice sur la loique la population suit dans son accroissement. *CorrespondanceMathematiqueet Physique, publieeparA. Quetelet*10, 113.

Verhulst, P. F., (1845). Recherchesmathematiquessur la loid'accroissement de la population.*Nouveaux Memoires de l'Academie Royale des Sciences, des Lettreset des Beaux-Arts de Belgique*18, 1-32.

Warner, B., Misra, M., (1996)."Understanding Neural Networks as Statistical Tools." The American Statistician.Vol. **50**. No.**4** (November 1996): 284 - 293.

WHO (2001). Iron Deficiency Anaemia Assessment prevention and Control: A guide for program managers. Geneva, Switzerland: WHO

WHO, (1993).Prevention and Management of Severe Anaemia in Pregnancy. WHO, WHO/FHE/MSM/ 93-5. Jonathan P. R., (2006).

WHO, (2008). Worldwide prevalence of anaemia 1993-2005: WHO Global Database on Anaemia

WHO, (2005) *Global database on anaemia*. Geneva: World Health Organization

WHO/INICEF/UNU.2001. Iron Deficiency Anaemia Assessment, Prevention, and Control. A guide for programme managers.

Widrow, B., (1990) "30 Years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation." Proceedings of the IEEE.Vol. 78. No. **9** (September 1990): 1415 1442.

Willacy, H., (2011). Management of Sickle Cell Disease in Pregnancy, Royal College of Obstetricians and Gynaecologists (August 2011)