JIMMA UNIVERSITY
COLLAGE OF NATURAL SCIENCE
DEPARTIMENT OF STATISTICS

MODELING DETERMINANTS OF ANEMIC STATUS FOR WOMEN AMONG REPRODUCTIVE AGE IN ETHIOPIA

BY

ZELALEM ESTIFANOS

A THESIS SUBMITTED TO THE DEPARTMENT OF STATISTICS, COLLEGE OF NATURAL SCIENCE, JIMMA UNIVERSITY AS A PARTIAL FULFILLMENT FOR THE REQUIREMENT FOR THE DEGREE OF MASTERS OF SCIENCE IN BIOSTATISTICS

October, 2017
Jimma, Ethiopia

MODELING DETERMINANTS OF ANEMIC STATUS FOR WOMEN AMONG REPRODUCTIVE AGE IN ETHIOPIA

MSc. THESIS

BY

ZELALEM ESTIFANOS

ADVISORS: AGATAMUDI LAKSHMANA (PHD)

AKALU BANBETA (PHD FELLOW-)

October, 2017

Jimma, Ethiopia

DEPARTMENT OF STATISTICS, SCHOOL OF GRADUATE STUDIES

JIMMA UNIVERSITY

As thesis research advisors, we her by certify that we have read and evaluated the thesis prepare d by **Zelalem Estifanos** under our guidance, which is entitled **Modeling Determinants of Anemic Status for Women Among Reproductive Age in Ethiopia.** We recommend that the thesis be submitted as it fulfills the requirements for the degree of Master of Science.

Agatamudi Lakshmana (PhD.)	____________________	________________
Main Advisor	Signature	Date
Akalu Banbeta (PhD. Fellow.)	____________________	________________
Co-Advisor	Signature	Date

As the members of the board of examiners of MSc. thesis open defense examination of **Zelalem Estifanos Chutulo**, we certify that we have read and evaluated the thesis and examined the candidate. We recommend that the thesis be accepted as it fulfills the requirements for the degree of Master of Science in Biostatistics.

________________________	____________________	________________
Name of Chair Man	Signature	Date
________________________	____________________	________________
Name of Main Advisor	Signature	Date
________________________	____________________	________________
Name of Co-Advisor	Signature	Date
________________________	____________________	________________
Name of Internal Examiner	Signature	Date
________________________	____________________	________________
Name of External Examiner	Signature	Date

STATEMENT OF AUTHOR

I declare that this thesis is a result of my genuine work and all sources of materials used, for writing it, have been appropriately acknowledged. I have submitted this thesis to Jimma University in the partial fulfillment for the requirements of Degree of Master of Science in Biostatistics. The thesis can be deposited in the university library to be made available to borrowers for reference. I impressively declare that I have not so far submitted this thesis to any other institution anywhere for that award of any academic degree, diploma or certificate.

Concise quotations from this thesis are allowed without requiring special permission if an accurate acknowledgement of the source is made. Requisites for extended quotations for the reproduction of the thesis in whole or in part may be granted by the head of the department of statistics when in his or her judgment the proposed use of the material is for a scholarly interest. In all other instances, however, permission must be obtained from the author.

Zelalem Estifanos

Date: _____________

Signature: _________

Jimma, Ethiopia

ACKNOWLEDGEMENTS

First and above all I would like to thank the almighty God for his boundless grace, wisdom, favor, faithfulness protection and for seeing me through all my life. I would like to acknowledge the support of all who contributed in various ways to make this study success. My deep gratitude goes to my advisor, Dr. Agatamudi Lakshmana for his valuable suggestions and comments to the successful realization of this study.

Secondly, I gratefully acknowledge my co-advisor Mr. Akalu Banbeta for his continuous supportive comments and constant guidance in the course of my thesis work. It is a great pleasure working with him. I can only hope that his cooperation will keep on going in the future.

Thirdly, my heartfelt thanks go to my parents for their support in the period of study and my thesis work. Finally, I am also very grateful thanks to all my friends whose presence and support have always been important to me.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGUERS

LIST OF ABBREVIATIONS

AIC	Akaike's Information Criterion
AIDS	Acquire Immunodeficiency Syndrome
BIC	Bayesian Information Criterion
BMI	Body Mass Index
CI	Confidence Interval
DHS	Demographic Health Survey
EDHS	Ethiopia Demographic Health Survey
ENA	Essential Nutrition Action
GEE	Generalized Estimating Equation
GLM	General Linear Model
GLMM	General Linear Mixed Model
Hb	Hemoglobin
HIV	Human Immunodeficiency Virus
IDA	Iron Deficiency Anemia
INACG	International Nutritional Anemia Consultative Group
LM	Linear Model
LMM	Linear Mixed Model
LRT	Likelihood Ratio Test
MI	Micronutrient Initiative
ML	Maximum Likelihood
OR	Odds Ratio
SNNP	South Nation Nationality and People
REML	Restricted Maximum Likelihood
UNFPA	United Nations Population Fund
WHO	World Health Organization

ABSTRACT

Background: Anemia is a problem characterized by insufficient red blood cell volume and a low concentration of hemoglobin in the blood (WHO, 2014). Anemia is a global public health problem that affects low, middle and high-income countries and has significant adverse health consequences, as well as adverse impacts on social and economic development (WHO, 2015). According to the 2011 EDHS, 17 percent of Ethiopian women age 15-49 are anemic. A higher proportion of anemic women were pregnant (22%) compared to others.

Objective: The objective of this study is to model anemic status with associated factors in women among reproductive age in Ethiopia. In addition, we assess the significant variations of anemic patients within and between regions and estimate the prevalence of anemia in women among reproductive age in Ethiopia.

Method: A cross-sectional but cluster study carried out based on the secondary data of the Ethiopia Demographic Health Survey 2011. For the categorized response variable, generalized linear model, generalized estimating equations, and generalized linear mixed models are compared to model anemic status of women among reproductive age in Ethiopia to identify the most candidate predictors. That shows us each of them estimates parameters from among different statistical models and comment on the interpretation of parameters and the statistical properties of the methods involved. Data is mainly analyzed using SAS 9.3 and R 3.41 version software offers for the analysis of binary responses for correlated data and both marginal and cluster-specific effects take into account.

Results: The study shows that 19.9% of women in the reproductive age group are anemic. The generalized estimating equation is best fits the data for population-averaged effects for given factors of anemic status in women among reproductive age than that of the two models. Generalized linear mixed model with two random intercepts revealed that there is variation between and within regions of anemic status. The result of best model revealed that the variables: Occupation women employed (odds ratio OR=0.6808, 95% CI: 0.6266, 0.7396), higher education level women attain: primary education level (OR=0.7014, 95% CI: 0.6336, 0.7764), secondary education level (OR=0.567, CI: 0.4642, 0.6962) and Higher education level (OR=0.7167, 95% CI: 0.5689, 0.9028), marital status: married (OR=1.5498, 95%CI: 1.6398, 1.7535), living with partner (OR=1.4521, 95%CI: 1.1513, 1.8316), widowed (OR=1.7521, 95%CI: 1.3982, 2.1957), divorced (OR=1.3524, 95%CI: 1.1025, 1.6589) and separated (OR=1.5484, 95%CI: 1.1599, 2.067), wealth index: rich (OR=0.8642, 95%CI: 0.7825, 0.9544), contraceptive use method: modern (OR=0.4966, 95%CI: 4389, 0.5697), pregnant women (OR=1.3338, 95CI: 1.1799, 1.5334) and BMI of women: with $18.5 \leq BMI < 25$ (OR=0.7773, 95%CI: 0.7101, 0.8508) and $BMI \geq 25$ (OR=0.6849, 95%CI: 0.5686, 0.8249) are significant determinant factors for anemic status at 5% level of significance.

Conclusion: Anemia in women among reproductive age is the final outcome of the collective effects of health, socio-demographic and economic factors. GEE model with measure of association exhibited the best fit for this data than GLM and GLMM models. The GLMM provided interesting relationships that would not be evident from a standard logistic model. This concluded that there is heterogeneity of anemic status between and within regions.

Keywords: *Anemic Status; Generalized Estimating Equation; Regional variation; Determinant factors*

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Anemia is a case developed by insufficient red blood cell volume and a low concentration of hemoglobin in the blood. Most of the time, anemia is the final outcome of a nutritional deficiency of iron, folate, vitamin B_{12}, and other nutrients. Although several causes of anemia have been identified (such as hemorrhage, infection, genetic disorders, and chronic disease), nutritional deficiency, primarily due to a lack of dietary iron, accounts for most cases (WHO, 2014).

Anemia is a global public health problem that affects low, middle and high-income countries and has significant adverse health consequences, as well as adverse impacts on social and economic development. It occurs at all stages of the life cycle, but is more prevalent in women of reproductive age and young children. In 2011, it is estimated that roughly 43% of children, 38% of pregnant women, 29% of non-pregnant women and 29% of all women of reproductive age have anemic globally. This corresponds to 273 million children, 496 million non-pregnant women and 32 million pregnant women are anemic (WHO, 2015).

According to WHO the estimated prevalence of anemia in developed and developing countries in pregnant women is 14 percent and 51 percent, respectively. Recent findings suggest that there is a decline in the prevalence of iron deficiency anemia among industrialized regions but the global prevalence is still higher in the developing world compared to the expected decline rate (Mumbare et al, 2012).

In developing countries, the cause of anemia in women among reproductive age is multi-factorial and includes nutritional deficiencies and also parasitic diseases, such as malaria and hookworm. In Sub-Saharan Africa, iron and folate deficiencies are the most common causes of anemia of women in reproductive age (DeMaeyer, E. M., et al,. 1989).

For pregnant women the prevalence is slightly higher than non-pregnant; however, its prevalence distribution varies greatly by geographical location, season, and dietary practice. The highest prevalence is in Africa (57.1%) and in South-East Asia (48.2%), followed by the Eastern Mediterranean (44.2%), Western Pacific (30.7%), and the European Americas regions, 25% and 24.1% respectively. Overall, 56.4 million pregnant women are anemic (41.8% prevalence globally). In addition to that, the prevalence of anemia in non-pregnant women is slightly lower than in pregnant women. Overall, 468.4 million non-pregnant women are anemic (30.2% prevalence globally). The highest prevalence is found in Africa (47.5%) and in South-East Asia

(35.7%). In the Eastern Mediterranean region, the prevalence is 32.4%, 20.5% in the Western Pacific region, 19% in the European region, and 17.8% in the Americas (WHO, 2017).

In demographic and health study conducted by line minister of Ethiopia and its stakeholders in 2005, found that 27% of women in age group 15-49 are infected by anemia and equal amount are malnourished with significant regional variations. There is a shortage of information to show the cause or iron deficiency anemia in the entire country, and some are misleading and non-conclusive, despite the problem being among the 10 top morbidities (Federal Ministry of Health, 2004).

According to the 2011 EDHS, 17 percent of Ethiopian women age 15-49 are anemic. A higher proportion of anemic women were pregnant (22%) than women who are breastfeeding (19%) and women who are neither pregnant nor breastfeeding (15%). The prevalence of anemia is not the same across the world countries due to different factors. Even if the prevalence of anemia in women decrease by time but the case still now a serious public health issues that porn women to different infectious disease and significantly leads to maternal mortality in developing counters like Ethiopia. Anemia mostly occurs in women of reproductive age due to menstrual, closed pregnancy and other socio-economic and demographic factors.

Identifying the magnitude of anemia and its determinants in high-risk groups, such as women of child bearing age, would be essential for evidence-based intervention modalities, particularly in developing countries, such as Ethiopia, where the social conditions pose serious challenges to women. The nutritional status of women in Ethiopia, as in other developing countries, is low, and their daily workload is often enormous because of reproducing and ensuring the survival of their children (Berhane et.al, 2001).

Anemia is known to have detrimental health implications, particularly for mothers and young children. Complicated pregnancy outcomes have been reported to be more common in anemic women than non-anemic women (INACG, 1989). Women with severe anemia can experience difficulty meeting oxygen transport requirements near and at delivery, especially if significant hemorrhage occurs. This may take us the main cause of maternal death and antenatal and prenatal infant mortality (Omar et al., 1994).

Anemia is the result of a wide variety of causes that can be isolated, but more often coexist. Globally, the most significant contributor to the onset of anemia is iron deficiency so that Iron Deficiency Anemia (IDA) and anemia are often used synonymously, and the prevalence of anemia has often been used as a proxy for IDA. It is generally assumed that 50% of the cases of anemia is due to iron deficiency (WHO, 2015), but the proportion may vary among population groups and in different areas according to the local conditions. The main risk factors for IDA include a low intake of iron, poor absorption of iron from diets high in phytate or phenolic compounds, and period of life when iron requirements are especially high (i.e. growth and pregnancy).

This study focus on social, economic, geographical and demographical determinants for the prevalence of anemia and rank the significant determinants based on the comparisons of Generalized Linear Model, Marginal model and Generalized Linear Mixed model with mainly using statistical software R and SAS. Therefore, the outcomes of the findings can help in evidence-based decision to develop and control intervention strategies to improve the health status of the women of reproductive age as well as familiarize appropriate statistical methods for clustered data set.

The main motivation to conduct this thesis is from just reading international and national medical reports, newspaper, and from examining the environment we live by them complicated by various health problems. From different medical reports and studies as described among leading top ten killer diseases, anemia is the one cause of maternal mortality. Furthermore, this case frequently occurred in women's among reproductive age mainly in pregnant women's and infant children. In addition to this, it leads to other infectious disease and consequence is a significant effect on the community as well as counters development. Indeed the importance of conducting this study is to model the anemic status with some candidate factors among women in the reproductive age in Ethiopia, to show the seriousness of the problem and to compare the efficiency of various categorical general linear models using clustered dataset. The appropriate well-describing models for this kind of data are General Linear Model (GLM), General Estimating Equation (GEE) and General Linear Mixed Model (GLMM). The reasons to take these models into considerations are because: the data structure is non-normal responses, clustered, hierarchal and assuming there is correlation within and between clusters.

1.2 Statement of the Problem

Anemia in women causes significant maternal morbidity and mortality in the developing countries including Ethiopia especially at time of pregnancy. The burden and underlying factors are varied even within countries in such a way that anemia in women among reproductive age is related to different socio-demographic, dietary and economic factors. According to WHO's estimate, the global prevalence of anemia in pregnant women is 29%. In Africa its prevalence is estimated to be 57.1%. In Ethiopia, anemia is the severe public health problem affecting 22% of pregnant mothers and 15% non-pregnant women. This study, therefore, try to identify determinant factors of the case of anemia among pregnant women in Ethiopia and considering clustered data from EDHS, 2011.

Hierarchically clustered data with binary responses are now widespread among socio-demographic studies and standard statistical analysis methods have become inadequate in the answering of socio-demographic hypotheses. Instead of such conventional approaches, statisticians have started proposing better techniques, such as the Generalized Estimating Equations (GEE) approach and Generalized Linear Mixed Models (GLMM) technique. This study would attempt to model the anemic status by choosing the best statistical model from GLM, GEE, and GLMM. This could make difference from other research works that analyzed by Chi-square, simple descriptive statistics, and logistic regression models. After classifying the health status of women in the reproductive age (15-49) into anemic and non-anemic based on the concentration level of hemoglobin in the individual's blood, this study attempts to answer the following questions:

- Which model is best fit for the data of anemia among women in reproductive age in Ethiopia?
- Which covariates are the most determinant factors for anemic women among reproductive age in Ethiopia?
- Is there a significant variation within and between regions suffering from anemic women among reproductive age in Ethiopia?

1.3 Objectives of the study

1.3.1 General objective of the study

The main objective of this study is modeling determinants of anemic status for women in the reproductive age with some candidate sets of explanatory variables in Ethiopia.

1.3.2 Specific Objectives

The specific objectives of the study which should be accomplished to achieve the general objective stated above are:

1. To find the most parsimonious model that can fit the anemic status of women among GLM, GEE and GLMM.
2. To identify the potential factors affecting anemic status among women of reproductive age in Ethiopia.
3. To test the significance of variations within and between the regions in suffering from anemia among women in reproductive age.

1.4 Significance of the study

The significance of this study is to create awareness about how the anemia case is serious in women of reproductive age and to identify the common risk factors associated with anemia using the most appropriate statistical model. It also uses to assess socio-economic and demographic differentials on anemic status among women of reproductive age in Ethiopia. This study helps to the stakeholders to reduce maternal and infant mortality rate due to the severity of anemia, and clarifying the main determinant factors that significantly affect women in reproductive age due to compression of statistical models using EDHS (2011) data. Generally, this research will expect to give an idea to those focuses on:

- To familiarize different statistical model for analyzing socio- economic and demographic factors for health staffs as well as related researchers.
- To provide local and regional planners useful information for policy making, monitoring and evaluating the activities for the government, non-government and different concerned bodies that could encourage optimal intervals.
- The result of this study can provide better opportunity for further study in future.

CHAPTER TWO

LITERATURE REVIEW

2.1 Causes and Consequence of Anemia

Causes of Anemia

The most common cause of anemia worldwide is iron deficiency, resulting from prolonged negative iron balance, caused by inadequate dietary iron intake or absorption, increased needs for iron during pregnancy or growth periods, and increased iron losses as a result of menstruation and helminths (intestinal worms) infestation. An estimated 50% of anemia in women worldwide is due to iron deficiency (Stevens G, 2013). Other important causes of anemia worldwide include infections, other nutritional deficiencies (especially folate and vitamins B12, A, and C) and genetic conditions including sickle cell disease, thalassemia (one of several hereditary abnormalities of synthesis of the globin chains of hemoglobin leading to severe anemia).

Consequence of anemia

Anemia has a significant effect on physical, social and economic on women. It affects the sense of well-being, resulting in fatigue, stress, and a decrease in word capacity. Most maternal mortality and morbidity occurs in developing countries is the consequence of anemia (Harrison & Rossiter, 1985; WHO, 1992). Anemia is attributed as a direct or indirect cause of about 26% of maternal deaths in Africa. Severe anemia may cause cardiac failure and death, whereas chronic anemia is considered to be contributory, especially in cases of hemorrhage and infection. Furthermore, anemic women are poor anesthetic and operative risks, as anemia may lead to poor healing of the wound and to increased susceptibility to infections (Brabin, 2001).

Anemia is a detrimental consequence in individuals as well as the country by affecting prominently women of reproductive age and children's. If a person is anemic their heart has to work harder to pump the quantity of blood needed to get adequate oxygen around the body (IPHN, 2007). It is not a disease in itself, but it is a result of a malfunction somewhere in the body. Therefore, anemia devastating effect on the health physically, mental productivity and affect the quality of life of human being. Anemia also plays an important role in the economic development of a nation (Brick and Peters, 2014).

2.2 Overview of Determinant of Anemia in Women among Reproductive Age

According to the result obtained from WHO, the prevalence of anemia is high in developing countries due to the socio-economic and health development. Africa and South East Asia countries are highly affected by anemia (WHO, 2008).

A cross-sectional study was conducted in urban areas in Malaysia to determine the Hb levels of antenatal mothers and their association with various socio-economic characteristics. A structured self-administered questionnaire was used to identify the factors influencing Hb level among the antenatal mothers and used frequencies and percent method to describe the study. This study found that 73 out of 217 mothers (33%) were anemic. The findings also suggested that Hb levels among the antenatal mothers were influenced by various factors such as education level, occupation, and family income. Anemia was generally more common among antenatal mothers who had a lower level of education and who were from a background of lower family income (Kim Lam Soh et al, 2015).

Anemia during pregnancy is a major cause of morbidity and mortality of pregnant women in developing countries and has both maternal and fetal consequences. Institution-based, cross-sectional study was conducted from February 16 to April 8, 2015, among 332 pregnant women who attended antenatal care at government health institutions of Arba Minch town. Bivariate and multivariate logistic regressions were used to identify predictors of anemia. Low average monthly income of the family, having birth interval less than two years, iron supplementation, and family size >2 were found to be independent predictors of anemia in pregnancy (Alemayehu Bekele, 2016).

Information about demographics, socioeconomic and anemia status on 5,666 married women ageing between 13 and 40 years were collected from a nationally representative cross-sectional survey Bangladesh Demographic and Health Survey (BDHS 2011). Data were analyzed using cross-tabulation, chi-square tests and multiple logistic regression methods. Logistic regression showed statistically significant association with anemia and type of residency, wealth status, educational attainment and household food insecurity. Women who reported food insecurity were about 1.6 times more likely to suffer from anemia compared to their food secure counterparts (Bishwajit Ghose et al., 2016).

A retrospective case-control study was conducted on 1221 women who delivered between 37 and 42 weeks of gestation between July 2014 and January 2015. Data on the subjects' socioeconomic and demographic characteristics, pregnancy outcomes, and hemoglobin levels within 24h prior to delivery were collected. The prevalence of pre-delivery anemia was estimated, and antenatal predictors of anemia were determined using multivariate logistic regression analysis. The prevalence of anemia in women attending our center for delivery was 41.6%. After multivariate logistic regression analysis, parity >3, illiterate and primary educational level, household monthly income per person < 250 Turkish liras, first admission at second and third trimester, number of antenatal visits <5 and 5–10, duration of iron supplementation < 3 months and 3–6 months, and occurrence of were independently associated with anemia (Cüneyt Eftal et al., 2015).

A quantitative cross-sectional study was carried out for correlates of anemia among women of reproductive age based on secondary data obtained from Ethiopian DHS 2005 and using binary logistic regression. This method was employed to control potential predictors and to explore associations between the dependent variable (anemic status) and a wide range of the aforementioned independent variables. Rural residence, poor educational and economic status, 30-39 years of age and high parity were key factors predisposing women to anemia. Utilizing maternity services, taking iron and vitamin A supplement during pregnancy and the postpartum period didn't have a significant effect on reducing the burden of anemia (Gebremedhin and Enquselassie, 2011).

Using EDHS 2005 data, the study by Wondu and Bijlsma (2012) has shown that women's educational status, grouped altitude of residential places and household wealth index categories have a significant impact on the prevalence of anemia. The prevalence of anemia was positively associated with past five year's fertility level. Unavailability of toilet facilities, being a resident of the rural area and not using contraceptive methods were also associated with the prevalence of anemia among women.

A cross-sectional community-based study was conducted by Haider in nine of the 11 regions of Ethiopia to assess the magnitude of anemia, deficiencies of iron and folic acid and compare the factors responsible for anemia among anemic and non-anemic women of childbearing age (15-49 years) (Haider, 2010). To identify the effect and predict the most important determinants of anemia, a stepwise logistic regression analysis was performed. Women having two or more children, using an open field as a toilet, suffered from chronic illnesses, and who had intestinal parasites were positively associated with anemia. Women with no formal education and who did not use contraceptives were negatively associated with anemia. The major determinants identified for anemia were chronic illnesses (AIDS, diabetes, cancer).

Based on the study conducted on prevalence and predictors of paternal anemia during pregnancy in Gondar, northwest Ethiopia a multiple logistic regression analysis, controlling the possible confounders, low monthly family income, large family size, hookworm infection and HIV sero-positivity were identified as significant predictors of anemia (Melku et al., 2014).

A study conducted in India to determine the prevalence of anemia among ever-married women of reproductive age and to explore some factors commonly associated with anemia. Background characteristics such as age, place of residence, nutritional status, number of children ever born, pregnancy status, educational achievement, and economic status were considered in the study. As a response variable, anemia level was taken as a dichotomous variable. The findings of the study revealed that the predictors such as pregnancy status, nutritional status, economic status, education level and the habit of cigarette smoking/pan/bidi/gutka were found to be statistically significant. About 49.6% of the women were anemic. Women in the age groups 20-24 years were at high risk of anemia. Women who were pregnant and undernourished were at high risk of being anemic; urban women and with high education level were at low risk of anemia. The habit of cigarette smoking/pan/bidi/gutka etc. also increased the risk of anemia (Sanku et. al., 2010).

A study at Bushulo health center, southern Ethiopia, showed that Age, residence, occupation, income family, religion, marital status have significance on the prevalence of anemia (Tadege, 2009).

A hospital-based study conducted in Tibet to study the levels of Hemoglobin and anemia on pregnant women in the highlands of Tibet. The result showed that gestational age, ethnicity, residence, and income were significantly associated with the hemoglobin concentration and prevalence of anemia. Especially, the hemoglobin concentration of pregnant women decreased with increase in gestational age (Yuan et.al, 2009).

Based on the data obtained from socio-demographic and socio-economic survey on the anemia and associated risk factors among pregnant women in gilgel gibe dam area, southwest Ethiopia; place of residence has significant factor for being anemic, rural women were highly affected due to lack of knowledge of anemia and due to smaller number health facilities in rural residences (Getachew et al., 2012).

A study was conducted in Jaipur city, India on the prevalence of anemia and socio-demographic factors associated with anemia among pregnant women attending antenatal hospital. The results showed that overall prevalence of anemia among pregnant women was found to be sixty-three percent. Factors such as level of education and socioeconomic status were significantly associated with the prevalence of anemia (Priyanka et.al, 2011).

A study based on the data obtained from primary healthcare clinic attendees was undertaken in Trinidad and Tobago to investigate the relationship of anemia with abortion, parity and child spacing. Logistic regression showed relationships between anemia with the variables parity, gravidity hence, previous spontaneous abortions were direct. Clinic attendees, age were not associated with the severity of anemia (Uche-Nwachi et al., 2010).

A study conducted on prevalence and risk factors of anemia among women of reproductive age in Bursa, Turkey revealed that more than 2 sanitary pads during menstruation and more than five days of menstrual bleeding was found to risk a factor for anemia. There was no significant association between anemia and age, education, marital status, job, parity, BMI, regularity of cycle and length of cycle (Kayihan and Nilgun, 2008).

A study in the Niger Delta, Nigeria showed that prevalence of anemia was 66.7% among pregnant women. None of the women had severe anemia. Age and occupation were significantly associated with the risk of anemia. Marital status, educational level, social class, and parity did not significantly affect the incidence of anemia in pregnancy (Ibrahim et.al, 2012).

A cross-sectional study was carried out on the prevalence and socio-demographic factors associated with anemia during pregnancy based on the data obtained from primary health center 13 in Rivers State, Nigeria. A Chi-square test showed that anemia was observed to be least prevalent among women at the extremes of reproductive age (≤20 years and 36–40 years). There was no statistically significant association between age, educational level, and marital status. The association of anemia with social class was statistically significant. Severe anemia was significantly associated with educational status and socio-economic status (Ndukwu and Dienye, 2012).

A cross-sectional study using nationally representative data over the 7-year period showed that anemia prevalence increased significantly from 51.3% to 56.1% among Indian women. This corresponded to a 1.11-fold increase in anemia prevalence after adjustment for age and parity, and 1.08-fold increase after adjustment for wealth, education, and caste. There was marked state variation in anemia prevalence in only 4 of the 25 states was a decline in anemia prevalence. Anemia was socially patterned, being positively associated with lower wealth status, lower education and belonging to scheduled tribes and scheduled castes. In this context socioeconomic inequalities in anemia by wealth, education and caste have narrowed significantly over time of the study (Balarajan et.al, 2013).

A cross-sectional secondary analysis of data pooled from two rounds of the 2005 and 2011 Ethiopian Demographic and Health Survey (EDHS) was statistically analyzed by multilevel logistic regression. The multivariate statistical model showed that having a husband who had attended primary education, working during the 12 months preceding the survey, having a normal maternal body mass index, being in the middle wealth quintile or rich wealth quintile, having ever used family planning, having attended antenatal care (ANC) for the indexed pregnancy four times or more, having experienced time variation between the two surveys, and breastfeeding for 2 years were factors associated with lower odds of having anemia in lactating mothers (Lakew Y., 2014).

The study carried out to determine risk factors for anemia in pregnancy among women at primary care level and document the contribution of HIV/AIDS to anemia in pregnancy in low-risk pregnant women at primary care level. A prospective study carried out among pregnant women attending the booking clinics of primary health care center in Ibadan, Nigeria. HIV positive and HIV negative mothers were followed throughout pregnancy till delivery of their babies. History of use of iron, folate, Vitamin B complex and daraprim were obtained. Haemoglobin, malaria parasitemia, and HIV serostatus were determined. Use of iron ($P < 0.006$), folate ($P = 0.032$), vitamin B complex ($P = 0.001$) and treatment for malaria ($P = 0.05$) significantly reduced the risk for anemia in pregnancy. Malaria parasitemia ($P = 0.0001$) significantly increased the risk of anemia. However, use of daraprim and HIV seropositivity increased the risk of anemia in pregnancy but not significantly (Dario MD 2005).

The study aimed to determine the prevalence of maternal anemia and the risk factors associated with it in the rural health district of Hounde in Burkina Faso. This cross-sectional study conducted in 2010 had a sample of 3,140 pregnant women attending antenatal care in all the 18 primary health care facilities of the district. Thus, women who regularly consumed alcoholic beverages were significantly more anemic than those who did not. The prevalence of anemia was significantly higher among geophagic women. Similarly, a body mass index <19 kg/m2 a maternal weight <50 kg and brachial perimeter <24 cm were significantly and strongly associated with anemia. Among gynecological and obstetric factors potentially associated with anemia, only the number of pregnancies was statistically significant. Thus, anemia was significantly more prevalent among primigravidae compared to paucigest or multigest patients. Parity and abortion did not appear to be associated with maternal anemia. No association was found to be statically significant between anemia and helminthiasis, STDs, HIV) for malaria (Meda et al. 2016).

2.3 Generalized Linear Models, Generalized Estimating Equation and Generalized Linear Mixed Model

Generalized linear models GLMs extend ordinary regression models to encompass non-normal response distributions and modeling functions of the mean. Three components specify a generalized linear model: A random component identifies the response variable Y and its probability distribution; a systematic component specifies explanatory variables used in a linear predictor function, and a link function specifies the function of $E(Y)$ that the model equates to the systematic component. Nelder and Wedderburn (1972) introduced the class of GLMs, although many models in the class were well established by then (Agresti, 2002).

When interest is in the first-order marginal parameters, McCullough and Nelder (1989) have shown that a full likelihood procedure can be replaced by quasi-likelihood based methods. Wedderburn (1974) shows the likelihood and quasi-likelihood theories coincide for exponential families and that the quasi-likelihood estimating equations provide consistent estimates of regression parameter. In any generalized linear model, even for choices of link and variance functions that do not correspond to exponential families. Consequently, Liang and Zeger (1986) proposed the method of generalized estimating equations (GEE) as an extension of GLM to accommodate correlated data using quasi-likelihood approach. Rather than assuming a particular distribution for the response, GEE method requires a correct specification of the mean as well as how the variance depends on the mean. One of the desirable properties of the GEE method is that it yields consistent and asymptotically normal solutions even with the misspecification of the covariance structure (Liang and Zeger, 1986).

Over the past 20 years, the GEE approach has proven to be an exceedingly useful method for the analysis of longitudinal or cluster data, especially when the response variable is discrete (e.g., binary, ordinal, or a count). Correlated data are modeled using the same link function and linear predictor setup (systematic component) as the independence case. The random component is described by the same variance functions as in the independence case, but the covariance structure of the correlated measurements must also be modeled. The focus is on estimating the average response over the population (population-averaged effects) rather than the regression parameters that would enable prediction of the effect of changing one or more components of X on a given individual.

Generalized Estimating Equation (GEE) is a general statistical approach to fit a marginal model for longitudinal/clustered data analysis, and it has been popularly applied in clinical trials and biomedical studies (J. W. Hardin and J. M. Hilbe, 2003). There are two types of approaches, mixed-effect models, and GEE, which are traditional and are widely used in practice now. Of note is that these two methods have different tendencies in model fitting depending on the study objectives. In particular, a mixed-effect model is an individual-level approach by adopting random effects to capture the correlation between the observations of the same subject (M. Crowder, 1995). On the other hand, GEE is a population-level approach based on a quasi-likelihood function and provides the population-averaged estimates of the parameters (R. W. Wedderburn, 1974). As is well known, GEE has several defining features.

1. The variance-covariance matrix of responses is treated as nuisance parameters in GEE and thus this model fitting turns out to be easier than mixed-effect models (P. McCullagh and J. A. Nelder, 1989). In particular, if the overall treatment effect is of primary interest, GEE is preferred.
2. Under mild regularity conditions, the parameter estimates are consistent and asymptotically normally distributed even when the "working" correlation structure of responses is miss-specified, and the variance-covariance matrix can be estimated by robust "sandwich" variance estimator (G. Fitzmaurice, 2008).
3. GEE relaxes the distribution assumption and only requires the correct specification of marginal mean and variance as well as the link function which connects the covariates of interest and marginal means (D. Hedeker and R.D. Gibbons, 2006).

Crowder addressed some issues on the inconsistent estimation of within-subject correlation coefficient under a miss-specified "working" correlation structure based on asymptotic theory (M. Crowder, 1995). In addition, the estimation of the correlation coefficients using the moment-based approach is not efficient; thus the correlation matrix may not be a positive definite matrix in certain cases. Also, Liang and Zeger did not incorporate the constraints on the range of correlation which was restricted by the marginal means because the estimation of the correlation coefficients was simply based on Pearson residuals.

Furthermore, for discrete random vectors, the correlation matrix was usually complicated, and it was not easy to attain multivariate distributions with specified correlation structures. These limitations lead researchers to actively work on this area to develop novel methodologies. Several alternative approaches for estimating the correlation coefficients have been proposed; for example, one method was based on "Gaussian" estimation (S. R. Lipsitzetl, 2000). The basic idea was to estimate the correlation coefficients based on multivariate normal estimating equations. The feature was that this estimation can ensure the estimated correlation matrix was positive-definite. Wang and Carey proposed to estimate the correlation coefficients by differentiating the Cholesky decomposition of the working correlation matrix (Y.G. Wang and V. J. Carey, 2004). In particular, for binary longitudinal data, the estimation of the correlation coefficients was proposed based on conditional residuals (Y. Lee and J. A. Nelder, 2009). Nevertheless, in this paper, the above issues are not discussed in great depth, and the assumption that, under the regular mild conditions, the consistency of parameter estimates as well as within-subject correlation coefficient estimate holds is satisfied.

Generalized linear mixed models (GLMMs) (Breslow and Clayton, 1993) are obtained from generalized linear models (GLMs) (Mc Cullagh and Nelder.1986) by incorporating random effects into the linear predictors, and include the well-known linear mixed models (LMMs) for normal responses (Laird and Ware,1982) as a special case. These models are useful for modeling the dependence among response variables inherent in longitudinal or repeated measures studies, for accommodating over-dispersion among binomial or Poisson responses, and for producing shrinkage estimators in multi-parameter problems. Due to the wide range of applications of GLMMs, these models have received substantial attention during the last decade and are available in the major software packages. The computational burden associated with high dimensional numerical integration has limited past studies of GLMMs to the case of simplified models (e.g., random intercept models), to tractable random effects distributions (e.g., the Gaussian and conjugate distributions such as the beta-binomial and negative binomial models), or to conditional inference for the regression coefficients, conditioning on the random effects (Zeger and Karim, 1991).

A variety of novel approaches have been proposed to overcome the computational difficulties, with the goal to improve inference and estimation procedures for the fixed effects in GLMMs. These include Gibbs sampling (Zeger and Karim, 1991), penalized quasi-likelihood and marginal quasi-likelihood (Breslow and Clayton, 1993), pseudo-likelihood based on approximate marginal models (Wolfinger and O'Connell, 1993), and maximum likelihood with Monte Carlo versions of EM, Newton-Raphson and simulated maximum likelihood algorithms (McCulloch, 1997), among many others (Jiang, 1998). These approaches typically require Gaussian distribution assumptions for the random effects.

One approach to account for the within-subject association is via the introduction of random effects in generalized linear models. This leads to a class of models known as generalized linear mixed models (GLMMs). GLMMs are an extension to GLMs that includes random effects in the linear predictor, giving an explicit probability model that explains the origin of the correlations. The resulting subject-specific parameter estimates are suitable when the focus is on estimating the effect of changing one or more components of the predictor on a given individual. In statistics, a generalized linear mixed model (GLMM) is a particular type of mixed model. Fitting such models by maximum likelihood involves integrating over these random effects.

CHAPTER THREE

DATA AND METHODOLOGY

3.1 Study Design

The design of 2011 Ethiopian Demographic Health Survey to provide estimates for the health and demographic variables of interest is based on the cross sectional study.

3.1.1 Source of Data

The source of data for this study is based on secondary data obtained from the 2011 Ethiopia demographic and health survey (EDHS, 2011) which was conducted by the Central Statistical Agency with the support of the Ministry of Health and other donor agencies. It was the third survey conducted in Ethiopia as part of the worldwide Demographic and Health Surveys project. The survey was primarily designed to collect data on marriage, fertility, family planning, maternal and child health, HIV/AIDS, malaria, anemia, nutrition and gender.

3.1.2 Study Area

The 2011 EDHS study would be conducted in national level including all the regions and city administrations in Ethiopia. Ethiopia has great geographical diversity; its topographic features range from the highest peak at Ras Dashen, 4,550 metres above sea level, down to the Affar Depression, 110 metres below sea level (CSA, 2009). The climate varies with the topography, from as high as 47 degrees Celsius in the Affar Depression to as low as 10 degrees Celsius in the highlands. Ethiopia's total surface area is about 1.1 million square kilometers. Djibouti, Eritrea, the Republic of the Sudan, the Republic of the Southern Sudan, Kenya, and Somalia border the country. In Ethiopia there are nine regions and two city administrations. At present Ethiopia is administratively structured into nine regional states: Tigray, Affar, Amhara, Oromiya, Somali, Benishangul-Gumuz, Southern Nations Nationalities and Peoples (SNNP), Gambela, and Harari and two city administrations: Addis Ababa and Dire Dawa Administration Councils.

3.1.3 Study Population

The study population for this study (anemic case) would be conducted on women among reproductive age 15-49 for nine regions and two city administrations in Ethiopia according to the survey of Ethiopia Demographic and Health survey 2011. **Target Population** for this study is Ethiopian women in reproductive age.

3.1.4 Sampling Procedures of EDHS 2011

The 2007 Population and Housing Census, conducted by the CSA, provided the sampling frame from which the 2011 EDHS sample was drawn. Administratively, regions in Ethiopia have been divided into zones, and zones, into administrative units called weredas. Each wereda was further subdivided into the lowest administrative unit, called Keble. During the 2007 Census, each Keble was subdivided into census enumeration areas (EAs) or clusters, which were convenient for the implementation of the census. The 2011 EDHS sample was selected using a stratified, two-stage cluster sampling design. Enumeration areas are the sampling units for the first stage. The sample included 624 clusters (EAs), 187 in urban areas and 437 in rural areas. Households comprised the second stage of sampling. In the second stage, a fixed number of 30 households were selected for each cluster. A complete listing of households was carried out in each of the selected clusters from September 2011 through January 2011.

3.1.5 Sampling Size Determination

Sample size is the number of observation included in the sample. All women aged 15-49 and all men aged 15-59 were eligible for interview. In the interviewed households 17,385 eligible women were identified for an individual interview; complete interviews were conducted for 16,515, yielding a response rate of 95%. There are cases in which information on the relevant variables was missing and these cases were excluded from the analysis. Thus, the analysis presented in this study on the risk factors of Anemia status is based on the 15,338 women aged 15-49 years.

3.1.6 Data Collection Tolls

Under 2011 EDHS used three questionnaires: the Household Questionnaire, the Woman's Questionnaire, and Man's Questionnaire to collect the available data from the target population. These questionnaires were adapted from model survey instruments developed for the measure DHS project to reflect the population and health issues relevant to Ethiopia. In addition to English, the questionnaires were translated into three major local languages-Amharigna, Afan Oromo, and Tigrigna.

For the anemia testing, blood drops were taken from each eligible woman, men, and child by pricking the finger (or heel in the case of very young children). The blood drop was tested using the Hemo Cue system (photometer and microcuvette), which assesses the level of hemoglobin in the blood. Results were provided immediately following the anemia testing both verbally and in writing for each of the individuals who were tested. The non-pregnant women were referred to the nearest health center for follow-up care if their hemoglobin level was below 7g/dl, and pregnant women and men were referred if their hemoglobin level was below 9 g/dl (EDHS, 2011).

3.1.7 Inclusion and Exclusion Criteria

In this study, it could access and further screened these summary data using our exclusion criteria. A woman was eligible if she was resident in Ethiopia and lies between 15 and 49 years of age (reproductive age group of women). When these data is obtained, a women's of Ethiopia not exist in Ethiopia is although excluded from the survey. Consistent with our inclusion and exclusion criteria, we excluded summarized data sources if:

- The women from abroad that means nationality of the women in household is not Ethiopian.
- The women blood hemoglobin concentration was not measured, and another measure such as serum ferritin, haematocrit, or previously diagnosed anemia was reported;
- It had access to the same data as individual-level records;
- They were not representative of the general population (e.g., were refugees), or non-random sampling methods were used, or sampling methods were not adequately described;
- They were representative of fewer than three areas within a country;
- Data for women of reproductive age were combined with data for children under 10 years of age without reporting summary statistics in smaller age bands; or
- The study did not have data on hemoglobin concentration or anemia prevalence in children aged 6-59 months or women aged 15-49 years.

3.2 Variables

3.2.1 Dependent variable

Anemia testing was included in the two rounds of the Ethiopian Demographic and Health Survey (EDHS). However, little information is available on the socio-demographic factors associated with anemia in reproductive women. As demonstrated in the literature review the prevalence of anemia was positively associated with maternal and infant death in Ethiopia. A study conducted in various countries to determine the prevalence of anemia among different classes' women of reproductive age and to explore some factors commonly associated with anemic status. This study aimed to identify factors associated with anemic status in women among reproductive age in Ethiopia using the 2011 EDHS data.

Often in many epidemiologic, biomedical and related fields of studies, the outcome of interest is a binary variable such as anemic versus non-anemic. In this condition, it is possible to employ plausible statistical methods for estimating the magnitude of the association between the response variable of interest as a function of independent predictor variables. The association provides information about the risk of developing an outcome. In practical, an advantage of using statistical methods for a binary response over statistical methods for continuous or another categorical response variable in epidemiologic research is that parameter estimates of the possible risk factors can be directly converted to an odds ratio, which is interpretable. Additionally, the use of binary outcome for defining anemic status and its severity at the population level, as well as the chronology of their founding allows the identification of populations at Sevier risk of anemia and priority areas for taking action, especially when resources are inadequate. In view of the above, the response or outcome variables of this study is anemic status in women of reproductive age, which gives a binary response of women, belongs to anemic and non-anemic categories in time of data collection. This indicates that the individual women in reproductive age and the region in which the women belongs.

Table 3.1: Description of the dependent variable

Response variable	Value of the levels	Category type
Anemic status	0 = non-anemic 1 = anemic	Binary

3.2.2 Independent variables

The explanatory variables that would be related to anemia among women of reproductive age based on various literature reviews and theoretical aspect as the determinant factors and covariates picked from the reviewed are described in table below:

Table 3.2: Description and Coding of explanatory or independent variables

Covariates	Description	Categories
Age	Age of women	1 = (15-19), 2 = (20-24), 3 = (25-29), 4 = (30-34), 5 = (35-39), 6 = (40-44), 7 = (45-49)
Region	Region of women live	1 = Tigray, 2 = Afar, 3 = Amhara, 4 = Oromiya, 5 = Somali, 6 = B.Gumz, 7 = SNNP, 12 = Gambela , 13 = Harar, 14 = Addis Ababa, 15 = Dire dawa
Residence	Residence women live	1 = Urban 2=Rural
Education level	Highest level Education of women	1= Non-educated, 2= Primary, 3 = Secondary, 4 = Higher
Occupation	Occupation women leveled	0 = Non-employed, 1= Employed
Marital status	Marital status of women	0 = Never in union, 1 = Married, 2 = Living with partner, 3 = Widowed, 4 = Divorced, 5= No longer living together/separated
Wealth index	Wealth index of women	1= Poor, 2= Middle and 3 =Rich
Parity	Number of childes for women ever born	1= No child, 2=1-2, 3=3-5, 4= 6 and above
Pregnancy status	Pregnancy status of women	0= No or unsure, 1= Yes
Contraceptive	Contraceptive use method of women	1= No method, 2= Folkloric and 3 = traditional methods,
Breast Feeding Status	Breast feeding status of women	0 = No, 1 = Yes
BMI	Body Mass Index of women	1= Less than 18.5, 2=18.5-24.9, 3= ≥25.0
Cigarette Smoking	Cigarette smoking habit	0=No, 1=Yes
HIV status	HIV Status of women	0=HIV-,1=HIV+

3.3 Statistical Models

3.3.1 Generalized Linear Model

The statistical model used for seak of comparision is using the generalized linar model. Generalized linear models (GLMs) extend ordinary regression models to encompass non normal response distributions and modeling functions of the mean (Agresti, 2007). In this section powerly discuss about three components of generalized linear model in subsection one. Generalized linear models for binary response data are explained in sub section two, goodness of fit of GLM model in subsection three and model diagnosis is deminostrate in final subsectio.

3.3.1.1 Components of Generalized Linear Models

The random component of a GLM consists of a response variable Y with independent observations (y_1, y_2, … y_N) from a distribution in the natural exponential family. This family has probability density function or massfunction of form:

$$\boldsymbol{f(y_i;\theta_i,\emptyset) = \exp\{[y_i\theta_i - b(\theta_i)]\,/a(\emptyset) + c(y_i,\emptyset)\}}. \tag{3.1}$$

This is called the exponential dispersion family and $\emptyset$ is called the dispersion parameter (Jorgensen 1987). The parameter θ_i is the natural parameter. When $\emptyset$ is known, (3.1) simplifies to the form (3.2) for the natural exponential family, which is

$$\boldsymbol{f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp[y_iQ(\theta_i)]}. \tag{3.2}$$

In case of this study, the response of dependeint variable is binary distributed with anemic statis of women categoriezed with anemic and non-anemic class. It identify $Q(\theta)$ here with $\theta/a(\emptyset)$ in (3.1), $a(\emptyset)$ with $\exp[-b(\theta)\,/a(\emptyset)]$ in (3.1), and $b(y)$ with $\exp[c(y_i,\emptyset)]$ in (3.1).

The systematic component of a GLM relates a vector $(\eta_1,\dots,\eta_N)$ to the explanatory variables through a linear model. Let x_{ij} denote the value of predictor j $(j = 1,2,\dots p)$ for women i (Agresti, 2007). Then

$$\eta_i = \sum_j \boldsymbol{\beta_j x_{ij}}, \qquad \boldsymbol{i = 1,\dots,N}. \tag{3.3}$$

This linear combination of explanatory variables is called the linear predictor. Usually, one $x_{ij} = 1$ for all women i, for the coefficient of an intercept (often denoted by α) in the model (McCullagh, P. 1989).

The third component of a GLM is a link function that connects the random and systematic components. Let $\mu_i = E(y_i), i = 1,\dots N$. The model links μ_i to η_i by $\eta_i = g(\mu_i)$, where the link function g is a monotonic, differentiable function (Faraway, 2006). Thus, g links E(Y) to explanatory variables through the formula

$$\boldsymbol{g(\mu_i) = \sum_j \beta_j x_{ij}, \qquad i = 1, \dots, N.} \tag{3.4}$$

The link function $g(\mu)$ called the identity link, has η_i to μ_i. It specifies a linear model for the mean itself. This is the link function for ordinary regression with normally distributed Y. The link function that transforms the mean to the natural parameter is called the canonical link. For it, $g(\mu_i) = Q(\theta_i), and\ Q(\theta_i) = \sum_j \beta_j x_{ij}$.

In summary, a GLM is a linear model for a transformed mean of a response variable that has distribution in the natural exponential family. We now illustrate the three components by introducing the key GLMs for discrete response variables.

3.3.1.2 Generalized Linear Models for Binary Data

Many categorical response variables have only two categories: for example, like in case of this study anemic status of women in reproductive age wheather women is anemic or non-anemic. Denote a binary response variable by Y and its two possible outcomes by 1 ("success") as anemic and 0 ("failure") as non-anemci. The distribution of Y is specified by probabilities $P(Y = 1) = \pi$ of women getting anemic and $P(Y = 0) = (1 - \pi)$ of women getting non-anemic. Its mean is $E(Y) = \pi$. For independent observations, the number of successes has the binomial distribution specified by the index n and parameter π (Agrist, 2007). This study discribes GLMs for binary responses.

3.3.1.2.1 Logistic Regression Model

Relationships between $\pi(x)$ (probability of women getting anemic with some covariate) and x (covariates) are usually nonlinear rather than linear. A fixed change in x (covariates) may have less impact when π is near 0 or 1 than when π (probability of women getting anemic with some covariate) is near the middle of its range. In practice, $\pi(x)$ often either increases continuously or decreases continuously as x increases. The S-shaped curves are often realistic shapes for the relationship. The most important mathematical function with this shape has formula using the exponential function:

$$\boldsymbol{\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} = \frac{e^{(\alpha+\beta x)}}{1+e^{(\alpha+\beta x)}}.} \tag{3.5}$$

This is called the logistic regression function, that the corresponding logistic regression model form is:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x. \quad (3.6)$$

The logistic regression model (3.7) is a special case of a GLM. The random component for the (success, failure) or anemc, non-anemci in case of this study outcomes has a binomial distribution. The link function is the logit function $log[\pi/(1-\pi)]$ of π, symbolized by "$logit(\pi)$." Logistic regression models are often called logit models. Whereas π (probabilty of getting anemic) is restricted to the $0 - 1$ range, the logit can be any real number. The real numbers are also the potential range for linear predictors (such as $\alpha + \beta x$) that form the systematic component of a GLM, so this model does not have the structural problem that the linear probability model has. The parameter β in equation (3.7) determines the rate of increase or decrease of the curve. When $\beta > 0$, $\pi(x)$ increases as x increases, when $\beta < 0$, $\pi(x)$ decreases as x increases. The magnitude of β determines how fast the curve increases or decreases. As $|\beta|$ increases, the curve has a steeper rate of change. When $\beta = 0$, the curve flattens to a horizontal straight line (Agrist, 2007).

3.3.1.2.2 Probit Regression Model

Another model that has the S-shaped curves is called the probit model. The link function for the model, called the probit link, transforms probabilities to z-scores from the standard normal distribution. The probit model has expression:

$$\boldsymbol{probit}[\boldsymbol{\pi}(\boldsymbol{x})] = \boldsymbol{\alpha} + \boldsymbol{\beta x}. \quad (3.7)$$

The probit link function applied to $\pi(x)$ probabilts of women getting anemic gives the standard normal z-score at which the left-tail probability equals $\pi(x)$. In practice, probit and logistic regression models provide similar fits. If a logistic regression model fits well, then so does the probit model, and conversely for binary distributed anemci status data (Agrist, 2002).

3.3.1.2.3 Complementary Log-Log Model

Complementary log-log model says: -

$$\boldsymbol{log}\{-\,\boldsymbol{log}[\mathbf{1} - \boldsymbol{\pi}(\boldsymbol{x})]\} = \boldsymbol{X}^{T}_{\boldsymbol{pxn}}\boldsymbol{\beta}_{\boldsymbol{px1}}. \quad (3.8)$$

The expression on the left-hand side is called Complementary Log-Log transformation. Like the logit and the probit transformation, the complementary log-log transformation takes a response restricted to the (0,1) interval and converts it into something in (−∞,+∞) interval. Here, we need mentioned that the log of 1-π(x) is always a negative number. This is changed to a positive number before taking the log a second time. We can also write the model down like form as $\pi(x) = 1 - exp[-exp(X^{T}_{pxn}\beta_{px1})]$.

Both logit and probit links have the same property, which is $link[\pi(x)] = -link[1 - \pi(x)]$. This means that the response curve for π(x) has a symmetric appearance about the point π(x) = 0.5 and so π(x) has the same rate for approaching 0 as well as for approaching 1. When the data given is not symmetric in the [0, 1] interval and it increase slowly at small to moderate value but increases sharply near one. The logit and probit models are inappropriate. However, in this situation, the complementary log-log model might give a satisfied answer. Unlike logit and probit the complementary log-log model is asymmetrical, it is frequently used when the probability of an event is very small or very large. Under the assumption that the general features are not lost for GLMs (Zeger, S. and Karim, M., 1991)

3.3.1.3 Goodness of Fit of GLM

Once the model is fitted, we would be interested to know how effective the model is in describing the outcome variable. This is referred to as goodness-of-fit.

3.3.1.3.1 Wald Test

A Wald test is used to test the statistical significance of each coefficient (β) in the model. If the Wald test is significant for a particular explanatory variable, then we would conclude that the parameter associated with this variable is not zero so that the variable should be included in the model otherwise it should be omitted from the model (Agresti, 1996).

The hypothesis to be tested is:

$$H_0: \beta_j = 0 \quad Vs \quad H_A: \beta_j \neq 0.$$

The Wald test statistic, Z, for this hypothesis is

$$\boldsymbol{Z^2 = \frac{\hat{\beta}_j^{\,2}}{var(\hat{\beta}_j)} \sim \chi^2(1).} \tag{3.9}$$

Where, $\hat{\beta}_j$ is the estimated regression coefficient and $var(\hat{\beta}_j)$ is the variance of $\hat{\beta}_j$.

3.3.1.3.2 The likelihood ratio test

The likelihood ratio test statistic (G^2) is the test statistic commonly used for assessing the overall fit of the GLM. The likelihood ratio test is computed based on -2Logliklehood. The likelihood ratio statistic is obtained by subtracting the two times log likelihood (-2LL) for the full model from the log likelihood for the intercept only model. This log likelihood-ratio test uses the ratio of the maximized value of the likelihood function for the intercept only model over the maximized value of the likelihood function for the full model. The likelihood test statistic is given by:

$$\mathbf{G^2} = -\mathbf{2}\,\mathbf{log}\left(\frac{\mathbf{L_0}}{\mathbf{L_1}}\right) = -\mathbf{2}[\mathbf{log}(\mathbf{L_0}) - \mathbf{log}(\mathbf{L_1})] = -\mathbf{2}[\mathbf{LL_0} - (-\boldsymbol{LL_1})]\,. \quad (3.10)$$

Where LL_0 the log likelihood value of the model which is have the intercept term only and LL_1 is the log likelihood value of the full model. The likelihood ratio statistic has a chi-square distribution and it tests the null hypothesis that all logistic regression coefficients except the constant are zero. The degrees of freedom are obtained by differencing the number of parameters in the both model. It is compared with chi-square value at the difference between degree of freedom of both models. If p-value is less than 5 % level of significance leads the rejection of the null hypothesis that all the predictor effects are zero. When this likelihood test is significant, at least one of the predictors is significantly related to the response variable (Hosmer and Lemeshow, 2000).

3.3.1.3.3 Hosmer-Lemeshow Test

The Hosmer-Lemeshow test is used to check the overall model fit. In this approach, data are divided into ten groups. From each group, the observed and expected numbers of events are computed. Then, the Hosmer-Lemeshow test statistic is given by:

$$\widehat{\boldsymbol{C}} = \frac{\sum_j^g (\boldsymbol{O_j} - \boldsymbol{E_j})^2}{\boldsymbol{V_j}}. \quad j = 1, 2, \dots 10 \quad (3.11)$$

Where, $E_j = np_j$, $V_j = np_j(1 - P_j)$, g is the number of groups,
O_j is observed number of events (women getting anemic) in the j^{th} group
E_j is expected number of events (women getting anemic) in the j^{th} group and
V_j is a variance correction factor for the j^{th} group.

If the observed number of events (women getting anemic) differs from what is expected by the model, the statistic $\hat{C}$ will be large and there will be evidence against the null hypothesis that the model is adequate to fit the data. This statistic has an approximate chi-square distribution with (g-2) degrees of freedom (Agresti, 1996).

3.3.1.4 Model Diagnosis

Regression model building is often an iterative and interactive process. The first model try to fit may prove to be inadequate. Regression diagnostics are used to detect problems with the model and suggest improvements. There are three ways that an observation can be considered as unusual, namely outlier, influence, and leverage. In logistic regression, observations whose values deviate from the expected range, produce extremely large residuals, and may indicate a sample peculiarity called **outliers**. These outliers can unduly influence the results of the analysis and lead to incorrect inferences. An observation said to be **influential** if removing the observation substantially changes the estimate of coefficients. Influence can be thought of as the product of leverage and outliers. An observation with an extreme value on a predictor variable is called a point with high leverage.

Leverage is a measure of how far an independent variable deviates from its mean. In fact, the leverage indicates the geometric extremeness of an observation in the multi-dimensional covariate space. These leverage points can have an unusually large effect on the estimate of logistic regression coefficients (Cook, 1977).

Multicollinearity: is a statistical phenomenon in which predictor variables in a logistic regression model are highly correlated. It is not uncommon when there are a large number of covariates in the model. Multicollinearity has been the thousand pounds monster in statistical modeling. Taming this monster has proven to be one of the great challenges of statistical modeling research. Multicollinearity can cause unstable estimates and inaccurate variances which affects confidence intervals and hypothesis tests. The existence of collinearity inflates the variances of the parameter estimates, and consequently incorrect inferences about relationships between explanatory and response variables. Examining the correlation matrix may be helpful to detect multicollinearity but not sufficient. Much better diagnostics are produced by logistic regression with the option tolerance, VIF, condition indices and variance proportions. For moderate to large sample sizes, the approach to drop one of the correlated variables was established entirely satisfactory to reduce multicollinearity. On the light of different collinearity diagnostics, we may safely conclude that without increasing sample size, the second choice to omit one of the correlated variables can reduce multicollinearity to a great extent.

To identify if an observation is outlier or influential, and multicolleinarity problem the following rules of thumbs were employed in this study.

Residuals: Standardized, Standard, deviance and Pearson residuals are obtained using different software. Observations with values larger than three in absolute values are considered as outliers (Agresti, 2007).

Leverage Values (Hat Diag): Measure of how far an observation is from the others in terms of the levels of the independent variables. Observations with values larger than **2p⁄n** are considered potentially highly influential.

Cook's D: Measures of aggregates impacts of each observation on the group of regression coefficients, as well as the group of fitted values. In logistic regression, a case is identified as influential if its Cook's distance is greater than 1.0 (Hosmer and Lemeshow, 2000).

Multicollinarty: A commonly given rule of thumb is that VIFs of 10 or higher (or equivalently, tolerances of .10 or less) may be reason for concern. This is, however, just a rule of thumb; Allison says he gets concerned when the VIF is over 2.5 and the tolerance is under .40. if simple correlation coefficient between two regressors is greater than 0.8 or 0.9, the multicollinearity is a serious problem.

Myers suggests a tolerance value below 0.1 indicates serious collinearityproblem and Menard suggests that a tolerance value less than 0.2 indicates a potential collinearity problem. As a rule of thumb, a tolerance of 0.1 or less is a cause for concern.

Like tolerance there is no formal cut off value to use with VIF for determining the presence of multicollinearity., In weaker models, which is often the case in logistic regression; values above 2.5 may be a cause for concern (Allison, 2001).

3.3.2 Marginal Model

As with independent observations, with clustered observations models focus on how the probability of a particular outcome (e.g., "success") depends on explanatory variables (Agresti, A., 2002). In marginal models, the main scientific objective is to analyze the population averaged effects of the given factors in the study on the binary response variable of interest. This means that the covariates are directly related to the marginal expectations.

3.3.2.1 The Generalized Estimating Equations (GEE) Approach

For binary data, recently, Balemi and Lee (1999) obtaine finite expansion bias and efficiency of the estimates from GEE approaches with miss-specified correlation matrices. The main findings are:

i. Bias and efficiency depend on the combination of a number of characteristics of the data cluster size, intra cluster correlation covariates, intra cluster correlation response variable, variability of cluster size and the relative response association and
ii. The performance of GEE is excellent for moderate degree of response correlation small clusters.

GEE is non-likelihood method that uses correlation to capture the association within the clusters or subjects in terms of marginal correlations (Molenberghs & Verbeke, 2005). For clustered as well as repeated data, (Liang & Zeger 1986) proposed GEE which require only the correct specification of the univariate marginal distributions provided one is willing to adopt "working" assumptions about the correlation structure. The "working" assumptions as proposed by Liang and Zeger included independence, unstructured, exchangeable and auto-regressive AR (1). Independence and exchangeable working assumptions can be used in virtually all applications, whether longitudinal, clustered, multivariate, or otherwise correlated. Auto regressive AR(1) and unstructured correlation structures are less relevant for clustered data, studies with unequally spaced measurements and/or sequences with differing lengths (Molenberghs and Verbeke, 2005).

3.3.2.1.1 Independence Structure

In GEE, the model assumes this correlation by default. With this structure the correlations between subsequent measurements are assumed to be zero or measurements are independent to each other within individuals in the given cluster regions (Molenberghs and Verbeke, 2005).

3.3.2.1.2 Autoregressive (AR1)

Box et al. (1994) described the family of correlation structure, which includes different classes of linear stationary models: autoregressive models, moving average models, and mixture of autoregressive-moving average models. Autoregressive models express the current observation as a linear function of previous observation plus a homoscedasticty noise term (Molenberghs and Verbeke, 2005).

3.3.2.1.3 Exchangeable correlation structure (compound symmetry)

It assumes the correlations between subsequent measurements are assumed to be the same, irrespective of the cluster data. Generally, assuming no missing data, the J x J covariance matrix y is modeled as:

$$\boldsymbol{V_i} = \Phi\, \boldsymbol{A_i}^{1/2} \boldsymbol{R_i} \boldsymbol{A_i}^{1/2}. \tag{3.12}$$

Where Φ is a GLM dispersion parameter which is assumed 1 for binary categorical data, A_i is a diagonal matrix of variance functions, and R_i is the working correlation matrix of Y. Generalized estimating equations (GEEs) can be used to model correlated data with the variance covariance matrix V by iteratively solving the quasi score equations. The score function of a GEE for has the form:

$$\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta_i}\right) \boldsymbol{V_i^{-1}}(\boldsymbol{Y_i} - \boldsymbol{\mu_i}) = \mathbf{0}. \tag{3.13}$$

Where μ_i is the fitted mean, which is given by $g(\mu_{it}) = X_{it}\beta$ for covariates $X = x_{i1}, x_{i2}, \dots, x_{im}$ and regression parameteers $\beta = \beta_1, \beta_2, \dots, \beta_p$ starting R_i as the identity matrix and $\Phi = 1$, the parameters β are estimated by solving equations as follows.

i.e. in normal case $\mu_i = X_i\beta \; and \; \frac{\partial \mu_i}{\partial \beta_i} = x_i, \;\; V_i = \widehat{\Phi} R_i$

$$\sum_{i=1}^{N} (\boldsymbol{x_i^t})\, \boldsymbol{R_i^{-1}}(\boldsymbol{Y_i} - \boldsymbol{\mu_i}) = \mathbf{0}. \tag{3.14}$$

More generally, because solution only depends on the mean and variance of y, these are quasi likelihood estimates. The estimates from a GEE analysis are robust to miss-specification of the covariance matrix (Liang & Zeger, 1986), so, the regression parameter estimates are consistent even for independent covariance matrix. Upon convergence, in order to perform hypothesis tests and construct confidence intervals, it is of interest to obtain standard errors associated with the estimated regression coefficients. These standard errors are obtained as the square root of the diagonal elements of the matrix β .

Two models are compared using generalized Wald test for GEE and likelihood ratio test for GLMM (Patetta ,2002).

Let $Y_j = (y_{j1}, \dots, y_{jn})'$ be the response values of observations from j[th] cluster (region), $j = 1, \dots, 11$ follows a binomial distribution i.e $Y_j \sim Bin(n_j, \pi_j)$ that belongs to the exponential family with the density function of the form. Then to model the relation between the response and covariates, one can use a regression model similar to the generalized linear models given by:

$$\boldsymbol{g(\pi_j) = logit(\pi_j) = X_j'\beta.} \qquad (3.15)$$

Where, $g(\pi_j) = logit\ link\ function,$

$X_j = (n_j \times p)$ dimensional vector of known covariates.

$\beta = (1 \times p)$ dimensional vector of unknown fixed regression parameter to be estimated

$E(Y_j) = \pi_j$ is expected value of the response variable.

3.3.2.2 Parameter Estimation for GEE

Here GEE is not likelihood approach, rather it is quasi-likelihood based and estimates $\hat{\beta}$ by solving estimating equations which consist of the working covariance matrix V_j. The score equation used to estimate the marginal regression parameters while accounting for the correlation structure is given by:

$$\boldsymbol{s(\beta) = \sum_{j=1}^{m} \frac{\partial \pi_j}{\partial \beta'} \left[A_j^{1/2} R_j A_j^{1/2} \right]^{-1} (Y_j - \pi_j) = 0.} \qquad (3.16)$$

Where R_j is working correlation matrix, and the covariance matrix of Y_j is decomposed in to $A_j^{1/2} R_j A_j^{1/2}$ with A_j the matrix with the marginal variance on the main diagonal and zeros elsewhere, Y_j multivariate vector of asymptotically normal response variables with mean vector π_j i.e $Y_j \sim N(X_j\beta, V_j)$. An advantage of the GEE approach is that it yields a consistent estimator of

$\hat{\beta}$, even when the working correlation matrix R_j is misspecified. However, severe misspecification of working correlation may seriously affect the efficiency of the GEE estimators (Zeger et al 1993).

3.3.2.3 Model Building and Variable Selection for GEE

Model selection is an important issue in almost any practical data analysis. A common problem is variable selection in regression given a large group of covariates (including some higher order terms) one needs to select a subset to be included in the regression model. Model selection is data analysis strategy, which leads to a search of best model. With this, we mean selecting the best subset of the covariates from the available covariates in the data.

To select candidate variables, firstly under the GEE, model building strategy started by fitting a model containing all possible covariates in the data. This is done by considering two working correlation assumptions (exchangeable and independence). In order to select the important factors related to the response variable, the backward selection procedure was used. The strategy is called backward because we were working backward from our largest starting model to a smaller final model. In this case, the procedure is used to remove covariates with non-significant p-values. This means that variables that did not contribute to the model based on the highest p-value would be eliminated sequentially and each time a new model with the remaining covariates were refitted, until we remained with covariates necessary for answering our research question.

3.3.2.4 Model Comparison Technique

Quasi Information Criterion (QIC): In a condition, when the likelihood function cannot be fully specified, such as in the GEE case, the Akaike's Information Criterion (AIC) cannot be directly applied to select either the optimal set of explanatory variables or correlation matrix. As an alternative, one can use the modified Akaike's Information Criterion called Quasi Information Criteria (QIC), which is based on the quasi-likelihood function (Pan, 2001). QIC is derived from the AIC and conceptually similar. The quasi-likelihood function takes the following form (McCullagh & Nelder, 1989).

$$Q(\pi) = \int_{y}^{\pi} \frac{y-t}{\phi_v(t)} dt$$

Where $\pi = E(y),\ v(y) = \phi_v(\pi)\ and\ \phi$ is the disspersion parameter. An equation for the QIC is $QIC = -2(\hat{\pi}, I) + 2trace[(\Omega_{\mathrm{I}}^{-1}\hat{V}_R]$ where I represient the independent correlation structure (diagonal matrix) and R is the speccified working correlation structure.

The p-dimensional matrices Ω_{I}^{-1} and $\hat{V}_R$ are variance estimators of the regression coefficients under the correlation structure I and R respectively. The QIC value would computed based on the quasi-likelihood estimate $\hat{\pi}$ and Ω_{I}^{-1} will be used to select the candidate explanatory variables. The model with the smallest QIC value for all correlation structures would considered as the best candidate model.

3.3.2.5 Model checking technique

Preisser and Qaqish (1996) further generalize regression diagnostics to apply to models for correlated data fitted by GEEs, where the influence of entire clusters of correlated observations is measured. The diagnostic measures proposed for marginal models were similar to those that exist for generalized linear models: DFBETAC, Cluster Cooks 'D, Cluster leverage and Cluster DFFIT. The diagnostic purpose of each measure is similar as well. DFBETAC is a measure of the influence that any cluster has on each $\hat{\beta}$ (Belsley et al., 1980); Cluster Cooks' D is a measure of the influence of any cluster on the overall fit of the model (Cook, 1977); Cluster leverage is a measure of how extreme cluster is with respect to the predictors (Belsley et al., 1980). Cluster DFFIT represents the studentized Cook distance type statistic to measure the influence of deleting cluster on the overall model fit. DFBETAC, Cluster Cooks'D and Cluster DFFIT are referred to as deletion diagnostics because the magnitude of each is related to changes in the fit of the model after a particular cluster is removed compared to the fit of the model on the full data. Let n_i be the number of responses for cluster i, and $N = \sum_{i=1}^{k} n_i$ the total number of observations. A_i is n_i x n_i diagonal matrix. Let B NxN diagonal matrix and let B_i the n_i x n_i diagonal matrix corresponding to cluster i. Let $Q_i = X_i(X'X)^{-1}X_i'$ where X_i is the n_i x p design matrix corresponding to cluster i. The adjusted residual vector is defined as $E = B(y - \hat{\pi})$ and $E_i = B_i(Y_i - \hat{\pi}_i)$ the estimated residual for the ith cluster (regions).

CLEVERAGE

The leverage of cluster i is contained in the matrix $H_i = Q_i$ and is summarized by the trace of H_i, where H_i is the hat matrix of cluster i.

$CLEVERAGE_i = tr(H_i)$

The leverage value greater than one for the ith cluster indicates that cluster is influential (Belsley et al., 1980).

DFBETAC

The effect of deleting cluster on the estimated parameter vector is given by the following one-step approximation for $\hat{\beta} - \hat{\beta}_{[i]}$:

$DFBETAC_i = (X'X)^{-1}X_i'(I - Q_i)^{-1}E_i$

If $DFBETAC_i$ is less than unity, this implies no specific impact of cluster on the coefficient of a particular predictor variable, while $DFBETAC_i$ of i^{th} cluster (region) greater than 1.0, implies the cluster is an outlier (Cook and Weisberg, 1982).

DFBETACS

DFBETACS is the standardized DFBETAC. The cluster deletion statistic DFBETAC can be standardized by dividing the components of DFBETAC by its standard error.

CLUSTERCOOKSD

Let $DCLS_i$ be the cluster-level Cook's D for cluster i, which can be calculated as $DCLS_i = E_i'(I-Q_i)^{-1}Q_i(I-Q_i)^{-1}E_i/p\,\hat{\phi}$ where p is the number of predictors in the model and $\hat{\phi}$ is dispersion parameter. The suggested cut off values for i^{th} cluster (region) to be influential is, if $DCLS_i$ is greater than "one" (Preisser and Qaqish ,1996).

CLUSTRDFFIT

Let $MCLS_i$ be the cluster-level DFFIT for cluster i which can be calculated as $MCLS_i = E_i'(I-Q_i)^{-1}H_iE_i/p\,\hat{\phi}$. The suggested cut off values for i^{th} cluster (region) to be influential is, if $MCLS_i$ is greater than "one" (Preisser and Qaqish, 1996).

3.3.3 Generalized Linear Mixed Model (GLMM)

The generalized linear mixed model, denoted by GLMM, is a further extension that permits random effects as well as fixed effects in the linear predictor in generalized linear model. Denote the random effect for cluster i by u_i for i^{th} women. We begin with the most common case, in which u_i is an intercept term in the model (Agresti, A., 2002).

An alternative way to fit a longitudinal (cluster) model to non-normal response data is to fit a generalized linear mixed model. These models are similar to the ones fit in GEE because the normality assumption regarding the error terms is relaxed. Some of the error distributions supported by generalized linear mixed models include the binomial, Poisson, gamma e.t.c. These models also support a large variety of link functions, which include the logit, log, and reciprocal. The type of response variable determines the distribution and link function for the model. Since the response variable for GLMM was categorical, binary data the logit link function was used to identify the associated factors of good control of anemia. However, unlike the models fit in GEE, generalized linear mixed models have the flexibility to specify random effects and also to generate subject-specific parameter estimates (Verbeke & Molenberghs, 2005). Let denote the response for i^{th} individual at j^{th} cluster (region), is categorical response variable with each follows a binomial distribution.

Let y_{it} denote observation (women) t in cluster (region) i, $t = 1, .., T_i$. As in the GEE analyses, the number of observations may vary by cluster. In a longitudinal/clustered study, even if clusters have equal size, many of them may have missing observations. Let x_{it} denote a column vector of values of explanatory variables, for fixed effect model parameters. Let u_i denote the vector of random effect values for region i. This is common to all observations in the cluster (regions). Let z_{it} denote a column vector of their explanatory variables. Often, it the random effect is univariate (Molenberghs and Verbeke, 2005).

Conditional on u_i, a GLMM resembles an ordinary GLM. Let $\mu_{it} = E(Y_{it}|u_i)$. The linear predictor for a GLMM has the form:

$$\boldsymbol{g(\mu_{it}) = X'_{it}\beta + Z'_{it}u_i.} \tag{3.17}$$

for link function $g(.)$. The random effect vector u_i is assumed to have a multivariate normal distribution $N(0, \sum)$. The covariance matrix $\sum$ depends on unknown variance components and possibly also correlation parameters (Belsley et al., 1980).

Denote $var\ (Y_{it}|u_i) = \emptyset_{it} v(\mu_{it})$, where the variance function $v(.)$ describes how the (conditional) variance depends on the mean. Often, $\emptyset_{it} = 1\ or\ \emptyset_{it} = \emptyset/\omega_{it}$, where ω_{it} is a known weight (e.g., number of trials for a binomial count) and $\emptyset$ is an unknown dispersion parameter. Conditional on u_i, the model treats $\{y_{it}\}$ as independent over i and t. The variability among u_i induces a non-negative association women i among the responses, for the marginal distribution averaged over the subjects. This is caused by the shared random effect u_i for each observation in acluster (region).

In equation (3.18), the random effect enters the model on the same scale as the predictor terms. This is convenient but also natural for many applications. For instance, random effects sometimes represent heterogeneity caused by omitting certain explanatory variables. Consider the special case with univariate random effefct and $z_{it} = 1$. With u_i replaced by $u_i * \sigma$ where $\{u_i\ *\}$ are N(0,1), the GLMM has the form:

$$\boldsymbol{g(\mu_{it}) = X'_{it}\beta + U^*_i\sigma.} \tag{3.18}$$

This has the form of an ordinary GLM with unobserved values $u_i *$ of a particular covariate. Thus, random effects models relate to methods of dealing with unmeasured predictors and other forms of missing data. The random effects part of the linear predictor reflects terms that would be in the fixed effects part if those explanatory variables had been included. Random effects also sometimes represent random measurement error in the explanatory variables. If we replace a particular predictor x_{it} by $x_{it} * +\in_i$, with $x_{it} *$ the true value and $\in_i$ the measurement error, then $\in_i$ times the regression parameter can be absorbed in the random effects term. Related to these motivations, random effects also provide a mechanism for explaining overdispersion in basic models not having those effects (Breslow and Clayton 1993).

3.3.3.1 Parameter Estimation for GLMM

It is the integral over the unobserved random effects of the joint distribution of the data and random effects. The problem in maximizing is the presence of m integrals over the q-dimensional random effects b_j. With Gaussian data, the integral has a closed form solution and relatively simple methods exist for maximizing the likelihood or restricted likelihood (Molenberghs and Verbeke, 2005). With non-linear models, numerical techniques are needed.

$$\boldsymbol{L(\beta, D, \phi) = \prod_{j=1}^{m} f_j\left(\frac{Y_j}{\beta}, D, \phi\right) = \prod_{j=1}^{m} \int \prod_{j=1}^{nj} f_{ij}\left(\frac{Y_{ij}}{b_j}, D, \phi\right) f_j\left(\frac{b_j}{D}\right) db_j.} \quad (3.19)$$

3.3.3.2 Model Building for GLMM

Under the GLMM for clustering, random effects are included the model to address the between region and within-region variations. These will be introduced in the generalized linear mixed model due to the fact that, the probability of women in reproductive age with the anemia are possibly varies for individuals within the same regions as well as individuals in different regions.

3.3.3.3 Model Comparison in GLMM

This study can used Likelihood ratio test and Information criteria to select the best model based on the values of asymptotic estimations. In order to decide on the best of the two random effects models, two models will be fitted, one with the two random intercepts (between and within clusters variations) and another one with one random intercept (within cluster variation). One can use the approximate restricted maximum likelihood ratio test (LRT) to compare these two models (Myers et al., 2010).

Let $LR_{full} = -2loglikelihood$ value for the full model and $LR_{redu} = -2loglikelihood$ value for reduced model. Then, the likelihood ratio test statistic, is given by

$$\lambda = LR_{full} - LR_{redu}.$$

The asymptotic null distribution of the likelihood ratio test statistic λ, is a chi-square distribution with degrees of freedom equal to the difference between the numbers of parameters in the two models.

Akaike's information criterion (AIC)

AIC is a measure of goodness of fit of an estimated statistical model. It is not a test on the model in the sense of hypothesis testing; rather it is a tool for model selection. The AIC penalizes the likelihood by the number of covariance parameters in the model, therefore

$$AIC = -2\log(L) + 2p.$$

Where, L is the maximized value likelihood function for the estimated model and p is the number of parameters in the model. The model with the lowest AIC value is preferable (McCullagh, 1989) .

3.3.3.4 Model Checking Technique

In GLMM, it is assumed that the random effects are normally distributed and uncorrelated with the error term. Normality of the random effects is assessed using normal plot of each random effect. Normal Q-Q plot of estimated random effects is an important method for checking the marginal normality and identify outlier (Myers et al., 2010).

3.4 Ethical Consideration

Ethical clearance which is used to be provided previously by the Ethiopian Health and Nutrition Research Institute (EHNRI) Review Board, the National Research Ethics Review Committee (NRERC) at the Ministry of Science and Technology, the Institutional Review Board of ICF International, and the Centers for Disease Control and Prevention (CDC) currently conferred to Jimma University. Accordingly, the Research Ethics Review Board of Jimma University has provided an ethical clearance for the study. The data for analysis was brought from EDHS, and to do so the department of statistics asked to write an official co-operation letter to the Central Statistical Agency of Ethiopia from where data would obtained. In this research, individuals have not forced to give information at all levels without his/her informed consent. Moreover, the confidentiality of information secured from respondents is kept and information obtained from respondents will not be given to third party.

CHAPTER FOUR

DATA ANALYSIS AND RESULTS

4.1 Summary of Descriptive statistics

Before any statistical analysis, it is better to examine the overall pattern of the data. Table 4.1 below presents basic descriptive information that summarizes the associations between the determinant factors and anemic status of women in reproductive age. The total of 15388 women in reproductive age from nine regional states and two city administrations in Ethiopia are eligible for this study. Among these eligible mother, 3063 (19.9%) women in reproductive age are anemic where as 12325 (80.1%) are non anemic.

As it presented in (table 4.1) below, the basic descriptive statistics presents the information that summarizes the associations between the determinant factors and anemic status of women among reproductive age. The percentage of anemic status of women in reproductive age is relatively larger which is (22.7%) and (22.1%) for age groups (40-44) and (30-34) respectively as compared to the young (15-19), youth (20-24), (25-29), (35-39) and relatively older age group (45-49). Whereas the percentage is lower in young age group (15-19) which is 16.4% exposed to the anemia.

The anemic status of women also varied in the regions that is the percentage of anemic women in reproductive age in each region is different to some extent. For instance, the high percentage of anemic status for women is recorded in Somali, Affar, and Dire Dawa with 43.6%, 36.3%, and 31% respectively and the low percentage of anemic status recorded in Addis Ababa city administration 10.2%. Similarly, the anemic status of women is varied with a place of residence, as it can present in the table below, the high proportion of anemic women among reproductive age is a remarkable variation of anemic status due to a place of residence women live. Here, the high percentage anemic status of women among reproductive age in rural is (22.4%) compared to that of (14.4%) in urban.

There is variation between anemic status among women in different Parity or number of a child ever born. The highest percentage of women, who were ever born child greater than or equal to six and the lowest percentage of women never born child is 24.6% and 15.4% respectively. The anemic status of women also varied with occupation, Wealth index and highest education level, that is the percentage of anemic women who employed are (16.4%) and (28.6%) for none employed. The percentage of the anemic status of women with poor wealth index is (21.8%), (20.6%) for middle and (15.8%) for richer wealth index. Educational status of women in reproductive age is erratic proportion to anemic status from higher to lower level. The percentage of anemic status is (25.1%) for non-educated women, (15.6%) for primary educated women, (11.4%) for secondary educated and (12.6%) for higher educated women.

Similarly anemic status of women in reproductive age also varied with the marital status of women, the highest proportion of anemic status is in widowed marital status (25.0%) compared to that of the others and the smaller percentage is women never in a union or single (13.9%). The anemic status of women in reproductive age also varied with current pregnancy, Brest feeding status, and smoking status, that is the percentage of anemic women of pregnant is (28.8%) and (19.1%) for non-pregnant, the percentage of anemic status for women of Brest feeding is (22.8%) and (18.7%) for non-breastfeeding. The anemic status for women according to a habit of smoking cigarettes showed that (19.7%) for smoking women and (19.9%) of nonsmoking women.

Anemic status is higher in BMI <18.5 compared to that of BMI between 18.5 to 25 and greater than 25. Also anemic status is slightly similar in both HIV test result among women in reproductive age.

Table 4.1: Descriptive statistics for anemic status in women among reproductive age

Variables	Levels	Non Anemic(%)	Anemic (%)	Total
Age group	15-19	2996(83.6%)	589(16.4%)	3585
	20-24	2295(81.4%)	524(18.6%)	2819
	25-29	2303(77.9%)	654(22.1%)	2957
	30-34	1523(78.5%)	416(21.5%)	1939
	35-39	1429(78.3%)	395(21.7%)	1824
	40-44	956(77.7%)	274(22.3%)	1230
	45-49	823(79.6%)	211(20.4%)	1034
	Total	12325(80.1%)	3063(19.9%)	15388
Regions	Tigray	1455(87.5%)	207(12.5%)	1662
	Affar	798(63.7%)	454(36.3%)	1252
	Amhara	1619(82.4%)	34617.6%)	1965
	Oromiya	1663(80.8%)	394(19.2%)	2057
	Somali	453(56.4%)	350(43.6%)	803
	Benishangul-G	968(80.9%)	229(19.1%)	1197
	SNNP	1702(88.6%)	219(11.4%)	1921
	Gambela	855(80.0%)	214(20.0%)	1069
	Harari	779(80.4%)	190(19.6%)	969
	Addis Ababa	1350(89.8%)	153(10.2%)	1503
	Dire Dawa	683(69.0%)	307(31.0%)	990
	Total	12325(80.1%)	3063(19.9%)	15388

Residence	Urban	4047(85.6%)	680(14.4%)	4727
	Rural	8278(77.6%)	2383(22.4%)	10661
	Total	12325(80.1)	3063(19.9%)	15388
Pairty/number of child ever born	0	4352(84.6%)	795(15.4%)	5147
	1-2	2851(80.3%)	699(19.7%)	3550
	3-5	2800(77.5%)	813(22.5%)	3613
	>=6	2322(75.4%)	756(24.6%)	3078
	Total	12325(80.1)	3063(19.9%)	15388
Occupation	non employed	5713(76.4%)	1767(23.6%)	7480
	Employed	6612(83.6%)	1296(16.4%)	7908
	Total	12325(80.1)	3063(19.9%)	15388
Highest education	No education	5834(74.9%)	1956(25.1%)	7790
	Primary	4638(84.4%)	857(15.6%)	5495
	Secondary	1113(88.6%)	143(11.4%)	1256
	Higher	740(87.4%)	107(12.6%)	847
	Total	12325(80.1)	3063(19.9%)	15388
marital status	Never in union	3477(86.1%)	560(13.9%)	4037
	Married	6907(77.3%)	2028(22.7%)	8935
	Living with partner	551(82.6%)	116(17.4%)	667
	Widowed	405(75.0%)	135(25.0%)	540
	Divorced	690(81.4%)	158(18.6%)	848
	NLLT/separated	295(81.7%)	669(18.3%)	361
	Total	12325(80.1)	3063(19.9%)	15388
Wealth index	Poor	4414(75.2%)	1454(24.8%)	5868
	Middle	1710(79.4%)	443(20.6%)	2153
	Riche	6201(84.2%)	1166(15.8%)	7367
	Total	12325(80.1%)	3063(19.9%)	15388
Currently pregnant	No or unsure	11457(80.9%)	2712(19.1%)	14169
	Yes	868(71.2%)	351(28.8%)	1219
	Total	12325(80.1)	3063(19.9%)	15388
Breast feeding	No	8841(81.3%)	2035(18.7%)	10876
	Yes	3484(77.2%)	1028(22.8%)	4512
	Total	12325(80.1)	3063(19.9%)	15388
Contraceptive Use method	No method	9899(78.3%)	2743(21.7%)	12642
	Folkloric method	5(71.4%)	2(28.6%)	7
	Traditional method	108(83.1%)	22(16.9%)	130
	Modern method	2313(88.7%)	296(11.3%)	2609
	Total	12325(80.1%)	30603(19.9%)	15388

Smoking	No	12276(80.1%)	3051(19.9%)	15327
	Yes	49(80.3%)	12(19.7%)	61
	Total	12325(80.1)	3063(19.9%)	15388
Body Mass Index	BMI < 18.5	3267(76.4%)	1007(23.6%)	4274
	18.5≤BMI<25	8032(81.0%)	1879(19.0%)	9911
	BMI ≥ 25	1026(85.3%)	177(14.7%)	1203
	Total	12325(80.1%)	3063(19.9%)	15388
Blood test result	HIV –ve	12079(80.1%)	2999(19.9%)	15078
	HIV +ve	246(79.4%)	64(20.6%)	310
	Total	12325(80.1%)	3063(19.9%)	15388

Source: 2011, EDHS descriptive statistics for each independent variables with anemic categories.

The overlook for distribution of anemic status in regions numerically described above expressively shown in the bar graph below to visualize with an overall pattern of anemic status. The figure tells us the severity of anemia is higher in Affar, Somali and Dire Dawa regions with the comparison of the count of none anemic to the anemic proportion in the individuals.

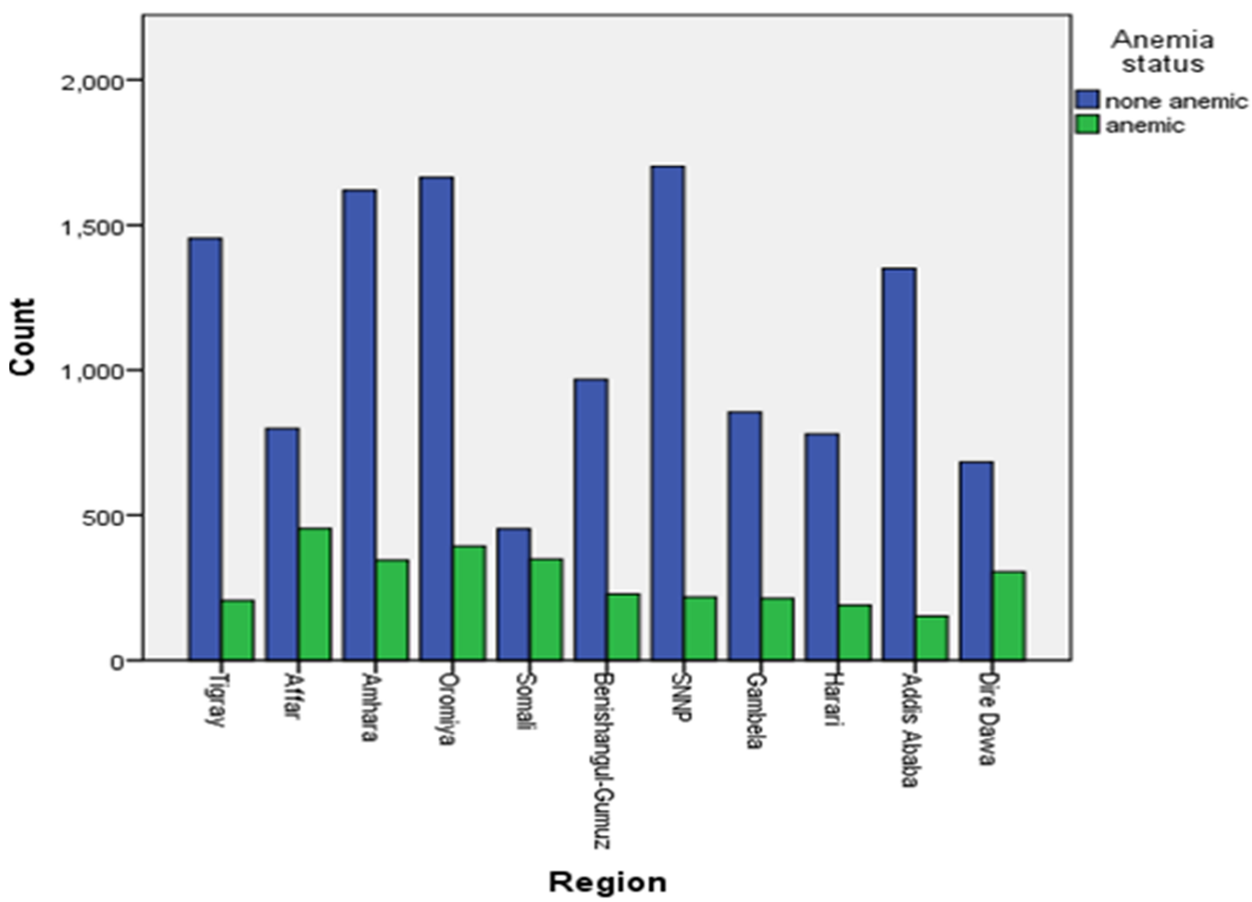


Fig 4.1: Bar graph of anemic status in eleven administrative regions from 2011 EDHS.

4.2 Statistical Analysis

4.2.1 Generalized Linear Model (GLM)

Generalized linear models (GLMs) extend ordinary regression models to encompass non normal response distributions and modeling functions of the mean. In this section we beable to see the fit of generalized linear model with appropreate link function for the seak of comparison to the generalized estimation equation approach and generalized linear mixed models in section one. In section two assesse the goodness of fit test for the GLM using testing significant effect of each individual predictors on response, likelihood ratio test and hosemer lemshew test of the model. Lastely in section three diagonsie the model to examine the adequency of our model fit well.

4.2.1.1 Analysis of Data using Generalized Linear Model (GLM)

The appropreat link function for binary response data selected by comparison of the three link functions logit, probit, and cloglog with their magnitude, direction, and simplicty to interpreation logit is selected as precious to fit for this study. In addition to that logit is popular to use. There are two important reasons that make logistic regression popular: 1) The range of the logistic function is between 0 and 1; that makes it suitable for use as a probability model, representing individual risk. 2) The logistic curve has an increasing S-shape with a threshold; that makes it suitable for use as a biological model, representing risk due to exposure. To get best model for GLM, stepwise elimnation method used to select variables for the model.

The difference of generalized linear model from generalized estimating equation and generalized mixed model is that the correlation structure and random effect does not take in to account in GLM. However, the regression coefficient interprtation is the same as to GEE and GLMM, since both of them are interpreated in terms of their odds ratio.

The table (4.2) below describes the parameter estimates and their corresponding standard errors beside the p-values for GLM model. Each parameter β_j reflects the effect of factor X_j on the log odds of the probability of women getting anemic, statistically controlling all the other covariates in the model. A negative sign in column labeled Estimate indicates an inverse relationship of explanatory variable with the log odds of the dependent variable. In contrast a positive coefficient column labeled Estimate indicates a positive relationship to the log odds of the dependent variable. Therefore, we can use odds ratio $\exp(\beta_j)$ of predictors estimate to interpreat the effect on response.

Table 4.2: Parameter estimates and standard error among GLM for the best model

Effects	Estimates	S. error	z value	Pr(>|z|)	OR	95% CI of OR	
Intercept	-1.0259	0.0735	-13.954	< 2e-16	0.3585	0.3102	1.1846
Age							
15 – 19	®						
20 – 24	0.0204	0.0761	0.268	0.7890	1.0206	0.8789	1.1846
25 – 29	0.0837	0.0793	1.057	0.2906	1.0874	0.9310	1.2704
30 – 34	-0.0285	0.0879	-0.324	0.7458	0.9719	0.8178	1.1547
35 – 39	-0.0149	0.0897	-0.167	0.8676	0.9852	0.8263	1.1744
40 – 44	-0.0572	0.0988	-0.579	0.5624	0.9444	0.7776	1.1456
45 – 49	-0.2499	0.1067	-2.343	0.0191	0.7788	0.6312	0.9591
Occupation							
Not employed	®						
Employed	-0.3792	0.0425	-8.913	< 2e-16	0.6844	0.6295	0.7438
Education							
Not educated	®						
Primary	-0.3736	0.0535	-6.975	3.06e-12	0.6882	0.6194	0.7642
Secondary	-0.5898	0.1025	-5.755	8.68e-09	0.5544	0.4522	0.6759
Higher	-0.3635	0.1192	-3.051	0.00228	0.6952	0.5482	0.8749
Marital Status							
Never in union	®						
Married	0.4521	0.0757	5.972	2.35e-09	1.5716	1.3552	1.8236
Living partner	0.3834	0.1254	3.058	0.00223	1.4672	1.1442	1.8710
Windowed	0.6457	0.12699	5.085	3.68e-07	1.9074	1.4843	2.4423
Divorced	0.3166	0.11046	2.866	0.00415	1.3725	1.1031	1.7012
Separated	0.4429	0.15250	2.904	0.00368	1.5572	1.1476	2.0881
Wealth Index							
Poor	®						
Middle	-0.09419	0.06270	-1.502	0.13301	0.9101	0.8043	1.0284
Riche	-0.14318	0.05089	-2.814	0.00489	0.8666	0.7843	0.9574
Contraceptive							
No method use	®						
Folkiloric	0.24610	0.84638	0.291	0.77123	1.2790	0.1809	6.0749
Traditional	-0.01982	0.24057	-0.082	0.93435	0.9804	0.5972	1.5403
Modern	-0.72137	0.07040	-10.246	< 2e-16	0.4861	0.4228	0.5573
Pregnancy							
No / unsure	®						
Yes	0.25583	0.07200	3.553	0.00038	1.2915	1.1206	1.4862
BMI							
BMI < 15	®						
15 ≤ BMI < 25	-0.25824	0.04619	-5.591	2.26e-08	0.7724	0.7057	0.8458
BMI ≥ 25	-0.36438	0.09539	-3.820	0.000133	0.6946	0.5749	0.8357
Criterion	**AIC: 14752**			**BIC: 14935.81**		**Deviance: 14704**	

Source: 2011, EDHS data; final model for GLM; NB: BMI = body mass index, OR=odds ratio, CI=Confidence Interval, ® =stands for reference categories.

The odds ratio indicates the effect of each explanatory variable directly on the odds of being women gettine anemic rather than on the odds of non-anemic. Estimates of odds greater than 1.0 indicate that the risk of women getting anemic is greater than that for the reference category. Estimates less than 1.0 indicate that the risk of death of wonen getting anemic is less than that for the reference category of each variable. Therefore, the final model presented in Table 4.2 is interpreted in terms of odds ratio as follows.

Women whose age group within 45-49 were 77.88% (OR = 0.7788) less likely to be getting anemic compared to women who age group within 15-19 controlling for other variables in the model. Similarly, Women form wealth index rich were 86.66% (OR = 0.8666) less than likely to getting anemic compared to women from poor wealth index family controlling for other variables in the model.

The odds of occupation of women who was from employed categories were 68.44% (OR = 0.6844) times less than the odds of anemic status of women from non-employed family. Similarly women from primary educate family were 68.82% (OR =0.6882) less likely getting anemic, women from secondary educate categories were 55.44% (OR =0.5544) less likely getting anemic and women from higher educate categories were 69.52% (OR= 0.6952) less likely getting anemic compared to women from non-educated categories controlling for other variables in the model.

The odds of marital status of women who is marred were 57.16% (OR = 1.5716) times higher than the odds of anemic status of women who were single and similarly women who were marital status living with partner is 46.72% (OR =1.4672) more likely getting anemic, women from widowed marital status were 90.74% (OR =1.9074) more likely getting anemic, women from divorced marital status were 37.25% (OR= 1.3725) more likely getting anemic, and women from separated marital status were 55.72% (OR =1.5572) more likely getting anemic compared to women from single or never in union marital status categories controlling for other variables in the model.

Women who were use modern contraceptive method is 48.61% (OR = 0.4861) less likely getting anemic compared to women not use any contraceptive use method, controlling for other variables in the model.

Women who were pregnant is 29.15% (OR = 1.2915) more likely getting anemic compared to women who were non-pregnant or unsure, controlling for other variables in the model.

Women who BMI is $15 \leq \text{BMI} < 25$ were 77.24% (OR = 0.7724) less likely getting anemic compared to women BMI less than 15, controlling for other variables in the model. Similarly, women who $\text{BMI} \geq 25$ were 69.46% (OR = 0.6946) less likely getting anemic compared to women who were BMI is less than 15, controlling for other variables in the model.

4.2.1.2 Assessing Goodness of Fit for GLM

4.2.1.2.1 Test of significance of each predictor individually

The table below is testing the significance association of each predictor individually related to the response variable.

Table 4.3: Testing the significance of each predictor individually

Covariates	DF	Deviance	AIC	Chi-Square	Pr>Chi-Sq
Age	6	14714	14750	14.047	0.029116
Occupation	1	14783	14829	82.619	<2.2e-16
Education	3	14766	14808	65.373	4.175e -14
Marital status	5	14742	14780	41.681	6.835e -08
Wealth index	2	14709	14753	8.461	0.014548
Contraceptive Use	3	14818	14860	117.105	< 2.2e -16
Pregnancy Status	1	14713	14759	12.358	0.000439
BMI	2	14735	14779	34.795	2.783e -08

As seen, from the above table since p-value is less than 5% level the association of each individual predictor to response shows that there is significant relation between predictors and response on anemic status of women in reproductive age.

4.2.1.2.2 Likelihood ratio test of overall GLM model

Table 4.4: Result of Model fit likelihood ratio test for Full and Reduced model

Model	AIC	BIC	-2 logLik	Chi-Square	Pr>ChiSq
Intercept model	15362	15369.46	-7679.909	-	-
Intercept with covariates	14748	14931.74	-7350.174	659.47	2.2e-16

The above table of output for likelihood ratio test for goodness of fit the model describes that there is evidence against null hypothesis which is no difference between model without predictors and model with predictors. This implies that the fit is adequate and at least one of the predictors is significantly related to the response variable.

4.2.1.2.3 Hosmer-Lemeshow Goodness of Fit Test.

The Hosmer-Lemeshow goodness-of-fit test compares the observed and expected frequencies of events and non-events to assess how well the model fits the data. Use the goodness of fit tests to determine whether the predicted probabilities deviate from the observed probabilities in a way that the binomial distribution does not predict. If the p-value for the goodness of fit test is lower than your chosen significance level, the predicted probabilities deviate from the observed probabilities in a way that the binomial distribution does not predict.

Table 4.5: Hosmer-Lemeshow Test for GLM

Chi-square	DF	Pr>ChiSq
11.52	8	0.1739

The Hosmer and Lemeshow Goodness-of-Fit Test, tests the hypothesis:

H_0: The model is a good fit, **vs.**
H_A: The model is not a good fit

The value of the Hosmer-Lemeshow goodness-of-fit statistic is X^2=11.52 with the corresponding P-value of 0.1739 (Table 4.5). A large p-value (i.e. p-value >0.05) suggests that the fitted model is an adequate model. Since the p-value = 0.1739 is greater than 0.05, we do not reject the null hypothesis that the model is a good fit. Therefore, GLM with logit link function for anemic status of women in reproductive age is fitted the data very well.

4.2.1.3 Model Diagnostics

Model diagnostics are used to detect problems with the model and suggest improvements. A failure to detect outliers and influential cases can have severe distortion on the validity of the inferences. It would be reasonable to use diagnostics to check if the model is adequate or not. The adequacy of the fitted model was checked for possible presence and treatment of outliers and influential values. The diagnostic test results for detection of outliers and influential values are presented using figure below. The residuals like Studentized, Pearson and standardized residuals are also use in order to check model diagnostic. The Cook's distance less than unity showed each observation had no impact on the group of regression coefficients. A value of the leverage statistic show's that no observation is far apart from the others in terms of the levels of the independent variables. Thus, from the above goodness of fit tests and diagnostic checking with graphs below, we can say that our model is adequate.

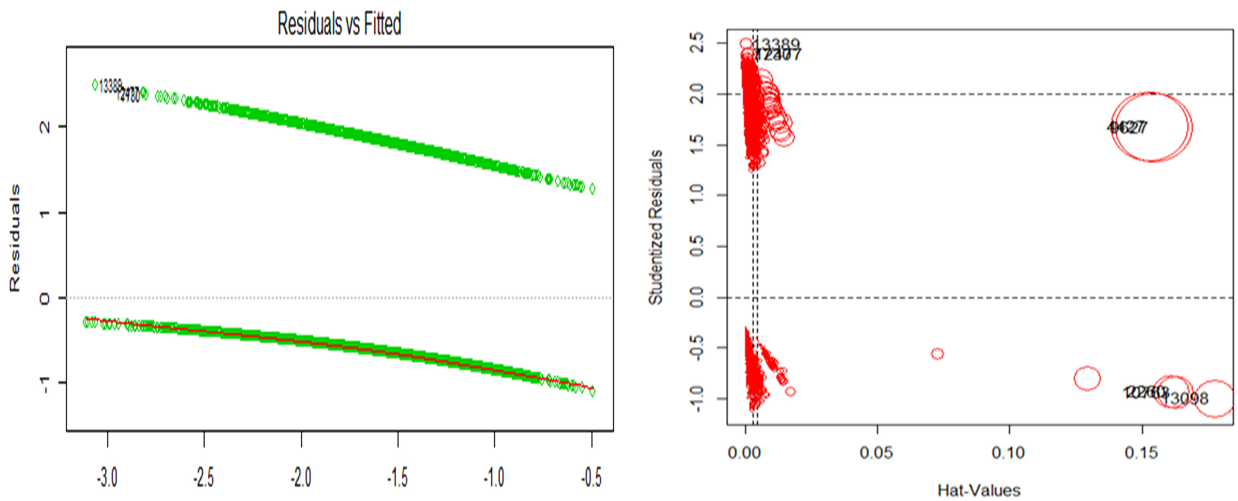


Fig 4.2: plot of residual versus fitted values and Studentized residual versus Hat-values.

In the above figure the first panel shows that residual versus fitted value plot for final GLM. According to rule of thumb observations between absolute value of three, it does not show any systematic pattern to point out the model fits the data well. In second panel we examine that influence point plots are useful because they display the studentized residuals, hat values, and Cook's distances all on the same plot. According to figure 4.2 there are the eight extreme observations of the fitted model; these are observation number 2260 and 10763 from Amhara, 4427 and 12477 from Affar, 9267 from Oromya, 9627 from Tigray, 13098 from SNNP and 13389 from Benshengul-Gumuz region. The studnetized residual is less than three and the hat-values less than one, indicates that the eight influential observations not influence the adequacy of the model.

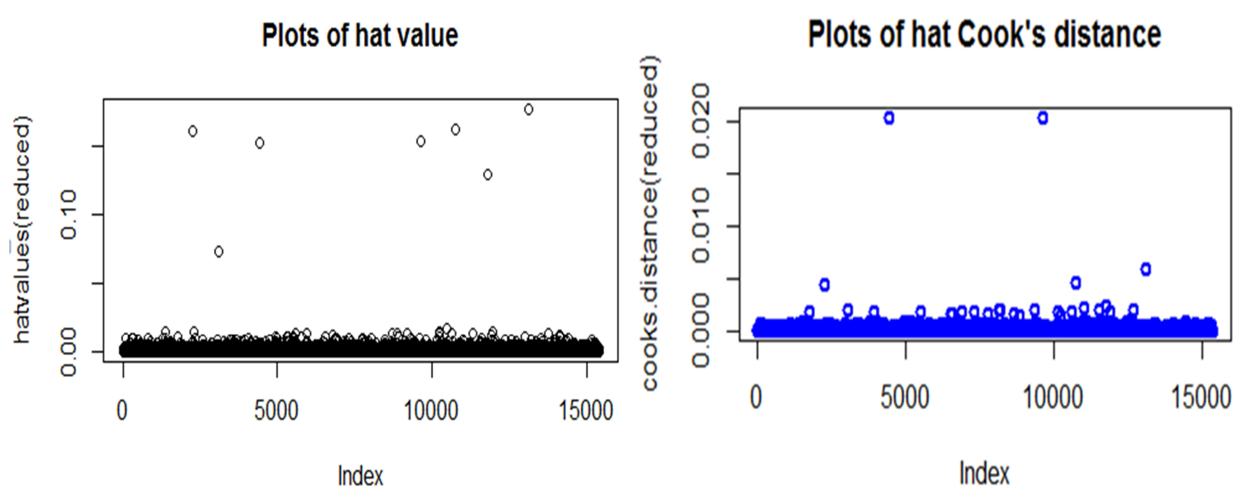


Fig 4.3: plot of index hat values and cook's distance.

In the above figure the first panel elaborate the leverage of an observation measures its ability to move the regression model all by itself simply moving in the response direction. The leverage always takes values between 0 and 1. A point with zero leverage has no effect on the regression model. If a point has leverage equal to 1 the line must follow the point perfectly. The second panel in the above figure shows that two observations are large cook's distance from others. But, all not point within 0.25 inches. Furthermore, from the figure in appendix the residual versus each categorical predictor recommended that there is a uniformity of residuals across each level of covariates specifies that homogeneity of error variances.

Multicollinearity: is a statistical phenomenon in which predictor variables in a logistic regression model are highly correlated. It is not uncommon when there are a large number of covariates in the model. The following table tells us variance inflation factor, toleranc and R-square:

Table 4.6: Multicollinearity checking for covariates in GLM

Covariates	VIF	Tolerance	R-square
Age	1.244370	0.8036196	0.19638038
Residence	1.034623	0.9665359	0.03346406
Marital Status	1.350878	0.7402591	0.25974092
Higher Education level	1.145897	0.8726785	0.12732155
Wealth Index	1.237656	0.8079790	0.19202097
Contraceptive Use	1.045948	0.9560702	0.04392983
Pregnancy	1.051300	0.9512033	0.04879665
BMI	1.083079	0.9232940	0.07670604

Since, according to rule of thumb for VIF less than 2.5, tolerance near 1 and residual square getting to 0, we conclude that there is no multicollinearity problem among candidate covariates and the model fits adequate.

4.2.2 Analysis of data using Generalized Estimating Equation (GEE)

The most important part of GEE is instead of attempting to model the within-subject covariance structure, to treat it as a nuisance and simply model the mean response. In this section, we observe building the model by selection of candidate predictor variables, selection of appropriate working correlation structures, and precious standard error in subsection one. Parameter interpretation for GEE with selected covariates, correlation structures, and precious standard error present in subsection two. Subsection three tells about the model diagnosis for detect the problems in the model and suggest improvements.

4.2.2.1 Model Building for Generalized Estimating Equation (GEE)

With this analysis, GEE has considered two correlation structures such as independent and exchangeable correlation structures and compared with their QIC values. Before selecting the correct correlation structure, consider the model building strategy (variable selection). Under the GEE, model building strategy is started by fitting a model containing all possible covariates in the data. This was done by considering two different working correlation assumptions (exchangeable and independence). In order to select the important factors related to anemic status, the backward elimination procedure was used. The full model for the probability of getting anemic status of i^{th} women from j^{th} cluster (region) is fitted as:

$$logit(\pi_{ij}) = \beta_0 + \sum_{i=1}^{7} \beta_i Age_{group.i} + \sum_{j=1}^{2} \beta_j Recidence_j + \sum_{k=1}^{4} \beta_k Pairty_k + \sum_{l=0}^{1} \beta_l Occupation_l + \sum_{m=0}^{3} \beta_m Education_m + \sum_{n=0}^{5} \beta_n Marital.S_n + \sum_{p=1}^{3} \beta_p Wealth.i_p + \sum_{q=0}^{3} \beta_q Contraceptive_q + \sum_{r=0}^{1} \beta_r Preginancy_r + \sum_{s=0}^{1} \beta_s Breast.f_s + \sum_{x=0}^{1} \beta_x Smoking_x + \sum_{y=1}^{3} \beta_y BMI_y + \sum_{z=0}^{1} \beta_z HIV_z .$$

After fitting the full model, covariates with the largest p-value are removed and the model was refitted with the rest of the covariates sequentially. Then, Age of the women, residence of women, parity, breast feeding status of women, smoking status of women and HIV test result are the covariates excluded from the model with p-value for the given covariates are large (P-value > 0.05) is found in the appendix A table A1. Independent and exchangeable correlation structures were considered and compared to select best correlation structure depending on the QIC value.

Table 4.7: Comparison of independent and exchangeable correlations using QIC values

Correlation structure	QIC value	QICu value
Independent	14860.7095	14754.1186
Exchangeable	14987.7843	14847.2775

As it can be seen from (table 4.7), the QIC value of the model with independent is less than that of exchangeable correlation structure. Then now let's compare the empirical and model based standard error of independent correlation structure to fit the data set in smaller standard error estimation. However, the best correlation with smaller standard error has been selected by comparing the standard error for both of the model based and empirical standard error of fitting the model.

Table (4.8) below shows that, the standard error of the Model-based Standard Error Estimates is relatively less as compared to Empirical-based Standard Error Estimate. Thus the independent correlation structure with model based standard error gives us best fit of the model. Therefore, the parameter estimates and their corresponding model based corrected standard errors with the p-values is parsimonious and given in (table 4.9).

Table 4.8: Empirical and model based standard errors for two proposed working correlation.

Covariates	Categories	Independent		
		Estimate	Empirical (SE)	Model based (SE)
Intercept		-1.0338	0.2315	0.0706
Occupation	Rural	-0.3845	0.1083	0.0423
Education Level	Primary	-0.3547	0.0820	0.0519
	Secondary	-0.5673	0.1440	0.1021
	Higher	-0.3331	0.1637	0.1178
Marital Status	Married	0.4381	0.0648	0.0630
	Living with partner	0.3730	0.1016	0.1184
	Windowed	0.5608	0.1859	0.1151
	Divorced	0.3019	0.1074	0.1042
	Separated	0.4372	0.1277	0.1474
Contraceptive Use Method	Folkloric method	-0.0984	0.0791	0.0626
	Traditional method	-0.1459	0.0714	0.0507
	Modern method	0.2081	0.7925	0.8469
Wealth Index	Middle	-0.0104	0.3022	0.2403
	Riche	-0.7000	0.1116	0.0701
Pregnancy	Yes pregnant	0.2880	0.0986	0.0712
BMI	$18.5 \leq$ BMI < 25	-0.2519	0.0748	0.0461
	BMI ≥ 25	-0.3785	0.1287	0.0950

Source: 2011, EDHS data; Standard error comparison for the two empirical and model based; NB: BMI = body mass index, SE=Standard Error.

Finally, as a customary, comparison of empirical and model based standard errors for the parameter estimates obtained based on the given working correlation is performed using selected predictors. The correlation structure that the model based and empirical standard errors are closest to each other is referred to be the best assumption correlation structure. Therefore, the final proposed generalized estimating equation model for anemic status of women among reproductive age in Ethiopia is given as:

$$\begin{aligned} \boldsymbol{logit}(\boldsymbol{\pi_{ij}}) = & -\mathbf{1.0338} - \mathbf{0.3845}\boldsymbol{Occupation_e} - \mathbf{0.3547}\boldsymbol{Education_p} \\ & - \mathbf{0.5673}\boldsymbol{Education_s} - \mathbf{0.3331}\boldsymbol{Education_h} + \mathbf{0.4381}\boldsymbol{Marital.S_m} \\ & + \mathbf{0.3730}\boldsymbol{Marital.S_{lwp}} + \mathbf{0.5608}\boldsymbol{Marital.S_w} + \mathbf{0.3019}\boldsymbol{Marital.S_d} \\ & + \mathbf{0.437}\boldsymbol{Marital.S_s} - \mathbf{0.1459}\boldsymbol{Contraceptive_t} - \mathbf{0.7}\mathbf{Welath\ index_r} \\ & + \mathbf{0.2880}\boldsymbol{Preginancy_y} - \mathbf{0.2519}\boldsymbol{BMI_{(18.5-25)}} - \mathbf{0.3785}\boldsymbol{BMI_{(\geq 25)}}. \end{aligned}$$

Parameter estimates and their corresponding empirically corrected standard errors alongside the p-values for the final GEE model is presented in table below.

Table 4.9: Parameter estimates of Model based standard errors for GEE model.

Parameter	Categories	Estimate	Standard Error	95% Confidence Limits		Pr > |Z|
Intercept		-1.0338	0.0706	-1.1722	-0.8955	<.0001
Occupation	Urban ®					
	Rural	-0.3845	0.0423	-0.4674	-0.3017	<.0001
Education Level	None educated ®					
	Primary	-0.3547	0.0519	-0.4564	-0.2531	<.0001
	Secondary	-0.5673	0.1021	-0.7674	-0.3673	<.0001
	Higher	-0.3331	0.1178	-0.5639	-0.1023	0.0047
Marital Status	Never in union ®					
	Married	0.4381	0.0630	0.3147	0.5616	<.0001
	Living with partner	0.3730	0.1184	0.1409	0.6052	0.0016
	Windowed	0.5608	0.1151	0.3352	0.7865	<.0001
	Divorced	0.3019	0.1042	0.0976	0.5062	0.0038
	Separated	0.4372	0.1474	0.1483	0.7261	0.0030
Contraceptive Use Method	No method ®					
	Folkloric method	-0.0984	0.0626	-0.2212	0.0243	0.1161
	Traditional method	-0.1459	0.0507	-0.2453	-0.0466	0.0040
	Modern method	0.2081	0.8469	-1.4518	1.8679	0.8059
Wealth Index	Poor ®					
	Middle	-0.0104	0.2403	-0.4813	0.4606	0.9656
	Riche	-0.7000	0.0701	-0.8373	-0.5627	<.0001
Pregnancy	No ®					
	Yes pregnant	0.2880	0.0712	0.1484	0.4275	<.0001
BMI	BMI < 18.5 ®					
	18.5 ≤ BMI < 25	-0.2519	0.0461	-0.3423	-0.1616	<.0001
	BMI ≥ 25	-0.3785	0.0950	-0.5646	-0.1924	<.0001

Source: 2011, EDHS data; final model for GEE approach; NB: BMI = body mass index, ® =stands for reference categories.

4.2.2.2 Parameter Interpretation of GEE model

The interpretation of the parameters in the marginal (population averaged) effects model is analogous to the standard logistic regression model, however there are differences (as noted above) in how we adjust for the correlations. Therefore the sentence would be the typical sentence describing strength, direction, and p-value/confidence limit of the association.

The parameter estimates for GEE stand for the effect of the predictors averaged across all individuals with the same predictor values. Like standard normal logistic regression, the interpretation of the parameters in the marginal (population average) model would be interpreted in terms of odd ratio. Each parameter β_j reflects the effect of factor X_j on the log odds of the probability of women in reproductive age being anemic, statistically controlling all the other covariates in the model. Then, the odds ratio of variables were calculated as the exponent of β_j i.e. odds ratio = exp (β_j).

Therefore, the GEE fit from table (4.9) shows that, Occupation status of women in reproductive age is statistically significant on anemic status with 5% significance level. Thus, the odds ratio of getting anemic for women whose occupation status is employed is estimated to be exp (β_1) = exp (-0.3845) = 0.6808 (95% CI OR: 0.6266, 0.7396) times decrease to have an outcome of anemic compared to those not employed. The women whose occupation is employed would be 68.08% less likely anemic than women whose occupation is not employed after adjusting for correlated outcome data and controlling for other covariates.

Likewise, education is one of the factors that related to anemic status of women, which means that the average effect of women among reproductive age who have highest educated level is primary has exp(β2) = exp(-0.3674) = 0.7014 (95% CI OR: 0.6336, 0.7764) times lower than those women of reproductive age have no education level, which means that the probability that the reproductive women who have primary educated and getting anemic are 70.14% less likely than those who has not educated after adjusting for correlated outcome data and controlling for other covariates. For secondary level educated women similarly the average effect is exp(β3) = exp(-0.5673) = 0.567 (95% CI OR: 0.4642, 0.6962) times lower than those reproductive women who have no education level, which means that the probability that the women in reproductive age have secondary educated and getting anemic is 56.7% less likely than those who are not educated after adjusting for correlated outcome data and controlling for other covariates. For higher educated level women the average effect is exp(β4) = exp(-0.3331) = 0.7167 (95% CI OR: 0.5689 , 0.9028) times lower than those women getting anemic who have no education level, which means that the probability that the reproductive women who educated higher and being anemic is 71.67% less likely than those who are not educated after adjusting for correlated outcome data and controlling for other covariates.

Marital status also has significantly associated with the anemic status of women in reproductive age. The odds ratio of women in reproductive age getting anemic who were married is exp (β5) =exp (0.4381) = 1.5498 (95% CI OR: 1.6398, 1.7535) times higher than those women getting anemic who never in a union. This implies that the probability of women in reproductive age who is married and getting anemic is 54.98% times more likely than women who have never in a union after adjusting for correlated outcome data and controlling for other covariates. The odds ratio of women in reproductive age getting anemic who was living with a partner is exp (β6) =exp (0.3730) = 1.4521 (95% CI OR: 1.1513, 1.8316) times higher than women among reproductive age getting anemic who have never in a union. This means that the probability of reproductive women who are living with a partner and getting anemic is 45.21% times more likely than those women who did never in a union after adjusting for correlated outcome data and controlling for other covariates. The average effect of women among reproductive age getting anemic who were widowed is exp (β7) = exp (0.5608) = 1.7521 (95% CI OR: 1.3982, 2.1957) times higher than those reproductive women getting anemic who never in a union. The probability of reproductive women who are windowed and getting anemic is 75.21% times more likely than those women who did never in a union after adjusting for correlated outcome data and controlling for other covariates. The odds ratio of reproductive women getting anemic who were divorced is exp (β8) = exp (0.3019) = 1.3524 (95% CI OR: 1.1025, 1.6589) times higher than those reproductive women getting anemic who never in a union. That means that the women marital status is windowed and getting anemic is 35.24% times more likely than that of women who did never in a union after adjusting for correlated outcome data and controlling for other covariates. Although, the risk of women among reproductive age is being anemic and who were no longer living together is exp (β9) = exp (0.4372) = 1.5484 (95% CI OR: 1.1599, 2.067) times higher than those reproductive women getting anemic who never in a union. This implies that woman no longer living together and getting anemic is 54.84% times more likely than women who did never in a union after adjusting for correlated outcome data and controlling for other covariates.

Wealth index of the women is a significant effect on the anemic status of women among reproductive age. The odds ratio of the anemic status of women in reproductive age who were rich wealth index is exp (β11) = exp (-0.1459) = 0.8642 (95% CI OR: 0.7825, 0.9544) times lower than women among reproductive age who were poor wealth index level. This implies that the probability of women among reproductive age who were in rich wealth index and getting anemic is 86.42% times less likely than that of women in poor wealth index after controlling for other covariates. In similar fashion, some part of contraceptive use method is statistically significant on the anemic status of women in reproductive age since not all contraceptive use method is statistically significant. Thus the odd ratio of the women in reproductive age getting anemic and whose contraceptive use method is modern use is exp (β14) = exp (-0.7) = 0.4966

(95% CI OR: 0.4389, 0.5697) times lower than women among reproductive age who is getting anemic and who are no contraceptive use methods. This means that probability of women among reproductive age who were using a modern contraceptive method and being anemic is 49.66% times less likely than those women who did not use any contraceptive method after adjusting for correlated outcome data and controlling for other covariates.

Finally, the odd ratio of the anemic status of women in reproductive age who had pregnant is exp (β_{15}) = exp (0.288) = 1.3338 (95% CI OR: 1.1799, 1.5334) times higher than women in reproductive age those had not pregnant/unsure. The probability that the women who are pregnant and getting anemic is 33.38% times more likely than the reference group (those who were not pregnant or unsure) after adjusting for correlated outcome data and controlling for other covariates. BMI status also has significant effect on anemic status of women in reproductive age, that is the odds ratio of the women getting anemic and who were BIM is between 18.5 and to 25 is exp (β_{16}) = exp (-0.2519) = 0.7773 (95% CI OR: 0.7101, 0.8508) times lower than those who were BMI less than 18.5. The averaged probability of women in reproductive age getting anemic and their BMI between 18.5 and to 25 is 77.73% less likely anemic than those who were BMI less than 18.5 after adjusting for correlated outcome data and controlling for other covariates. Similarly for BMI greater than 25, the odds ratio of women becomes exp (β_{17}) = exp (-0.3785) = 0.6849 (95% CI OR: 0.5686, 0.8249). The probability of women in reproductive age getting anemic and BMI greater than 25 is 68.49% less likely anemic than those who were BMI less than 18.5 after adjusting for correlated outcome data and controlling for other covariates.

4.2.2.3 Model diagnostic for Marginal Model GEE

In time of model diagnosing for marginal model GEE plots used to see the patterns of all cases are DFBETA, COOK'S distance, LEVERAGE and CLUSTER DFFIT value as a function of ordered cluster.

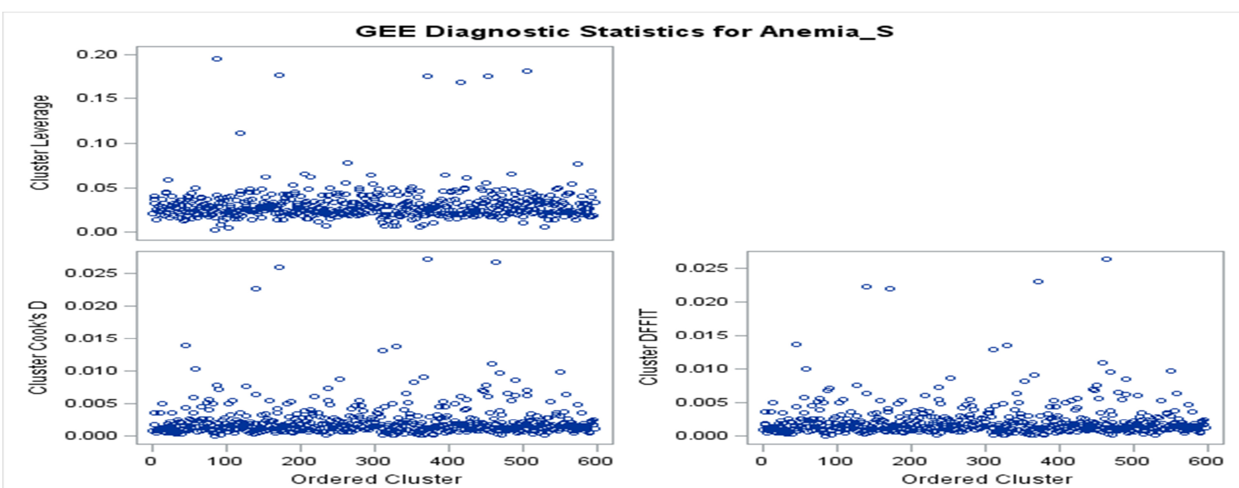
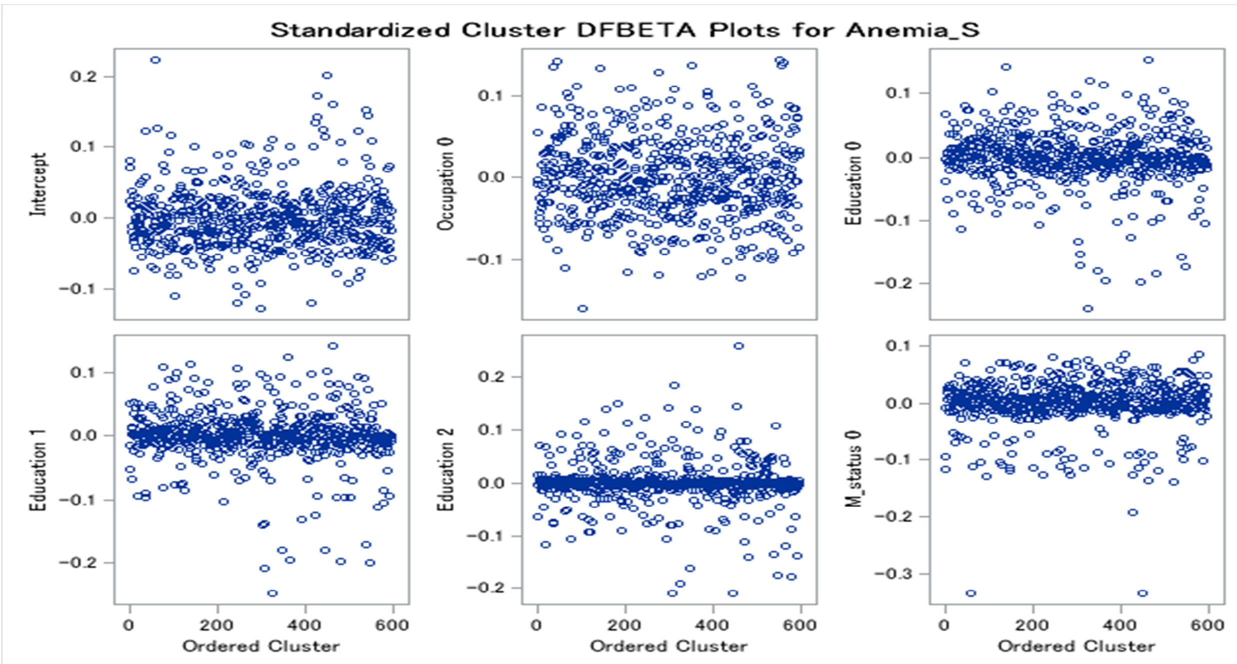
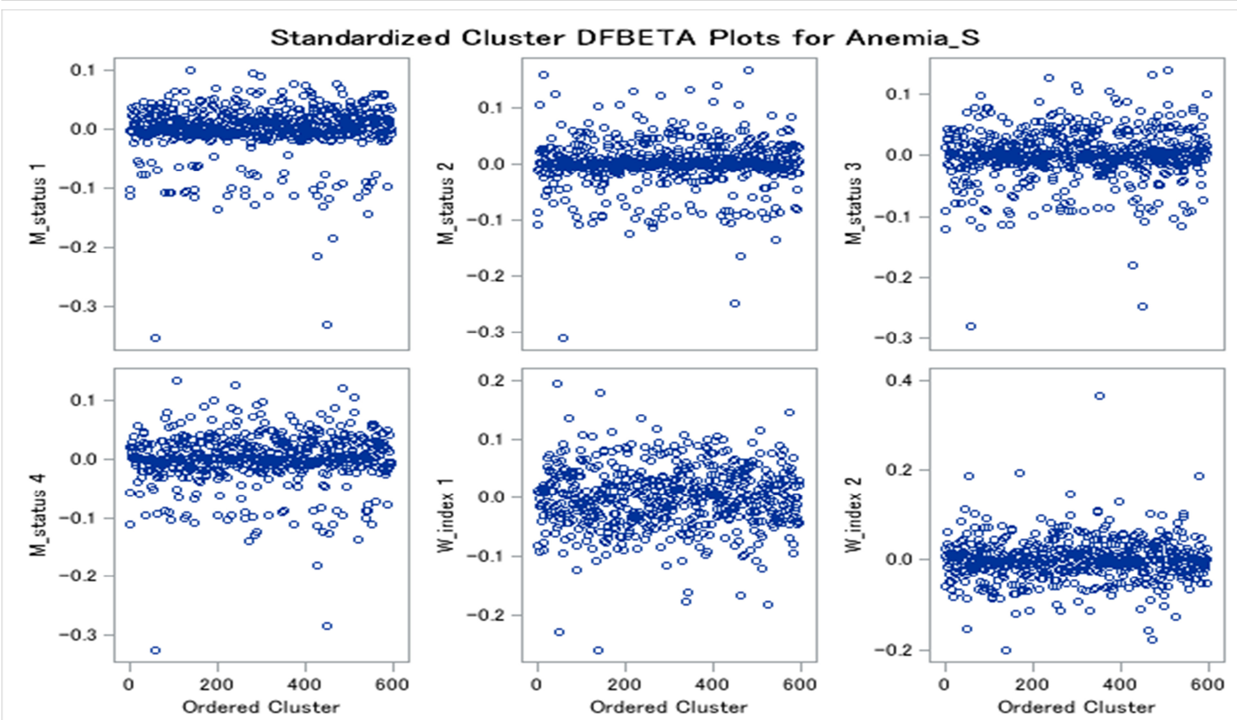


Fig 4.4: Plots of cluster leverage cluster cook's *D* and cluster DFFIT versus ordered cluster.

In above Figure (4.3) the top left panel is cluster leverage value versus the ordered cluster of all clusters. It was observed that leverage values of the above plots are less than one. This implies that there are no outliers. The bottom left panel of the above figure also shows the plot of Cook's D statistic versus the ordered cluster of all cluster. There are clusters a little far away from the others but these are not influential clusters since all Cook's D statistic are less than one. Similarly for cluster DFFIT all the influential points less than one, therefore the model is adequate.





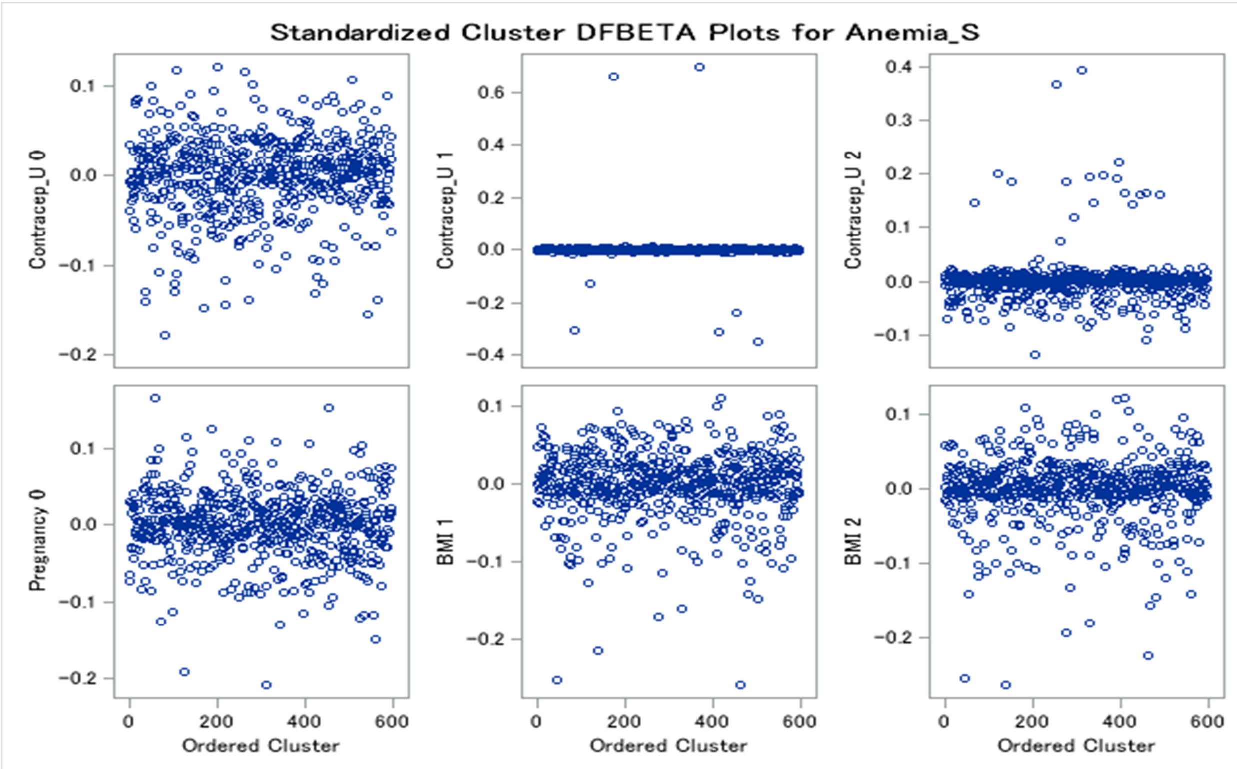


Fig 4.5: Plots of DFBETACS versus ordered cluster for all predictors in the fitted model.

The above figure Plots of DFBETACS of all explanatory variables vs. order cluster are shows that all the DFBETACS of all explanatory variables are less than one. This is an indication that there is no serious problem with the fitted model.

4.2.3 Analysis of Generalized Linear Mixed Model (GLMM)

For a study with repeated measurement or cluster of subjects, for instance, in case of this study, a regional cluster is a set of observations for a particular subject (women), and the model contains a random effect term for each subject. The random effects refer to a sample of clusters from all the possible clusters. The generalized linear mixed model is a further extension that permits random effects as well as fixed effects in the linear predictor. In this section, we can observe model building by selection of significant predictor variables and comparison of one intercept and two intercept model to elaborate with and between the regional variations of anemic status in women among reproductive age in subsection one. Parameter interpretation for GLMM with selected covariates present in subsection two and finally discusses model diagnosis for detecting the problems in the model and suggest improvements in section three.

4.2.3.1 Model Building in GLMM

Under the GLMM, model fitting began by adoption of the marginal model covariates. Additionally, the model also included the random effects in this case, random intercepts to address the between and within-regional variations. First, main effect covariates and the two random intercepts model were fitted and as usual, non-significant covariates were removed sequentially starting from variables with highest p-value for fixed effect covariates. The saturated models for GLMM with two random intercept b_j and b_{ij} fitted as follows:

$$\begin{aligned} logit(\pi_{ij}) = \beta_0 &+ \sum_{i=1}^{7} \beta_i Age_{group.i} + \sum_{j=1}^{2} \beta_j Recidence_j + \sum_{k=1}^{4} \beta_k Pairty_k + \sum_{l=0}^{1} \beta_l Occupation_l \\ &+ \sum_{m=0}^{3} \beta_m Education_m + \sum_{n=0}^{5} \beta_n Marital.S_n + \sum_{p=1}^{3} \beta_p Wealth.i_p \\ &+ \sum_{q=0}^{3} \beta_q Contraceptive_q + \sum_{r=0}^{1} \beta_r Preginancy_r + \sum_{s=0}^{1} \beta_s Breast.f_s \\ &+ \sum_{x=0}^{1} \beta_x Smoking_x + \sum_{y=1}^{3} \beta_y BMI_y + \sum_{z=0}^{1} \beta_z HIV_z + b_j + b_{ij}. \end{aligned}$$

In order to decide on the better of the two random effects models, two models were fitted, one with the two random intercepts that is saturated model to estimate between and within regional variations and the other with one random intercept model to estimate within regional variation. AIC and Likelihood ratio test (LRT) were used to compare the two models to select most parsimonious model.

Table 4.10: AIC and LRT for comparison of one and two random intercept models.

Models	AIC	BIC	LLD	σ_W	σ_B	P
Model with one Random intercept	14759.4	15005.9	-7348.7	2.011e-06	---	---
Model with two Random intercept	14327.7	14580	-7130.8	3.218e-05	5.137e-01	< 2.2e-16

Here, σ_W and σ_B are within and between regional standard deviation respectively. As we have seen from table 4.10, the AIC of model with two random intercept is reduced from 14759.4 to 14327.7 and the deviance reduced from 14697 to 14262. The p-value of the log likelihood ratio test < 2.2e-16 also indicates that the model with two random intercept is parsimonious model. P is the p-value of the log likelihood ratio test of the two models. Also when considered a model without random effects (i.e. the generalized linear model), it gives AIC value of 14760 which is large as compared to the above two random intercept model.

Next, the covariates for the fixed effect were assessed and the candidate covariates were selected by removing covariates starting from with highest p-value sequentially. Then the first removable covariate is smoking cigarette status of women with the highest p-value 0.506 and refitted the reduced model with the remaining covariates. The AIC is reduced from 14328 to 14326.1 and the p-value of log likelihood ratio test (p=0.4974) supports the reduced model is preferable one. The next removable variable is HIV test result of women with p-value (p=0.377) and refitted the reduced model. The AIC is reduced from 14326.1 to 14324.9 and the p-value of log likelihood ratio test (p=0.3826) supports the reduced model is preferable. The next removable variable is parity with p-value (p=0.346) and refitted the reduced model. The AIC is reduced from 14324.9 to 14321.0 and the p-value of log likelihood ration test (p=0.5501) support the reduced model is preferable.

Still us continued with highest p-value, the next removable variable is breast feeding status of women with p-value (p=0.194) and refitted the reduced model. For this model AIC is reduced from 14321.0 to 14320.7 and the p-value of log likelihood ratio test (p= 0.1947) supports the reduced model is preferable. The other variable next to breast feeding status with highest p-value is wealth index with p-value (p=0.161) and when refitted the reduced model the AIC is reduced from 14320.7 to 14319.0 and the p-value of log likelihood ratio test (p= 0.319) supports the reduced model is preferable. Finally the last variable with highest p-value is age of women with p-value (0.0844) and when refitted the reduced model the AIC is reduced from 14319.0 to 14313.7 and the p-value of log likelihood ratio test (p= 0.3466) supports the reduced model is preferable. In addition, the model with small number of covariates is considered to be preferable. Therefore, the final proposed GLMM for anemia status of women in reproductive age is given as:

$$\begin{aligned} \boldsymbol{logit(\pi_{ij})} &= \mathbf{-1.5922 + 0.4070Residense_r} - \mathbf{0.1113}\boldsymbol{Occupation_e} - \mathbf{0.1935}\boldsymbol{Education_p} \\ &- \mathbf{0.3690}\boldsymbol{Education_s} + \mathbf{0.3831}\boldsymbol{Marital.S_m} + \mathbf{0.4415}\boldsymbol{Marital.S_{lwp}} \\ &+ \mathbf{0.5793}\boldsymbol{Marital.S_w} + \mathbf{0.2966}\boldsymbol{Marital.S_d} + \mathbf{0.4196}\boldsymbol{Marital.S_s} \\ &- \mathbf{0.5892}\boldsymbol{Contraceptive_t} - \mathbf{0.7Welath\ index_r} + \mathbf{0.2733}\boldsymbol{Preginancy_y} \\ &- \mathbf{0.1752}\boldsymbol{BMI_{(18.5-25)}} - \mathbf{0.4526}\boldsymbol{BMI_{(\geq 25)}}. \end{aligned}$$

The parameter estimate and standard error of GLMM are presented in table 4.11 of below.

Table 4.11: Fit the parameter estimates and standard errors with p-value for GLMM.

Effect	Categories	Estimate	Standard Error	Pr > |z|	95% CI for β Lower	Upper
Intercept		-1.5922	0.1796	< 2e-16	-2.0064	-1.1733
Residence	Urban ®					
	Rural	0.4070	0.0672	1.43e-09	0.2671	0.5310
Occupation	Unemployed ®					
	Employed	-0.1113	0.0459	0.0153	-0.1947	-0.01478
	Not educated ®					
	Primary	-0.1935	0.0539	0.0003	-0.2976	-0.08644
Education	Secondary	-0.3690	0.1065	0.0005	-0.5747	-0.1570
Level	Higher	-0.1829	0.1224	0.1349	-0.4189	0.06127
	Single ®					
	Married	0.3831	0.0643	2.54e-09	0.1978	0.4683
	LWP	0.4415	0.1208	0.0003	0.1534	0.6361
Marital	Widowed	0.5793	0.1175	8.19e-07	0.3353	0.7966
Status	Divorced	0.2966	0.1060	0.0051	0.06734	0.4849
	No LLT	0.4196	0.1492	0.0049	0.1019	0.6888
Contraceptive	No ®					
Use Method	Folkloric	0.4733	0.8642	0.5839	-1.2431	2.1220
	Traditional	-0.0059	0.2429	0.9804	-0.4710	0.4815
	Modern	-0.5892	0.0711	< 2e-16	-0.7220	-0.4429
Pregnancy	Not pregnant ®					
	Pregnant	0.2733	0.0731	<.0001	0.1689	0.4696
	BMI < 18.5 ®					
	18.5-25	-0.1752	0.0477	0.0002	-0.2723	-0.08518
BMI	≥ 25	-0.4526	0.0984	4.26e-06	-0.6402	-0.2541

Source: 2011, EDHS data; final model for GLMM; NB: LWP = Living with Partner, LLT = Longer Livening Together, CI=Confidence Interval, ® =stands for reference categories.

4.2.3.2 Parameter Interpretation of GLMM

Unlike in the marginal model GEE, where parameters are treated as population averages, in the GLMM analysis, parameter interpretation is based on specific subjects or cluster. The parameter interpretation is conditional on the random effects, which is common for all individual women in the same cluster (region).

Given the same random intercept bj, the estimated odds of anemic status of women in reproductive age live in urban is exp(0.4070) = 1.5023 (95% CI: 1.3062 , 1.7006) times higher for women live in rural when compared to women live in urban in the same cluster keeping constant the other fixed effect variable in the model. This implies the probability of anemic status of women is 50.23% higher likely for women live in urban compared to women live in rural in the same region at the given random effect.

In the same way, the estimated odds that a reproductive women work status is employed is exp (-0.1113) = 0.8946 (95% CI: 0.8231, 0.9853) times lower to have anemic compared to the reference group in the same cluster. This shows that the probability of getting anemic is reduced by 89.46% for women whose occupation status is employed compared with women whose belongs to none employed occupation in the same region.

The estimated odds that reproductive women from education level primary is exp (-0.1935) = 0.8241 (95% CI: 0.7426, 0.9172) times lower to have anemic compared to reference of no educated level. This means percentage of anemic is decreased by 82.41% for women educated primary level compared to women not educated in the same cluster. The estimated odds that women from education level secondary is exp (-0.3690) = 0.6914 (95% CI: 0.5629, 0.8547) times lower to have anemic when compared to reference education level. This means percentage of being anemic is decreased by 69.14% for women belongs to secondary education level compared to women who not educated in the same cluster. In the same manner, estimated odds that women from education level higher is exp (-0.1829) = 0.8328 (95% CI: 0.6578, 1.0632) times lower to have anemic when compared to reference education level. This means percentage of being anemic is decreased by 83.28% for women belongs to higher education level compared to women who not educated in the same cluster.

The odds ratio of reproductive women being anemic who were married is $exp(0.3831) = 1.4668\ (95\%CI\colon 1.2187\ ,\ 1.5973)$ times higher than those reproductive women being anemic who never in union. Equivalently, the probability of reproductive women who are married and being anemic is 46.68% times more likely than those women who did never in union in the same cluster. The odds ratio of reproductive women being anemic who were living with partner is $exp(0.4415) =\ 1.555\ (95\%CI\colon 1.1305\ ,\ 1.8157)$times higher than those reproductive women being anemic who never in union. Equivalently, the probability of reproductive women who are living with partner and being anemic is 55.5% times more likely than those women who did never in union in the same cluster.

The odds ratio of reproductive women being anemic who were widowed is$exp(0.5793) = 1.7848\ (95\%CI\colon 1.3984\ ,\ 2.2179)$times higher than those reproductive women being anemic who never in union. The probability of women among reproductive age who are windowed and being anemic is 78.48% times more likely than women who were single the same cluster. The odds ratio of reproductive women being anemic who were divorced is exp (0.2966) = 1.3453 (95% CI: 1.0697, 1.6240) times higher than those reproductive women being anemic who never in union. That means that women windowed and being anemic is 34.53% times more likely than those women who did never in union from the same cluster. Additionally, the odds ratio of reproductive women being anemic who were no longer living together is $exp(0.4196) =$

$1.5214\ (95\%CI: 1.1073\ ,\ 1.9913)$ times higher than those reproductive women being anemic who never in union in the same cluster. This implies that woman no longer living together to partner and being anemic is 52.14% times more likely than those women who did never in union from the same cluster.

In similar fashion, some part of contraceptive use method is statistically significant on anemic status of reproductive women since not all contraceptive use method is statistically significant. Thus the odd ratio of the reproductive women being anemic and whose contraceptive use method is modern use method is $exp(-0.5892) =\ 0.5547\ (95\%CI: 0.4857\ ,\ 0.6422)$ times lower than women among reproductive age who is anemic and no contraceptive use methods in the same cluster. Equivalently the probability that reproductive women who were use modern contraceptive method and being anemic is 55.47% times less likely than those women who were not use any contraceptive method from the same cluster.

Likewise, the odd ratio of the anemic status of reproductive women who were pregnant is $exp(0.2733) =\ 1.3143\ (95\%CI: 1.184\ ,\ 1.5993)$ times higher than reproductive women those who were not pregnant/unsure in the same cluster. The probability that the women who are pregnant and being anemic is 31.43% times more likely than the reference group (those who were not pregnant or unsure). BMI status also has significant effect on anemic status of reproductive women, that is the odds ratio that the women being anemic and who were BMI is between 18.5 and to 25 is $exp(-0.1752) =\ 0.8393\ (95\%CI: 0.7616\ ,\ 0.9183)$ times lower than those who were BMI less than 18.5 in the same cluster. The probability that the women in reproductive age getting anemic and BMI between 18.5 and to 25 is 83.93% less likely anemic than those who were BMI less than 18.5 form the same cluster. Similarly for BMI greater than 25, odds ratio women becomes exp(-0.4526) = 0.636 (95% CI:0.5268 , 0.7756) times lower than those who were BMI less than 18.5 in the same cluster. The probability that the woman among reproductive age is being anemic and BMI greater than 25 is 63.6% less likely anemic than those who were BMI less than 18.5 in the same cluster.

4.2.3.3 Model diagnostic for GLMM

Residuals versus observation cluster level identity number plot panel one, suggested that the residuals are symmetric around zero (i.e. positive and negative residuals are almost equal). Q-Q plots for normality of random effects at regional and cluster levels are also given in the figure at panel two and three, and illustrates that the random effects are normally distributed with mean zero and variance covariance matrix D. Thus, the fitted GLMM model is fine for the given data.

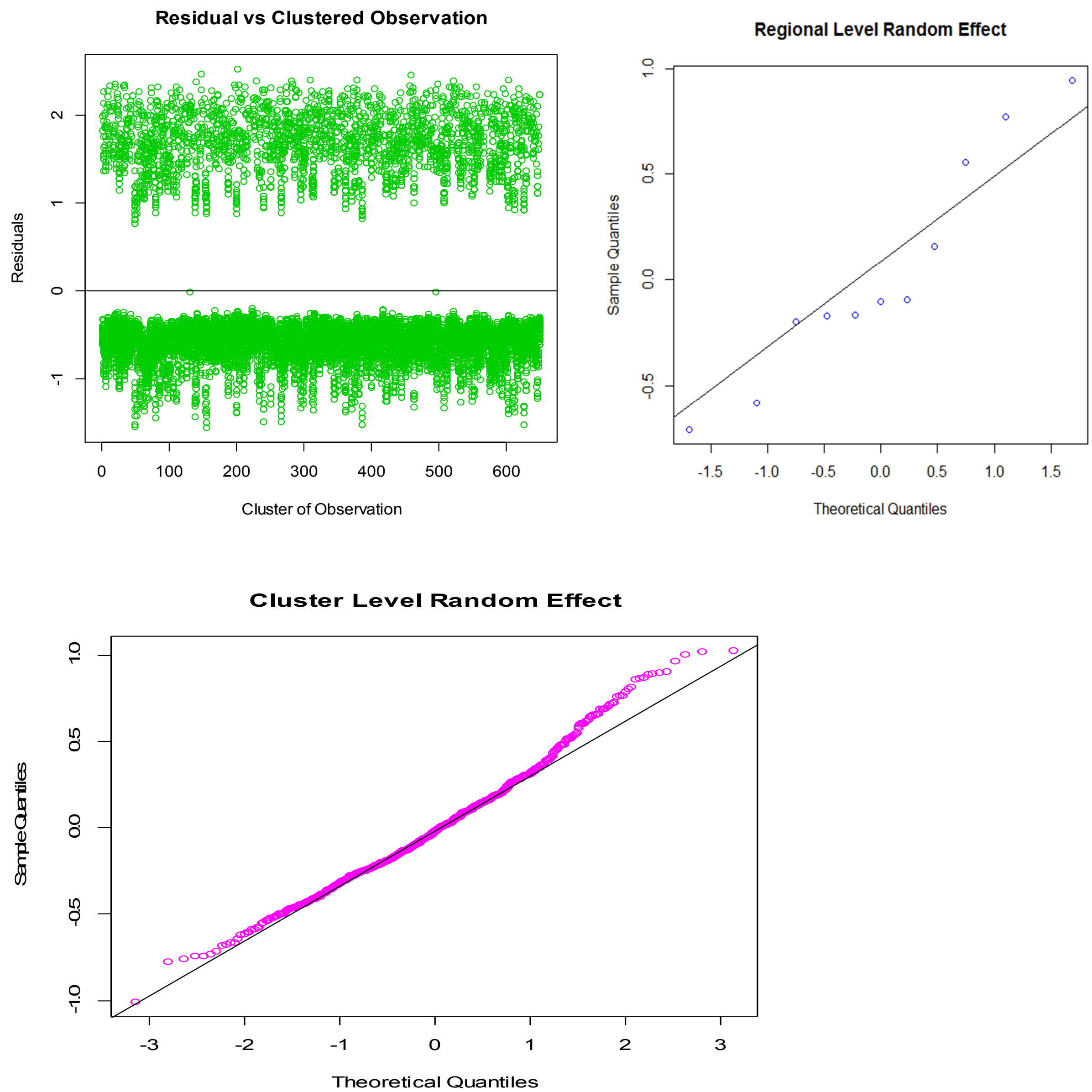


Fig 4.6: Diagnosis plots for the generalized linear mixed model.

Although, the sub options of the RESIDUALPANEL request produce two panels in model diagnosis method. The panel of conditional residuals is constructed from $y - x'\hat{\beta} - z'\hat{\gamma}$ (Figure A2) in appendices. The panel of marginal residuals is constructed from $y - x'\hat{\beta}$ (Figure A3) in appendices. Note that these residuals are deviations from the observed data, because the model is a normal linear mixed model, and hence it does not involve pseudo-data. Whenever the random-effects solutions $\hat{\gamma}$ are involved in constructing residuals, the title of the residual graphics identifies them as conditional residuals (Figure A3) attached in Appendix.

4.2.4 Comparison of the three models GLM, GEE and GLMM

The three models described in previous section build the best model with different information criteria's explained in each section and here compare the best model among the three models using their information criteria and standard errors for the models. The comparison of the best fit of GLM model and GLMM with Akaike's Information Criteria, GLMM with two random intercept model is better fit with AIC of 14313.7 than that of GLM with AIC of 14748.35. Taking GLMM with regard to precision as compared to GLM for estimates for subject specific effect interpretations.

The results of comparison between the GEE and GLMM are shown in Table 4.12 below. Differences between regression coefficients and between standard errors from marginal and random effects model are expected. The coefficients estimates from the GEE are lower in magnitude than corresponding coefficients estimates from the GLMM (Fitzmaurice, et al., 2011). Furthermore, standard errors from the GLMM are larger than those from the GEE. Therefore, GEE is the best or robust parameter estimates for population average interpretations by smaller standard error to fit the model.

Table 4.12: Parameter estimates, standard errors for three models GLM, GEE and GLMM.

Effect	Categories	GEE		GLMM	
		Estimate	SE	Estimate	SE
Intercept		-1.0338	0.0706	-1.5922	0.1796
Occupation	Employed	-0.3877	0.0423	-0.1113	0.0459
Education Level	Primary	-0.3547	0.0519	-0.1935	0.0539
	Secondary	-0.5673	0.1021	-0.3690	0.1065
	Higher	-0.3331	0.1178	-0.1829	0.1224
Marital Status	Married	0.4381	0.0630	0.3831	0.0643
	LWP	0.3730	0.1184	0.4415	0.1208
	Widowed	0.5608	0.1151	0.5793	0.1175
	Divorced	0.3019	0.1042	0.2966	0.1060
	No LLT	0.4372	0.1474	0.4196	0.1492
Contraceptive Use Method	Folkloric	0.2081	0.8469	0.4733	0.8642
	Traditional	-0.0104	0.2403	-0.0059	0.2429
	Modern	-0.7000	0.0701	-0.5892	0.0711
Pregnancy	Pregnant	0.2880	0.0712	0.2733	0.0731
BM	18.5≤BMI25	-0.2519	0.0461	-0.1752	0.0477
	BMI≥ 25	-0.3785	0.0950	-0.4526	0.0984
Information Criteria		QIC =14877.4129		AIC =14313.7	

Source: 2011, EDHS data; Comparison of GEE vs. GLMM; NB: LWP = Living with Partner, LLT = Longer Livening Together, CI=Confidence Interval, ® =stands for reference categories.

4.3 Discussion

This study was aimed at modeling the determinants of anemic status among reproductive women in Ethiopia. As a preliminary analysis, assortments of summary statistics were employed to explore the association between the response variable of interest and available covariates. It should be well-known that there is inconsistency in the conclusion from the analysis of various summary statistics, which might be due to the fact that they make use of varying amount of information, which determines the power of their inferences. Thus, the analysis was extended to other statistical methods to account for the clustered nature of correlated observations. The data were then analyzed using three model families one with generalized linear mode without consideration of correlation among clusters, the second with marginal models GEE, and the other is random effects model (Generalized linear mixed model). Through the application of the three methods, we are able to show that GEE and GLMM are more flexible than GLM for investigating dichotomous correlated outcomes and modeling a variety of correlation patterns between clustered measures.

There are other important differences among the three approaches. Besides the noted differences, the three methods were used to answer different research questions. With this two stage stratified cluster sampled data collected from every individual from clustered regions and EAs, we wanted to make statistical inferences regarding the change in mean response over time (a population averaged inference) or the individual trajectory over time (a subject-specific inference).

In GEE a model, the regression coefficient describes how the average rates for any variable may be changed in the study population. The exponential of an estimate parameter represents a population-averaged odds ratio for the response and relates to the sub-population that includes the covariate concerning the sub-population not including the covariate. GLMM by contrast, could identify subject-specific effects of covariates on the changes in the response over time. The exponential of the estimate parameter is an odds ratio for a person that has a covariate, when compared to the same person not having a covariate (Fitzmaurice, et al., 2011). Therefore, GLMM would be helpful where an intervention is likely to affect some individuals differently than others as compared to GEE which do not take individual response into account in their interpretations. GLMM could allow for a more various analysis of individuals of this sub population such as predicting individual risk of complications.

Generalized linear model was fitted for dichotomous response variable to compare to other models which are suitable to dichotomous response variable and clustered data structure. As seen from the previous section of data analysis, the three component of generalized linear model that fulfills the assumptions are the distribution of the response variable is binary distribution and linear predictors are the function of possible factors they have significant effect to the response variable. The link function used to link response to systematic component is logit link function. As compared to the generalized linear mixed model by comparing the AIC of each model the GLMM model is the best model with smaller AIC of 14313.7 to describe the clustered data than GLM with AIC of 14748.35.

Two proposed working correlation structures, exchangeable and independence correlation assumptions were taken for the comparison, in GEE model-building strategy. The model with independent working correlation structure with QIC value of 14860.71 is found to be better fits the data than exchangeable with QIC value of 14987.7843. This does not support the consideration of dependency within the clustering nature of the data analysis and the dependency of individuals in the same group for the given data.

The purpose of GLMM was to evaluate within and between regional variations of anemic status of women among reproductive age in Ethiopia. Two models was fitted one with only one random intercept model to assess only within regional variation and other with two random intercepts model, in order to account within and between regional variations. Additionally, generalized linear model was fitted as the sake of comparison whether including random effects in the analysis is important or not. The two models were compared using the AIC value followed by likelihood ratio test and we got a model with two random intercept was favorable. This demonstrates that, accounting within and between regional variations for the analysis of anemic status of women among reproductive age should be vital and, indicates within and between regional heterogeneity in anemic status of women among reproductive age. This finding is supported by the explanation or suggestion of Antonio & Beirlant (2011). Even though the two model families are different and their comparability may not be meaningful as they have different parameter interpretations and estimations, parameter estimates obtained from GLMM are generally bigger in absolute values than those from marginal models (GEE) similar with Agresti (2007).

In this study we considered the explanatory variables Age, Residence, Parity (number of child ever born), Occupation, Education level, Marital status, Wealth index, Contraceptive Use method, Pregnancy status, Breast feeding status, smoking status, BMI and HIV test result of women among reproductive age. The three models are leads to slightly different conclusion in model building with selection of candidate significant covariates. The covariates: occupation of women, highest education level, marital status, contraceptive use method, pregnancy status and BMI of women are commonly significant at 5% level of p-value for three models. Whereas, age is significant in GLM, residence is significant in GLMM and wealth index is significant both in GLM and GEE separately at 5% level of p-value.

Age of women is not significant variable in this study with consideration of correlations and variations for both marginal and subject specific models. This result is consistent with the result obtained by Uche-Nwachi et al (2010); Kayihan and Nilgun (2008). However, age of women is significant for model without consideration of correlations or any variation within and between the subjects and cluster, the finding is consistent with the result obtained by Ibrahim et al. (2012), Tadege (2009); Judith et al. (2008); and Yuan et al. (2009).

Place of residence was found as a significant determinant of anemia levels. Women live in rural were more likely to be anemic than urban women. This result is consistent with studies conducted by; Bishwajit Ghose (2016) and Gebremedhin and Enquselassie (2011) showed that rural residence was key factor predisposing women to anemia. Getachew et al. (2012) found that rural women were highly affected by anemia. Tadege (2009) and Yuan et al. (2009) found that residence is a significant determinant of anemia. Sanku et al. (2010) showed urban women were at lower risk of anemia. Wondu and Bijlsma (2012) found the same result.

In this study as seen parity (number of children women ever born) was not statistically significant to anemic status women in reproductive age. The finding is consistent result by Sanku et al. (2010) and Ibrahim et al. they showed that parity was not significant factor to anemia. In contrast, the result founded by Gebremedhin and Enquselassie (2011) showed that high parity was a key factor predisposing to anemia. Haider (2010) also showed that women having more than two children are associated with anemia. Uche-Nwachiet a. (2010), Balarajan et al. (2013) and Judith et al. (2008) showed that anemia is associated with high parity.

Occupation is significant determinant of anemia in this study which is not consistent with the study conducted by Tadege (2009). However the result is the same to study obtained by LakewY (2014).

Educational level has significant effect in this study, this showed that illiterate women were at higher risk of being anemic, consistent with the results obtained by Cüneyt Eftal et al,. (2015), Bishwajit Ghose et al., (2016), Kim Lam Soh et al,, (2015), Lakew Y (2014), Balarajan et al. (2013), Ndukwu and Dienye (2012) and Suegaet al. (2002) who found that educational level is a significant predictor of anemia and severe anemia. The study results obtained by Ibrahim et al. (2012); Kayihan and Nilgun (2008) conclude the opposite.

The current study identified marital status is a significant predictor of anemia. This finding is similar to the result by Tadege (2009) and not consistent with the result obtained by Kayihan and Nilgun (2008) and Ibrahim et al. (2012). The study explore that wealth index is a significant predictor of anemia status in GEE model. This finding is consistent with the findings by Bishwajit Ghose et al., (2016) and Balarajan et al. (2013).

Wealth index significantly determines the anemia status in women among reproductive age according to this study. This result is consistent with the finding by Balarajan et al. (2013) result revealed that wealth index is a significant determinant of anemia.

Women in reproductive age who did not use any contraceptive method were at higher risk of being anemic than that of women use traditional contraceptive use method. This disagrees with the result obtained by Haider (2010) which concluded the contrary. However, there is the same findings to this study is Wondu and Bijlsma (2012) who showed that women who did not use contraceptive methods were at higher risk of being anemic than who did use traditional methods.

The result of the current study showed that the chance of being anemic (relative to none anemic women) was higher by 37.6% compared to women who were not pregnant. The finding is similar to those by Melku et al. (2014); Getachew et al. (2012); Tadege (2009); Priyanka (2011); Suega et al (2002); Judith et al. (2008); Ibrahim et al. (2012); Ndukwu and Dienye (2012); Yuan et al. (2009), and Sanku et al. (2010) which showed that pregnancy is a significant determinant of anemia.

Breast feeding is statistically not significant effect on anemic status among women among reproductive age in Ethiopia in this study. The finding is not consistent to Lakew Y, et al,. (2014) that revealed breast feeding mothers are significantly prevalence to anemia.

In this study body mass index was found to be highly related with prevalence of anemia. Women with low body mass index were more likely to be anemic. This result is consistent with the result of studies by Melku et al. (2014), Lakew Y (2014) and Meda et al., (2016). Whereas, the finding is not consistent to the result found by Kayihan and Nilgun (2008) and Thomson et al, (1986). Women with BMI <18.5 are 37% more likely to be in severe anemia than those who have BMI >25. This result is similar to the finding by Aabroo et al. (2012) that showed those suffering from severe anemia had BMI <17.

The finding of this study revealed that smoking habit of women among reproductive age in Ethiopia is statistically not significant effect on anemic status. This result is consistent with the study conducted by Adamu AL, (2017), and Taner CE, (2015). This disagrees with the result obtained by Okumbe OT, (2016), and Sanku et.al, (2010).

HIV status of women among reproductive age in this study revealed that there is no statistical association to anemic status. This finding is consistent to the result obtained by Meda et al., (2016) and Dario MD (2005), both showed that HIV test result among women is not statistically significant factor for being anemic.

CHAPTER FIVE

CONCLUSION AND RECOMENDETION

5.1 Conclusion

In this study, we compared both methods and found that the regression parameters from GEE are smaller than those from GLMM and standard errors from GEE are smaller than that of GLMM. We conclude that, GEE model with measure of association exhibited the best fit for this data than GLM and GLMM models, even if both GEE and GLMM are their own aims to conduct the data analysis. The GLMM provided interesting relationships that would not be evident from a standard logistic model. For this study GLMM, with two random intercept models is found to be parsimonious for the analysis of within and between regional variations for anemic status of women among reproductive age in Ethiopia.

This concluded that there is heterogeneity of anemic status between and within regions. Even if significant variation of anemic status between regions the highest prevalence are occurred in Somali, Affar and Dire Dawa regions.

Anemia has modest public health effect in women among reproductive age in Ethiopia. The results of this study suggested that a women among reproductive age lived in rural areas, the three higher educational level attained (primary, secondary and higher), all categories of marital status (married, living with partner, windowed, divorced and separated), being use traditional contraceptive method, being from rich wealth index, being pregnant and both of BMI's that are $18.5 \leq BMI < 25$ and ≥ 25 make women among reproductive age prone to anemia.

Generally, socio demographic characters (occupation, education and marital status), health factors (contraceptive use method, pregnancy status and BMI) and economic factor wealth index of women belongs are major determinates of significant contribution to the hazard of anemia.

5.2 Recommendation

This study has identified a number of important factors that significant effect on anemic status of women among reproductive age in Ethiopia. Strategy to reduce anemic status in Ethiopia focus has to be given awareness on nutrition intake, iron and vitamins supplementation for women especially during pregnancy time. It is suggested that programs that work to reduce the rate of anemic status of women should focus on improving feeding style with getting iron, foalet and vitamins to make hemoglobin concentration in our body normal, while addressing the intractable socio-economic disparities that continue to indirectly contribute to the incidence of anemic status in women. Socio-economic factors influenced the increasing rate of anemia in women typically at time of pregnancies. Most women lacked knowledge of the pregnancy risk factors that adversely affect anemia, and the exact mechanisms by which the risk factors act to cause the adverse effects. Intervention programs and behavior change communication nutrition intake on women should focus on significant risk factors associated with anemia, and target poorest wealth index women at risk. Health education for pregnant women should be strengthened to promote care seeking and demand for skilled care at all stages of anemic maternity. This way healthy reproductive women are produced who have a better chance of change world and becoming tomorrow's wealth.

Generally, based on the results of this study in addition to the above recommendations the following recommendations are suggested.

- The generalized estimating equation model provides better predictions to the marginal effect and generalized linear mixed model for conditional subject specific effect of two stage clustered data sets like anemia status in women. So, it is advisable to future researchers should have to use the generalized estimating equation model for clustered data.
- Since the prevalence of anemia differs among regions, the Ministry of Health should give special attention to regions Somali, Affar and Dire Dawa.
- As seen illiterate women were at higher risk of anemia, we recommend that Ethiopia government facilitating assess of schools and engaged of women participation in teaching learning plan.
- Uses of contraceptive methods decrease the risk of anemia. Hence, concerted effort should be made to create the awareness and understanding the use of contraceptive methods.
- The government and other concerned bodies should pay attention to the above and make appropriate efforts to tackle problems that contribute to anemia.

REFERENCES

A Aabroo T., Khand A., Laghari, A. (2012). Prevalence of anemia in adolescent girls, Department of Physiology, University of Sindh, Jamshoro, Pakistan.Pak J Physiol, 8(Suppl 1) P-BN-01. http://www.pps.org.pk/PJP/8-1S.pdf.

Agresti, A. (1996). An introduction to categorical data analysis (Vol. 135). New York: Wiley.

Agresti, A. (2002). Inference for contingency tables. Categorical Data Analysis, Second Edition, 70-114.

Agresti, A. (2007). An introduction to categorical data analysis, 2nd edn. Hoboken.

Antonio, K., & Beirlant, J. (2007). Actuarial statistics with generalized linear mixed models. Insurance: Mathematics and Economics, 40(1), 58-76.

Balarajan, Y. S., Fawzi, W. W., & Subramanian, S. V. (2013). Changing patterns of social inequalities in anaemia among women in India: cross-sectional study using nationally representative data. BMJ open, 3(3), e002233.

Balemi, A., & Lee, A. (1999). Theory & Methods: Some Properties of the Liang-Zeger Method Applied to Clustered Binary Regression. Australian & New Zealand Journal of Statistics, 41(1), 43-58.

BAMLAKU, T. (2009). DETERMINANTS OF ANAEMIA IN PREGNANT WOMEN WITH EMPHASIS ON INTESTINAL HELMINTHIC INFECTION AT BUSHULO HEALTH CENTER IN SOUTHERN ETHIOPIA.

Bekele, A., Tilahun, M., & Mekuria, A. (2016). Prevalence of anemia and its associated factors among pregnant women attending antenatal care in health institutions of Arba Minch Town, Gamo Gofa Zone, Ethiopia: a cross-sectional study. Anemia, 2016.

Belsley D. A., E. Kuh and R. E. Welsch, (1980) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, New York.

Berhane, Y., Gossaye, Y., Emmelin, M., & Hogberg, U. (2001). Women's health in a rural setting in societal transition in Ethiopia. Social Science & Medicine, 53(11), 1525-1539.

Brabin, B. J., Hakimi, M., & Pelletier, D. (2001). An analysis of anemia and pregnancy-related maternal mortality. The Journal of nutrition, 131(2), 604S-615S.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. Journal of the American statistical Association, 88(421), 9-25.

Brick T and Peters MJ. (2014). Kathryn Maitland on treating severe anaemia in sub-Saharan Africa; BMC Medicine 2014, 12:68.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52(3), 345-370.

Chaganty, N. R., & Joe, H. (2006). Range of correlation matrices for dependent Bernoulli random variables. Biometrika, 93(1), 197-206.

Cook, R. D. (1977). Detection of influential observation in linear regression. Technometrics, 19(1), 15-18.

Crowder, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. Biometrika, 82(2), 407-410.

Dairo, M. D., Lawoyin, T. O., Onadeko, M. O., Asekun-Olarinmoye, E. O., & Adeniji, A. O. (2005). HIV as an additional risk factors for anaemia in pregnancy: evidence from primary care level in Ibadan, Southwestern Nigeria. African journal of medicine and medical sciences, 34(3), 275-279.

Davison, A. C., & Snell, E. J. (1991). Residuals and diagnostics. Statistical Theory and Mod.

DeMaeyer, E. M., Dallman, P., Gurney, J. M., Hallberg, L., Sood, S. K., Srikantia, S. G., & World Health Organization. (1989). Preventing and controlling iron deficiency anaemia through primary health care: a guide for health administrators and programme managers.

Demographic, E. (2011). Health Survey: Addis Ababa. Ethiopia and Calverton, Maryland, USA: central statistics agency and ORC macro.

Diggle, P. (2002). Analysis of longitudinal data. Oxford University Press.

Faraway, J. J. (2005). Linear models with R. series ed. Chapman and hall. New York Washington. D.C.

Federal Ministry of Health (2004), National guideline for control and prevention of micronutrient deficiencies. Addis Ababa: Family Health Department, Federal Ministry of Health, Government of Ethiopia, p. 26.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (Eds.). (2008). Longitudinal data analysis. CRC Press.

Fitzmaurice GM, Laird NM, Ware JH. (2011). Applied longitudinal analysis. 2nd edition. New Jersey: John Wiley & Sons Inc.

Fleming, A. F. (1987). Maternal anaemia in northern Nigeria: causes and solutions.
Gebremedhin S., Enquselassie F. (2011) Correlates of anemia among women of reproductive age in Ethiopia: Evidence from Ethiopian DHS 2005. Ethiop J Health Dev; 25(1):23-30

Getachew, M., Yewhalaw, D., Tafess, K., Getachew, Y., & Zeynudin, A. (2012). Anaemia and associated risk factors among pregnant women in Gilgel Gibe dam area, Southwest Ethiopia. Parasites & vectors, 5(1), 296.

Ghose, B., Tang, S., Yaya, S., & Feng, Z. (2016). Association between food insecurity and anemia among women of reproductive age. PeerJ, 4, e1945.

Gibson, M. A., & Mace, R. (2006). An energy-saving development initiative increases birth rate and childhood malnutrition in rural Ethiopia. PLoS Medicine, 3(4), e87.

Haidar, J. (2010). Prevalence of anaemia, deficiencies of iron and folic acid and their determinants in Ethiopian women. Journal of health, population, and nutrition, 28(4), 359.

Haidar, J. A., & Pobocik, R. S. (2009). Iron deficiency anemia is not a rare problem among women of reproductive ages in Ethiopia: a community based cross sectional study. BMC Hematology, 9(1), 7.

Haidar, J., Nekatibeb, H., & Urga, K. (1999). Iron deficiency anaemia in pregnant and lactating mothers in rural Ethiopia. East African medical journal, 76(11), 618-622.

Haidar, J., Muroki, N. M., Omwega, A. M., & Ayana, G. (2003). Malnutrition and iron deficiency anaemia in lactating women in urban slum communities from Addis Ababa, Ethiopia. East African medical journal, 80(4), 191-194.

Harrison, K. A., & Rossiter, C. E. (1985). Child-bearing health and social priorities: a survey of 22774 consecutive hospital births in Zaria northern Nigeria. 12. Maternal mortality. British Journal of Obstetrics and Gynaecology, 92, 100-15.

Hedeker, D., & Gibbons, R. D. (2006). Longitudinal data analysis (Vol. 451). John Wiley & Sons.

Hosmer, D. W., & Lemeshow, S. (1989). Applied regression analysis. New York, John Willey.

Hosmer, W.D. and Lemeshow, S. (2000), Applied Logistic Regression. 2nd Ed., John Wiley and Sons, New York.

Hurtado, A., Merino, C., & Delgado, E. (1945). Influence of anoxemia on the hemopoietic activity. Archives of Internal Medicine, 75(5), 284-323.

Ibrahim Isa., Kemebradikumo P., Dennis A. (2012) .The Burden of Anaemia among Pregnant Women at Booking in the Niger Delta of Nigeria. Online Journal of Medicine and Medical Science Research ISSN 2277-0879; Volume 1 Issue 5, Pages 91-95.Available Online at http://www.onlineresearchjournals.org/JMMSR.

INACG (2002) .Why iron is important and what to do about it: A new perspective. From: http://www.idpas.org/pdf/2071 Why Iron is Important and what to DO.pdf Date of access [cited 21/06/2010].

International Nutritional Anemia Consultative Group (INACG) 1979, Iron deficiency in infancy and childhood. Geneva, Switzerland: INACG, World Health Organization.

International Nutritional Anemia Consultative Group (INACG) 1989, Iron deficiency in women. Geneva, Switzerland: INACG, World Health Organization.

IPHN (2007). National Strategy for anaemia prevention and control in Bangladesh, Ministry of Health and Family Welfare Government of the People's Republic of Bangladesh.

J. W. Hardin and J. M. Hilbe (2003). Generalized Estimating Equations. Chapman and Hall/CRC Press, Boca Raton, Fla, USA.

Kayihan P., Nilgun D. (2008), Prevalence & risk factors of anaemia among women of reproductive age in Bursa, Turkey. Indian J Med Res128, 282-286.

Koller O. (1982). The clinical significance of hemodilution during pregnancy, Obstetrical and Gynecological Survey, 37:649–652.

Kullback, S., & Gokhale, D. V. (1978). Information analysis of linear interactions in contingency tables: Information analysis of linear interactions. Communications in Statistics-Theory and Methods, 7(1), 27-45.

K.Y.Liang and S.L.Zeger (1986). "A comparison of two bias-corrected covariance estimators for generalized estimating equations," Biometrika, vol.73, pp.13–22.

Lakew, Y., Biadgilign, S., & Haile, D. (2015). Anaemia prevalence and associated factors among lactating mothers in Ethiopia: evidence from the 2005 and 2011 demographic and health surveys. BMJ open, 5(4), e006001.

Lee, Y. and Nelder, J.(1996). Hierarchic9al generalized linear models. J. R: Statist. Soc. B, Vol.**8**, pp. 619-678.

Lee and J. A. Nelder, (2009). Likelihood inference for models with unobservable: another view, Statistical Science, vol. **24**, no. **3**, pp. 255–269.

Littell, R. C., Stroup, W. W., & Freund, R. J. (2002). SAS for linear models. SAS institute.

Magder, L. S., & Zeger, S. L. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. Journal of the American Statistical Association, 91(435), 1141-1151.

Mayers, R. H. (1990). Classical and Modern Regression with Applications: PWSKENT Publ. Co.: Boston.

McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability.

McCullagh, P., & Nelder, J. A. Chapman and Hall; London, UK: 1989. Generalized linear models.

Melku, M., Addis, Z., Alem, M., & Enawgaw, B. (2014). Prevalence and predictors of maternal anemia during pregnancy in Gondar, Northwest Ethiopia: an institutional based cross-sectional study. Anemia, 2014.

Menard, S. (2002). Applied logistic regression analysis (Vol. 106). Sage.

Michael, E. Griswold & Scott, L. Zeger (2004). on marginalized multilevel models and their computation, Johns Hopkins Bloomberg School of Public Health.

Micronutrient Initiative (2006), Ethiopian country profile [document on the internet]. The institute; Available from: http://www.micronutrient.org.

Molenberghs, G., & Verbeke, G. (2005). Models for discrete longitudinal data.

Mumbare, S. S., Maindarkar, G., Darade, R., Yenge, S., Tolani, M. K., & Patole, K. (2012). Maternal risk factors associated with term low birth weight neonates: a matched-pair case control study. *Indian pediatrics*, *49*(1), 25-28.

Myers H., Montgomery C., Vining, G. & Robinson, J. (2010). Generalized Linear Models with Applications in Engineering and the Sciences, Second Edition

Ndukwu, G. U., & Dienye, P. O. (2012). Prevalence and socio-demographic factors associated with anaemia in pregnancy in a primary health centre in Rivers State, Nigeria. African Journal of Primary Health Care and Family Medicine, 4(1), 1-7.

Nordenberg, D., Yip, R., & Binkin, N. J. (1990). The effect of cigarette smoking on hemoglobin levels and anemia screening. Jama, 264(12), 1556-1559.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. Biometrics, 57(1), 120-125.

P. D. Allison, 2001, Logistic Regression Using the SAS system: Theory andApplications, Cary, NC: SAS Institute Inc, 2001.

Preisser, J. S., & Qaqish, B. F. (1996). Deletion diagnostics for generalised estimating equations. Biometrika, 83(3), 551-562.

Priyanka Sh, S. M., & Ranjana, N. (2013). Prevalence of anemia and socio-demographic factors associated with anemia among pregnant women attending antenatal Hospital in Jaipur City. India. IOSR-JPBS, 6(3), 01-5.

Qu, A., & Lindsay, B. G. (2003). Building adaptive estimating equations when inverse of covariance estimation is difficult. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(1), 127-142.

Sabo, R. T., & Chaganty, N. R. (2010). What can go wrong when ignoring correlation bounds in the use of generalized estimating equations. Statistics in medicine, 29(24), 2501-2507.

Sanku D. Sankar G., Madhuchhanda G. (2010). Prevalence of anaemia in women of reproductive age in Meghalaya: A logistic regression analysis. Turk J Med Sci, 40(5): 783-789.

Scrimshaw N.S. (1984). Functional consequences of iron deficiency in human populations Journal of Nutritional Science and Vitaminology 30: 47-63.

Sharman, Almaz. 2000. Anemia testing in population-based surveys:General information and guidelines for country monitors and program managers.Calverton, Maryland USA: ORC Macro.

Soh, K. L., Tohit, E. R. M., Japar, S., Geok, S. K., Ab Rahman, N. B., & Raman, R. A. (2015). Anemia among Antenatal Mother in Urban Malaysia. Journal of Biosciences and Medicines, 3(03), 6.

Stevens, G. A., Finucane, M. M., De-Regil, L. M., Paciorek, C. J., Flaxman, S. R., Branca, F., ... & Nutrition Impact Model Study Group. (2013). Global, regional, and national trends in haemoglobin concentration and prevalence of total and severe anaemia in children and pregnant and non-pregnant women for 1995–2011: a systematic analysis of population-representative data. The Lancet Global Health, 1(1), e16-e25.

Stoltzfus, R. J. (2011). Iron interventions for women and children in low-income countries. The Journal of nutrition, 141(4), 756S-762S.

Sutradhar, B. C., & Das, K. (1999). Miscellanea. On the efficiency of regression estimators in generalised linear models for longitudinal data. Biometrika, 86(2), 459-465.

Taner, C. E., Ekin, A., Solmaz, U., Gezer, C., Çetin, B., Keleşoğlu, M., ... & Özeren, M. (2015). Prevalence and risk factors of anemia among pregnant women attending a high-volume tertiary care center for delivery. Journal of the Turkish German Gynecological Association, 16(4), 231.

Thomsen, K., Riis, B., Krabbe, S., & Christiansen, C. (1986). Testosterone regulates the haemoglobin concentration in male puberty. Acta Paediatrica, 75(5), 793-796.

Uche-Nwachi, E. O., Odekunle, A., Jacinto, S., Burnett, M., Clapperton, M., David, Y., ... & Seereeram, R. (2010). Anaemia in pregnancy: associations with parity, abortions and child spacing in primary healthcare clinic attendees in Trinidad and Tobago. African health sciences, 10(1), 66.

UNFPA (2012). Mortality In-depth Analysis of the EDHS 2000-2011, December Trends in Maternal Health in Ethiopia, December 2012 Addis Ababa.

Wang, Y. G., & Carey, V. J. (2004). Unbiased estimating equations from working correlation models for irregularly timed repeated measures. Journal of the American Statistical Association, 99(467), 845-853.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. Biometrika, 61(3), 439-447.

WHO (2014). WHA Global Nutrition Target 2025, Anaemia policy Brief, (www.who.int/nutrition/topic/nutrition global target 2025/en/, accessed 6 October 2014.

WHO (2017), Vitamin and Mineral Nutrition Information System (VMNIS); Worldwide prevalence on anemia 1993-2005; 2017

Wolfinger, R., & O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. Journal of statistical Computation and Simulation, 48(3-4), 233-243.

Wolfinger, R. D. (1996). Heterogeneous variance: covariance structures for repeated measures. Journal of Agricultural, Biological, and Environmental Statistics, 205-230.

Wondu, T., & Bijlsma, M. (2012). "The hidden hunger": Understanding the burden of anaemia and its determinants among pregnant and non-pregnant women in Ethiopia. African Journal of Food, Agriculture, Nutrition and Development, 12(7).

World Health Organization. (2001). Iron deficiency anaemia: assessment, prevention and control: a guide for programme managers.

World Health Organization. (2008). Worldwide prevalence of anaemia 1993-2005: WHO global database on anaemia.

World Health Organization. (1993). Prevention and management of severe anaemia in pregnancy: report of a technical working group, Geneva, 20-22 May 1991.

World Health Organization. (2015). The global prevalence of anaemia in 2011.

World Health Organization. (2002). The world health report 2002: reducing risks, promoting healthy life. World Health Organization.

APPENDIX

Table A1: Full model fit for GEE model

Empirical Standard Error Estimates							
Covariates	Categories	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	β_0	-1.0710	0.3466	-1.7504	-0.3917	-3.09	0.0020
Age group	15 – 19	0.0191	0.0689	-0.1158	0.1541	0.28	0.7812
	20 – 24	0.0682	0.1071	-0.1417	0.2780	0.64	0.5244
	25 – 29	-0.0841	0.1366	-0.3519	0.1837	-0.62	0.5381
	30 – 34	-0.1117	0.1669	-0.4388	0.2155	-0.67	0.5035
	35 – 39	-0.1801	0.1616	-0.4968	0.1367	-1.11	0.2652
	40 – 44	-0.3919	0.2380	-0.8583	0.0746	-1.65	0.0996
Residence	Rural	0.0365	0.1843	-0.3248	0.3977	0.20	0.8432
Parity	1 - 2	0.0386	0.0938	-0.1453	0.2225	0.41	0.6808
	3 - 5	0.0512	0.0966	-0.1381	0.2405	0.53	0.5960
	≥ 6	0.1966	0.1394	-0.0766	0.4699	1.41	0.1584
Occupation	Employed	-0.3774	0.1027	-0.5786	-0.1761	-3.68	0.0002
Higher Education Level	Primary	-0.3624	0.0996	-0.5577	-0.1672	-3.64	0.0003
	Secondary	-0.5588	0.1898	-0.9308	-0.1868	-2.94	0.0032
	Higher	-0.3306	0.1855	-0.6941	0.0329	-1.78	0.0747
Marital Status	Single	0.4428	0.1072	0.2327	0.6529	4.13	<.0001
	Married	0.3870	0.1185	0.1549	0.6192	3.27	0.0011
	Windowed	0.6506	0.1713	0.3148	0.9864	3.80	0.0001
	Divorced	0.3289	0.1352	0.0639	0.5938	2.43	0.0150
	Separated	0.4551	0.1648	0.1321	0.7781	2.76	0.0058
Wealth Index	Middle	-0.0930	0.0742	-0.2385	0.0524	-1.25	0.2100
	Rich	-0.1241	0.0805	-0.2818	0.0337	-1.54	0.1232
Contraceptive Use Method	Folkilorc	0.2721	0.8093	-1.3141	1.8583	0.34	0.7367
	Traditional	-0.0202	0.2991	-0.6063	0.5660	-0.07	0.9462
	Modern	-0.7227	0.1153	-0.9487	-0.4967	-6.27	<.0001
Pregnancy	Yes	0.2371	0.1096	0.0223	0.4520	2.16	0.0305
Breast feeding	Yes	-0.0463	0.0413	-0.1273	0.0347	-1.12	0.2628
Smoking	Yes	-0.2549	0.4110	-1.0605	0.5507	-0.62	0.5352
BMI	18.5≤BMI<25	-0.2600	0.0759	-0.4088	-0.1113	-3.43	0.0006
	BMI ≥ 25	-0.3521	0.1239	-0.5950	-0.1092	-2.84	0.0045
HIV	Yes	0.0808	0.1548	-0.2227	0.3842	0.52	0.6019

GEE Fit Criteria	
QIC	14880.6618
QICu	14759.3880

Table A2: Empirical standard error estimate for independent correlation structure of GEE.

Covariates	Categories	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-1.0338	0.0706	-1.1722	-0.8955	-14.65	<.0001
Occupation	Rural	-0.3845	0.0423	-0.4674	-0.3017	-9.10	<.0001
Highest Education Level	Primary	-0.3547	0.0519	-0.4564	-0.2531	-6.84	<.0001
	Secondary	-0.5673	0.1021	-0.7674	-0.3673	-5.56	<.0001
	Higher	-0.3331	0.1178	-0.5639	-0.1023	-2.83	0.0047
Marital Status	Never in union	0.4381	0.0630	0.3147	0.5616	6.96	<.0001
	Married	0.3730	0.1184	0.1409	0.6052	3.15	0.0016
	Windowed	0.5608	0.1151	0.3352	0.7865	4.87	<.0001
	Divorced	0.3019	0.1042	0.0976	0.5062	2.90	0.0038
	Separated	0.4372	0.1474	0.1483	0.7261	2.97	0.0030
Wealth Index	Middle	-0.0984	0.0626	-0.2212	0.0243	-1.57	0.1161
	Rich	-0.1459	0.0507	-0.2453	-0.0466	-2.88	0.0040
Contraceptive Use Method	Folkloric	0.2081	0.8469	-1.4518	1.8679	0.25	0.8059
	Traditional	-0.0104	0.2403	-0.4813	0.4606	-0.04	0.9656
	Modern	-0.7000	0.0701	-0.8373	-0.5627	-9.99	<.0001
Pregnancy	Yes	0.2880	0.0712	0.1484	0.4275	4.04	<.0001
BMI	$18.5 \leq BMI < 25$	-0.2519	0.0461	-0.3423	-0.1616	-5.47	<.0001
	$BMI \geq 25$	-0.3785	0.0950	-0.5646	-0.1924	-3.99	<.0001

Table A3: Model based standard error estimate for independent correlation of GEE model.

Covariates	Categories	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-1.0338	0.2315	-1.4876	-0.5800	-4.47	<.0001
Occupation	Rural	-0.3845	0.1083	-0.5969	-0.1722	-3.55	0.0004
Highest Education Level	Primary	-0.3547	0.0820	-0.5155	-0.1939	-4.32	<.0001
	Secondary	-0.5673	0.1440	-0.8495	-0.2852	-3.94	<.0001
	Higher	-0.3331	0.1637	-0.6539	-0.0123	-2.04	0.0418
Marital Status	Never in union	0.4381	0.0648	0.3111	0.5652	6.76	<.0001
	Married	0.3730	0.1016	0.1739	0.5721	3.67	0.0002
	Windowed	0.5608	0.1859	0.1965	0.9252	3.02	0.0026
	Divorced	0.3019	0.1074	0.0914	0.5124	2.81	0.0049
	Separated	0.4372	0.1277	0.1869	0.6874	3.42	0.0006
Wealth Index	Middle	-0.0984	0.0791	-0.2535	0.0566	-1.24	0.2135
	Rich	-0.1459	0.0714	-0.2859	-0.0060	-2.04	0.0410
Contraceptive Use Method	Folkloric	0.2081	0.7925	-1.3452	1.7614	0.26	0.7929
	Traditional	-0.0104	0.3022	-0.6027	0.5819	-0.03	0.9726
	Modern	-0.7000	0.1116	-0.9188	-0.4812	-6.27	<.0001
Pregnancy	Yes	0.2880	0.0986	0.0947	0.4812	2.92	0.0035
BMI	$18.5 \leq BMI < 25$	-0.2519	0.0748	-0.3985	-0.1053	-3.37	0.0008
	$BMI \geq 25$	-0.3785	0.1287	-0.6307	-0.1263	-2.94	0.0033

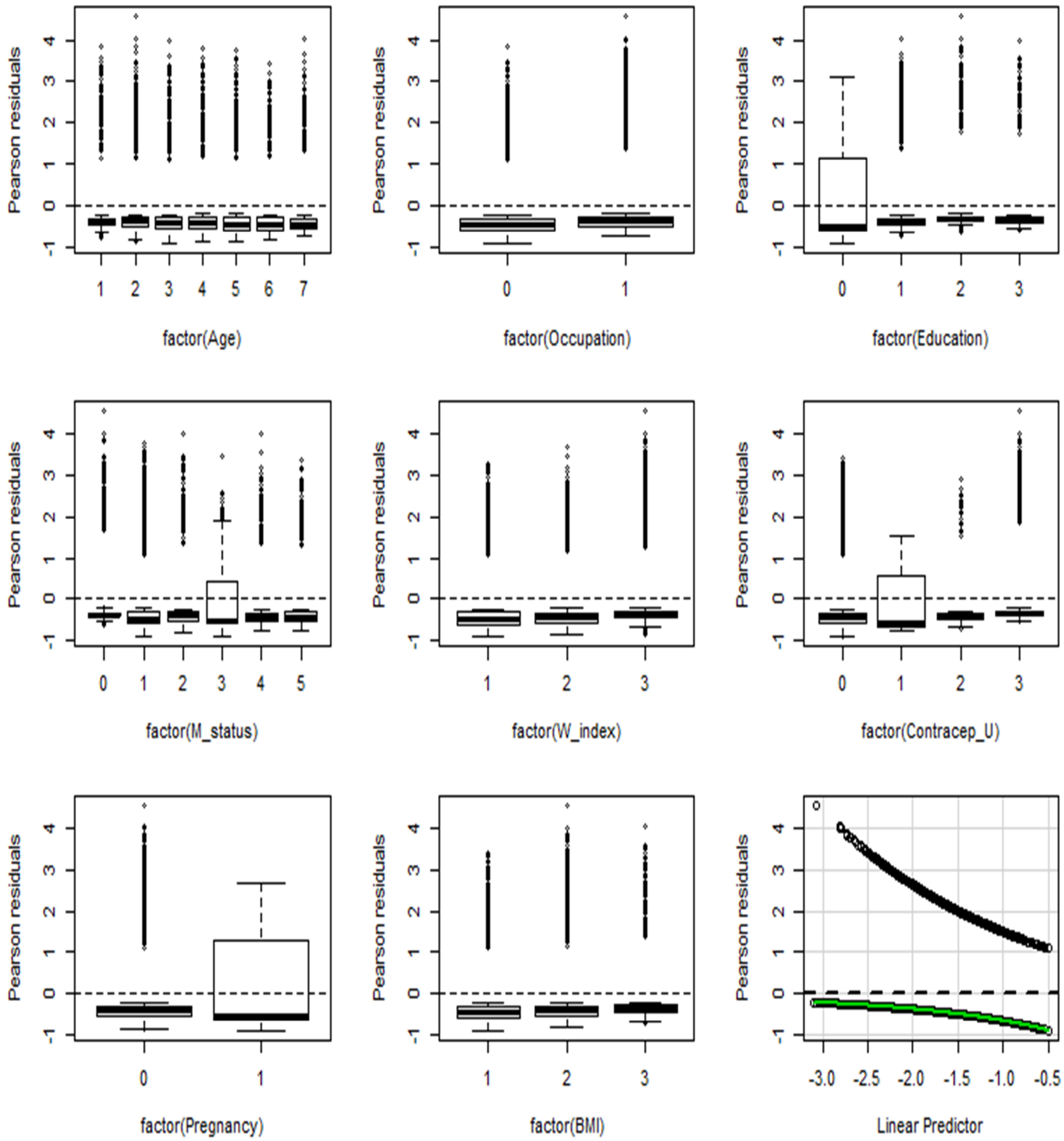


Fig A7: Plots of residual versus each categorical predictor

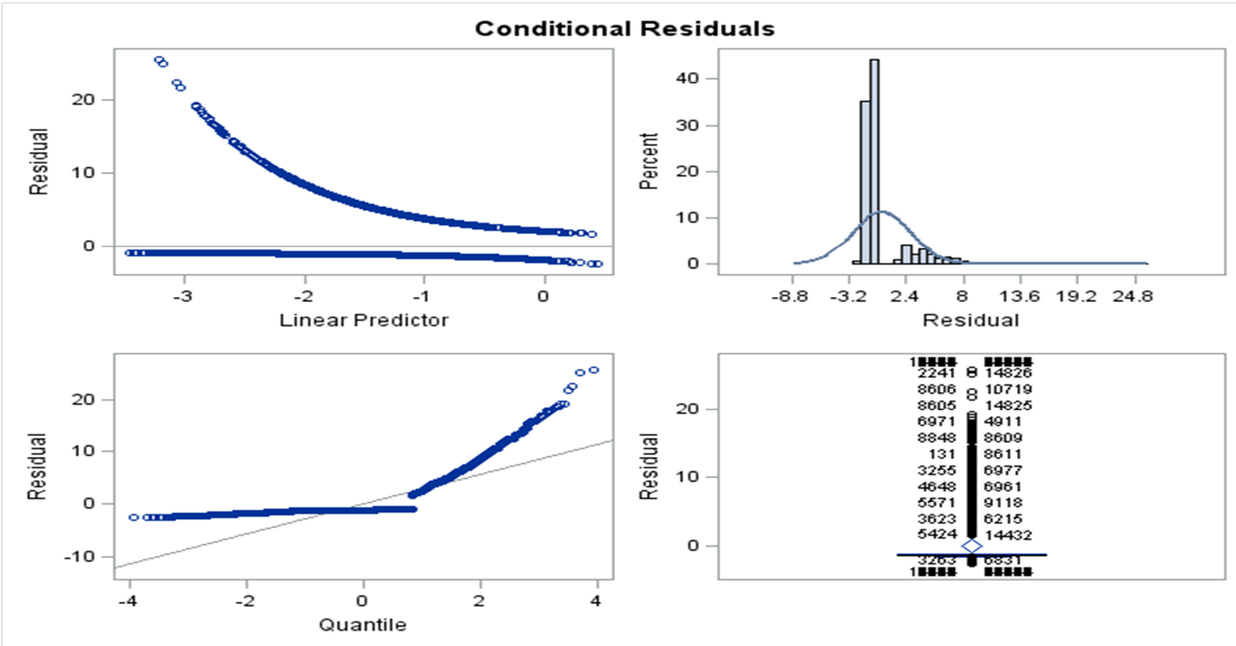


Fig A8: Conditional Residuals of GLMM.

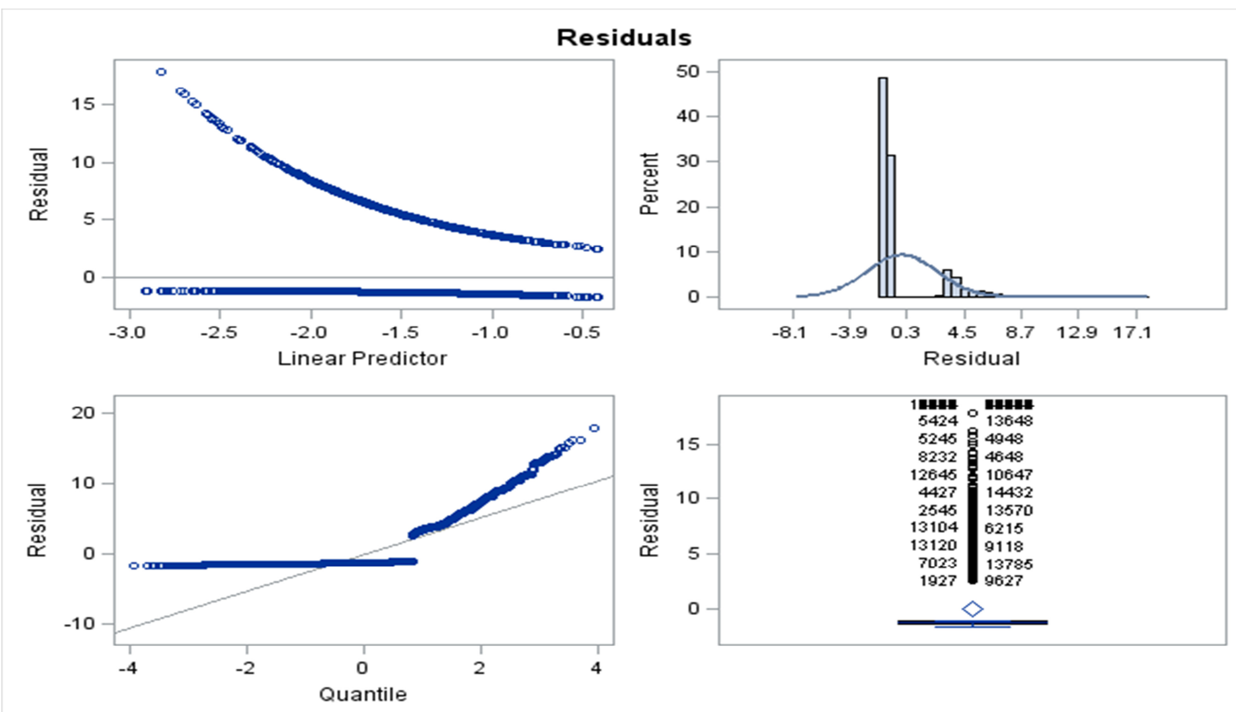


Fig A9: Marginal Residual of GLMM.