KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, KUMASI, GHANA

Factors Explaining the Observed Pattern in Under-Five Mortality in the Ghana Demographic and Health Survey 2008

By

OPOKU GYABAAH (B.Sc.)

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FUFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF M.PHIL MATHEMATICAL STATISTICS

May 28, 2014

Declaration

I hereby declare that this submission is my own work towards the award of the M. Phil degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgement had been made in the text.

Opoku Gyabaah (PG8065312) ....

Student Signature Date

Certified by:

Nana Kena Frempong ....

Supervisor Signature Date

Certified by:

Prof. S. K. Amponsah ....

Head of Department Signature Date

Dedication

This work is dedicated to my late mum for her love and care

Abstract

The study examined the risk factors of under-five mortality in the Ghana Demographic and Health Survey (DHS) 2008 data. The risk factors were investigated statistically by fitting a logistic regression model and controlling for household effects. The data was analyzed using the SAS 9.1 version for windows. Three factors (age of respondent, number of people in the household and ethnicity) were found to be the significant risk factors of under-five mortality in the Ghana DHS 2008 data. The age of respondent was positively related to the risk of under-five mortality while the number of people in the household was negatively related to the risk of under-five mortality. Respondents from the Akan, the Ewe, the Gruma and the Mande ethnic groups were significant in terms of the effect of ethnicity on under-five mortality. The risk of under-five mortality is higher in the Gruma and the Mande ethnic groups as compared to the ”other” ethnic group while it was lower in the Akan and the Ewe ethnic groups as compared to the ”other” ethnic group. It was also found that, household effect was not a significant contributing risk factor for under-five mortality in the Ghana DHS 2008 data.

Acknowledgement

I am most grateful to the Almighty God for His guidance in the realization of this thesis. I also wish to express my heartfelt appreciation to my supervisor, Nana Kena Frempong for his mentorship and tutelage throughout the period.

It is my pleasure to thank Mr. Emmanuel Nakuah, lecturer at the School of Medical Sciences, KNUST and Mr. Emmanuel Harris, lecturer at the Department of Mathematics, KNUST for their support during the research period.

Finally, I would like to appreciate the support of my colleagues; Daniel Yeboah, Daniel Lipi and Mr. Peter Kwasi Sarpong for their encouragement and assistance.

Contents

List of Abbreviations

ALR	Alternating Logistic Regression
DHS	Demographic and Health Survey
GEE	Generalized Estimating Equation
GHS	Ghana Health Service
GLM	Generalized Linear Model
GOG	Government of Ghana
GSS	Ghana Statistical Service
MDG	Millennium Development Goal
NDPC	National Development Planning Commission
NGO	Non Governmental Organization
NMIMR	Noguchi Memorial Institute for Medical Research
UNICEF	United Nations International Children's Emergency Fund
UN	United Nations
USAID	United States Agency for International Development

List of Tables

Chapter 1

Introduction

1.1 Background to the Study

Under-five mortality rate is one of the key indicators for monitoring the progress of a country towards global goals such as the Millennium Development Goals (MDGs). The estimate for under-five mortality is a measure of child survival in every country and the world as a whole. The estimate also reflects the social, economic, health and environmental conditions in which children live. As part of the MDGs for health, nations pledged in 2000 to ensure a two-thirds reduction in under-five mortality between 1990 and 2015 (UNICEF, 2013a). The pledge to reduce under-five mortality by two-thirds was specifically labeled by the United Nations (UN) as the MDG 4. There is a substantial effort by governments, the UN, the private sector and civil society groups towards achieving the MDGs including the pledge to reduce under-five mortality by two-thirds.

According to the UN/MDG (2012), under-five mortality is higher in Sub-Saharan Africa and Southern Asia as compared to other parts of the world. Although the progress to reduce under-five mortality in these developing parts of the world has accelerated, it is still unlikely to meet the MDG 4 target by 2015. The report continues that, some parts of the world including Northern Africa have already achieved the MGD 4 target by bringing down the child mortality rate by 67 percent and Eastern Asia is close with a 63 percent decline. Sub-Saharan Africa has achieved reductions of only around 30 percent, which is less than half of what is required to reach the 2015 target. Although under-five mortality is declining worldwide, stakeholders are still anxious on whether they would be able to meet the 2015 deadline. If the current trends continue, the world would not be able to meet the MDG 4 target of reducing under-five mortality by two-thirds until 2028 (UNICEF, 2013a).

In Ghana, rates of mortality among children continue to decline gradually (NDPC/GOG, 2012) which follows the global trend. Although evidence shows that there has been a significant reduction in under-five mortality rates in Ghana over the past years (GSS/ICF, 2010), it is unlikely that the 2015 target of reducing the under-five mortality rates by two-thirds would be met unless the coverage of effective child survival interventions is vigorously increased. The GSS/GHS/ICF (2009) report shows only a 30 percent reduction in the under-five mortality rate in Ghana. The UNICEF (2013a) placed Ghana at the 36th position among over 190 countries in a descending order of the under-five mortality rate league table 2012 with under-five mortality rate of 72 deaths per 1,000 live births in 2012. This clearly shows that, there is still much work to do in order to meet the MDG 4 target in Ghana. Hence, there should be a concerted effort from all stakeholders to help reduce under-five mortality drastically.

The results of childhood mortality estimates including under-five mortality are used as a basis for formulating policies on the growth and development of children. Improvement in child survival to meet the MDG 4 thereby reducing childhood mortality as a whole has been one of the major considerations by policy formulators worldwide. Government institutions, donor agencies, civil society groups and all other stakeholders are interested in reducing under-five mortality which is one of the key indicators of the quality of the nation's health system. This is because the future of every nation depends on its young population. For example, high childhood mortality rate has the implication for a rapid population growth because parents who fear that their newborn babies would not survive to older ages are likely to have more children (Agha, 2000). Trends in the under-five mortality rate can be used to track the overall development as well as specific successes and failures in the implementation of child survival interventions over time.

The Ghana Demographic and Health Survey is a national survey which is designed to collect data on population and health as part of the global Demographic and Health Survey (DHS). The Ghana DHS started in 1988 and there have been five successive surveys since 1988. Childhood mortality estimates in the Ghana DHS 2008 report measure the risk of a newborn dying from birth through to the age of five years. Childhood mortality has been categorized as:

- Neonatal mortality : the probability of a newborn baby dying between birth and the first month of life
- Post neonatal mortality : the difference between infant mortality and neonatal mortality
- Infant mortality : the probability of a newborn baby dying between birth and exact age one (12 months)
- Child mortality : the probability of a baby dying between exact age one (12 months) and five (60 months)
- Under-five mortality : the probability of a newborn baby dying between birth and exact age five (60 months)

(GSS/GHS/ICF, 2009).

Results from all the five successive Ghana DHS since 1988 show a consistent decrease in all the childhood mortality estimates in Ghana including under-five mortality. According to the GSS/GHS/ICF (2009), the under-five mortality rate in Ghana as at 2008 is 80 deaths per 1000 live births. It means that one in every thirteen Ghanaian children died before reaching five years old. This is an improvement in the mortality estimates for the past five surveys where the under-five mortality rates of 111 deaths per 1000 live births, 108 deaths per 1000 live births, 119 deaths per 1000 live births and 155 deaths per 1000 live births were recorded for the 2003, 1998, 1993 and 1988 Ghana DHS respectively [(GSS/MI, 1989), (GSS/MI, 1994), (GSS/MI, 1999), (GSS/NMIMR/ORC, 2004)].

There have been a lot of research works on under-five mortality in the past years. Most of these investigations make use of the logistic regression as the method for determining the factors that affect the probability of a child dying before five years old due to the binary nature of the outcome variable. Several other methods have been discussed in literature for the analysis of such binary outcomes. Among these alternatives for the logistic regression includes the Cox regression, Poisson regression, Log-binomial regression among others. The authors that contributed to the discussion on the alternatives for logistic regression in cross sectional studies have not reached a conclusion on which method is the most appropriate approach and only a few publications make some comparison among the available alternatives (Barros and Hirakata, 2003).

The logistic regression is used to analyze the relationship between predictor variables and a response variable that is dichotomous in nature (Reed and Wu, 2013). In fact, this type of regression has become very popular and it has increasingly become a statistical tool that is employed for the analysis of a binary response variable by many researchers across different study disciplines. The logistic regression is widely used in developing models in study areas such as the physical sciences, economics, political science, medicine, epidemiological studies among many others. The major advantage of the logistic regression over ordinary regression is its ability to deal with variables that are dichotomous and categorical in nature.

In spite of the several advantages for the use of the logistic regression, one important concern that has been discussed in literature is the relationship between the sample size and the calculation of the odds ratio. The size of the odds ratio that is derived from the logistic regression is inversely related to the sample size. This means that, as the sample size decreases the bias in the odds ratio becomes larger and thus, it gives a poor estimate of the actual population effect (Reed and Wu, 2013). This problem of overestimation of the odds ratio is minimized when the sample size is greater than about 500 (Kojo, 2012).

Clustered data arise in a wide variety of applications including demographic surveys where samples are arranged in a hierarchy. The Ghana DHS 2008 data as an example contains clusters in the various households where respondents were selected and interviewed. Such clustered data are often correlated which violates the statistical assumption of independence of observations in the case of the logistic regression method. A classic example in the Ghana DHS 2008 data describes respondents within the same household where there is the likelihood for these respondents within the same household to share similar characteristics which may affect the survival of their children. Statistical models for clustered data must account for the intra cluster correlation at each level analysis else it could lead to misleading inferences. If the intra cluster correlation is not properly accounted for in the analyses, standard errors of the parameters may be biased (Ghisletta and Spini, 2004). There are several discussions in literature on time dependent correlated data but very little has been done on the options that are available when these correlations are not time dependent (Staley, 2013) as in the case of the Ghana DHS 2008 data.

One method which allows researchers to analyze correlated data is the Generalized Estimating Equations (GEE). This method is an extension of the Generalized Linear Models (GLMs) to accommodate correlated data with its main advantage in the unbiased estimation of regression coefficients despite a possible misspecification of the structure of the correlation (Ghisletta and Spini, 2004). The GEE is particularly effective for modeling clustered binary or count data (Wang, 2010). The two basic advantages for the GEE method when dealing with correlated data are as follows. Firstly, it could be used regardless of the nature of the response variable whether it is continuous, dichotomous, polychotomous, ordinal or an event-count. Secondly, the method allows for a variety of specifications of the correlation patterns within the clusters (Zorn, 2001). One attractive property of the Generalized Estimating Equations is that, one can use a working correlation structure that may be specified wrongly while the resulting regression coefficient estimate is still consistent and asymptotically normal (Pan and Connett, 2002).

1.2 Statement of the Problem

Ghana is a signatory to the Millennium Declaration in 2000 with the aim of meeting all the eight Millennium Development Goals including the reduction of under-five mortality by two-thirds before the 2015 deadline. Although the country has made a substantial progress over the past years in achieving these goals, there are still many areas to improve such as the MDG 4. The attention and effort to meet these goals is a matter of priority to every government and the success to achieve the goals depends greatly on commitment and the available resources in the country. Ghana is doing very well in other areas but there is still anxiety on whether the country would meet the MDG 4 by the 2015 deadline even when there is a decline in under-five mortality in the past years.

Although under-five mortality estimates continue to decline in Ghana over the years, it is still a matter of great concern among stakeholders as the decline is not enough to reduce under-five mortality by two-thirds by 2015. Several studies have concluded a decline in under-five mortality in Ghana but there is not much investigation into the extent to which the proposed factors have an effect on the decline. Some parts of the country seem to have considerably low under-five mortality rates while it is the opposite in other places. Also, under-five mortality estimates vary with respect to the various households in the Ghana DHS 2008 data.

It has therefore become necessary to examine the factors that influence under-five mortality estimates and investigate the extent to which the proposed factors have an effect on the decline of under-five mortality. It is also important to find out whether there exist household effects on under-five mortality by considering the clusters in the various households found in the Ghana DHS 2008 data. This is because, respondents within the same household may be related and this could have some effects on parameter estimates produced by the binary logistic regression method. If the within cluster correlation is not accounted for in the analysis, the standard errors of the parameters may be biased which will lead to a misleading inference when the biased parameters are used for the interpretation of results.

In summary, under-five mortality is a matter of great concern to all stakeholders in Ghana being the government, civil society groups, parents, families and international organizations. Due to this important viewpoint, any factor that affects under-five mortality in Ghana poses a greater concern of which there is a greater interest to investigate in order to determine the cause. It is in the light of these, that there is the need for a deeper research to investigate the factors that affect under-five mortality in the various households of the Ghana DHS 2008 data so that more effective measures would be taken to help curb this problem.

1.3 Objectives of the study

1.3.1 General Objective

The general objective of this study is to investigate the risk factors and their effects on under-five mortality in Ghana.

1.3.2 Specific Objectives

Specifically, the study aims to:

- identify the significant factors that affect under-five mortality by fitting a logistic regression model
- investigate the effect of households as a cluster on under-five mortality by considering GEE1 and alternating logistic regression.

1.4 Methodology

The study uses a secondary data from the Women's Questionnaire of the Ghana DHS 2008 which is a national survey covering all the ten regions of the country. The survey was carried out by the Ghana Statistical Service (GSS) and the Ghana Health Service (GHS) with support from the ICF Macro which provided financial and technical assistance through the USAID-funded MEASURE DHS programme. The survey was designed to assist developing countries like Ghana to collect data on fertility, family planning, maternal and child health among others.

According to the GSS/GHS/ICF (2009), data for the survey was collated through the use of a "Verbal Autopsy Questionnaire". In half of the households selected for the survey, all eligible women of age 15-49 were interviewed with the Women's Questionnaire. A total of 5,096 eligible women of age 15-49 were identified and interviews were conducted for 4,916 of these women due to unavailability of some individuals at home despite repeated visits to the households by enumeration officials. Data collection took place over a three-month period from early September, 2008 to late November, 2008.

The analysis of the data is based on the ever-born women of reproductive age 15-49 years who have had at least one live birth in history. Since the general objective of this study is to investigate the effects of socioeconomic and demographic variables on the level of under-five mortality in Ghana, the dependent variable is the number of children who had died before age of five years (0 to 59 months). The logistic regression model is fitted to identify significant factors and their effects on under-five mortality in Ghana. In addition, the alternating logistic regression, a GEE method is used to investigate the effects of the factors that affect under-five mortality by considering clusters in the various household of the Ghana DHS 2008 data.

The SAS statistical software version 9.1 for windows was used for the analysis of the data.

1.5 Justification of the Work

The research findings will enable stakeholders to have much insight into the significant factors of under-five mortality in Ghana in order to reduce the scope of intervention policies as a way to accelerate the success of achieving the MDG4. It is also amazing to note that, this study will serve as a guide in formulation and implementation of child survival policies in the country. In addition to these, the outcome of this study would provide an additional source of information to the literature on under-five mortality in Ghana.

The study is very significant because, it uses the descriptive information from the Ghana DHS 2008 report to draw conclusions about the general population of Ghanaian children based on the information obtained from the sample. In this regard, the research findings will be a contribution to knowledge and would be a basis for further research into under-five mortality in Ghana and the world as a whole.

1.6 Limitations of the Study

Like any research work, this study is not without limitations. The Ghana DHS 2008 report outlined only the estimates of the deaths of under-five children without the causes of these deaths. Furthermore, the issue of missing information in data resulted in the dropping of some respondents from the final data thereby decreasing the sample size used for the final analysis.

1.7 Organization of the Thesis

The thesis consists of five chapters. Chapter one gives a general background of the thesis and how it was carried out. It includes a general introduction and trends of under-five mortality in Ghana and the world as whole, statement of the problem under study, objectives of the study, methodology, justification and limitations of the study. In chapter two, selected research works that are related to under-five mortality, the logistic regression and the alternating logistic regression are reviewed. Chapter three discusses the data and an in depth explanation of the methods used. In chapter four the data is explored, analyzed and discussed using the SAS statistical software to get results. The last chapter presents the conclusions and recommendations of the study.

Chapter 2

Literature Review

2.1 Global Analysis of Under-Five Mortality

The issue of child survival over the past years is one of significant progress worldwide including Ghana and it is an unfinished business so far as under-five mortality targets are to be achieved universally. According to the UNICEF (2013a), there is much to celebrate as the world is currently reducing under-five deaths faster than at any other period in the past twenty years. The global number of under-five deaths has roughly halved from 90 deaths per 1000 live births in 1990 to 48 deaths per 1000 live births in 2012. The estimated annual rate of under-five deaths has reduced from 12.6 million in 1990 to 6.6 million in 2012. This means that, 17,000 fewer under-five children die each day in 2012 than in 1990. The report continues that, all regions have shown steady reduction in childhood mortality estimates over the past two decades. In fact, the progress in reducing childhood mortality has accelerated with the annual rate of decline in the worldwide under-five mortality rate increasing from 1.2 percent in 1990-1995 to 3.9 percent in 2000-2011. Under-five mortality according to the report is hugely concentrated in Sub-Saharan Africa because all regions with the exception of Sub-Saharan Africa have halved their under-five mortality rates since 1990 (UNICEF, 2013b).

Furthermore, it is evident in the UNICEF (2012a) that, emerging figures have shown alarming disparities in childhood mortality estimates at the subnational levels in many countries across the world. The UNICEF analysis of the International Household Survey Data shows that children born to poorest homes are almost twice as likely to die before they reach five years old as compared to their counterparts in wealthy homes. In fact, the under-five mortality rate based on 39 countries analyzed by the UNICEF using wealth quintiles revealed the following: 121 deaths per 1000 live births in the poorest wealth quintile, 114 deaths per 1000 live births in the second wealth quintile, 101 deaths per 1000 live births in the middle wealth quintile, 90 deaths per 1000 live births in the fourth wealth quintile and 62 deaths per 1000 live births in the richest wealth quintile. It has also been revealed that, poverty is not the only determinant of under-five mortality worldwide (UNICEF, 2012b). It is evident in the report that, children in rural areas have a greater risk of dying than those in the urban areas. In a similar analysis of 45 countries by UNICEF, there were 114 deaths and 67 deaths per 1000 live births in rural and urban areas respectively.

The UN/MDG (2012) also corroborates the progress in reducing under-five mortality worldwide. From the report, under-five mortality rate declined by 35 percent from 97 deaths per 1,000 live births in 1990 to 63 deaths per 1,000 live births in 2010 in developing countries. The progress in reducing childhood mortality is gaining momentum but despite the determined progress, a greater proportion of the deaths occur in Sub-Saharan Africa. The report states that, childhood mortality is more likely to ”strike children in rural areas” than those in urban areas based on the analysis of 82 developing countries with data on under-five mortality rate by residence which accounts for 75 percent of the total births in developing countries in 2010. Also, analysis of 73 developing countries with data on under-five mortality rate by household wealth quintiles accounting for 71 percent of total births in developing countries in 2010 shows that, ”children born into poverty are almost twice as likely to die” before they reach five years as those from wealthy families. In addition to these, mother’s access to education is a ”survival factor for under-fives”. This is evident in the analysis of 78 developing countries with data on under-five mortality rates by mother’s education, accounting for 75 percent of total births in developing countries in 2010 (UN/MDG, 2012).

According to the NDPC/GOG (2012), childhood mortality estimates have witnessed considerable progress globally with an under-five mortality declining by a third from 89 deaths per 1,000 live births to 60 deaths per 1,000 live births between 1990 and 2009. Like the other reports discussed earlier, under-five mortality rate still remains highest in Sub-Saharan Africa with 129 deaths per 1,000 live births in 2009, followed by South Asia with 122 deaths per 1,000 live births. The report continues that, the global trend of decline in childhood mortality is not different from the estimates in Ghana where the under-five mortality rate is 80 deaths per 1,000 live births, which is an improvement from the previous years. After a successive decline from 122 deaths per 1,000 live births in 1990 to 98 deaths per 1,000 live births in 1998, the under-five mortality rate stagnated at 111 deaths per 1,000 live births during the period of 2003 and 2008.

2.2 The Ghana Demographic and Health Survey

The Ghana Demographic and Health Survey is the major source of data for analyzing childhood mortality including under-five mortality in Ghana since the past twenty years. The Ghana DHS 2008 is the fifth Demographic and Health Survey to be undertaken in Ghana since 1988. All mortality rates in the Ghana DHS 2008 report are expressed as ”per 1000 live births” with the exception of child mortality which is expressed as ”per 1000 children surviving to the age of 12 months”. Results from the five successive Ghana DHS surveys showed a marked decline in childhood mortality over the past twenty years. This decline appears to have halted briefly during the period 1999-2003 and continued further between 2003 and 2008 (GSS/GHS/ICF, 2009).

As an example, the under-five mortality rate decreased from 111 deaths per 1,000 live births for the period 0-4 years preceding the 2003 GDHS to 80 deaths per 1,000 live births in 2008. This means that as at 2008, one in every thirteen Ghanaian children died before the fifth birthday while one in every nine Ghanaian children died before the fifth birthday by 2003. Under-five mortality estimates as at 1998 was 108 deaths per 1,000 live births and 119 deaths per 1,000 live births for the period preceding 1993 [GSS/NMIMR/ORC (2004), GSS/MI (1999), GSS/MI (1994)].

According to the GSS/GHS/ICF (2009), several factors determine the level of under-five mortality in Ghana. Among these factors are socioeconomic variables which include type of residence, region of residence, mother's education and household wealth status (quintile). Mortality levels in rural areas are consistently higher than those in urban areas. In the ten year period before the 2008 survey, the under-five mortality rate was 90 deaths per 1,000 live births in rural areas and 75 deaths per 1,000 live births in urban areas. Furthermore, childhood mortality as evident in the 2008 GDHS report shows different estimates for the various regions in Ghana with some regions having considerably low rates while others have high rates. For example, the under-five mortality estimate ranges from a lowest rate of 50 deaths per 1,000 live births in the Greater Accra region to a highest rate of 142 deaths per 1,000 live births in the Upper West region. The report continues that, the education of the mother is inversely related to the risk of death of her child. This means that, children born to mothers with higher educational levels have lower risk of dying and vice versa. The under-five mortality rate among children of mothers with no education was 102 deaths per 1,000 live births which was substantially higher than the under-five mortality among children of mothers with primary, middle/JSS and above secondary education levels with rates of 88 deaths, 68 deaths and 64 deaths per 1,000 live births respectively (GSS/GHS/ICF, 2009).

In addition to the socio economic factors, the income level of the household played a major role in determining under-five mortality estimates in Ghana. Children in households which are in the highest wealth quintile have the lowest mortality rate of 60 deaths per 1000 live births for under-five mortality. On the contrary, children in households which are in the lowest, second, middle and fourth wealth quintiles recorded under-five mortality estimates of 103 deaths, 79 deaths, 102 deaths and 68 deaths per 1,000 live births respectively.

The demographic factors that are strongly related to childhood mortality in Ghana as revealed by the Ghana DHS 2008 report include sex of the child, age of mother at birth, birth order, previous birth interval and size of child at birth. Childhood mortality is generally higher for males than females in all the categories. The under-five mortality rate for male children was 93 deaths per 1,000 live births and that of female children was 76 deaths per 1,000 live births. Moreover, findings from the survey shows that, children born to young mothers who are under 20 years old and older mothers of more than 35 years are at a higher risk of dying. The results of the Ghana DHS 2008 report confirm this assertion with under-five mortality for children born to mothers under twenty years old given as 109 deaths per 1,000 live births as compared to 73 deaths per 1,000 live births and 95 deaths per 1,000 live births for children born to mothers between 20-29 years and 30-39 years respectively.

Children of first order and higher-order births have a greater risk of dying with under-five mortality rates of 110 deaths per 1000 live births for birth order above 7 and 84 deaths per 1000 live births for first birth order. Mortality among children as reported in the Ghana DHS 2008 report is negatively related to the length of previous birth interval. It is evident from the report that, children born with lesser previous birth intervals have higher mortality rates in all the childhood mortality categories. As an example, the under-five mortality rates were 131 deaths, 86 deaths, 83 deaths and 58 deaths per 1000 live births for children born with previous birth interval less than 2 years, 2 years, 3 years and more than 3 years respectively. Lastly, a baby's size at birth has often been found as an important factor which affects the chance of his or her survival. According to the Ghana DHS 2008 report, majority of births in Ghana occur outside of a health facility and most of these babies are not weighed. The Ghana DHS 2008 survey however used the mothers' assessment of the size of the baby as small/very small or average/larger at birth as a proxy for birth weight. The rate for under-five mortality was given as an asterisk which indicates that "the rate is based on fewer than 250 unweighted exposed persons and has been suppressed". Nevertheless, the mortality rate was higher in the remaining two categories (infant and neonatal mortalities) which were not suppressed. The infant mortality rates were 84 deaths and 42 deaths per 1000 live births for children born with small/very and average/larger birth sizes respectively. The neonatal mortality rate was 49 deaths per 1000 live births for children born with small/very small birth size and 25 deaths per 1000 live births for children born with average/larger birth size (GSS/GHS/ICF, 2009).

2.3 Review of Literature on Study Variables

The estimates for childhood mortality including under-five mortality have changed over the years. In this regard, there is a continuous research year after year to investigate this phenomenon in order to determine the risk factors that affect this problem and suggest possible ways of reducing it to the barest minimum. Several research works have investigated and proposed the factors that affect childhood mortality in Ghana and the world as a whole. Examples of the risk factors which have been discussed in literature include maternal education, mother's age, occupation of parents, type of residence, region of residence, wealth quintile or income level of the household, size of the family and sanitation facility. Others include sex of the baby, birth weight of the baby, birth order, previous birth interval, antenatal care, breastfeeding and immunization.

Folasade (2000) did an investigation to determine the relative significance of environmental and maternal factors on childhood mortality in two contrasting towns in southwestern Nigeria. While most studies have focused only on the effect of maternal factors on childhood mortality on one hand, others have suggested the effects of only environmental factors. Folasade (2000) therefore integrated the separate influence of environmental and maternal factors on childhood mortality. The conclusion was that, domestic environmental conditions were stronger predictors of child mortality in the more developed study town of Ota than in the more traditional town of Iseyin. However in both towns, the age of mother at marriage, the age of mother at first childbirth and parity were statistically significant in the more urban center and generally remains inconsistent in its relationship with child mortality. It was continued in the analysis that, child mortality rates continued to be a function of an environmental factor which could either be the source of drinking water and a child care behavior factor or where the child was kept when the mother was at work especially in the market environment.

Lin (2006) in ”The effects of economic instability on infant, neonatal and post neonatal mortality rates: Evidence from Taiwan” used a data set comprising 23 cities in Taiwan for the years 1979-2002 and a fixed-effects model to find evidence of the effect of economic instability on infant, neonatal and post neonatal mortality rates. In addition, the effects of income, demographic factors and the availability of medical resources were also examined in relation to childhood mortality. The most important finding in the work was that, infant, neonatal and post neonatal mortality rates move counter-cyclically with the city unemployment rate in Taiwan. Also, the provision of national health insurance was found to have a positive impact on the health of infants in Taiwan. Furthermore, it was found that, the impact of economic instability on the infant, neonatal and post neonatal mortality rates was found to be the strongest in the eastern part of Taiwan where there are few health care resources.

Agha (2000) also investigated the determinants of infant mortality in Pakistan. The study examined the factors which are associated with the survival of infants in Pakistan using a data from the Pakistan Integrated Household Survey 1991. The study stated that, the infant mortality rate was still very high in Pakistan until the early 1990s at 100 deaths per 1000 live births. It showed that, there is no evidence of a secular decline in infant mortality during the

1980s. It was therefore concluded that, the underlying cause of the stagnation of infant mortality in Pakistan is due to the large differentials in infant survival by socio-economic factors, access to water and sanitation. The low social, economic and legal status of women is intimately linked to the well-being of their children. Agha (2000) therefore recommended for a design of health intervention policies in Pakistan to reach the most under-served women and children. The investigation also added that, there should be a systematic evaluation of the health intervention policies in order to make informed decisions about the health investments in the future.

Suwal (2001) in "The main determinants of infant mortality in Nepal" investigated the factors that contribute to either low or high rate of infant mortality in Nepal. The study used data from the 1991 Demographic Health Survey of Nepal. It was concluded after fitting a logistic regression model that, among the entire variables used for the analysis in the study, place of residence, parity, immunization and ethnicity were the most influential risk factors of infant mortality in Nepal.

Sastry (1997) in "What Explains Rural-Urban Differentials in Child Mortality in Brazil" analyzed the differentials in child survival by rural-urban place of residence in Brazil. The work also examined the hypothesis that observed mortality differentials by place of residence are merely manifestations of the underlying differences in socioeconomic status, demographic and reproductive behavior. The study used data from the 1986 Demographic and Health Survey of Brazil and supplementary community level variables obtained from a database assembled by the Brazilian Federal Statistical Agency. The conclusion was that, child mortality was substantially and significantly lower in urban areas of Brazil. The results however suggested that, the urban advantage does not simply reflect the differences in socioeconomic and behavioral characteristics at the individual and household levels but rather community variables appear to play an independent and important role. It was also found that, the effects of community characteristics on child survival are moderated by household socioeconomic factors especially maternal education. Differences in socioeconomic characteristics are therefore important in explaining rural-urban child mortality differentials which has not been hypothesized by previous researchers.

Rahman and Sarkar (2009) in "Determinants of Infant and Child Mortality in Bangladesh" investigated the mortality of children under five years old using information from women's birth histories pertaining to children born during the 10 years before the 1999-2000 Bangladesh Demographic and Health Survey. The work specifically provided information on the levels, trends and differentials in neonatal, post-neonatal, infant and child mortality and assessed the effects of socio-economic, demographic and mother's health-care characteristics on infant and child mortality. They concluded that residence, education of father and mother, preceding birth interval, family size, toilet facility, delivery place and antenatal care are the major contributors of infant and child mortality.

Mondal et al. (2009) in "Factors Influencing Infant and Child Mortality: A Case Study of Rajshahi District, Bangladesh" examined the factors that influence infant and child mortality of suburban and rural areas of the Rajshahi District in Bangladesh. Their results revealed that, the most significant predictors of neonatal, post-neonatal and child mortality levels are immunization, ever breastfeeding, mother's age at birth and birth interval. Again, the risk of child mortality was 78.20% lower among the immunized child than never immunized child. Also, the risk of neonatal mortality was 57.70% lower after a birth interval of 36 months and above than under 18 months. However, parents' education, toilet facilities and treatment places are significant predictors during neonatal and childhood period but father's occupation is significant at post-neonatal periods. For instance, the risk of neonatal mortality is 31.40% lower among the women having primary education and 52.30% lower among the women having secondary and higher education than those having no education. They also observed that, the risk of child mortality was 32.00 lower among the household having hygienic toilet facility than those who do not have such facilities. Also, the risk of child mortality decreased with increased female education and wider access to safe treatment places. It was their recommendation that, attention should be given to female education and expansion of public health systems for reducing the risk of infant and child mortality.

Foggin et al. (2001) also investigated the risk factors and child mortality among the Miao in Yunnan, Southwest China. They stated that, environmental factors and lifestyle of communities in developing countries as in the industrialized world have a great impact on their health status. The study revealed important links between child mortality and a number of risk factors among the Miao of Yunnan Province in Southwest China. These factors include lifestyle variables such as geographic mobility, age of weaning and religious belief. In addition to these, the use of available health care facilities was another explanatory factor. Although the research did not reveal a significant statistical relationship between traditional practices and child mortality, the authors have observed qualitatively that birth customs play an important role in explaining the perinatal component of child mortality.

In addition to these studies done outside Ghana, Osei-Kwakye et. al. (2010) investigated the determinants of under-five mortality in the Builsa District in Upper East Region of Ghana. A case control study was used to collect data from mothers of 60 cases and 120 controls matched for age, sex and place of residence. According to the studies, even though 70% of the mothers were illiterates, the educational level of mothers did not influence the child's risk of death. Children of mothers who had had previous child deaths were about 8 times more likely to die while those who had not had vitamin A supplementation were about 10 times more likely to die. It continued that, over 90% of mothers had an insecticide treated bed net and more than 50% of them exclusively breastfed their children for the first 6 months of life. Protective risk factors identified during the investigation included exclusive breastfeeding, the use of an insecticide treated bed net, the number of live children a mother had and immunization. They recommended that, all health staff responsible for health programmes should be trained to identify and pay attention to mothers with previous child deaths in order to educate and assist them to prevent future deaths of their children.

Kojo (2012) in his thesis titled "Modeling the Risk Factors of neonatal mortality in Ghana" sought to analyze the risk factors of neonatal deaths in Ghana and to suggest some interventions that can be used in order to improve the survival of newborn babies in Ghana. He developed three models each for mother level factors, child level factors and environmental level factors. The results of the research revealed that, for the mother level factors, it was found that the age of the mother and the wealth index were the main factors causing neonatal deaths in Ghana. Also, sex of the baby and whether the baby is a twin were not significant causes of neonatal deaths in Ghana. He added that, for the environmental level factors, only the region of delivery was a significant cause of neonatal mortality in Ghana. He however recommended for further research to be conducted using other factors of neonatal mortality.

2.4 Review of Literature on Study Methods

Most of the investigations into the causes of under-five mortality make use of the logistic regression as the method for determining the risk factors that affect the probability of a child dying before five years old due to the binary nature of the response variable.

Reed and Wu (2013) in "Logistic regression for risk factor modelling in stuttering research" investigated the uses of the logistic regression in stuttering. The work outlined the steps, assumptions, limitations and the principles of the logistic regression model as applied in the stuttering field. They concluded that, the logistic regression provides an employed and a recognized approach to allow for the prediction of dichotomous outcomes.

In a similar work, Peng et al. (2002) in "An Introduction to Logistic

Regression Analysis and Reporting” outlined the guidelines in using the logistic regression. The article demonstrated the application of the logistic regression method with an illustration to a hypothetical data set. It was established that, traditional ordinary least squares regression or linear discriminant function analysis were found to be less ideal for handling binary responses due to their strict statistical assumptions. It was concluded that, the logistic regression could be a powerful analytical technique when the response variable is dichotomous and that, the method has gained popularity because it is easy to access sophisticated statistical software that perform detailed analyses using the method.

When there are clusters in the data, it is very appropriate to use a method that would cater for the correlation among variables within the same cluster. Zorn (2001) in ”Generalized Estimating Equation Models for Correlated Data: A Review With Applications” reviewed the Generalized Estimating Equation (GEE) method for dealing with a correlated data. The investigation continued that, GEEs offer a number of advantages for modeling correlated data including its applicability to continuous, dichotomous, polychotomous, ordinal and count response variables. Furthermore, it was revealed in the research that GEEs allow for a range of correlation patterns within clusters and offer valuable insights into the dynamics of such correlation. Concluding, Zorn (2001) however noted that, the usefulness of GEE models depends on the nature of data being analyzed as the method offer a flexible and easily implemented means of estimating models with correlated data.

An investigation titled ”The Case for Generalized Estimating Equations in State-level Analysis” by Staley (2013) revealed that, researchers are often confronted with correlated data when working in the real world. While there are several discussions in literature on time dependent correlated data, little has been done on the options that are available when these correlations are not time dependent. Staley (2013) further reported that, one technique which allows researchers to handle such forms of correlation is the GEE method. It was also noted in the research that, the GEE is best suited for instances where the researcher is interested in making comparisons across groups or sub populations as opposed to the effects within an individual observation.

In a related work by Ghisletta and Spini (2004) titled ”An Introduction to Generalized Estimating Equations and an Application to Assess Selectivity Effects in a Longitudinal Study on Very Old Individuals”, it was revealed that, correlated data are very common in educational and more generally in social science research. The work aimed to provide a non-statistical introduction to the GEE method using an analysis of the selectivity effects in the Swiss Interdisciplinary Longitudinal Study. The two common analytical situations in which data are correlated are longitudinal and hierarchically organized or clustered data. They continued that, GEEs represent an extension of GLMs to accommodate correlated data. Ghisletta and Spini (2004) continued that, time invariant covariates such as data from cross-sectional studies may require some special care when applying GEEs.

As noted by Carey et al. (1993) in ”Modelling Multivariate Binary Data with Alternating Logistic Regressions”, the first order generalized estimating equations (GEE1) introduced by Liang and Zeger (1986) is easy to implement when modeling the relationship between the response and the explanatory variables of marginal models for multivariate binary data. This method gives efficient estimates of the regression parameters but estimates of the association among the dichotomous outcomes could be inefficient. The research continued that, when estimates of the association is of interest, the simultaneous modeling of the responses and all pairwise products using the second order generalized estimating equations (GEE2) gives more efficient estimates of the association parameters. It was however pointed out that, GEE2 can become computationally impractical as the cluster size gets large. In furtherance of the above limitations of GEE1 and GEE2, they proposed the Alternating Logistic Regression (ALR) as an alternative approach for regressing the response variable on explanatory variables as well as finding the association among the response variables in terms of the pairwise odds ratios. In analysis of neuropsychological tests on patients with epileptic seizures, it was concluded that, the alternating logistic regression showed reasonably efficient estimates as compared to solutions from the GEE2.

Chapter 3

Methodology

3.1 Introduction

The advancement in statistical computing has made it not very difficult to analyze data using any of the available statistical software packages with some knowledge on the theoretical concepts that underline the method used in getting the results. Contrary to this phenomenon, one needs to acquire the best of knowledge and understanding on the theoretical background of every method used for data analysis in order to help for an effective interpretation of the results produced by the software. This chapter therefore focuses on a detailed and comprehensive understanding of the logistic regression and the alternating logistic regression which is a type of the generalized estimating equations method as the basic methods used in this research work.

3.2 Data Description

This study uses a secondary data from the Women's Questionnaire of the Ghana DHS 2008 which is a national survey covering all the ten regions of the country. Data for the survey was collated through the use of a "Verbal Autopsy Questionnaire". In about half of the households selected for the survey, all eligible women of age 15-49 were interviewed with the Women's Questionnaire (GSS/GHS/ICF, 2009). Questions on whether the women have ever given birth were asked and a follow up question on whether the child was alive or dead was also asked in order to identify the adequate number of under-five deaths.

Information on the ever born women and their last five births were extracted from the Ghana DHS 2008 data and used for the analysis. Respondents (women) who have never given birth as well as those whose children were above five years old and who fall outside the period of the survey from January, 2004 to December, 2008 were excluded from the data. The number of respondents in the data was therefore reduced from 4,916 as found in the women's questionnaire of the Ghana DHS 2008 data to 2,170.

3.3 Study Design

According to the GSS/GHS/ICF (2009), the sample data was selected using a two stage sampling design in the selection of clusters, households and respondents as the hierarchical order. The first stage involved selection of clusters from an updated master sampling frame constructed from the 2000 Ghana Population and Housing Census. A total of 412 clusters were selected from the master sampling frame using a systematic sampling procedure. The second stage of the selection involved the systematic sampling of 30 households listed in each cluster. Out of the targeted 12,360 households, a total sample of 12,323 households were selected of which 11,913 were occupied at the time of the field work because some of the selected structures were found to be vacant or destroyed. A total of 11,778 households were interviewed yielding a household response rate of 99 percent. In about half of the households, a total of 5,096 eligible women were identified for the Women's Questionnaire and interviews were completed with 4,916 of these women yielding a response rate of 97 percent for the Women's Questionnaire.

The primary sampling unit for the Ghana DHS 2008 was the cluster which consisted of one or more enumeration areas from the 2000 Population and Housing Census. In each region, the enumeration areas were stratified and the clusters were selected systematically. In each of the selected cluster, a household listing operation was carried out from June-July 2008 and the household samples were selected after which respondents in the selected households were interviewed. The sampling of clusters, households and respondents for the Ghana DHS 2008 was a combination of sampling techniques such as cluster and systematic sampling.

3.4 Software Used

The analysis of the data was done using the SAS version 9.1 for windows.

3.5 Study Methods

The two main methods used in this thesis are the logistic regression and the alternating logistic regression which is a type of the generalized estimating equations.

3.5.1 The Binary Logistic Regression Model

The binary logistic regression model also known as the logit model is a type of generalized linear model used for modeling a binary response variable. This type of regression is very appropriate to determine the effect of categorical, continuous or both categorical and continuous independent variables on a dichotomous dependent variable (Reed and Wu, 2013). This defeats the crucial limitation of the ordinary linear regression which cannot deal with categorical variables. For example, the probability that a child would die before the fifth birthday might be predicted by independent variables such as the sex of the child, weight at birth, age of the mother, occupation of the mother and many others. Larger samples are needed when using the logistic regression because the odds ratio reported by the logistic regression is overestimated when the sample size is less (Kojo, 2012).

3.5.2 The Logistic Function

The logistic function describes the mathematical form on which the logistic model is based. It predicts the probability of the dependent variable which lies on the probability scale (between 0 and 1 inclusive) rather than the value of the response as in simple linear regression. It gives each predictor (independent) variable a coefficient which measures its independent contribution to the variation in the response (dependent) variable. Let f(x) be the logistic function with an input x. The input x takes values from negative infinity to positive infinity ($-\infty$ to $+\infty$) whereas the output f(x) confines the values to the probability scale between 0 and 1 inclusive.

Mathematically;

$$f(x) = \frac{1}{1+e^{-x}}$$

If x=$-\infty$, f(x)=0 and if x=$+\infty$, f(x)=1.

3.5.3 General Mathematical Representation of the Binary Logistic Regression

For a binary response variable Y of binomial distribution $Y_i\sim$ B(n_i,π_i) for n_i (i=1, 2, $\cdots$, N) independent observations, let $\pi(x_i)$ denote the probability of success (probability of a child dying before the fifth birthday) and 1-$\pi(x_i)$ denote the probability of failure (probability of a child not dying before the fifth birthday).

The logistic regression model or the logit (natural log of the odds) is given by;

$$logit[\pi(x_i)] = log_e\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \tag{3.1}$$

In terms of the odds;

$$\frac{\pi(x_i)}{1-\pi(x_i)} = exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) \tag{3.2}$$

The probability $\pi(x_i)$ is also given as;

$$\pi(x_i) = \frac{exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})} \tag{3.3}$$

where β_0 is the intercept and β_j (for j=1, 2, $\cdots$, k) are the regression coefficients of the independent variables x_j (for j=1, 2, $\cdots$, k) respectively.

The intercept is the value of the logit model when the values of all the independent variables are zero (no predictor variables). The regression coefficients β_i describe the amount of contribution of the independent (predictor) variables on the logit model. A positive value of the regression coefficient means the independent (predictor) variable increases the probability of the model and a negative value means the independent (predictor) variable decreases the probability of the model (Agresti, 2002).

3.5.4 The Maximum Likelihood Estimation

The unknown parameters β_i (for i=0,1, $\cdots$, k) for the logit model are estimated using the maximum likelihood estimation method based on the statistical assumption of independence of observations. With respect to this thesis, the various respondents in the Ghana DHS 2008 data are assumed to be independent of each other. The maximum likelihood estimate is the parameter value for which the probability of the observed data takes its greatest value (Agresti, 2007). The method assumes a particular probability distribution for the dependent variable Y_i in order to determine its likelihood function using the natural log.

Maximum Likelihood Estimation for the Exponential Family

Generally, a random variable Y that belongs to the exponential family has a probability density function of the form;

$$f(y;\theta,\phi) = exp\left[\frac{y\theta - \psi(\theta)}{\phi} + c(y,\phi)\right] \tag{3.4}$$

for a given set of unknown natural parameter θ and scale parameter ϕ for known functions $\psi(.)$ and c(. , .). From the first and second moments of the function, mean μ=E(Y)= $\psi'(\theta)$ and variance $\sigma^2 = Var(Y) = \phi\psi''(\theta)$

To estimate the regression parameter β using the maximum likelihood estimation by assuming independence of observations, the log-likelihood of the exponential family is given by;

$$l(f(y)) = \frac{1}{\phi}\sum_i [y_i\theta_i - \psi(\theta_i)] + \sum_i c(y_i, \phi) \tag{3.5}$$

For a fixed ϕ treated as a nuisance parameter, the score equation obtained from equating the first order derivative (with respect to β) of the log-likelihood to zero is given by;

$$S(\beta) = \sum_i \frac{\partial\theta_i}{\partial\beta}[y_i - \psi'(\theta_i)] = 0 \tag{3.6}$$

Since the mean function $\mu_i = \psi'(\theta_i)$ and the variance function $v_i = v(\mu_i) = \psi''(\theta_i)$

$$\frac{\partial\mu_i}{\partial\beta} = \psi''(\theta_i)\frac{\partial\theta_i}{\partial\beta} = v_i\frac{\partial\theta_i}{\partial\beta}$$

which implies that

$$\frac{\partial\theta_i}{\partial\beta} = v_i^{-1}\frac{\partial\mu_i}{\partial\beta}$$

The general score equation in matrix form then becomes;

$$S(\beta) = \sum_i \left(\frac{\partial\mu_i}{\partial\beta}\right)^T v_i^{-1}(y_i - \mu_i) = 0 \tag{3.7}$$

Iterative methods such as weighted least squares, Newton-Raphson or Fisher scoring algorithm could be used to solve for the parameter estimates [(Molenberghs and Verbeke, 2005), (Agresti, 2002)].

Maximum Likelihood Estimation for the Binomial Distribution

The probability distribution of a binomial response variable $Y_i \sim B(n_i, \pi_i)$ is given by;

$$P(Y_i = y_i) = \binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i} \tag{3.8}$$

for $y_i = (0, 1, 2, \cdots, n_i)$ where $\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}$ is the probability of obtaining y_i successes and $n_i - y_i$ failures with a combinatorial coefficient $\binom{n_i}{y_i}$ which is the number of ways of obtaining y_i successes in n_i trials. The mean of Y_i is given as $E(Y_i) = n_i\pi_i$ and the variance of Y_i is $Var(Y_i) = n_i\pi_i(1 - \pi_i)$.

If the combinatorial term is treated as a constant, the likelihood function is given as;

$$L(\pi_i) = \prod_{i=1}^{N} \pi_i^{y_i}(1 - \pi_i)^{n_i - y_i} \tag{3.9}$$

Taking natural logs on both sides, the log likelihood is also given as;

$$logL(\pi_i) = log\left[\prod_{i=1}^{N} \pi_i^{y_i}(1 - \pi_i)^{n_i - y_i}\right] \tag{3.10}$$

$$logL(\pi_i) = \sum_{i=1}^{N}[y_i log(\pi_i) + (n_i - y_i)log(1 - \pi_i)] \tag{3.11}$$

Rearranging the terms in equation (3.11) above, we obtain equation (3.12) below

$$logL(\pi_i) = \sum_{i=1}^{N} y_i log\left(\frac{\pi_i}{1 - \pi_i}\right) - \sum_{i=1}^{N} n_i log\left(\frac{1}{1 - \pi_i}\right) \tag{3.12}$$

The parameters can therefore be estimated by finding score equations and using an iterative method such as the Fisher scoring algorithm or Newton-Raphson method to solve for their values (Dobson, 2002).

3.5.5 Statistical Inference

Wald Test

The Wald test is used to find the statistical significance of the respective coefficients in the model. The z-statistic is calculated as;

$$z = \frac{\hat{\beta}_i}{Se_{\hat{\beta}_i}}$$

The z value is squared to yield a Wald statistic with a Chi-Squared distribution which has an approximately normal distribution for large samples.

Confidence Interval

The Wald test can be used to determine a $100(1-\alpha)\%$ confidence interval for $\hat{\beta}_i$ to show that the true parameter lies in the interval with boundaries;

$$\hat{\beta}_i \pm z_{1-\alpha/2} Se_{\hat{\beta}_i}$$

where $z_{1-\alpha/2}$ is the critical value for the two sided normal distribution of size α.

In terms of the odds ratio, confidence intervals are formed by finding exponents of the boundaries.

$$[exp(\hat{\beta}_i \pm z_{1-\alpha/2} Se_{\hat{\beta}_i})]$$

$$[exp(\hat{\beta}_i - z_{1-\alpha/2} Se_{\hat{\beta}_i}), exp(\hat{\beta}_i + z_{1-\alpha/2} Se_{\hat{\beta}_i})]$$

3.5.6 Goodness-of-fit Tests

Goodness of fit tests are used to determine how well the proposed model fits the data using the logistic model. A model is poorly fit if either the model's residual variation is large or it does not follow the variability postulated by the model (Hallet, 1999). Examples of these tests are the deviance statistic, Pearson's Chi-Squared test and Hosmer-Lemeshow test.

Pearson Chi-Square Test

The Pearson's chi-square test is a goodness of fit test for a given model versus the full model by testing the difference between the expected (fitted) value of response and the observed one. The test statistic is approximately chi-squared distribution provided that sample sizes are large enough. For a binomial data, it is given as;

$$X_p^2 = \sum \frac{n_i(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(n_i - \hat{\mu}_i)}$$

Each term is the squared difference of the observed and fitted values divided by the variance of y_i which is equal to $\frac{\hat{\mu_i}(n_i - \hat{\mu_i})}{n_i}$

Hosmer-Lemeshow Goodness-of-fit Test

The Hosmer and Lemeshow goodness-of-fit statistic recommends forming G (usually 10) equally sized groups according to their predicted probabilities. It is a measure of lack of fit for a model.

$$G^2_{HL} = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - E_j/n_j)} \sim X^2_g$$

where

n_j is the number of observations in the jth group

$O_j = \sum y_{ij}$ is the number of cases in the jth group

$E_j = \sum P_{ij}$ is the expected number of cases in the jth group

A model with $G^2_{HL} \sim X^2_g$ value greater than the p-value (0.05) shows no evidence of lack of fit based on the Hosmer and Lemeshow statistic.

3.5.7 Odds Ratio

The logistic regression reports the odds ratio which is a measure of association between categorical outcomes like a binary response variable. The odds of an event is the probability that the event will occur relative to the probability that it will not occur. In this thesis, the odds refer to the probability that a child will die before the fifth birthday π relative to the probability that the child will not die before the fifth birthday $(1 - \pi)$. For a probability of success π and the probability of failure $(1 - \pi)$, the odds is given as;

$$Odds = \frac{\pi}{1 - \pi}$$

The odds ratio is defined as the ratio of the odds for group 1 to the odds for group 2. For example, to compare the odds of a child dying before the fifth birthday between female children (group 1) and male male (group 2), the odds ratio is given by;

$$OddsRatio = \frac{odds1}{odds2}$$

From the logistic regression model in equation (3.1), the odds

$$\frac{\pi(x_i)}{1-\pi(x_i)} = exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

$$\frac{\pi(x_i)}{1-\pi(x_i)} = (e^{\beta_0})(e^{\beta_1})^{x_{i1}}(e^{\beta_2})^{x_{i2}} \cdots (e^{\beta_k})^{x_{ik}}$$

The odds from the logistic regression model is equal to e^{β_i} (for i=0, 1, 2, $\cdots$, k). This means that, for every unit increase in the variable x_i (for i=1, 2, $\cdots$, k), the odds multiply by the value of the respective β_i. If $\beta = 0 (e^{\beta} = 1)$, the odds do not change as the variable changes. If the odds $e^{\beta_i} < 1$, a unit increase in the value of the variable correspond to a decrease in the odds of a child dying before the fifth birthday. If the odds $e^{\beta_i} > 1$, a unit increase in the variable correspond to an increase in the odds of a child dying before the fifth birthday.

3.5.8 The Generalized Estimating Equations (GEE)

The GEE is a method of parameter estimation which represents an extension of the generalized linear models to accommodate correlated data. For example, respondents from the same household in the Ghana DHS 2008 data. The method models the marginal expectations of the outcome variable such as the probability of a child dying before the fifth birthday. As proposed by (Liang and Zeger, 1986), the GEE method is a non-likelihood approach which estimates the parameters associated with the expected value of binary responses (Molenberghs and Verbeke, 2005). The method requires the correct specification of the marginal distributions and combines the estimating equations for the regression parameters with moment-based estimation of the correlation parameters entering the working assumptions. For an assumed working correlation matrix, the GEE method uses the data to estimate a consistent and unbiased standard errors of parameters even when the correlation structure is specified wrongly. The GEE method provides a computationally simple alternative to the maximum likelihood for clustered categorical data of high dimensional vectors.

When using the original GEE method also known as the first order generalized estimating equations (GEE1), correlation is used rather than information on the association structure to estimate the effect of the main regression parameters. The GEE1 yields consistent main effect estimators even when the association structure is wrongly specified. (Prentice, 1988) amended GEE1 to allow for estimation of both the regression coefficients and the association (odds ratio) parameters in the marginal response model and the pairwise correlations respectively. Further studies into the method proposed by (Prentice, 1988) yielded the second order estimating equations (GEE2) that included the marginal pairwise association. The GEE2 was proposed by (Zhao and Prentice, 1990) using correlations and (Liang et al., 1992) using odds ratios. However, (Carey et al., 1993) proposed a conditional probability idea which is known as the alternating logistic regression to estimate the parameters of a clustered data such as the clusters (respondents) found in the same household as in the Ghana DHS 2008 data.

3.5.9 Marginal Models

Marginal models also known as population averaged models has the primary scientific objective of analyzing the marginal expectation of responses for a given explanatory variables. The population averaged parameters represent the average effects of a unit change in the explanatory variables for the whole population rather than individual subjects. In other words, marginal models are used to model the population averaged expectations of the dependent variable as a function of the independent variables across the entire population but not individual observations as in conditional models.

Consider a binary data obtained in m clusters of n sizes for clusters

$i = 1, 2, \cdots, m$ and let $Y_i = (Y_{i1}, \cdots, Y_{in_i})^T$ be an $n_i \times 1$ response vector with mean $E(Y_i) = \mu_i$ and ψ_{ijk} be the odds ratio between the responses Y_{ij} and Y_{ik} for $(1 \leq j < k \leq n_i)$.

$$\psi_{ijk} = \frac{P(Y_{ij} = 1, Y_{ik} = 1)P(Y_{ij} = 0, Y_{ik} = 0)}{P(Y_{ij} = 1, Y_{ik} = 0)P(Y_{ij} = 0, Y_{ik} = 1)} \tag{3.13}$$

A marginal model can be specified as follows;

1. $h(\mu_i) = (x_{ij})^T \beta_j$

   $h(.)$ is a known link function (a logit link in the case of a binary response),

   x_{ij} is a $p \times 1$ vector of explanatory variables associated with the response Y_{ij} and

   β_j are regression coefficients to be estimated.

2. $log\psi_{ijk} = (z_{ijk})^T \alpha$ where

   z_{ijk} is a $q \times 1$ vector of covariates which specifies the association between Y_{ij} and Y_{ik},

   α is a $q \times 1$ vector of association parameters to be estimated using odds ratio

(Carey et al., 1993).

3.5.10 Quasi-Likelihood Estimation

The GLM method assumes a distributional form (such as Binomial, Poisson, Normal etc) for the dependent variable Y_i in order to derive a likelihood function based on the statistical assumption of independence of observations. For a correlated data such as respondents in the same household as found in the Ghana DHS 2008 data, the maximum likelihood estimation which assumes independence of observations to estimate parameters becomes impracticable. The quasi likelihood estimation was therefore developed to be used in such situations. It specifies only the relationship between the mean $E(Y_i)$ and the variance $Var(Y_i)$ of the observations (Fan et al., 2009).

For a given dependent variable Y_i with mean $E(Y_i) = \mu_i$ and covariance matrix $v(\mu_i)$, then for each observation the quasi-likelihood function is defined as;

$$\frac{\partial l(Y_i; \mu_i)}{\partial \mu_i} = \frac{Y_i - \mu_i}{v(\mu_i)}$$

$$\frac{\partial l(Y_i; \mu_i)}{\partial \mu_i} = v(\mu_i)^{-1}(Y_i - \mu_i)$$

[Wedderburn (1974), McCullagh (1983)].

3.5.11 The Standard GEE Theory

The GEE method uses the quasi-likelihood function for estimating regression coefficients β in situations when the association (odds ratio) parameter α is treated as a nuisance factor. The joint distribution of the response vector does not need to be specified as that could be unnecessarily cumbersome; instead the marginal distribution could be used where only the relationship between the mean and the variance of the dependent variable needs to be specified. Respondents within the same household may be correlated to each other but the households are assumed to be independent across the clusters they belong to as found in the Ghana DHS 2008 data. To establish the notation, consider a binary data obtained in m clusters of n sizes for clusters $i = 1, 2, \cdots m$ and let $Y_i = (Y_{i1}, \cdots, Y_{in})^T$ be $n \times 1$ response vector.

The score equation to be solved for the maximum likelihood estimates of the exponential family as stated earlier in equation (3.4) is given by;

$$S(\beta) = \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T (v_i)^{-1}(y_i - \mu_i) = 0 \tag{3.14}$$

where

$v_i = (A_i)^{1/2} I_n (A_i)^{1/2}$

$A_i = A_i(\beta)$ is a diagonal matrix with marginal variances along the main diagonal

I_n is an identity matrix.

An extension of the above equation to account for the correlation structure obtained by replacing the identity matrix I_n with a correlation matrix $R_i = R_i(\alpha)$ for the m number of clusters is given as;

$$S(\beta, \alpha) = \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T \left[(A_i)^{1/2} R_i(\alpha) (A_i)^{1/2} \right]^{-1} (y_i - \mu_i) = 0 \tag{3.15}$$

The GEE estimate of β is obtained by solving the above estimating equation iteratively.

(Liang and Zeger, 1986), showed by method of moments that when the marginal mean μ_i has been correctly specified as $h(\mu_i = X_i\beta)$ with mild regularity conditions and the association (odds ratio) parameter α is treated as a nuisance, the estimator $\hat{\beta}$ obtained is consistent and asymptotically normal with mean β and covariance $Var(\hat{\beta}) = I_o^{-1} I_1 I_o^{-1}$ where

$$I_o = \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} \frac{\partial \mu_i}{\partial \beta}$$

and

$$I_1 = \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} Var(Y_i) V_i^{-1} \frac{\partial \mu_i}{\partial \beta}$$

In a case where the correlation matrix $R_i = R_i(\alpha)$ is correctly specified, $Var(Y_i) = V_i$ and $I_1 = I_o$ which makes $Var(\hat{\beta}) = I_o^{-1}$. Hence, whether the working correlation matrix is correctly specified or not the point estimates and standard errors based on $Var(\hat{\beta})$ are generally correct.

When the association (odds ratio) parameter α is the scientific focus, (Prentice, 1988) proposed the expansion of GEE1 to allow for the estimation of both the regression parameter β and the association parameter α using two sets of estimating equations. The method expands the estimating equations to model the response variable Y_i and ${}_nC_r$ cross products of

$$W_i = (Y_{i1}Y_{i2}, Y_{i1}Y_{i3}, Y_{i1}Y_{i4}, \cdots, Y_{i1}Y_{in}, \cdots, Y_{i,n-1}Y_{in})^T$$

in order to augment the efficiency of the standard GEE method. The resulting estimating equation will have the form;

$$S(\beta, \alpha) = \sum_{i=1}^{m} \left[\frac{\partial(\mu_i, V_i)^T}{\partial(\mu_i, \alpha)^T} \right]^T Cov^{-1}(Y_i, W_i)(y_i - \mu_i, w_i - v_i)^T = 0 \tag{3.16}$$

where $v_i = E(W_i)$

(Liang et al., 1992) showed that, $(\hat{\beta}, \hat{\alpha})$ the solution of $S(\beta, \alpha) = 0$ is highly efficient for both β and α. The problem it faces is that, for a cluster of size n, the matrix $B = cov(Y, W)$ has dimensions $(n + {}_nC_2) \times (n + {}_nC_2)$ before it is inverted. For an illustration, if

$n = 2, B$ is 3×3

$n = 3, B$ is 6×6

$n = 5, B$ is 15×15

$n = 10, B$ is 55×55

$n = 60, B$ is 1830×1830

It is therefore very clear that, the solution of $S(\beta, \alpha) = 0$ above is computationally difficult for cases of larger cluster sizes as can be found in real life situations. The GEE2 is therefore impractical for real life applications where the size of cluster could be very large in situations where the association parameter α is of interest (Carey et al., 1993).

3.5.12 Working Correlation Matrix

The GEE approach yields a consistent estimator for the regression coefficients β even when the working correlation matrix is wrongly specified (Liang and Zeger, 1986). Let $R_i(\alpha)$ be a symmetric $n \times n$ correlation matrix for the clustered measurements from respondents $Y_i = (Y_{i1}, Y_{i2}, Y_{13}, \cdots, Y_{in})^T$ from the Ghana DHS 2008 data. One advantage of the GEE method is the wide range of options that are available for specifying the structure of the within-cluster correlation (Zorn, 2001) for respondents in the same household. The common specifications of the working correlation matrix for these respondents in the same household of the Ghana DHS 2008 data would include; the independent, exchangeable and the unstructured working assumptions.

Independent Working Assumption

The independent working assumption assumes that, there is no correlation between the respondents within the same household. The non-existence of the within-cluster correlation reduces the working correlation to an identity matrix ($R_i(\alpha) = I$) with 1 as diagonal entries and 0 elsewhere. No estimate of α is obtained since the intra-cluster correlation is assumed to be zero.

$$Corr(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}$$

$$R_{ind} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Exchangeable Working Assumption

The exchangeable working assumption assumes that, there is a constant correlation between the respondents within the same household. The values of Y_i are assumed to vary equally across all respondents within the same household where α is a scalar which is estimated by the model. The working correlation matrix reduces to a form with 1 as diagonal entries and α elsewhere.

$$Corr(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \alpha & j \neq k \end{cases}$$

$$R_{exc} = \begin{pmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \alpha & \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{pmatrix}$$

Unstructured Working Assumption

The unstructured working assumption also known as the pairwise working assumption assumes that, there are different correlations between any two respondents within the same household. No constraints are placed on the correlations and every element of the correlation matrix is estimated separately [(Zorn, 2001), (Isaac, 2011)]. This type of correlation structure has no explicit form, only that every correlation coefficient is allowed to be different (Jang, 2011).

$$Corr(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \rho_{ik} & j \neq k \end{cases}$$

$$R_{uns} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2k} \\ \rho_{31} & \rho_{32} & 1 & \cdots & \rho_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \rho_{k3} & \cdots & 1 \end{pmatrix}$$

3.5.13 Inference

The standard errors associated with the regression estimates are needed to make inferences such as hypothesis testing and confidence intervals. These standard errors are given as the square root of the diagonal entries of the variance matrix $Var(\hat{\beta})$. The GEE approach provides two versions of the variance estimator (Ghisletta and Spini, 2004) which are the naive or model-based estimator and the robust or empirical or sandwich estimator.

Model-Based Estimator

It is also known as the naive estimator. The model-based estimator requires the correct specification of the working correlation matrix. As stated earlier, $Var(\hat{\beta}) = I_o^{-1} I_1 I_o^{-1}$ where

$$I_o = \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} \frac{\partial \mu_i}{\partial \beta}$$

and

$$I_1 = \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} Var(Y_i) V_i^{-1} \frac{\partial \mu_i}{\partial \beta}$$

If the working correlation matrix $R_i = R_i(\alpha)$ is correctly specified, $Var(Y_i) = V_i$ implying that $I_1 = I_o$ hence,

$$Var(\hat{\beta}) = I_o^{-1} = \left[\sum_{i=1}^{m} (\frac{\partial \mu_i}{\partial \beta})^T V_i^{-1} \frac{\partial \mu_i}{\partial \beta} \right]^{-1}$$

This is the GEE equivalent of the inverse of the Fisher information matrix which is used in GLMs (Stokes et al., 2000).

Empirical Estimator

It is also known as the robust or the sandwich estimator. This type of estimator is robust to wrongly specifying the working correlation matrix. It has the property of being a consistent estimator of the variance matrix $Var(\hat{\beta})$ even if the working correlation matrix is not specified correctly. For a misspecification of the working correlation matrix $R_i = R_i(\alpha)$, $Var(Y_i) = (y_i - \mu_i)(y_i - \mu_i)^T$ and

$Var(\hat{\beta}) = I_o^{-1} I_1 I_o^{-1}$. Therefore;

$$Var(\hat{\beta}) = \left[\sum_{i=1}^{m} (\frac{\partial \mu_i}{\partial \beta})^T V_i^{-1} \frac{\partial \mu_i}{\partial \beta}\right]^{-1} \times \left[\sum_{i=1}^{m} (\frac{\partial \mu_i}{\partial \beta})^T V_i^{-1} (y_i - \mu_i)(y_i - \mu_i)^T V_i^{-1} \frac{\partial \mu_i}{\partial \beta}\right] \times \left[\sum_{i=1}^{m} (\frac{\partial \mu_i}{\partial \beta})^T V_i^{-1} \frac{\partial \mu_i}{\partial \beta}\right]^{-1}$$

More generally, the empirical estimator provides a consistent estimator of the variance matrix $Var(\hat{\beta})$ even if the working correlation matrix $R_i(\alpha)$ is not the true correlation of y_i.

3.5.14 The Alternating Logistic Regression

The alternating logistic regression (ALR) method combines GEE1 for the regression parameters β with a modified logistic regression for estimating the association parameter α. In the standard GEE, the association parameter is a nuisance parameter which is not the case in the ALR where the marginal models are fitted based on the odds ratio such that inferences can be made not only about marginal regression parameters but on the pairwise association of the responses as well (Molenberghs and Verbeke, 2005). The ALR method proposed by Carey et al. (1993) used another approach to estimate the regression coefficients β and the association parameter α which can be efficient for both sets of parameters and avoids the computational problems of the standard GEE methods. The method iterates between a logistic regression using GEE1 to estimate regression parameters β and a logistic regression of each response on others from the same cluster (household) using an appropriate offset to update the odds ratio parameters α.

For a response variable Y_{ij} denoting the probability of a child dying before the fifth birthday for jth respondent from the ith household, let X_{ij} be a known set of covariates with $\mu_{ij} = P(Y_{ij} = 1)$ for $\mu_{ij} = E(Y_{ij})$;

$$logitP(Y_{ij} = 1) = X_{ij}^T \beta \qquad (3.17)$$

The logit model above shows that, the effects of the covariates X_{ij} on the response Y_{ij} are averaged over all the households. The marginal model therefore resembles a multiple logistic regression except that the parameter estimate β as well as the estimated standard errors of the parameters are corrected for by clustering for respondents in the same household as found in the Ghana DHS 2008 data.

Let ψ_{ijk} be the log odds ratio between the responses Y_{ij} and Y_{ik} for jth and kth respondents from the same ith household. In addition, let $v_{ijk} = P(Y_{ij} = 1, Y_{ik} = 1)$ for $v_{ijk} = E(Y_{ij}Y_{ik})$. Rather than specifying the marginal description of the association, a logistic model for a response Y_{ij} conditional upon another response Y_{ik} is presented as;

$$logitP(Y_{ij} = 1 \mid Y_{ik} = y_{ik}) = \psi_{ijk}y_{ik} + log\left(\frac{\mu_{ij} - v_{ijk}}{1 - \mu_{ij} - \mu_{ik} + v_{ijk}}\right) \tag{3.18}$$

Assuming $\psi_{ijk} = \alpha$, the pairwise log odds ratio α is the regression coefficient in the logistic regression of Y_{ij} given Y_{ik}. Instead of an intercept as found in the ordinary logistic regression model, there is an offset $log\left(\frac{\mu_{ij}-v_{ijk}}{1-\mu_{ij}-\mu_{ik}+v_{ijk}}\right)$ which is a constant term free of the unknown parameters β and α.

To be specific, the ALR method iterates between solving the two logistic models for regression parameters β and association parameters α using the GEE approach. The iteration requires solving these two estimating equations which depend on β and α simultaneously.

$$S(\beta) = \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T (v_i)^{-1}(Y_i - \mu_i) = 0 \tag{3.19}$$

$$S(\alpha) = \sum_{i=1}^{m} \left(\frac{\partial \gamma_i}{\partial \alpha}\right)^T (W_i)^{-1}(Y_{ijk} - \gamma_{ijk}) = 0 \tag{3.20}$$

where

γ_i is a vector of elements $\gamma_{ijk} = P(Y_{ij} = 1 \mid Y_{ik} = y_{ik})$ and

W_i is a diagonal matrix with elements $\gamma_{ijk}(1 - \gamma_{ijk})$.

To conclude, the alternating logistic regression algorithm alternates between these two steps;

- for a given association parameter α which is treated as a nuisance, estimate the regression coefficient β as a parameter in a marginal logistic regression using GEE1
- for a given regression coefficient β, estimate the association parameter α using a logistic regression of Y_{ij} on each $Y_{ik}(k > j)$ with an offset that involves $\mu_{ij} = E(Y_{ij})$ and $v_{ijk} = E(Y_{ij}Y_{ik})$.

[(Molenberghs and Verbeke, 2005), (Carey et al., 1993)]

Chapter 4

Data Analysis and Results

4.1 Introduction

This chapter looks at the selection and exploratory analysis of study variables, the ordinary logistic regression analysis and the alternating logistic regression analysis of under-five mortality in the Ghana DHS 2008 data. The sample for this thesis was taken from the Women's Questionnaire of the Ghana DHS 2008 data. During the data collection process, questions on whether the respondents (women) have ever given birth were asked and a follow up question on ”whether the child was alive” was also asked in order to identify the adequate number of under-five deaths. Information on the ever born women and their last five births were extracted as the sample for the analysis.

4.2 Exploratory Analysis of Study Variables

The data was explored using contingency tables to find the frequencies and percentages of the dependent variable as well as to investigate the association between under-five mortality and the selected independent variables. Summary statistics for the continuous independent variables were obtained, for example; the mean, the standard deviation, the minimum and the maximum observations of these variables. The association between the response variable and the nominal categorical variables were determined using the likelihood ratio test statistic and the Cochran-Mantel-Haenszel test was performed to test for the linear trend between the response variable and the ordinal categorical variables. All the independent variables that were not significant at 25% level of significance were dropped for the next stage of the analysis (Hosmer and Lemeshow, 2000).

The dependent variable is under-five mortality (0, indicator for child alive and 1, indicator for at least a child died) in the sample. It was formed from respondents who have experienced at least one under-five mortality in their last five births as found in the Ghana DHS 2008 data. Respondents who have never given birth as well as those whose children are over five years as at the time of the survey were dropped to yield a final sample of 2,170 respondents out of the 4,916 respondents found in the Women's Questionnaire of the Ghana DHS 2008 data. All respondents who have experienced at least one under-five mortality from their last five births were coded as 1 (for child died) and those who have not experienced under-five mortality in the last five births were assigned 0 (for child alive).

Table 4.1: Under-five mortality

Child Alive	**Frequency(%)**
	n(%)

Yes (0 for child alive)	1841 (84.84)
No (1 for child died)	329 (15.16)
Total	**2170 (100)**

From table 4.1, the observed data contains 2,170 respondents which is made up of 329 (15%) respondents who have experienced at least one under-five mortality from their last five births and 1,841 (85%) respondents who have never experienced under-five mortality from their last five births.

The independent variables were selected based on their statistical significance on under-five mortality as well as evidence from literature. The independent variables which were selected and used for the analysis are; age of respondent (AgeofResp), region of residence of respondent (RegofResid), type of residence of respondent (TypeofResid), highest educational level of respondent (EduLevel), religion, ethnicity, total number of household members (HHSize), literacy level of respondent (Lit), total number of children ever born (ChildEverBorn), BMI of respondent, marital status of respondent (MStatus), respondents occupation (RespOccu) and wealth index (WIndex).

The frequencies and percentages of the selected independent variables with their corresponding under-five mortality estimates are shown in table 4.2, table 4.3 and table 4.4.

Table 4.2: Frequencies and percentages of the independent variables

Variable	**n(%)**	**mortality(%)**
RegofResid		
ashanti	320(14.75)	45(13.68)
brong ahafo	205(9.45)	24(7.29)
central	161(7.42)	27(8.21)
eastern	189(8.71)	23(6.99)
greater accra	210(9.68)	27(8.21)
northern	310(14.29)	66(20.06)
upper east	182(8.39)	25(7.60)
upper west	209(9.63)	42(12.77)
volta	186(8.57)	21(6.38)
western	198(9.12)	29(8.81)
TypeofResid		
rural	1397(64.38)	211(64.13)
urban	773(35.62)	118(35.87)
EduLevel		
no education	788(36.31)	148(44.98)
primary	513(23.64)	77(23.40)
secondary	821(37.83)	100(30.40)
higher	48(2.21)	4(1.22)
Religion		
anglican	11(0.51)	2(0.61)
catholic	313(14.42)	48(14.59)
methodist	114(5.25)	17(5.17)
moslem	423(19.49)	78(23.71)
no religion	102(4.70)	18(5.47)
other	4(0.18)	1(0.3)
other Christian	210(9.68)	34(10.33)
pentecostal/charismatic	693(31.94)	87(26.44)
presbyterian	129(5.94)	9(2.74)
traditional/spiritualist	171(7.88)	35(10.64)

Table 4.3: Frequencies and percentages of the independent variables cont'd

Variable	n(%)	mortality(%)
Ethnicity		
akan	854(39.35)	120(36.47)
ewe	274(12.63)	24(7.29)
ga/dangme	100(4.61)	13(3.95)
gruma	123(5.67)	29(8.81)
grussi	116(5.35)	18(5.47)
guan	57(2.63)	9(2.74)
mande	17(0.78)	7(2.13)
mole-dagbani	553(25.48)	95(28.88)
other	76(3.50)	14(4.26)
Lit		
blind	3(0.14)	1(0.30)
illiterate	1514(69.77)	252(76.60)
literate	653(30.09)	76(23.10)
MStatus		
divorced	104(4.79)	15(4.56)
living together	406(18.71)	60(18.24)
married	1327(61.15)	216(65.65)
never married	139(6.41)	14(4.26)
not living together	131(6.04)	15(4.56)
widowed	63(2.90)	9(2.74)
WIndex		
poorest	219(10.09)	38(11.55)
poorer	471(21.71)	69(20.97)
middle	488(22.49)	86(26.14)
richer	535(24.65	74(22.49)
richest	457(21.06)	62(18.84)

Table 4.4: Frequencies and percentages of the independent variables cont'd

Variable	n(%)	mortality(%)
AgeofResp		
< 20	105(4.84)	5(1.52)
20-29	972(44.80)	127(38.60)
30-39	809(37.28)	136(41.34)
40-49	284(13.08)	61(18.54)
HHSize		
1-5	1185(54.61)	178(54.10)
6-10	851(39.22)	130(39.51)
> 10	134(6.17)	21(6.38)
ChildEverBorn		
1-5	1821(83.92)	267(81.16)
6-10	337(15.53)	56(17.02)
> 10	12(0.55)	6(1.82)
BMI		
< 20.00	285(13.32)	52(15.95)
20.00-29.99	1550(72.43)	234(71.78)
30-39.99	274(12.80)	39(11.96)
$\geq$40	31(1.45)	1(0.31)

There are four (4) continuous independent variables among the selected variables in the observed data. Each variable had a total number of 2,170 respondents except BMI which had 2,140 respondents due to missing values.

Table 4.5: Summary statistics for continuous independent variables

Variable	n	Mean	Std Dev	Min	Max
AgeofResp(years)	2170	30.29	7.35	15	49
HHSize	2170	5.72	2.77	1	22
ChildEverBorn	2170	3.40	2.15	1	13
BMI	2140	24.93	5.15	14.81	56.94

From table 4.5, the observed data consists of women from the ages of 15 years to 49 years with a mean age of 30 years and a standard deviation of about 7 years. The maximum number of persons in a household is 22 while there is a minimum number of 1 person in a household with an average membership of about 6 and a standard deviation of 3 members. In addition to these, the minimum and the maximum total number of children ever born by a respondent in the sample are 1 and 13 under-fives respectively. The average BMI for respondents in the observed data is 25 kg/m^2 with a minimum and maximum values of 15 kg/m^2 and 57 kg/m^2 respectively.

A pairwise correlation test was done to test for the correlation among the continuous independent variables shown in table 4.6 below. Among the pairwise correlations, only AgeofResp and HHSize showed a weak significant correlation of $0.2277(p - value < 0.0001)$.

Table 4.6: Pearson Correlation Coefficient (p-value) for continuous independent variables

	AgeofResp	**HHSize**	**ChildEverBorn**	**BMI**
AgeofResp	1	0.2277(< 0.0001)	0.0277(0.1964)	-0.0214(0.3227)
HHSize		1	0.0206(0.3368)	-0.0155(0.4740)
ChildEverBorn			1	0.0046(0.8327)
BMI				1

From table 4.7 below, he observed data contains nominal and ordinal categorical variables which were investigated by testing their association with the response variable using the p-values of their respective likelihood ratio and Cochran-Mantel-Haenszel tests at 25% level of significance.

Table 4.7: Likelihood Ratio test and Cochran-Mantel-Haenszel (CMH) test for categorical variables

Variables	**Likelihood Ratio Test**	**CMH Test**
Nominal Variables		
RegofResid	0.0212∗	0.0075
TypeofResid	0.9201	0.9200
Religion	0.0195∗	0.1275
Ethnicity	0.0010∗	0.1597
Lit	0.0071∗	0.0030
MStatus	0.3050	0.1470
RespOccu	0.7007	0.3195
Ordinal Variables		
WIndex	0.3114	0.3679
EduLevel	0.0014	0.2181∗

∗ *shows a significant p-value at* 25% *level of significance*

As shown in table 4.7, the variables that showed non significance at 25% level of significance (Hosmer and Lemeshow, 2000) were dropped while the rest (RegofResid, Religion, Ethnicity, Lit, EduLevel) were included for the model building procedure.

From table 4.8, the observed data contains respondents that were found in the same household. The maximum number of respondents in the same household is 3 and the minimum number is 1 respondent in the same household. Majority of the households contained only one (1) respondent, followed by those with two (2) respondents and few households contained three (3) respondents.

Table 4.8: Household Cluster levels

Household Cluster	**No. of Households**	**No. of Respondents** n(%)
1	1938	1938(89.30)
2	95	190(8.76)
3	14	42(1.94)
Total	**2047**	**2170(100)**

It is evident from the observed data that the number of households with 3 respondents are 14 representing about 42(2%), the number of households with 2 respondents are 95 representing about 190(9%) and the number of households with only 1 respondent are 1,938 representing 89% of the total respondents in sample.

4.3 Model Building Procedure

The model building procedure started from the model which includes all the independent variables which were found in literature or thought to be statistically important. The categorical variables which were seen to be most non-significant at 25% level of significance (Hosmer and Lemeshow, 2000) considering the p-values of their likelihood ratios and Cochran-Mantel-Haenszel tests were dropped. Correlation test was run to check for the pairwise correlation between the continuous independent variables.

The set of variables that were selected for the model building process after the univariate analysis were AgeofResp, HHSize, ChildEverBorn, BMI, RegofResid, Religion, Ethnicity, Lit and EduLevel. These variables were modeled using the SAS stepwise selection criteria until a final model which includes all the variables that are significant at 5%(0.05) level of significance.

The stepwise selection procedure is one of the automatic selection procedures in SAS. It was used to assess and drop insignificant variables for the final model. This type of selection procedure combines elements in the forward and backward selection procedures by allowing variables to be added or removed upon successive revisions of their p-values. The significance level for entry and exit of variables was specified as 5%(0.05).

Table 4.9: Stepwise selection procedure

	Step 1	**Step 2**	**Step 3**	**Step 4**
Effect	P-value	P-value	P-value	P-value
AgeofResp	< 0.0001 ∗∗			
RegofResid	0.0166∗	0.0305∗	0.4401	0.1647
EduLevel	0.0015∗	0.0253∗	0.2954	0.1588
Religion	0.0277∗	0.1005	0.6767	0.5144
Ethnicity	0.0005∗	0.0006∗∗		
HHSize	0.2274	0.0120∗	0.0010∗∗	
Lit	0.0080∗	0.0490∗	0.2888	0.1625
ChildEverborn	0.0240∗	0.0338∗	0.0537	0.0502
BMI	0.2786	0.3281	0.5779	0.5938

∗ *shows a significant p-value at* 5% *level of significance*
∗∗ *variable is selected for the model*

From table 4.9, at the end of the first stage of the selection, seven (7) variables (AgeofResp, RegofResid, EduLevel, Religion, Ethnicity, Lit and ChildEverborn) were significant at 5% level of significance. Out of the seven variables, AgeofResp had the minimum p-value and was subsequently included in the model. In the second stage, six variables (RegofResid, EduLevel, Ethnicity, HHSize, Lit and ChildEverborn) were significant at 5%. Ethnicity had the minimum p-value and was included in the model. In the next stage of the selection procedure, only HHSize was significant and in the final step, none of the remaining variables was significant at 5%. At the end of the selection process, three variables (AgeofResp, HHSize and Ethnicity) were significant at 5% level of significance.

4.3.1 Logistic Regression Model

For the outcome under-five mortality (UFMort), let Y_i be the binary outcome (child died/alive) for respondent i and π_i be the probability of a child dying before the age of five years old. For $Y_i \sim Binomial(\pi_i)$

$$logit(\pi_i) = \beta_o + \beta_1(AgeofResp)_i + \beta_2(HHsize)_i + \beta_3(akan)_i + \beta_4(ewe)_i + \beta_5(ga/dagme)_i$$
$$+ \beta_6(gruma)_i + \beta_7(grussi)_i + \beta_8(guan)_i + \beta_9(mande)_i + \beta_{10}(mole/dagbani)_i$$

Table 4.10: PROC LOGISTIC Parameter estimates (standard errors) and p-values of significant variables

Effect	Parameter	Estimate(Std Error)	P-value
Intercept	β_o	-2.574(0.282)	$< 0.0001*$
AgeofResp	β_1	0.049(0.009)	$< 0.0001*$
HHSize	β_2	-0.084(0.025)	$0.0010*$
Ethnicity akan	β_3	-0.332(0.132)	$0.0116*$
Ethnicity ewe	β_4	-0.833(0.213)	$< 0.0001*$
Ethnicity ga/dangme	β_5	-0.341(0.281)	0.2254
Ethnicity gruma	β_6	0.477(0.215)	$0.0261*$
Ethnicity grussi	β_7	-0.077(0.248)	0.7565
Ethnicity guan	β_8	-0.150(0.338)	0.6575
Ethnicity mande	β_9	1.190(0.451)	$0.0084*$
Ethnicity mole-dagbani	β_{10}	-0.033(0.140)	0.8147

$*$ *shows a significant p-value at* 5% *level of significance*

From the estimated model in table 4.10 above, the logit of a child dying before five years is negatively related to the number of people in the household (HHSize), akan, ewe, ga/dangme, grussi, guan and mole-dagbani ethnic groups. This means that, the larger the household size the less likely it is for a child before five years to die from that household. Also, respondents from the akan, ewe, ga/dangme, grussi, guan and mole-dagbani ethnic groups have a lesser risk of losing their under-five children. On the other hand, the logit of a child dying before five years is positively related to age of respondent (AgeofResp), gruma and mande ethnic groups. This means that, the higher the age of respondent the more likely it was for a child before five years to die from that household. Also, respondents from the gruma and mande ethnic groups have a greater risk of losing their under-five children.

$$
\begin{aligned}
logit(\pi_i) = -2.574 &+ 0.049(AgeofResp)_i - 0.084(HHsize)_i - 0.332(akan)_i - 0.833(ewe)_i \\
&- 0.341(gadangme)_i + 0.477(gruma)_i - 0.077(grussi)_i - 0.150(guan)_i + 1.190(mande)_i \\
&- 0.033(mole/dagbani)_i
\end{aligned}
$$

Odds Ratio Analysis

Table 4.11: Parameters, Odds Ratio (OR) and 95% Confidence Intervals (CI) of significant variables

Effect	**Parameter**	**OR(95% CI)**
Intercept	β_o	
AgeofResp	β_1	1.050(1.033; 1.068)
HHSize	β_2	0.920(0.875; 0.967)
Ethnicity akan vs other	β_3	0.651(0.351; 1,207)
Ethnicity ewe vs other	β_4	0.394(0.192; 0.812)
Ethnicity ga/dangme vs other	β_5	0.645(0.281; 1.477)
Ethnicity gruma vs other	β_6	1.461(0.709; 3.011)
Ethnicity grussi vs other	β_7	0.840(0.387; 1.821)
Ethnicity guan vs other	β_8	0.781(0.309; 1.973)
Ethnicity mande vs other	β_9	2.979(0.952; 9.323)
Ethnicity mole-dagbani vs other	β_{10}	0.877(0.469; 1.643)

The odds ratio of the age of respondent is 1.050 which means that, the multiplicative effect of a year increase in the age of a mother on the odds of losing a child below the fifth birthday is 1.050 when all other covariates as are held constant. The risk of under-five mortality is higher with a year increase in the age of a respondent because the odds ratio is greater than 1 ($1.050 > 1$). Also, the odds ratio of the household size is 0.920 which means that, the multiplicative effect of a unit increase in the number of people in the household on the odds of losing a child below the fifth birthday is 0.920 when all other covariates are held constant. This means that, the risk of under-five mortality is less with a unit increase in the household size since the odds ratio is less than 1 (0.920 < 1).

Respondents who belong to the gruma and mande ethnic groups were respectively 1.461 and 2.979 more likely to lose a child before the fifth birthday than the reference ethnic group (other). The risk of under-five mortality is higher in these ethnic groups as compared to the reference ethnic group (other)because they recorded odds ratios greater than 1 (1.462 > 1 and 2.979 > 1). On the other hand, respondents who were akan, ewe, ga/dangme, grussi, guan and mole-dagbani were respectively 0.651, 0.394, 0.645, 0.840, 0.781 and 0.877 less likely to lose an under-five child than respondents in the reference ethnic group (other). In these ethnic groups, the risk of under-five mortality is less since they recorded odds ratio less than 1 (0.651 < 1, 0.394 < 1, 0.645 < 1, 0.840 < 1, 0.781 < 1 and 0.877 < 1).

4.3.2 Model Diagnostics

Goodness-of-fit Test

Hosmer and Lemeshow test statistic showed a p-value of 0.8017 which was insignificant at 5% level of significance which suggests that the model was a good fit to the data. Hence, the fitted logistic model is appropriate for predicting under-five mortality in the Ghana DHS 2008 data. The Pearson Chi-Square test showed a p-value of 0.3053 confirming a good fit of the data.

Overall Model Evaluation

A comparison between the null model (intercept-only model), the reduced model showed an improvement in the AIC and the SC. The AIC reduced from approximately 1849 to 1806 for the null model and the reduced model respectively while the SC increased from approximately 1854 to 1868 for the null model and the reduced model respectively. There is therefore an indication that, the model fits well for the data.

4.4 Generalized Estimating Equation Analysis

Table 4.12: PROC GENMOD Parameter estimates (Standard errors) for the GEE exchangeable working assumption

Effect	**Parameter**	**Empirical**	**Model-Based**
Intercept	β_o	-2.477(0.398)	-2.477(0.396)
AgeofResp	β_1	0.049(0.009)	0.049(0.009)
HHSize	β_2	-0.084(0.026)	-0.084(0.026)
Ethnicity akan	β_3	-0.429(0.311)	-0.429(0.316)
Ethnicity ewe	β_4	-0.928(0.364)	-0.928(0.369)
Ethnicity ga/dangme	β_5	-0.441(0.423)	-0.441(0.424)
Ethnicity gruma	β_6	0.382(0.365)	0.382(0.369)
Ethnicity grussi	β_7	-0.173(0.389)	-0.173(0.396)
Ethnicity guan	β_8	-0.245(0.483)	-0.245(0.474)
Ethnicity mande	β_9	1.093(0.585)	1.093(0.582)
Ethnicity mole-dagbani	β_{10}	-0.129(0.317)	-0.129(0.320)
Ethnicity other		0.000(0.000)	0.000(0.000)

From table 4.12, the standard errors for the model-based and the empirical estimators show a slight difference, they are however approximately the same across the significant variables. This confirms that, the GEE produces consistent standard errors of the parameter estimates even when the working correlation structure is specified correctly (in the case of the model-based estimator) or specified wrongly (in the case of the empirical estimator).

Theoretically, the specification of the correlation structure does not affect the parameter estimates (Ghisletta and Spini, 2004). In research situations, empirical factors such as the nature of dependence, the number of clusters, the nature of missing values among others may have some influence on the correlation assumption. Hence, the slight differences in the parameter estimates and their respective standard errors of some of the variables.

4.5 Alternating Logistic Regression Analysis

Table 4.13: PROC GENMOD Parameter estimates (Standard errors) for the ALR exchangeable working assumption

Effect	Parameter	Empirical	Model-Based
Intercept	β_o	-2.477(0.398)	-2.477(0.396)
AgeofResp	β_1	0.049(0.009)	0.049(0.009)
HHSize	β_2	-0.084(0.026)	-0.084(0.026)
Ethnicity akan	β_3	-0.429(0.311)	-0.429(0.316)
Ethnicity ewe	β_4	-0.928(0.364)	-0.928(0.369)
Ethnicity ga/dangme	β_5	-0.440(0.423)	-0.440(0.424)
Ethnicity gruma	β_6	0.383(0.365)	0.383(0.370)
Ethnicity grussi	β_7	-0.173(0.389)	-0.173(0.396)
Ethnicity guan	β_8	-0.244(0.483)	-0.244(0.474)
Ethnicity mande	β_9	1.094(0.585)	1.094(0.583)
Ethnicity mole-dagbani	β_{10}	-0.129(0.317)	-0.129(0.321)
Ethnicity other		0.000(0.000)	0.000(0.000)
Association Parameter	α		0.212(0.745)0.776

The measure of association is of interest for the alternating logistic regression analysis. Table 4.13 shows that, the association parameter (log odds ratio) was 0.212 with a p-value of 0.776. There is the indication that, the association within the households is not significant at 5% level of significance, hence household effect is not a contributing risk factor for under-five mortality in the sample. This means that, under-five mortality is not significantly influenced by respondents found in the same household in the Ghana DHS 2008 data.

4.6 Comparison of the study methods

The parameter estimates and standard errors for the factors affecting under-five mortality in the GEE and ALR are closely similar while they are slightly different from that of the logistic regression. This is as a result of the fact that, both the GEE and ALR methods involve computations that alternate between estimating the regression parameters and the intra-cluster association. Both methods rely on the empirical standard error estimation to account for the intra-cluster association

Table 4.14: PROC GENMOD Parameter estimates (Standard errors) obtained using initial Logistic Regression, Generalized Estimating Equations (GEE), and Alternating Logistic Regression (ALR)

	Without Clustering	**Clustering**	
Effect	logistic	GEE	ALR
Intercept	-2.476(0.395)	-2.477(0.398)	-2.477(0.396)
AgeofResp	0.049(0.009)	0.049(0.009)	0.049(0.009)
HHSize	-0.084(0.025)	-0.084(0.026)	-0.084(0.026)
Ethnicity akan	-0.430(0.315)	-0.429(0.311)	-0.429(0.311)
Ethnicity ewe	-0.931(0.368)	-0.928(0.364)	-0.928(0.364)
Ethnicity ga/dangme	-0.439(0.423)	-0.441(0.423)	-0.440(0.423)
Ethnicity gruma	0.379(0.369)	0.382(0.365)	0.383(0.365)
Ethnicity grussi	-0.175(0.395)	-0.173(0.389)	-0.173(0.389)
Ethnicity guan	-0.248(0.473)	-0.245(0.483)	-0.244(0.483)
Ethnicity mande	1.092(0.582)	1.093(0.585)	1.094(0.585)
Ethnicity mole-dagbani	-0.129(0.317)	-0.129(0.317)	-0.129(0.321)
Ethnicity other	0.000(0.000)	0.000(0.000)	0.000(0.000)
Association Parameter			0.212(0.745)

to give results for risk factor model even when there is a misspecification of the association model. The ALR provides estimates of the pairwise odds ratio for interpreting the magnitude of intra-cluster association.

4.7 Discussion

Out of the thirteen (13) independent variables that were selected for the analysis, three (3) were significant as the risk factors that affect under-five mortality in the Ghana DHS 2008 data. Two of the significant risk factors are continuous (AgeofResp and HHSize) and one was categorical (Ethnicity). The age of respondent (AgeofResp) was positively related to the risk of under-five mortality which means that the higher the age of the respondent, the more likely it was for her to lose a child who is under-five and vice versa. The number of people in the household (HHSize) was negatively related to the risk of under-five mortality which means that the higher the number of people in the household, the less likely it was for an under-five child to die from that household. The only categorical variable that was significant was ethnicity which contains these categories; akan, ewe, ga/dangme, gruma, grussi, guan, mande, mole-dagbani and other. Four of the ethnic groups namely (akan, ewe, gruma and mande) were significant whiles the others were not.

Considering the respondents in the same household, the GEE and ALR were fitted. From the GEE, whether the working correlation working assumption is correctly specified or not, parameter estimates were consistent and their respective standard errors were efficient using the empirical estimator. Under the exchangeable working assumption where a constant correlation was assumed for respondents within the same household, some of the parameter estimates were significant at 5% level of significance while others were not. Taking association into account, the ALR was fitted where the association parameter was not significant. This means that, household effect is not a significant contributor on under-five mortality in the Ghana DHS 2008 data.

Chapter 5

Conclusion and Recommendations

5.1 Introduction

This chapter outlines the conclusion of the study which highlights the major findings and the recommendations based on the major findings. The study sought to investigate the risk factors and their effects on under-five mortality in the Ghana DHS 2008. The two main methods used in the study are the logistic regression which is a type of the generalized linear models and the alternating logistic regression which is a type of the generalized estimating equations. The data was analyzed using the SAS version 9.1 for windows. The analysis of the observed data started with the selection and exploration of the selected variables, model building procedure and the estimation of the parameter coefficients of the significant predictor variables as well as the association parameter.

5.2 Conclusion

Based on the summary of the results, it is concluded that; the significant risk factors of under-five mortality in Ghana DHS 2008 data are the age of respondent, the number of people in the household and ethnicity. The age of respondent was positively related to the risk of under-five mortality. The multiplicative effect of a year increase in the age of a respondent on the odds of losing a child before the fifth birthday is 1.050. Furthermore, the number of people in the household was negatively related to the risk of under-five mortality. The multiplicative effect of a unit increase in the number of people in the household on the odds of losing a child before the fifth birthday is 0.920.

Respondents from the Akan, the Ewe, the Gruma and the Mande ethnic groups were significant in terms of the effect of ethnicity on under-five mortality. The risk of under-five mortality is higher in the Gruma and the Mande ethnic groups as compared to the "other" ethnic group. Respondents who belong to the Gruma and the Mande ethnic groups were respectively 1.461 and 2.979 more likely to lose an under-five child than the reference ethnic group (other). Also, the risk of under-five mortality was lower in the Akan and the Ewe ethnic groups as compared to the "other" ethnic group. Respondents from the Akan and the Ewe ethnic groups were respectively 0.651 and 0.394 less likely to lose an under-five child than the reference ethnic group (other).

Furthermore, household effect is not a significant contributing risk factor for under-five mortality in the Ghana DHS 2008 data. This means that, factors that are similar in the same household have insignificant effect on under-five mortality in those households.

5.3 Recommendations

In furtherance of the findings from the investigation, it is recommended that;

- Education and health intervention policies should be designed to reach the various ethnic groups in Ghana

- very old women should be discourage from giving birth as they stand a greater risk of losing their under-five children

- further research should be conducted in subsequent Ghana Demographic and Health Surveys in order to monitor the causes and progress of under-five mortality.

REFERENCES

Agha, S. (2000). The determinants of infant mortality in pakistan. *Social Science & Medicine*, 51:199–208.

Agresti, A. (2002). *Categorical Data Analysis.* John Wiley & Sons, Inc., Hoboken, New Jersey.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* John Wiley & Sons, Inc.

Barros, A. J. D. and Hirakata, V. N. (2003). Alternatives for logistic regression in cross-sectional studies: am empirical comparison of models that directly estimate the prevalence ratio. *BMC Biomedical Research Methodology*, 3:3–13.

Carey, V., Zeger, S. L., and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80:517–526.

Dobson, A. . (2002). *An Introduction to Generalized Linear Models.* Chapman & Hall/CRC.

Fan, J., Wu, Y., and Fend, Y. (2009). Local quasi-likelihood with a parametric guide. *The Annals of Statistics*, 37:4153–4183.

Foggin, P., Armijo-Hussein, N., Marigaux, C., Zhu, H., and Liu, Z. (2001). Risk factors and child mortality among the miao in yunnan, southwest china. *Social Science & Medicine*, 53:1683–1696.

Folasade, I. B. (2000). Environmental fcators, situation of women and child mortality in southwestern nigeria. *Social Science & Medicine*, 51:1473–1489.

Ghisletta, P. and Spini, D. (2004). An introduction to generalized estimating equations and an application to assess selectivity effects in a longitudinal study on very old individuals. *Journal of Educational and Behavioral Statistics*, 29:421–437.

GSS/GHS/ICF (2009). Ghana demographic and health survey 2008. Technical report, Ghana Statistical Service (GSS), Ghana Health Service (GHS) and ICF Macro.

GSS/ICF (2010). Millennium development goals in ghana: A new look at data from the 2008 ghana demographic and health survey. Technical report, ICF Macro.

GSS/MI (1989). Ghana demographic and health survey 1988. Technical report, Ghana Statistical Service (GSS) and Macro Systems Inc.

GSS/MI (1994). Ghana demographic and health survey 1993. Technical report, Ghana Statistical Service (GSS) and Macro International Inc. (MI).

GSS/MI (1999). Ghana demographic and health survey 1998. Technical report, Ghana Statistical Service (GSS) and Macro International Inc. (MI).

GSS/NMIMR/ORC (2004). Ghana demographic and health survey 2003. Technical report, Ghana Statistical Service (GSS), Noguchi Memorial Institute for Medical Research (NMIMR), and ORC Macro.

Hallet, D. C. (1999). Goodness of fit tests in logistic regression. Master's thesis, University of Toronto.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression.* John Wiley & Sons, Inc.

Isaac, O.-D. (2011). Application of generalized estimating equation (gee) model on students academic performance: A case study of final year math four students at knust. Master's thesis, Kwame Nkrumah University of Science and Technology.

Jang, M. J. (2011). *Working correlation selection in Generalizeded Estimating equations.* PhD thesis, University of Iowa.

Kojo, K. (2012). Modelling the risk factors of neonatal mortality in ghana using logistic regression. Master's thesis, Kwame Nkrumah University of Science and Technology.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Liang, K.-Y., Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54:3–40.

Lin, S.-J. (2006). The effects of economic instability on infant, neonatal, and post-neonatal mortality rates: Evidence from taiwan. *Social Science & Medicine*, 62:2137–2150.

McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics*, 11:59–67.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.*

Mondal, M. N. I., Hossain, M. K., and Ali, M. K. (2009). Factors influencing infant and child mortality: A case study of rajshahi district, bangladesh. *J Hum Ecol*, 26:31–39.

NDPC/GOG (2012). 2010 ghana millennium development goals report. Technical report, UNDP and Government of Ghana/NDPC.

Pan, W. and Connett, J. E. (2002). Selecting the working correlation structure in generalized estimating equations with application to the lung health study. *Statistica Sinica*, 12:475–490.

Peng, C.-Y. J., Lee, K. L., and Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96:3–14.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44:1033–1048.

Rahman, K. M. M. and Sarkar, P. (2009). Determinants of infant and child mortality in bangladesh. *Pakistan Journal of Social Sciences*, 6(3):175–180.

Reed, P. and Wu, Y. (2013). Logistic regression for risk factor modelling in stuttering research. *Journal of Fluency Disorders*, 38:88–101.

Sastry, N. (1997). What explains rural-urban differentials in child mortality in brazil. *Soc. Sci. Med.*, 44:989–1002.

Staley, T. (2013). The case for generalized estimating equations in state-level analysis. In *State Politics and Policy Conference*.

Stokes, M. E., Davis, C. S., and Koch, G. G. (2000). *Categorical Data Analysis Using the SAS System*. SAS Institute Inc., Cary, NC.

Suwal, J. V. (2001). The main determinants of infant mortality in nepal. *Social Science & Medicine*, 53:1667–1681.

UNICEF (2012a). Committing to child survival: A promise renewed progress report 2012. Technical report, United Nations Children's Fund.

UNICEF (2012b). Levels & trends in child mortality report 2012. Technical report, United Nations Children's Fund, World Health Organization, The World Bank and United Nations.

UNICEF (2013a). Committing to child survival: A promise renewed progress report 2013. Technical report, United Nations Children's Fund.

UNICEF (2013b). Levels & trends in child mortality report 2013. Technical report, United Nations Children's Fund, World Health Organization, The World Bank and United Nations.

UN/MDG (2012). The millennium development goals report 2012. Technical report, United Nations.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61:439–447.

Zorn, C. J. W. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, 45:470–490.

Appendix A

SAS CODE

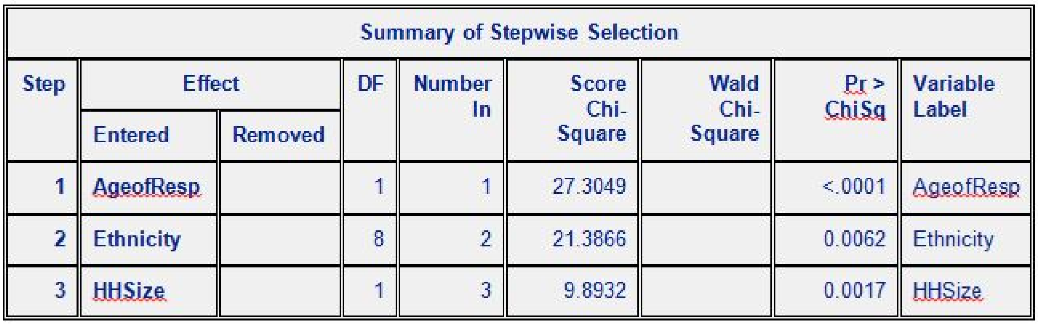
```
*library name GDHS  in mythesis folder;
libname GDHS 'C:\mythesis';

*import data;

*view imported data;
proc print data=GDHS.dhs; run;

*check content of data;
proc contents data=GDHS.dhs; run;

*frequency count for study variables;
proc freq data=GDHS.dhs; run;

*summary statistics for continous variables;
proc means data=GDHS.dhs;
var AgeofResp HHSize ChildEverBorn BMI; run;

*frequency count by under-five mortality;
proc sort data=GDHS.dhs; by  UFMort; run;
proc freq data=GDHS.dhs; by UFMort; run;

*Chi Square test for catedorical variables 2 by 2;
proc freq data=GDHS.dhs order=data;
tables RegofResid*UFMort /chisq norow nocol;
tables TypeofResid*UFMort /chisq norow nocol;
tables EduLevel*UFMort /chisq norow nocol;
tables Religion*UFMort /chisq norow nocol;
tables Ethnicity*UFMort /chisq norow nocol;
tables Lit*UFMort /chisq norow nocol;
tables MStatus*UFMort /chisq norow nocol;
tables RespOccu*UFMort /chisq norow nocol;
tables WIndex*UFMort /chisq norow nocol;
run;

* Stepwise model selection;
proc logistic data=GDHS.dhs;
class RegofResid EduLevel Religion Ethnicity Lit;
model UFMort=AgeofResp RegofResid EduLevel Religion Ethnicity HHSize Lit
ChildEverBorn BMI / selection=stepwise sle=0.05 sls=0.05 details;
run;

*proc LOGISTIC;
proc logistic data=GDHS.dhs descending;
class Ethnicity;
model UFMort=AgeofResp Ethnicity HHSize /scale=none aggregate lackfit;
run;

*proc GENMOD;
proc genmod data=GDHS.dhs descending;
class Ethnicity;
model UFMort=AgeofResp Ethnicity HHSize /dist=bin link=logit type3;
run;

*GEE;
```

```
proc sort data=GDHS.dhs; by HHCluster; run;

proc genmod data=GDHS.dhs descending;

class Ethnicity HHCluster;

model UFMort=AgeofResp Ethnicity HHSize /dist=bin link=logit type3;

repeated subject=HHCluster / type=exch covb corrw modelse;

run;


*ALR;

proc sort data=GDHS.dhs; by HHCluster; run;

proc genmod data=GDHS.dhs descending;

class Ethnicity HHCluster;

model UFMort=AgeofResp Ethnicity HHSize /dist=bin link=logit type3;

repeated subject=HHCluster / logor=exch covb corrw modelse;

run;
```

Appendix B

SELECTED SAS OUTPUT

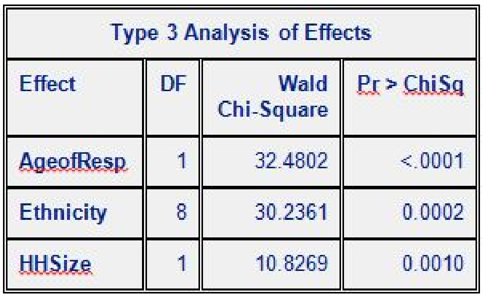
The stepwise selection



Summary of Stepwise Selection

Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
	Entered	Removed						
1	AgeofResp		1	1	27.3049		<.0001	AgeofResp
2	Ethnicity		8	2	21.3866		0.0062	Ethnicity
3	HHSize		1	3	9.8932		0.0017	HHSize

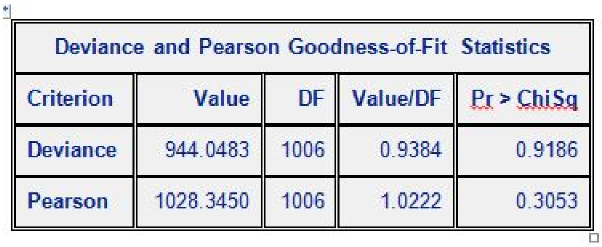
Analysis of significant variables



Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
AgeofResp	1	32.4802	<.0001
Ethnicity	8	30.2361	0.0002
HHSize	1	10.8269	0.0010

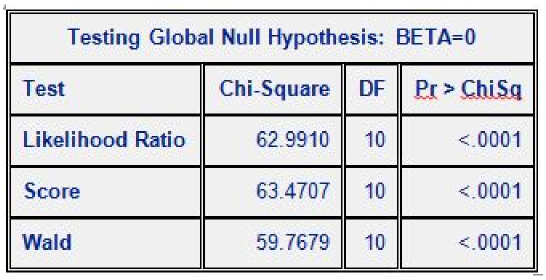
The Deviance and Pearson Goodness-of-fit test statistic



Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	944.0483	1006	0.9384	0.9186
Pearson	1028.3450	1006	1.0222	0.3053

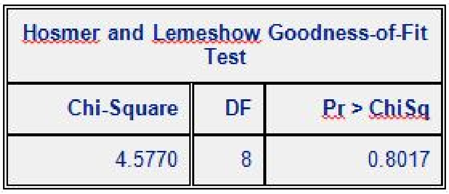
Test of global hypothesis



Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	62.9910	10	<.0001
Score	63.4707	10	<.0001
Wald	59.7679	10	<.0001

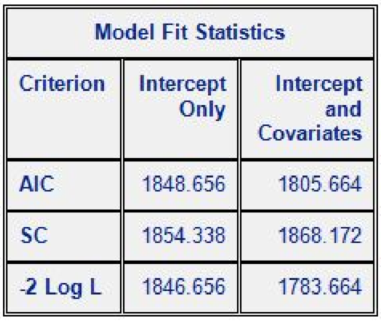
The Hosmer-Lemeshow test statistic



Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
4.5770	8	0.8017

Model fit statistics



Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1848.656	1805.664
SC	1854.338	1868.172
-2 Log L	1846.656	1783.664