Statistical models to study the BMI of under five children in Ethiopia

UNIVERSITY OF
KWAZULU - NATAL
INYUVESI
YAKWAZULU-NATALI

Ashenafi Argaw Yirga

August, 2018

Statistical models to study the BMI of under five children in Ethiopia

by

Ashenafi Argaw Yirga

A thesis submitted to the
University of KwaZulu-Natal
in fulfilment of the requirements for the degree
of
MASTER OF SCIENCE
in
STATISTICS

under the supervision of
Dr. Sileshi Fanta Melesse
Prof. Henry Godwell Mwambi
and
Dr. Dawit Getnet Ayele

UNIVERSITY OF
KWAZULU - NATAL
INYUVESI
YAKWAZULU-NATALI

UNIVERSITY OF KWAZULU-NATAL
SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE
PIETERMARITZBURG CAMPUS, SOUTH AFRICA.

Declaration

I, Ashenafi Argaw Yirga, declare that this dissertation titled ’Statistical models to study the BMI of under five children in Ethiopia’ and the work presented in it are my own original work. I confirm that:

- This thesis has not been submitted for any degree or examination at any other university.
- This thesis does not contain other persons’ data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
- This thesis does not contain other persons’ writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then
    (a) their words have been re-written but the general information attributed to them has been referenced, or
    (b) where their exact words have been used, then their writing has been placed in italics and referenced.
- This thesis does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the thesis and in the reference sections.

Ashenafi Argaw Yirga Date

Dr. Sileshi Fanta Melesse Date

Prof. Henry Godwell Mwambi Date

Dr. Dawit Getnet Ayele Date

Disclaimer

This document describes work undertaken as a Masters programme of study at the University of KwaZulu-Natal (UKZN). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institution.

Note

Articles that have been published from this thesis:

1. Application of Quantile Regression: Modeling Body Mass Index in Ethiopia (Yirga et al., 2018a).
2. Factors affecting child malnutrition in Ethiopia based on an ordinal response regression model (Yirga et al., 2018b). (Accepted for publication)

Acknowledgements

Glory be to God who is glorified in His Saints. Amen.

I am grateful to all of those with whom I have had the pleasure to work during this and previous Honors study. I would especially like to thank Dr. Sileshi F. Melesse, the main supervisor of this work. As my teacher and mentor, he has taught me, by his example, what a good researcher should be, more than I could ever give him credit for here. I am especially indebted to Prof. Henry G. Mwambi, who has been supportive of my career goals. His understanding, knowledge, patience and professionalism added considerable quality to my research work. I am also grateful to Dr. Dawit G. Ayele, for his immense intellectual guidance, advice, encouragement and robust contribution towards the successful completion of this work. Each of the members of this dissertation has provided me with extensive personal and professional guidance and taught me a great deal about both scientific research and life in general. It was a great privilege and honor to work and study under their guidance.

This work was supported through the DELTAS Africa Initiative. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Welcome Trust [Grant No. 107754/Z/15/Z- DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme] and the UK government. The views expressed in this work are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Welcome Trust or the UK government.

I would like to thank, with deep appreciation, ORC Macro and Measure DHS for giving us access to the data file. I would like to thank the staff members of the University of KwaZulu-Natal for the hospitable environment they provided during my study. I would also like to thank my fellow colleagues for their feedback and cooperation. Special thanks to Mr. Getachew Zenebe and Mr. Addis Habtamu for their encouragement and especially to my sister Tigist Argaw for her constant encouragement to continue my study.

Abstract

Maternal and child malnutrition has long and short-term consequences on the health status of the people and on the country's economy. It is among the major public health problems in Ethiopia. Worldwide, maternal and child malnutrition is an underlying cause for more than 3.5 million deaths each year. About 35% of the global disease burden is in under five children. Such a heavy burden requires an understanding of the nutritional status of the people, especially children under the age of five years and associated factors. Therefore, this study attempted to use possible statistical methods to estimate the effects of the risks related to the nutritional status of children. It also tried to identify the socio-economic and demographic factors that are associated with the BMI of under five children in Ethiopia. The study employed the 2016 Ethiopian Demographic and Health Survey data. A nationally representative sample of children under the age of five years was used to get information on weight and height measures of under five children.

The BMI of children under five years of age was used as a response variable to fit weighted quantile regression. The covariates, age of a child, sex and other relevant socio-economic and demographic factors were used in the study. Following the quantile regression, the generalized linear models such as logistic regression model was applied after categorizing the response variable, BMI of under five children, into two categories namely normal and malnourished. Following binary logistic regression, an attempt to fit ordinal logistic regression was made. That means nutritional status was considered as ordinal outcome with four categories namely underweight, normal, overweight and obese. The findings and comparison of estimates using these different statistical methods with and without complex survey design were presented. The results revealed that methods that take into account the complex nature of the design, perform better than those that do not take this into account. It has also been found that age of a child, weight of child at birth, mother's BMI, educational attainment of mother, region and wealth index were significantly associated with under five children's nutritional status. Furthermore, the results are discussed and then a conclusion is made in the context of policy implication for Ethiopia.

Contents

Page

List of Figures

List of Tables

Abbreviations

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BMI	Body Mass Index
BRR	Balanced Repeated Replication
CDC	United States Centers for Disease Control and Prevention
CED	Chronic Energy Deficiency
CI	Confidence Interval
CSA	Central Statistical Agency
CSD	Complex Survey Design
Deff	Design Effect
EA	Enumeration Area
EDHS	Ethiopian Demographic and Health Survey
GLM	Generalized Linear Model
HIV/AIDS	Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome
IRLS	Iterative Re-weighted Least Square
Kebele	Ethiopian (Amharic) term for a village or community
MDGs	Millennium Development Goals
ML	Maximum Likelihood
MoH	Ethiopian Ministry of Health
MOI	Ethiopian Ministry of Information
OR	Odds Ratio
POM	Proportional Odds Model
PSU	Primary Sampling Unit
ROC	Receiver Operating Characteristic Curve
SAS	Statistical Analysis System
SSU	Secondary Sampling Unit
SNNPR	Southern Nations, Nationalities and People's Region
SPSS	Statistical Package for Social Science
Stata	Statistical Software (Statistics and data)
UN	United Nations
UNAIDS	Joint United Nations Programme on HIV and AIDS
UNDP	United Nations Development Programme
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNICEF	United Nations Children's Fund
WHO	World Health Organization
WQR	Weighted Quantile Regression

In memory of my grand mother emama Lezeb, my sister Hana and nephew Biniam.

Chapter 1

Introduction

1.1 Country overview: Ethiopia

Ethiopia is located in the horn of Africa and it is the second most populous nation (after Nigeria) in Africa with around 106.6 million inhabitants (Worldometers, 2018). The total area of the country is estimated to be 1,104,300 km^2 (426,400 square miles). Ethiopia is bordered by Djibouti, Eritrea, Kenya, Somali, Sudan and South Sudan (Figure 1.1). Currently, Ethiopia is divided into nine regional states: Afar, Amhara, Benishangul-Gumuz, Gambella, Harari, Oromia, Somali, Southern Nation Nationalities and Peoples (SNNP) and Tigray. In addition, there are two administrative cities namely Addis Abeba (the capital and the largest city) and Dire Dawa. Amharic is the working language of the Federal government and there are over 80 different languages, consisting of Semitic, Cushitic, Omotic, and Nilo Saharan languages (MOI, 2004). The Ethiopian Highlands are the largest continuous mountain ranges in Africa. The highest peak is *Ras Dashin* at 4550m. The *Afar* depression at 110 m below sea level, is the hottest place in the country. Lake *Tana* in the north is the source of the Blue Nile (*Abbay*). Ethiopia is rich in faunal, floral and microbial diversity. It also has many species of endemic animals and plants. The *sof Omar* Caves contain the largest cave on the continent. The country is home to nine UNESCO Heritage sites. Anthropologists believe that East Africa's Great Rift Valley is the site of humankind origins; the valley traverses Ethiopia from southwest to northeast (CSA, 2012).

Ethiopia was a founding member of the UN, the organization of African unity, and the African base for many international organizations (CIA, 2018; CSA, 2012). The country is home to more than 80 ethnic groups, which vary in population size from more than 26 million people to fewer than 100 (CSA, 2012). Ethiopia has a unique alphabet, numerals and calendar that has existed for more than 3000 years. The official Ethiopian calendar (Ge'ez calendar) has twelve months of exactly 30 days each plus five or six epagomenal days. It is approximately seven years and three months behind the Gregorian calendar. This resulted from an alternate calculation in determining the date of the Annunciation of Christ Jesus. The first day of the Ethiopian year is 11 September. Time count is different in Ethiopia, one must add or subtract 6 hours to count as the western time (Molla, 2016; Tseday, 2008).

Ethiopia has suffered periodic droughts, poverty, political repression, and forced government resettlement that led to a long civil conflict in the 20th century. The current government is repeatedly reports that the country is on track to meet the Millennium Development Goal (MDGs) to eradicate extreme hunger and poverty and combat HIV/AIDS, malaria, tuberculosis (TB) and other diseases. However, Ethiopia remains one of the world's least developed countries with average per capita income less than half of the current sub-Saharan Africa average (Lives & Nations, 2015). Household food insecurity and undernutrition remain critical issues; undernutrition status of women and children, are consistent problems in the country. Undernutrition is an underlying cause of 53% of infant and child deaths in the country (USAID, 2014). The general measures of health and nutritional status are assessed at the population level through the Demographic and Health survey in the country (CSA, 2012).

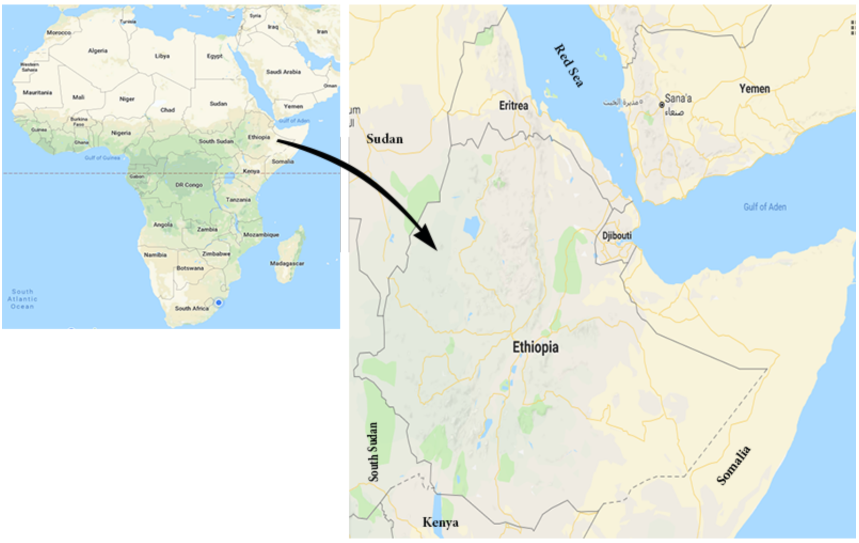

Figure 1.1 – Maps showing location of Ethiopia in Africa

Source:https://www.google.co.za/maps/place/Ethiopia/

1.2 Background

The health status of the people is the treasure of any country and nutrition is one of the most vital preconditions for good health. Child malnutrition is a very common public health problem in the world. Good nutrition is an essential determinant for children's well-being. The nutritional status of children under the age of five is an important outcome measure of children's health. For this reason, a national nutrition strategy and program has been developed and implemented by the government of Ethiopia. One of the objectives of the 2009 Ethiopian National Nutrition Strategy was to enhance good nutritional practices through health education, and treatment of micronutrients to the most vulnerable groups of the society, particularly under five children, pregnant and lactating mothers. The Health Sector Development plan IV (2010/11-2014/15) mainly strives to improve the nutritional status of mothers and children through the following programs: Enhanced Outreach Strategy (EOS) with Targeted Supplementary Food (TSF) and Transitioning of EOS into health extension programme (HEP), Health Facility Nutrition Services, Community Based Nutrition (CBN), and Micronutrient Interventions and Essential Nutrition Actions/Integrated Infant and Young Feeding Counseling Services (CSA, 2012). However, the poor nutritional status of children and women has been a severe problem in Ethiopia. In the 2016 Ethiopian Demographic and Health Survey (EDHS), children's nutritional status and health data were collected. In this nationally representative sample survey, measurements of children's weight and height were recorded. The purpose of taking these anthropometric measurements was to determine if children are growing ″normally″. A child's weight or size at birth is an important indicator of the child's vulnerability to the risk of childhood illnesses and the child's chances of survival. Children whose birth weight is less than 2.5 kilograms, or children reported to be ″very small″ or smaller than average, have a higher than average risk of early childhood death (CSA, 2016). Since most the births do not take place in health facilities, children are less likely to be weighed at birth in non-institutional settings. The mother's estimate of the baby's size at birth was used in the 2016 EDHS data. Only 5 percent of children in Ethiopia are weighed at birth (CSA, 2016).

The 2016 EDHS data were collected to calculate three indices of anthropometric indicators:- weight-for-age, height-for-age, and weight-for-height. The weight-for-height index measures body mass index in relation to body height; it describes children's nutritional status.

1.3 Body mass index of under five children (BMI-for-age)

Body mass index is defined as the ratio of weight (kg) to squared height (m^2). It is a measure of nutritional status (Keys et al., 1972). It provides a good indicator for levels of body fat. However, BMI is not a direct measure of body fatness. Having a BMI that is either too low or too high is associated with an increased risk of ill health. BMI is the most frequently used measure for assessing whether adults or children are underweight, a healthy weight, overweight, or obese. The BMI of adults remains relatively constant, regardless of age, unless they gain or lose a lot of weight. Assessing the BMI of children is more complicated than for adults because a child's BMI changes as they mature. Children's body fatness changes over the years as they grow. Also, girls and boys differ in their body fatness as they mature. This is why BMI for children, also referred to as BMI-for-age, is gender and age specific. If a person's BMI is out of the healthy BMI range, the risks of illness or death may increase significantly. A high amount of body fat in persons or children can lead to weight related diseases and other health issues. Being underweight can also put one at risk for health issues. Table 1.1 provides recommended BMI cutoffs for children under five years, based on many research studies.

Table 1.1: BMI-for-age cut-offs

Percentile range	Weights status
$< (5^{th})$ percentile	Underweight
(5^{th}) to (85^{th}) percentile	Normal or healthy weight
(85^{th}) to (95^{th}) percentile	Overweight
$\geq (95^{th})$ percentile	Obese

The expert committees' recommendations are to classify BMI-for-age at or above the (95^{th}) percentile as obese, between the (85^{th}) and (95^{th}) percentile as at risk of overweight and between the (5^{th}) and (85^{th}) percentile as normal (health weight) (Himes & Dietz, 1994). The cutoff for underweight of less than the (5^{th}) percentile is based on recommendations by the World Health Organization expert committee on physical status (World Health Organization, 1996).

According to the 2016 EDHS report, overall, male children are slightly more likely to be wasted (11%) than female children (8%); 38% of children under age of 5 are stunted or too short for their age, and 18% severely stunted. Ten percent are wasted or too thin for their height, including 3% who are severely wasted. Twenty-four percent of children under age of 5 are underweight or too thin for their age, with 7% severely underweight. The prevalence of overweight children remained low at 1%. Ten percent of children in rural areas are wasted, compared with 6% in urban areas. Wasting is most common in the children of mothers with BMI of less than 18.5 (15%), in those residing in the Somali region (22%), and in children whose mothers have no education (11%). Wasting, or low weight for height, is a strong predictor of mortality among children under five years of age. It is usually the result of acute significant food shortage and/or disease (CSA, 2016).

According to UNICEF-progress for children 2007 report, there were 24 developing countries with wasting rates of 10% or more, indicating a serious problem urgently requiring a response. The highest child malnutrition is found in the Sub-Saharan Africa countries. Ethiopia is among those countries with the highest rate of stunting in Sub-Saharan Africa. The proportion of underweight children is highest in the age range of 2 to 3 years (34%) and lowest among those under six months of age (10%). In general, 29% of children under the age of five are underweight, and 9% are severely underweight in Ethiopia. An estimated 159 million children under five years of age, or 23.8%, were stunting in 2016, 15.8% decrease from an estimated 255 million in 1990 worldwide (Achadi et al., 2016). Even though the occurrence of stunting and underweight among children under five years of age worldwide has decreased since 1990, overall improvement is unsatisfactory and millions of children remain at risk (De Onis et al., 2012).

In Ethiopia, babies reported as very small or small at birth are much more likely to be underweight later in life (39%) and (36%) respectively, than those reported as average or large at birth (25%). Children born to mothers who are thin (BMI less than 18.5) are more than three times as likely to be underweight (39%) as children born to mothers who are overweight/obese (12%). The proportion of underweight children is eight times higher for those born to uneducated mothers than for those whose mothers have higher than secondary education (32%) compared with 4%. Rural children are more likely to be underweight (30%) than urban children (16%). The proportion of underweight children varies by region. Amhara, Benishangul-Gumuz, Affar, and Dire Dawa are most highly affected by child stunting ($41 - 46\%$), whereas wasting imposes the heaviest burden in Somali, Affar, and Gambela, with rates of 23%, 18%, and 14%, respectively (CSA, 2016). The proportion of underweight children decreases as the wealth quintile of mother increases. Children born to mothers in the lowest wealth quintile are more than twice as likely to be underweight as children born to mothers in the highest wealth quintile (36%) compared with 15% (CSA, 2016). Overweight or obesity among children increases with increasing BMI of the mother, from 1% among children of mothers who are thin to 4% among children of mothers who are overweight/obese (BMI ≥ 25). Variation by region is minimal except for Addis Abeba, where 6% of children under five, the highest percentage in all regions, are overweight or obese. Globally, an estimated 43 million children under five years of age, or 7 percent, were overweight in 2011, a 54 percent increase from an estimated 28 million in 1990. Increasing trends in child overweight have been noted in most of the world's regions, not only in developed countries, where prevalence is highest (15%) in 2011. In Africa, the estimated prevalence of under five overweight increased from 4 percent in 1990 to 7 percent in 2011 (De Onis et al., 2012).

The nutritional status and/or weight status of under five children is a great concern. This is because the early years of life are very important for future growth and development. The children are the future citizens of the country; we have responsibility as parents to formulate and shape their present conditions in the best viable way. BMI is the most frequently used measure for assessing children's nutritional status and/or weight status. It is also related to health risks and can be a good indicator of the health status of individuals. Therefore, identifying factors that affect the BMI of under five children is very important for possible intervention activities. It can also assist policy makers to know and understand the areas that need considerable attention to enhance the planning and evaluation of health policies to prevent a child's death and to determine a child's health, diet and growth.

1.4 Objectives

The main objective of this study is to identify factors that affect under five children's nutrition status by fitting the most parsimonious statistical models, yet biologically reasonable models to describe the relationship between under five children's BMI and a set of independent variables such as: age, sex of a child and other relevant factors with reference to the 2016 Ethiopian DHS.

The specific objectives are:-

- To assess nutritional status and characteristics related to BMI in children under five years of age in Ethiopia.
- To examine the effects of selected socio-economic and demographic factors on BMI among children under five years of age in the country.
- To identify factors that contribute to the BMI in under five children using WQR model in the modeling of BMI of under five children.
- Applying logistic regression and OLR model without and with complex survey design, to identify factors that affect under five children BMI outcome.

1.5 Outline of the study

The thesis is divided into six chapters. This introduction section gave some background about BMI, objective of the study, advantages of studying BMI-for-age, and some existing information on under five children's BMI by reviewing research papers that have been done in this field. In Chapter 2, preliminary data analysis and description of the study variables are presented. In Chapter 3, we study weighted quantile regression where the response variable is under five children's BMI, which is a continuous variable. Statistical methods for binary outcome are used after categorizing the response variable under five children's BMI into two categories namely normal/malnutritioned in Chapter 4. The generalized linear models (GLMs) such as logistic regression model, which can be used to fit binary response are applied. Moreover, since EDHS is survey data, survey logistic regression is also applied in Chapter 4. A study of ordinal logistic regression model is made in Chapter 5. Ordinal logistic regression model considers any inherent ordering of the levels in the outcome variable, thus making complete use of the ordinal information. The comparison of results obtained from each model without and with complex survey design is also presented in Chapter 4 and in Chapter 5. Finally, in Chapter 6 the discussions and conclusions as well as possibilities for future research are presented.

1.6 Literature review

According to the Ethiopian Ministry of Economic Development and Cooperation (1999) study, 50 percent of the Ethiopian population are living below the food poverty line and cannot meet their daily minimum nutritional requirement of 2200 calories. Due to low dietary intakes, low agricultural production, inequitable distribution of food within the household, improper food storage and preparation, falling gross national product per capita, dietary taboos, infectious diseases and care, women in the reproductive age group and children in Ethiopia are the most vulnerable to improper nutritional status and/or weight status. Drought, civil war and political instability are also major contributing factors (Getahun et al., 2017). Investing in women's and children's nutrition will have both short-term and long-term effects on the social and economic wellbeing of not only the individual but the community and the nation at large (Garcia & Mason, 1992).

A recent study in Oromia region showed that 35 percent of non-pregnant women in this region had a BMI lower than 18.5, indicative of a high probability of getting underweighted children (Getahun et al., 2017). Underweight is commonly used as an indicator for malnutrition. It is influenced by the height and weight of a child/person and is thus a composite nature of stunting, and wasting makes interpretation complex (Kalhan et al., 2009). Malnutrition includes a wide range of nutrient-related deficiencies and disorders whether it is due to dietary deficiency, called under-nutrition (underweight and stunting), or to excess diet, called over-nutrition (overweight and obese) (Ratzan et al., 2000). The mean prevalence of underweight in developing countries is 31%, ranging from 6.5% in South America to around 51% in South Asia. A review of the trends of the nutritional status of Ethiopian children from 1983-1998 showed that the national rural prevalence of stunting increased from 60% in 1983 to 64% in 1992. The prevalence of underweight reported in the 1998 survey was 42%. This shows that the number of underweight children in Ethiopia is still higher than the mean for developing countries. For the country as a whole, the prevalence of underweight for under five children increased from 37.3% in 1983 to 46.9% in 1992. The prevalence of underweight was 47.1% from EDHS, 2000, showing that the situation is no different from the 1992-prevalence. In the 1998 survey, the prevalence of underweight is 54.4% for Tigray and 52.2% for Amhara region. From these observations, we can say that no progress was made in reducing the prevalence of underweight children in the last 17 years (Getahun et al., 2017).

Considering Ethiopia's position in the rate of stunting in sub-Saharan Africa, Ethiopia had the highest rate of stunting. Two countries, Nigeria and Ethiopia, accounted for about half (52%) of the stunted children in sub-Saharan Africa in 1995 (Getahun et al., 2017).

1.7 Review of the study variables

The socio-economic and demographic factors used in this study were supported by several researchers as most likely referred to as intermediate variables for the determinants of children's nutritional status (Hien & Hoa, 2009). According to different research studies, some of the common socio-economic and demographic factors that affect under five childrens BMI are reviewed below:

Wealth index

It serves as an indicator of level of wealth that consists of expenditure and income measures (Rutstein, 1999). The index is constructed using household assets data via a principal component analysis. The wealth index of a household is an indicator of access to adequate food supplies, use of health services, availability of improved water sources, and sanitation facilities, which are prime determinants of child and maternal nutritional status (Unicef et al., 1990). Wealth of households is calculated through household assets collected from DHS surveys -i.e., type of flooring; source of water; availability of electricity; possession of durable consumer goods. These are combined into a single wealth index (CSA, 2016).

WHO (2008) reported that low BMI (< 18.5) is common in women in low-income countries. A study in the Southern Nations, Nationalities and Peoples Region (SNNPR) of Ethiopia revealed that women belonging to low economic status households were mostly affected by malnutrition (Teller & Yimer, 2000). The prevalence of malnutrition and food insecurity increases as the household income decreases. The risk of being underweight is greater among children from households with a low or very low socio-economic status as compared to children from households with middle or upper socio-economic status. Comparative studies on child nutrition for many countries and results of some local studies in Ethiopia (Getaneh et al., 1998; Sommerfelt & Stewart, 1994; Genebo et al., 2017; Yimer, 2000) reveals that the level of child stunting is lower in households with a higher level of economic status.

Educational attainment of mother

Education is a key determinant of individual opportunities, attitudes, and economic and social status. It has a strong effect on reproductive behavior, fertility, infant and child mortality and morbidity, and attitudes and awareness related to family health, use of family planning, and sanitation (CSA, 2016). Education is one of the most important resource that enable women to provide appropriate care for their children, which is an important determinant of children's growth and development. The higher the level of a woman's education, the more awareness of how to utilize available resources for the improvement of their own nutritional status and that of their families (Engle et al., 1996).

Child's weight at birth

A child's weight/size at birth is an important indicator of the child's vulnerability to the risk of childhood illnesses and the child's chance of survival. According to the 2016 EDHS report, children whose birth weight is less than 2.5 kilograms, or children reported to be ”very small” or ”smaller than average” have a higher than average risk of early childhood death (CSA, 2016).

Gender and age of a child

Age and gender are important variables that are a primary basis for demographic classification in vital statistics censuses, and survey. They are also important variables for the study of mortality, nutrition status, fertility and marriage. For children, BMI is age and gender specific (Hammer et al., 1991; Pietrobelli et al., 1998), because BMI changes substantially as children age.

Mother's BMI

The Chronic Energy Deficiency (CED) or underweight is associated with impaired physical capacity, reduced economic productivity, increased mortality and poorer reproductive outcome. Some evidence in developing countries indicates that women with a body mass index below 18.5 show a progressive increase in mortality rate as well as increased risk of illness (Rotimi et al., 1999). Increased perinatal and neonatal mortality, a higher risk of low birth weight babies, stillbirths, and miscarriage are some of the consequence of malnutrition in women (Krasovec & Anderson, 1991).

Kulasekaran et al. (2012) showed a close correlation between mother's BMI with the incidence of anemia among the children. The women with Chronic Energy Deficit (underweight) had given birth to 77.6% of anemic children. Underweight among women contributed to more proportions of moderate and mild anemia among their children. Maternal low BMI has adverse effects contributing to poor fetal physical development, and low birth weight baby. The mortality rates are higher among low birth weight children in neonatal period and these children have more chance of developing non-communicable diseases such as type 2 diabetes and heart conditions in adulthood.

Place of residence

A study in the SNNPR of Ethiopia (Teller & Yimer, 2000) showed that compared to the urban women the rural women are more likely to suffer from Chronic Energy Deficiency. Several other local studies in Ethiopia also pointed out the higher rates of rural malnutrition (Taddese et al., 2017; Ferro-Luzzi et al., 1990). Children in rural areas are more likely to be stunted (46%) than those in urban areas (36%) (USAID, 2014).

Chapter 2

Preliminary data analysis

2.1 Introduction

For this study, the 2016 Ethiopian Demographic and Health Survey was used. The survey was carried out under the aegis of the Ministry of Health (MOH) and implemented by the Central Statistical Agency of Ethiopia (CSA). A total of 16,650 households, 5,232 in urban and 11,418 in rural areas were covered in the survey. The sample generated for women aged 15-49, 5,514 in urban and 11,149 in rural areas and 14,195 for men aged 15-59, with 4,472 in urban and 9,723 in rural areas. The EDHS have been conducted at five-year intervals since 2000. The primary objective of the 2016 EDHS was to provide up-to-date information for planning, policy formulation, monitoring, and evaluation of population and health programs in the country. For each EDHS, the key indicators were fertility, family planning behavior, child mortality, children's nutritional status, the utilization of maternal and child health services, knowledge of HIV/AIDS and other sexually transmitted infections (STIs). The sampling frame used for the 2016 EDHS is the Ethiopia Population and Housing Census (PHC), which was conducted in 2007 by the Ethiopia Central Statistical Agency. The 2016 EDHS sample was selected using a stratified, two-stage cluster design, and census enumeration area (EA) was the sampling units for the first stage. In the first stage, a total of 645 EA (202 in urban areas and 443 in rural areas) were selected with probability proportional to EA size (based on the 2007 PHC) and with independent selection in each sampling stratum. In the second stage of selection, a fixed number of 28 households per cluster were selected with an equal probability systematic selection from the newly created household listing (CSA, 2016).

2.2 Study variable

2.2.1 Response variable

The response variable is BMI of under five children in Ethiopia, which is a continuous variable. The response variable then categorized into two categories (normal/malnourished). Based on this binary outcome logistic regression for binary outcome was studied. Thereafter a four-category variable of nutrition status of under five children was also created, named as ”ordinal nutritional status”. Table 2.1 presents the under five children’s BMI cutoffs, recommended percentile range, and explicatory nutrition status with reference to the 2016 EDHS data. Based on recommendations by the World Health Organization Expert Committee on Physical Status based on children’s age and sex, children with BMI less than 5^{th} percentile are considered as underweight, children with BMI that falls between 5^{th} to 85^{th} percentile are considered as normal (healthy weight), children with BMI that falls between 85^{th} to 95^{th} percentile are considered as overweight and children with BMI that falls above the 95^{th} percentile are considered as obese.

Table 2.1: Summary of the response variable

Study result of BMI range	percentile range	nutrition status
< 12.5744	(5^{th}) percentile	underweight
$12.5744 - 17.0168$	(5^{th}) to (85^{th}) percentile	normal or healthy weight
$17.0168 - 18.4892$	(85^{th}) to (95^{th}) percentile	overweight
≥ 18.4892	(95^{th}) percentile	obese

2.2.2 Explanatory variables

The explanatory variables used in this study are the socio-economic, demographic and geographic factors. These are current age of a child, sex of a child, weight of a child at birth, mother’s age, mother’s BMI, educational attainment of mother, mother’s work status, religion, region, wealth index, place of residence (rural/urban) and marital status of mothers with reference to the 2016 Ethiopian DHS. Table 2.2 shows codes, labels and descriptions of the variables which are used in this study.

Table 2.3, Figure 2.1, and Figure 2.2 display socio-economic and demographic characteristics of the study variables. A nationally representative sample of under five children, for which the study observed their weight and height measures was studied. Table 2.3 show that 51.1% of the children were males and 48.9% of the children were females. From Table 2.3, it can be seen that 29.9%, 42.1% and 28% of weight

Table 2.2: Description of the study variables

Code	Label	Descriptions
B8	current age of child	children under five years
B4	sex of child	1=male, 2=female
M18	weight of child at birth	1=large, 2=average, 3=small
V012	mother's age	age of the mother during the survey (minimum=15 and maximum=49)
V439A	mother's BMI	body mass index of mother (mean=20.7225 and median=20.1252)
V149	educational attainment of mother	0=no education, 1=primary school, 2=secondary school and 3=higher
V714	mother work status	0= no, 1=yes
V130	religion	1=Orthodox, 2=Catholic, 3=Protestant, 4=Muslim and 5=other
V101	region	1=Tigray, 2=Afar, 3=Amhara, 4=Oromia, 5=Somali, 6=Benishangul, 7=SNNPR, 8=Gambela, 9=Harari, 10=Addis Abeba, 11=Dire Dawa
V190	wealth index	1=poor, 2=middle and 3=rich
V102	place of residence	1=urban, 2=rural
V501	marital status of mother	0=not married and 1=married
BMICHILD	under five children BMI	mean=15.3983, median=15.2554
Binary-nutritionstatus	Binary outcome	0=malnourished, 1=normal
Ordinal-nutritionstatus	Ordinal outcome	1=underweight, 2=normal(healthy weight), 3=overweight, 4=obese

of children at birth were large, average and small respectively. With regard to the educational attainment of mothers, 64% of mothers have no education (Figure 2.2). Furthermore, the majority of the mothers had no job (72.4%). Table 2.3 also show that the majority of the children were from Oromia (15.45%), followed by Somali (12.51%), SNNP (14.2%), Tigray (10.48%) and Amhara (9.91%) regions (Figure 2.1). Table 2.3 show that half of the mothers/households were either in poor economic category (53.8%) or in middle economic category (14.4%). An overwhelming majority of children's mothers were residing in rural areas (81.9%). Moreover, the results also show that the majority of the mothers were married (88.6%).

Table 2.3: Summary measures for the selected socio-economic and demographic characteristic of children

Characteristics	Frequency	Percentage(%)
Current age of child		
0	1948	21.71
1	1798	20.05
2	1738	19.37
3	1695	18.89
4	1792	19.98
Sex of a child		
male	4586	51.1
female	4385	48.9
Weight of child at birth		
large	2678	29.9
average	3781	42.1
small	2512	28
Religion		
Orthodox	2702	30.12
Catholic	55	0.61
Protestant	1610	17.95
Muslim	4447	49.57
Other	157	1.75
Mother work status		
No	6495	72.4
Yes	2476	27.6
Wealth index		
poor	4826	53.8
middle	1289	14.4
Rich	2856	31.8
Place of residence		
Urban	1623	18.1
Rural	7348	81.9
Current marital status		
not married	572	6.4
married	8399	93.6
Total	8971	100

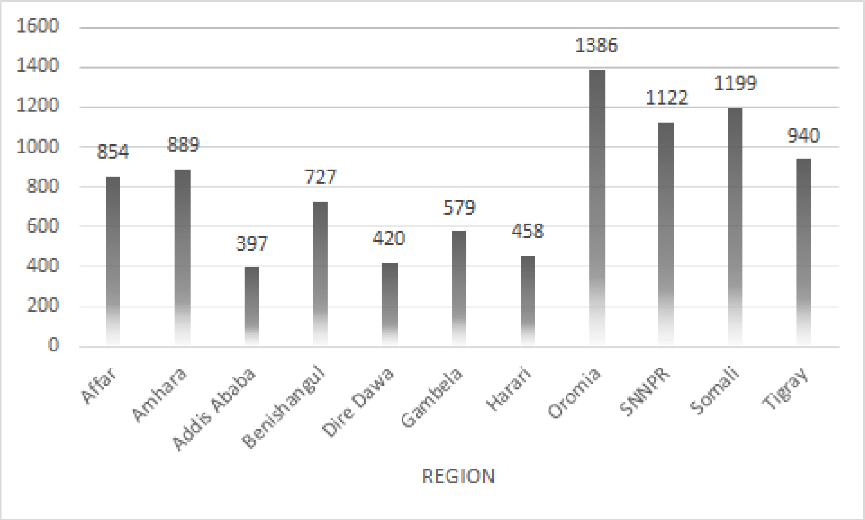

Figure 2.1 – Sample distribution of children according to region

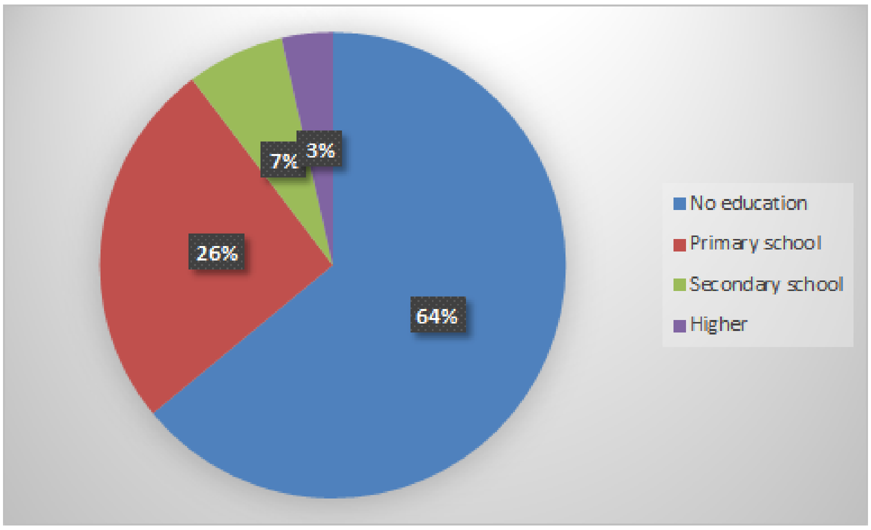

Figure 2.2 – Educational attainment of mother in Ethiopia

2.3 Summary

Descriptive statistics were used to describe the data at hand. The assessment of variables by measures of central tendency and measures of variability help us to understand different properties of the data being analyzed. The technique allows for the analysis of the relationships between the socio-economic and demographic factors and the response variable. The results of the analysis can be seen analytically and visually. Moreover, the results will be easier to interpret.

Overall, it has been observed that the number of under five children in different age and the number of male and female children were almost equal (Table 2.3). Majority of the respondents (84.2%) were residing in rural areas. About 93.6% of the respondents are married. More than half of the number of children (53.8%) were from poor economic class families. Most of the respondents were from Oromia followed by Somali, SNNP, Tigray and Amhara regions. Many of the respondents were Orthodox Christian and Muslim.

The continuous response variable (under five children's BMI) was recoded into a categorical variable that has two levels (Binary outcome) and thereafter four levels (Ordinal outcome). The recoding of the response variable into different levels directs us to use different statistical models to study the relationship between the response and a set of independent variables. The possible statistical methods to study for continuous and categorical data are discussed in the next chapters.

Chapter 3

Quantile regression

3.1 Introduction

The purpose of regression analysis is to analyze relationships between a response variable and predictor variables. It answers the question of how much the response variable changes with changes in the predictor variables and predicts the values of a response variable based on the values of the predictor variables. Since the response variable cannot be predicted exactly from the predictor variables in real applications, we use measure of central tendency, typically the mean, median and mode, to summarize the behaviour of the response for fixed values of the predictors. Regression analysis estimates the conditional mean of the response variable given the predictor variables. The idea of modelling and fitting the conditional mean function is the essence part of a broad family of regression modelling approaches, including the familiar simple linear regression model and multiple linear regression model.

The analysis of the conditional mean linear regression focuses on the mean response. The relationship between the response Y and k predictors $x_1, x_2, ..., x_k$ is described by the conditional mean of the response for given values of predictors. That is,

$$Y|x_1, x_2, ..., x_k = \boldsymbol{X\beta} + \epsilon \tag{3.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)^T$ is the unknown parameter vector.

The above equation specifies the change in the conditional mean of the response variable associated with a change in the explanatory variables.

The popular assumptions of conditional mean models are: they can provide a complete connection between the explanatory variables and the response distribution. Also, they lead to estimators like least-squares and maximum likelihood that are easy to compute and interpret. However, the conditional mean regression has inherent limitations. First, when summarizing the behavior of the response for given values of predictor variables, the conditional mean model cannot be readily extended to non-central locations. Second, when dealing with heavy-tailed distributions, the measure of central location under conditional mean models can be significantly affected by outliers and become inappropriate and misleading. Third, it is not readily direct for us to understand how the shape of underlying response distribution is affected by the changes in predictor variables (Huang et al., 2015).

The conditional median modelling or simply median regression is an alternative to conditional mean modelling. It addresses the issues of regression analysis regarding the choice of a measure of central tendency. Median regression modelling has the potential to be more useful when the distribution is highly skewed, where the mean can be challenging to interpret while the median remains highly informative (Hao & Naiman, 2007). Median regression is a special case of quantile regression in which the conditional 0.5^{th} quantile is modelled as a function of covariates. Note that median regression will be the same as mean regression if the response variable is from a symmetric distribution such as the normal distribution. More generally, other quantiles can be used to describe non-central positions of a distribution. Quantile regression provides a complete picture of the covariate effect when a set of percentiles is modelled. Quantile regression models can be easily fit by minimizing a generalized measure of distance using algorithms based on linear programming. Therefore, we are interested in estimating quantiles of the response distribution as a function of potential predictor variables. When the conditional densities of the response are heterogeneous, it is natural to consider whether weighted quantile regression might lead to efficiency improvements (Koenker, 2005). Weighted quantile regression improves the efficiency, if the design of the survey is taken into account.

Ordinary least squares regression models the relationship between one or more covariates X and the conditional mean of the response variable given $X = x$ or $E(Y|X = x)$. Quantile regression, which was first introduced by Koenker & Bassett Jr (1978), extends the regression model to conditional quantile of the response variable, such as 0.25 quantile or 25^{th} percentile and so on. Specifically the 0.25 quantile estimates the parameters that describe the 25^{th} percentile (first quantile) of the conditional distribution. Quantile regression is desired if conditional quantiles are of interest. It is a statistical technique intended to model and estimate conditional quantile functions. It is also particularly useful when the rate of change in the conditional quantile, expressed by the regression coefficients, depends on the quantile. Quantile regression gives complete information about the covariate effect if we model a set of percentiles. Median regression is a special case of quantile regression. This method helps to detect significant structures of the data that might be missed by models that average over the conditional distribution. Quantile regression allows for effects of the independent variables to differ over the quantiles. For example the effect of a covariate might be different at the tail of the distribution compared to the median effects and hence the interpretations are different. That is the effects of the independent variables may vary over quantiles of the conditional distribution. This is also an important advantage of quantile regression over mean regression (Chamberlain, 1994). One of the most important virtues of quantile regression is that it allows us to make inference on the entire conditional distributions of the response by estimating a number of different quantiles (Huang et al., 2015).

For a random variable Y with probability distribution function; $F(Y) = Pr(Y \leq y)$ the τ^{th} quantile of Y is defined as the inverse function:

$$Q_y(\tau) = inf\{Y : F(Y) \geq \tau\} = F^{-1}(\tau), \quad 0 < \tau < 1$$

The quantile regression model is described by the conditional τ^{th} quantiles of the response Y for given values of predictors $x_1, x_2, ..., x_k$. It is a natural extension of the traditional mean model in Equation (3.1):

$$Q_y(\tau|x_1, x_2, ..., x_k) = \beta_0^\tau + \beta_1^\tau x_1 + ... + \beta_k^\tau x_k, \quad 0 < \tau < 1 \tag{3.2}$$

where $\beta^\tau = (\beta_0^\tau, \beta_1^\tau, ..., \beta_k^\tau)$ is the unknown parameter vector.

Equation (3.2) gives the changes in the conditional quantiles. Because any τ^{th} quantile can be used, any predetermined situation of the distribution can be modelled. This is useful to obtain a more complete understanding about how the outcome distribution can be affected by the predictors. Hence, the method lets us select situations on the outcome distribution for their detailed inquiries.

For a random sample $\{y_1, ..., y_n\}$ of Y, it is well known that the sample median minimizes the sum of absolute deviations:

$$Median = argmin_{\xi \in R} \sum_{i=1}^{n} \rho_\tau |y_i - \xi|$$

where ρ_τ and ξ are explained below for any given quantile.

Likewise, the general τ^{th} sample quantile $\xi(\tau)$, which is the analogue of $Q(\tau)$, is formulated as the minimizer:

$$\xi(\tau) = argmin_{\xi \in R} \sum_{i=1}^{n} \rho_\tau(y_i - \xi)$$

where $\rho_\tau(Z) = Z(\tau - I(Z < 0))$, $0 < \tau < 1$, and where $I(\bullet)$ denotes the indicator function. The loss function ρ_τ assigns a weight of τ to positive residuals $y_i - \xi$ and a weight of $1 - \tau$ to negative residuals. Using this loss function, the linear conditional quantile function extends the τ^{th} sample quantile $\xi(\tau)$ to the regression setting in the same way that the linear conditional mean function extends the sample mean.

OLS regression estimates the linear conditional mean function $E(Y|X = x) = x'\beta$ by solving:

$$\hat{\beta} = argmin_{\beta \in R^P} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \tag{3.3}$$

The estimated parameter $\hat{\beta}$ minimizes the sum of squared residuals in the same way that the sample mean $\hat{\mu}$ minimizes the sum of squares:

$$\hat{\mu} = argmin_{\beta \in R} \sum_{i=1}^{n} (y_i - \mu)^2$$

Quantile regression also estimates the linear conditional quantile function, $Q(\tau|X = x) = x'\beta(\tau)$, by solving:

$$\hat{\beta}(\tau) = argmin_{\beta \in R^P} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta) \tag{3.4}$$

for any quantile $\tau \in (0, 1)$. The quantity $\hat{\beta}(\tau)$ is called the τ^{th} regression quantile. The case $\tau = 0.5$, which minimizes the sum of absolute residuals, corresponds to median regression, which is also known as L_1 regression. The set of regression quantiles $\{\beta(\tau) : \tau \in (0, 1)\}$ is referred to as the quantile process.

Quantile regression minimizes:

$$\sum_i \tau|\epsilon_i| + \sum_i (1 - \tau)|\epsilon_i|,$$

where $\sum_i \tau|\epsilon_i|$ is a sum that gives the asymmetric penalties $\tau|\epsilon_i|$ for under prediction and $(1-\tau)|\epsilon_i|$ for over prediction.

The SAS QUANTREG procedure computes the quantile function $Q(\tau|X = x)$ which the analyst or modeler can use to conduct statistical inference on the estimated parameters $\hat{\beta}(\tau)$.

The τ^{th} quantile regression estimator $\hat{\beta}(\tau)$ minimizes the objective function given by

$$Q(\beta_\tau) = \sum_{i:y_i \geq x_i'\beta}^{n} \tau|y_i - x_i'\beta_\tau| + \sum_{i:y_i < x_i'\beta}^{n} (1-\tau)|y_i - x_i'\beta_\tau| \tag{3.5}$$

where $0 < \tau < 1$, $i : y_i \geq x_i'\beta$ for under prediction, $i : y_i < x_i'\beta$ for over prediction. We have β_τ instead of β, because different choices of τ estimates different values of β.

3.1.1 Interpretation of quantile regression estimates

Since the τ^{th} conditional quantile of Y given x is given by $Q_\tau(y_i|x_i) = x_i'\beta_\tau$, its estimate is given by $\hat{Q}_\tau(y_i|x_i) = x_i'\hat{\beta}_\tau$. As one increases τ continuously from 0 to 1, one traces the entire conditional distribution of Y, conditional on x. As a note, various quantile regression estimates are correlated. The parameter estimates in quantile regression models have the same interpretation as those of any other linear model, as rates of changes. Therefore, in a similar way to the OLS model, the $\beta_{i(\tau)}$ coefficient of the quantile regression model can be interpreted as the rate of change of the τ^{th} quantile of the dependent variable distribution per unit change in the value of the i^{th} regressor; consider the partial derivative of the conditional quantile of y with respect to one of the regressors, say i, namely, $\partial Q_\tau(Y|X)/\partial X_i$. This derivative is to be interpreted as the marginal change in the τ^{th} conditional quantile due to marginal change in the i^{th} element of x. If x contains k distinct variables, then this derivative is given simply by $\beta_{i(\tau)}$, the coefficient on the i^{th} variable (Buchinsky, 1998).

Equivariance properties of quantile regression

One of the best advantages of quantile regression estimators is their behavior with respect to monotone transformations of the response variable. This behavior, named equivariance, refers to the ability to use the same interpretation rules when the data or the model is subjected to a transformation. According to Buchinsky (1998), some authors have proposed that the equivariance property can be exploited to speed up the estimation process by reducing the number of simplex iterations.

The quantile regression estimator has several important equivariance properties which help facilitate the computation procedure. According to the chosen transformation, the equivariance property can be distinguished into: scale equivariance, shift or regression equivariance, equivariance to reparametrization of design and equivariance to monotone transformations (Davino et al., 2013).

Practical implications of the equivariance properties of quantile regression

Let us consider the simplest quantile regression model with one explanatory variable and for a given quantile τ:

$$Q_\tau(Y|x) = \beta_{0(\tau)} + \beta_{1(\tau)}x$$

The scale equivariance property implies that, if the dependent variable is multiplied by a positive constant c, the coefficients of the new model can be easily obtained multiplying by c the coefficients in equation (above):

$$Q_\tau(cY|x) = c\beta_{0(\tau)} + c\beta_{1(\tau)}x$$

The shift equivariance property is also referred to as the regression equivariance because it denotes the effect of the dependent variable obtained as a linear combination, through the γ coefficients, of the explanatory variable. Such an effect holds when Y is subjected to a location shift:

$$Y^* = Y + \gamma Y$$

The QR estimator of Y^* on x results in: $Q_\tau(Y^*|x) = \beta_0 + [\beta_{1(\tau)} + \gamma]x$

The equivariance to reparametrization of design is derived from the effect of a nonsingular matrix A $(p \times p)$ introduced in the model:

$$Q_\tau(Y|X, A) = A^{-1}X\beta_\tau$$

where X is the matrix of p explanatory variables.

Finally, the equivariance to monotone transformations implies that if a nondecreasing function on R, $h(\bullet)$ is applied to the dependent variable, the quantiles of the transformed Y variable are the transformed quantiles of the original ones:

$$Q_\tau[h(y)|x] = h[\beta_{0(\tau)}] + h[\beta_{1(\tau)}]x$$

According to Davino et al. (2013), appropriate selection of the $h(\gamma)$ monotone function is very important in real data applications because it is necessary to manage and correct different kinds of skewness.

Details about quantile regression and its equivariance property can be found in different literatures (Davino et al., 2013; Koenker, 2005; Neter et al., 1996; Parzen et al., 1994; Sen & Srivastava, 2012; Weisberg, 2005).

3.2 Weighted quantile regression

The model for general linear regression is:

$$Y = A'\beta + \epsilon$$

where $Y = (y_1, ..., y_n)'$ is the $(n \times 1)$ vector of response, $A' = (x_1, ..., x_n)'$ is the $(n \times p)$ regressor matrix, $\beta = (\beta_1, ..., \beta_p)'$ is the $(p \times 1)$ vector of unknown parameters, and $\epsilon = (\epsilon_1, ..., \epsilon_n)'$ is the $(n \times 1)$ vector of unknown errors.

L_1 regression, also known as median regression, is a natural extension of the sample median when the response is conditioned on the covariates. In L_1 regression, the least absolute residuals estimate $\hat{\beta}_{LAR}$, referred to as the L_1-norm estimate, is obtained as the solution of the minimization problem: $\min_{\beta \in R^P} \sum_{i=1}^{n} |y_i - x_i'\beta|$. More generally, for quantile regression Koenker & Bassett Jr (1978) defined the τ^{th} regression quantile, $0 < \tau < 1$, as any solution to the minimization problem: Equation (3.5). The solution is denoted as β_τ, and the L_1-norm estimate corresponds to $\beta_{(1/2)}$. The τ^{th} regression quantile is an extension of the τ^{th} sample quantile $\xi_{(\tau)}$, which can be formulated as the solution of:

$$\min_{\xi \in R} \left[\sum_{i \in \{i : y_i \geq x_i'\xi\}} \tau |y_i - x_i'\xi| + \sum_{i \in \{i : y_i < x_i'\xi\}} (1 - \tau)|y_i - x_i'\xi| \right] \tag{3.6}$$

If we specify weights $w_i, i = 1, ..., n$, weighted quantile regression is carried out by solving:

$$\min_{\beta_w \in R^P} \left[\sum_{i \in \{i: y_i \geq x_i' \beta_w\}} w_i \tau |y_i - x_i' \beta_w| + \sum_{i \in \{i: y_i < x_i' \beta_w\}} w_i (1 - \tau) |y_i - x_i' \beta_w| \right] \tag{3.7}$$

Weighted regression quantiles β_w can be used for L-estimator (Koenker & Zhao, 1996).

Weighted quantile regression as an optimization problem

The traditional mean regression for conditional mean $\mu_{Y|X} = E[Y|X]$ is a solution of $\min_{\mu \in R} \sum_{i=1}^n (y_i - \mu_{Y|X})^2$. Assuming $\mu_{Y|X} = X^T \beta$, the least-squares estimator $\hat{\beta}$ is obtained from Equation (3.3). The classical quantile regression for the conditional quantile $\tau_{Y|X} = Q_y(\tau|x)$ is a solution of $\min_{q \in R} \sum_{i=1}^n \rho_\tau (y_i - q_{Y|X})$. Assuming $q_{Y|X} = X^T \beta_\tau$, the quantile regression estimator $\hat{\beta}_\tau$ is obtained by Equation (3.4). Therefore, the weighted quantile regression problem can be formulated as a linear programing problem:

$$\min_{(\beta, \boldsymbol{u}, \boldsymbol{v}) \in R^P \times R^{2n}} \{ \tau \boldsymbol{w}^T \boldsymbol{u} + (1 - \tau) \boldsymbol{w}^T \boldsymbol{v} | X \beta_\tau + \boldsymbol{u} - \boldsymbol{v} = \boldsymbol{y} \}, \tag{3.8}$$

where X denotes the $n \times p$ design matrix and $\boldsymbol{w}, \boldsymbol{u}, \boldsymbol{v}$ are $n \times 1$ three vectors with elements of w_i, u_i, v_i, respectively. w_i represent the weight where as u_i, v_i are $2n$ *'slack'* variables $\{u_i, v_i : i = 1, 2, ..., n\}$ to represent the positive and negative parts of the residual vector (see Huang et al., 2015).

The proposed weighted quantile regression for the conditional quantile $q_{Y|X} = Q_y(\tau|x)$ is a solution of $\min_{q \in R} \sum_{i=1}^n w_i \rho_\tau (y_i - q_{Y|X})$. Assuming $q_{y|x} = x^T \beta_\tau$, the weighted quantile regression estimator $\hat{\beta}_w(\tau)$ is obtained by:

$$\hat{\beta}_w(\tau) = argmin_{\beta \in R^P} \sum_{i=1}^n w_i(x_i, \tau) \rho_\tau (y_i - x_i^T \beta), \tag{3.9}$$

where $\hat{\beta}_w(\tau)$ is an extension of the loss function introduced by Koenker (2005) for the classical quantile regression; and $w_i(x_i, \tau)$ is defined as any uniformly bounded positive weight function independent of $y_i, i = 1, ..., n$ and

$$\rho_\tau(u) = u(\tau - I(u < 0)) = \begin{cases} u(\tau - 1), u < 0 \\ u\tau, u \geq 0. \end{cases}$$

Weighted quantile regression might lead to more efficient improvements than the conventional quantile regression (Koenker & Bassett Jr, 1978).

3.3 Analysis of data using weighted quantile regression

The data employed in this study is the 2016 Ethiopian Demographic and Health Survey data. The socio-economic, demographic and geographic factors presented in section 2.2.2 are considered in the modeling process. The weighted quantile regression model was applied in modeling the body mass index of under five children.

The parameter estimates at different quantile levels (Table 3.1) show how quantile regression allows us to study the impact of predictors on different quantiles of the response variable, and thus provides a complete picture of the relationship between the dependent and explanatory variables by taking the weight into account. Quantile regression analysis (Table 3.1) identified the significant predictor variables at different quantile levels. At 0.05 quantile: mother's BMI, sex of child, weight of a child at birth and region (Addis Abeba, Affar, SNNPR and Somali) were found to have significant effect on BMI of under five children. At 0.5 quantile: current age of child, mother's age, mother's BMI, sex of child, weight of child at birth and region (Affar, Gambela, SNNPR and Somali) were found to have significant effect on BMI of under five children. Similarly, at 0.85 quantile: current age of child, mother's current age, mother's BMI, sex of child, weight of child at birth and region (Addis Abeba and Somali) were found to significantly affect under five children's BMI. The findings using quantile regression at different quantile levels (0.25 quantile, 0.75 quantile and 0.95 quantile) were also presented (Table 3.1).

At 0.25 quantile, intercept = 13.7005, which is the predicted value of the 0.25 quantile under five children BMI when all the explanatory variables are zero. $\hat{\beta}_{1(0.25quantile)} = -0.1793$ indicates the rate of change of the 0.25 quantile (Q_1) of the explanatory variable per unit change of current age of child keeping all the other explanatory variables constant. In other words, the Q_1 regression coefficient indicates the 25^{th} percentile of the under five children BMI will decrease by 0.1793 for every one-unit change in current age of a child, setting all the other explanatory variables constant.

At 0.5 quantile, intercept = 14.9289, which is the predicted value of the 0.5 quantile under five children's BMI when all the explanatory variables are zero. $\hat{\beta}_{3(0.5quantile)} = 0.0730$ indicates the rate of change of the 0.5 quantile (Q_2) of the explanatory variable per unit change of mother's BMI keeping all the other explanatory variables constant. In other words, the Q_2 regression coefficient indicates the 50^{th} percentile of the under five children's BMI will increase by 0.073 for every one-unit increase in mother's BMI, setting all the other explanatory variables constant.

At 0.75 quantile, intercept = 16.2903, which is the predicted value of the 0.75 quantile under five children's BMI when all the explanatory variables are zero. $\hat{\beta}_{2(0.75quantile)} = -0.0181$ indicates the rate of change of the 0.75 quantile (Q_3) of the explanatory variable per unit change of mother's age keeping all the other explanatory variables constant. In other words, the Q_3 regression coefficient indicates the 75^{th} percentile of the under five children's BMI will decrease by 0.0181 for every one-unit increase in mother's age, setting all the other explanatory variables constant.

Table 3.1: Parameter estimates at different quantile levels

Parameter	**Estimate**	**P-value**	**Estimate**	**P-value**	**Estimate**	**P-value**	**Estimate**	**P-value**	**Estimate**	**P-value**	**Estimate**	**P-value**
Quantile level	0.05		0.25		0.5		0.75		0.85		0.95	
Intercept	11.7144	<0.0001	13.7005	<0.0001	14.9289	<0.0001	16.2903	<0.0001	16.5627	<0.0001	16.9252	<0.0001
Current age of child	0.0541	0.0935	-0.1793	<0.0001	-0.2908	<0.0001	-0.4013	<0.0001	-0.4648	<0.0001	-0.6750	<0.0001
Mother's age	-0.0102	0.2604	-0.0104	0.0104	-0.0149	0.0005	-0.0181	<0.0001	-0.0188	0.0018	-0.0112	0.4144
Mother's BMI	0.0604	<0.0001	0.0670	<0.0001	0.0730	<0.0001	0.0722	<0.0001	0.0875	<0.0001	0.1133	0.0011
Sex of child (ref. = male)												
Female	-0.3657	0.0006	-0.2722	<0.0001	-0.1699	0.0007	-0.2169	0.0004	-0.1899	0.0117	-0.1133	0.4826
Weight of child at birth (ref. = Small)												
Large	0.6257	<0.0001	0.4220	<0.0001	0.5169	<0.0001	0.4659	<0.0001	0.5101	<0.0001	0.4941	0.0233
Average	0.3596	0.0110	0.2413	0.0014	0.2263	0.0008	0.2648	0.0002	0.3791	<0.0001	0.4561	0.0541
Work status (ref. = Yes)												
No	0.1636	0.1400	0.0416	0.4779	0.0589	0.2715	0.0363	0.6589	0.0361	0.6956	0.0501	0.8058
Educational level (ref. = Sec. school)												
No education	-0.2849	0.3299	-0.2222	0.1429	-0.0252	0.8593	0.1337	0.3931	0.1749	0.3634	0.0809	0.8676
Primary school	-0.1431	0.6222	-0.1572	0.2843	-0.0361	0.7913	0.0491	0.7601	0.0196	0.9157	-0.0104	0.9837
Higher	-0.4693	0.3358	-0.1159	0.6162	0.1799	0.5129	-0.0570	0.8193	-0.0126	0.9583	-0.5030	0.5548
Marital status (ref. = Not married)												
Married	-0.0015	0.9946	0.0093	0.9403	0.0416	0.7664	0.2700	0.0291	0.2618	0.0662	0.3608	0.1500
Religion (ref. = Protestant)												
Orthodox	0.0542	0.7819	0.0464	0.6401	-0.0413	0.6707	-0.1707	0.1610	-0.2609	0.0607	0.0950	0.7899
Catholic	0.1356	0.8622	0.2755	<0.0001	-0.3314	0.3023	-0.5027	0.2939	-0.4242	0.5724	-0.3066	0.9195
Muslim	-0.1391	0.4574	0.0566	0.5438	-0.1484	0.1249	-0.1297	0.2913	-0.0986	0.5005	0.1518	0.7017
Other	-1.4190	0.1099	0.2181	0.4835	0.5450	0.0695	0.6020	0.1996	1.6959	0.0191	3.0352	0.1634
Region (ref. = Tigray)												
Addis Abeba	0.5076	0.0352	0.2490	0.0996	0.0533	0.7093	1.1381	0.4611	0.3122	0.1569	0.9173	0.0390
Affar	0.1048	<0.0001	-0.2568	0.0136	-0.2730	0.0034	-0.4732	0.0003	-0.4656	0.0034	-0.2932	0.3436
Amhara	0.2811	0.0854	0.0184	0.8477	-0.0096	0.9097	-0.1553	0.1071	-0.0749	0.4689	-0.3352	0.1592
Benishangul	0.2520	0.2309	0.0268	0.7772	-0.0851	0.3693	-0.2442	0.0390	-0.2967	0.0167	-0.0689	0.8186
Dire Dawa	-0.3238	0.3697	-0.2679	0.0693	-0.1933	0.1264	-0.5401	<0.0001	-0.5773	0.0005	-0.2942	0.4650
Gambela	-0.2240	0.3994	-0.2627	0.0367	-0.4479	0.0007	-0.5770	0.0003	-0.5453	0.0022	-0.2651	0.4330
Harari	-0.1805	0.5032	-0.0795	0.6127	0.0701	0.6096	-0.1211	0.3684	-0.1638	0.3471	0.0489	0.8822
Oromia	0.0365	0.8421	0.0396	0.6879	0.0267	0.7628	0.0672	0.5324	0.0733	0.5533	0.3069	0.2651
SNNPR	0.5568	0.0084	0.4065	0.0002	0.2809	0.0063	0.2034	0.1357	0.2229	0.1212	0.6302	0.0804
Somali	-0.4034	0.0323	-0.8043	<0.0001	-0.8032	<0.0001	-0.9451	<0.0001	-1.0115	<0.0001	-0.8831	0.0122
Place of residence (ref. = Urban)												
Rural	-0.0879	0.6539	0.0347	0.7682	-0.1274	0.2803	-0.2769	0.0366	-0.1580	0.2208	0.3011	0.3618
Wealth index (ref. = Rich)												
Middle	0.0734	0.7108	-0.0973	0.1930	-0.0936	0.2264	-0.1824	0.0713	-0.1926	0.1221	-0.7798	0.0036
Poor	-0.0415	0.7687	-0.1226	0.1075	-0.1063	0.1083	-0.1665	0.0396	-0.2529	0.0134	-0.6926	0.0060
Actual value of under five BMI	12.5744		14.1934		15.2554		16.3501		17.0168		18.4892	
Predicted value at mean	12.7647		14.4287		15.4281		16.4646		16.9088		17.6191	

Standard errors (SEs) for QR (Appendix A) are calculated in various ways. SAS PROC QUANTREG procedure provides three methods to compute SE and confidence intervals (CIs) for the QR parameter: sparsity, rank, and resampling. The resampling method, which uses the bootstrap technique, is more advantageous. Koenker (1994) considered a more interesting resampling mechanism, resampling directly from the full regression quantile process, which he called the Heqf bootstrap. In contrast with these bootstrap methods, Parzen et al. (1994) observed the τ^{th} regression quantile is a pivotal quantity for the τ^{th} QR parameter. The bootstrap method by Parzen et al. (1994) is much simpler but time consuming for larger data set (relatively $n > 5000$) and for high-dimensional data sets. The QUANTREG procedure implements a new, general resampling method developed by He & Hu (2002), which they referred to as, the Markov chain marginal bootstrap (MCMB). For QR, the MCMB method has the advantage that it solves p one-dimensional equations instead of p-dimensional equations. This improves the feasibility of the resampling method in computing SEs and CIs for regression quantiles.

3.4 Summary

Relative to the OLS regression, QR estimates are more robust against outliers in the response measurements by estimating various quantile functions at different parts of the distribution of the response variable (Koenker, 2005). Because it does not assume a particular distribution for the response, nor does it assume a constant variance for the response, unlike OLS regression, QR offers considerable model robustness. It also permits us to investigate the effect of explanatory variables on different quantiles of the outcome distribution. QR is very flexible for the reason that the model does not use a link function and distributional assumption that relates the variance and the mean of the outcome variable (SAS, 2014). Despite its bright future and usefulness in many important application areas, such as medicine and survival analysis, financial and economic statistics and environmental modelling, QR has some limitations. This is mainly because of the fact that the full implications of the estimation procedure are not always realized (McMillen, 2012). If we are interested in modelling the mean, the estimates of OLS should be preferred over the estimates of QR when estimates of OLS is more efficient than estimates of QR, such as when the error distribution is normal. QR methods are only applied to continuous-response data and we can possibly utilize them in the context of count data. It is also needs sufficient data; when n is small the usual linear regression is preferred. Unlike with some other SAS PROCs procedure, QUANTREG procedure does not have an option to change the reference level in the class statement yet (SAS, 2014).

Chapter 4

Logistic regression

4.1 Introduction

The general linear model requires that the response variable follows the normal distribution whilst the generalized linear model is an extension of the general linear model that allows the specification of models whose response variable follows the exponential family of distribution. The generalized linear model includes: logistic regression for binary response variable, multiple regression for continuous responses, poisson regression or negative binomial regression for count data, log-linear models for categorical data analysis, gamma regression for variance models, and exponential and gamma models for survival time models (Ayele et al., 2013). Among these models, logistic regression is the most popular modeling procedure used to analyze epidemiologic data when the outcome is dichotomous (Kleinbaum et al., 2002). The methods employed in an analysis using logistic regression follow the same general principles used in linear regression (Hosmer et al., 2000). However, it is not appropriate to use linear regression for binary data or to model probabilities. This is mainly because the response variables are not measured on the ratio scale and the errors terms are not normally distributed. In addition, the linear regression can generate predicted values that are any real values between negative infinity and positive infinity. On the other hand, probabilities have limited values between zero and one.

4.2 Generalized linear model (GLM)

The term general linear model usually refers to conventional linear regression model for a continuous response variable given continuous and/or categorical predictors. The form is $y_i \sim N(X_i^T\boldsymbol{\beta}, \sigma^2)$, where X_i contains known covariates and $\boldsymbol{\beta}$ contains the coefficients to be estimated. The parameters of these models are estimated by least squares and weighted least squares using statistical software.

The generalized linear model (GLM) is defined in terms of a set of independent random variables $Y_1, ..., Y_N$ each with a distribution from the exponential family of distributions (Dobson & Barnett, 2008). It is a generalization of linear regression that allows for response to have distribution other than normal distribution. The unknown parameters of GLM are estimated by using maximum likelihood estimation method.

4.2.1 Components of GLMs

All generalized linear models have three components. These are the **random component** which identifies the response variable Y and assumes a probability distribution for it. The **systematic component** specifies the explanatory variables $(x_1, x_2, ..., x_k)$ used as predictors in the model and the linear combination of the explanatory variables is called linear predictor. The linear predictor is given by

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = \textstyle\sum_{j=0}^{k} \beta_\mathrm{k}\mathrm{x}_\mathrm{k}.$$

The **link function** describes the functional relationship between the systematic component and the expected value of the random component. It reflects how the expected value of the response relates to the linear predictor of explanatory variables; $\theta = g(E(Y_i)) = E(Y_i)$ for linear regression or $\theta = logit(\pi)$ for logistic regression. For a general link function $g(\bullet)$, we have $g(E(Y_i)) = g(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$.

GLMs have many advantages over OLS regression. For GLMs we do not need to transform the response Y to approximate a normal distribution. The choice of the link function is separate from the choice of random component thus we have more flexibility in modeling; if the link produces additive effects, then we do not need constant variance, and inference tools and model checking like Wald and likelihood ratio tests, deviance, residuals, confidence intervals, and overdispersion can apply for GLMs. However, the GLMs have some limitations. The limitation includes the linearity assumption i.e., it can have only a linear predictor in the systematic component and responses must be independent (Science, 2017).

4.3 Odds and odds ratio

Probability is defined as the chances that an event will occur. In terms of numerical expression ranging from 0 to 1, with 1 meaning that the event will surely occur, and 0 meaning that the event will never occur hence referred to as the sure and null events respectively.

The odds can be defined in two ways. Odds in favor of an event or odds against an event. Odds in favor of an event is defined as the ratio of the probability of the occurrence of an event to the probability of non-occurrence the event. In other words, the odds in favor of an event is the ratio of the expected number of times that an event will occur to the expected number of times it will not occur. Odds against an event can be defined as the number of unfavorable outcomes divided by the number of favorable outcomes. In our case, we concentrate on the odds in favor of an event.

There is a simple relationship between probabilities and odds. Let P be the probability of an event and O be the odds of the event, then

$$O = \frac{P}{1-P} = \frac{probablity \quad of \quad event}{probablity \quad of \quad no \quad event}$$

or equivalently

$$P = \frac{O}{1+O}$$

The odds ratio (OR) is a ratio of two odds. It is widely used as a measure of the relationship between the response and the predictor variables. The odds ratio for a continuous variable in logistic regression represents how the odds change with a 1 unit increase in that continuous variable holding all other variables constant. For categorical predictors with two or more categories, the odds ratio is interpreted as the change in the odds of an event for each category compared to the odds of an event for the reference category. The odds ratio interpretation is always with reference to this category. Odds ratios (and various functions of them) are less sensitive to changes in the marginal frequencies than other measures of association. In this sense, they are frequently regarded as fundamental descriptions of the relationship between the variables of interest.

$$OR = \frac{Odds_1}{Odds_2} = \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}}$$

Like probabilities, odds ratios are bounded below by 0 but unlike probabilities, there is no upper bound on the odds ratios. That is, they can range from 0 to infinity. An OR of 1 indicates that the predictor variable has no effect on the odds of event. That is, the odds of an event is the same for the two categories of the predictor variable being compared. An OR less than 1 indicates that the odds of an event for the response variable with the higher value of x used in the numerator are less than the odds of an event for the response variable with the value of x used in the denominator. Thus, odds ratios are a more sensible scale for multiplicative comparisons.

4.4 Logistic regression model

The logistic function was discovered by Pearl and Reed in a study of the population growth in USA (1920) and developed by David Cox in 1958. The logistic regression model is designed to describe the probability of an event, which is always some number between zero and one (Melesse et al., 2016). Logistic regression is used to model the probability of an event of interest given the values of the predictor variables. In the logit model, the log odds of the outcome is modeled as a linear combination of the predictor variables. A major problem with the linear probability model is that probabilities are bounded by 0 and 1, but linear functions are inherently unbounded. The solution is to transform the probability so that it is no longer bounded between 0 and 1. Transforming the probability to the odds scale removes the upper bound. If we then take the logarithm of the odds, we also remove the lower bound, setting the $logit(\pi)$ to vary between minus infinity to plus infinity (Allison, 2012).

For k explanatory variables and i=1,...,n individuals, the logit model is

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \tag{4.1}$$

where π_i is the probability that $y_i = 1$, β_0 is the intercept parameter, $\beta_i(i = 1, 2, ..., k)$ are the slope parameters, and x_i stand for explanatory variables. The expression on the left-hand side is the logit or log-odds.

The equation for π_i is

$$\pi_i = P(Y_i = 1|X_i = x_i) = \frac{exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})} \tag{4.2}$$

Logistic regression is the appropriate regression model that can be used for binary response variable. Like all regression analysis, the logistic regression is a predictive model. It is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables.

4.4.1 Logistic regression model estimation

The goal of logistic regression is to estimate the probability of an event given the values of the predictor variables. To tie together the linear combination of variables and in essence the Bernoulli distribution we need a function that links them together or maps the linear combination of variables that could result in any value onto the Bernoulli probability distribution with a domain from 0 to 1. The natural log of the odds ratio, the logit, is that link function.

There are three available methods to estimate the coefficients in the logit model: ordinary least squares, weighted least squares, and maximum likelihood. To estimate a logit model by OLS, we would simply take the logit transformation of π, which is $log[\pi/(1-\pi)]$, and regress that transformation on characteristics of the independent variables and on the average characteristics of the dependent variable. A weighted least squares (WLS) analysis would be similar except that the data would be weighted to adjust for hetroscedasticity. However, for logistic regression, least squares estimation is not capable of producing minimum variance unbiased estimators for the actual parameters. Instead, maximum likelihood estimation is used to solve for the parameters that best fit the data (Czepiel, 2002). Maximum likelihood (ML) is the third method for estimating the logit model for grouped data and the only method in general use for individual-level data (Allison, 2012).

4.4.2 Maximum likelihood estimators

The maximum likelihood estimation, also known as the method of maximum likelihood, is a way of estimating the value of an unknown parameter. This method is widely regarded as the best method of point estimation. This is for various reasons. Some of those reasons are:- the method will gives us a feasible results, it can be used when we have censored or truncated data, and its asymptotic properties (Gujarati, 2014).

The reason for the popularity of maximum likelihood estimation is that it is often straightforward to derive ML estimators when there are no other obvious possibilities. One case that ML handles very nicely is data with categorical dependent variables (Allison, 2012).

In ML estimation, there are two steps: constructing the likelihood function based on the distributional assumptions of the data and maximization which requires an iterative numerical method, which means that it involves successive approximations.

Different iterative methods such as Iterative reweighted least square (IRWLS), Fisher scoring, Iterative weighted least square (IWLS), and Newton-Raphson are used to maximize the log-likelihood. All iterative methods give the same solution, but they differ in such factors as speed of convergence, sensitivity to starting values, and computational difficulty at each iteration. The most widely used iterative methods are the Newton-Raphson algorithm, and the Fisher scoring method. The Fisher scoring method is equivalent to the iterative reweighted least square (IRWLS). The Newton-Raphson method uses the standard least squares method to iteratively calculate the maximum likelihood estimates. Detailed discussion of Fisher's scoring and Newton-Raphson can be found in many literatures (Agresti & Kateri, 2011; Kutner et al., 2005; Allison, 1999; Efron & Hinkley, 1978; McCullagh & Nelder, 1989; Schabenberger & Pierce, 2001). The most popular SAS procedure for fitting logistic regression is PROC LOGISTIC. This procedure provides ML estimation of the logistic regression model, which uses Fisher's scoring method by default.

4.4.3 Assumptions of logistic regression

Logistic regression assumes that the dependent variable is a stochastic (randomly determined) event, the outcome must be discrete, there should be no outliers in the data and there should be no high intercorrelation (multicollinearity) among the predictors (Tabachnick & Fidell, 2007).

4.5 Binary logistic regression

4.5.1 Fitting the logistic regression model for binary response

The analysis in the subsequent section uses logistic regression, with the dichotomous outcome of nutritional status of under five children in Ethiopia. In this study, the probability of malnourishment of under five children is modeled as a function of pedictor variables which are stated in section 2.2.2. The categorized variable namely malnourished/normal (Table 2.2) is used as the response variable.

The AIC of the full model (contains intercept and covariates) is smaller compared to the AIC of the reduced model (contains intercept only); this indicates that the fitted model better explains the data (Table 4.1).

Table 4.1: Model fit statistics for binary logistic regression

Criterion	Intercept only	Intercept and Covariates
AIC	8990.746	8589.688
SC	8997.848	8958.979
-2 Log L	8988.746	8485.688

The Likelihood Ratio test (LRT) tests the overall significance of the logistic regression model. The value of likelihood ratio statistic is 503.0579 with P-value < 0.0001. The value of the score test is 504.8851 (P-value < 0.0001) and the Wald test 461.6973 (P-value < 0.0001) also support the results obtained using the likelihood ratio test (Table 4.2). For all three tests the P-value is less than 0.05. It means that the overall fitted logistic model is significant. There is a significant contribution of independent variables in predicting the probability of malnourishment of under five children. In other words, at least one of the parameters is significantly different from zero. The Hosmer and Lemeshow test for goodness of fit of this model is 13.9454 with P-value=0.0832, which shows that the model is a good fit to the data.

Table 4.2: Model evaluation for binary logistic regression

Model evaluation parameters	Chi-square	D.F	P-value
Overall significance			
Likelihood Ratio	503.0579	51	< 0.0001
Score	504.8851	51	< 0.0001
Wald	461.6973	51	< 0.0001
Goodness of fit test			
Hosmer and Lemeshow	13.9454	8	0.0832
Association of predicted probabilities and observed response			
Percent Concordant	65.3	Somers'D	0.307
Percent Discordant	34.7	Gamma	0.307
Percent Tied	0.0	Tau-a	0.098
Pairs	12897054	c	0.653

Another important aspect of the fitted logistic regression that needs to be checked is the validation of the model. The degree to which the predicted probabilities agree with the actual outcomes was expressed using a classification table with a cut-off point set at 0.5 (Melesse et al., 2016). For better prediction power, the c-statistics has to be greater than 0.5. But it ranges from 0 (no association) to 1 (perfect association). The c-statistic is a measure of the predictive accuracy of a logistic regression model. In this particular study, the c-statistic is 0.653 (Table 4.2). This result shows that there is a moderate (65.3%) association between the predicted probabilities and the observed responses (actual probabilities). In addition, Table 4.2 shows that the concordance rate was 65.3%; this value tells us the agreement between the logistic regression model and the observed outcomes. The Pairs indicates the number of pairs for 1 and 0. The Somer's D statistic is 0.307 suggesting that not all pairs are concordant. The Gamma statistic has a value of 0.307 which indicates a small positive association between variables.

Type 3 Analysis of Effects, (Table 4.3) shows the hypothesis tests for each of the variables in the binary logistic regression model individually using multiple degrees of freedom test for the overall effect of the categorical variables (for k categories we use $k - 1$ dummy variables). The Wald χ^2 test statistics and associated P-values are shown in Table 4.3. The results indicate that the continuous variables: mother's BMI and mother's age were found to have statistically significant effect on the response variable (P-value $<$ 0.001). Similarly, the overall effect of categorical variables: sex of child, weight of child at birth, mother's work status, educational attainment of mother, and region were found to have a statistically significant effect on the probability of malnourishment of under five children. However, the overall effect of current age of child, current marital status, religion, place of residence, and wealth index were found to have no significant effect on the probability of the event based on Ethiopian DHS, 2016 data. Moreover, four two-way interaction terms were found significant. These two-way significant interaction terms were: the interaction between current age of child and mother's BMI; current age of child and region; mother's BMI and region; and mother's BMI and weight of child at birth (Table 4.3).

Table 4.4 shows significant interaction effects between the socio-economic, demographic and geographic factors that have influence on the status of malnourishment of under five children. Thus, the effect of current age of child and mother's BMI was found to be positively associated with malnutrition of under five children (P-value=0.0002). The corresponding odds ratio was 0.9756. This implies that with a unit increase in the age of a child and mother's BMI, the odds of malnourishment of under five children increases by $(1 - 0.9756) \times 100\% = 2.44\%$. The interaction between current age of child and Addis Abeba region was found to be positively associated with malnutrition of under five children (P-value=0.0078). The corresponding odds ratio was 1.312. The odds of malnourishment of a child from Addis Abeba region is 1.312 times higher than the odds of malnourishment of a child of the same age from Oromia region. Similarly, the odds of malnourishment of a child from

Table 4.3: Type 3 analysis of effects for the binary logistic model

Main effect	**DF**	**Wald χ^2**	**P-value**
Current age of child	1	3.4020	0.0651
Sex of child	1	11.1631	0.0008
Weight of child at birth	2	7.0563	0.0294
Mother's current age	1	11.8260	0.0006
Mother's BMI	1	14.0367	0.0002
Mother's work status	1	4.4548	0.0348
Educational attainment of mother	3	14.8606	0.0019
Current marital status	1	0.0766	0.7820
Religion	4	5.8852	0.2079
Region	10	21.0452	0.0208
Place of residence (rural/urban)	1	0.0245	0.8756
Wealth index	2	2.5340	0.2817
Significant interaction effect			
Current age of child and mother's BMI	1	13.4686	0.0002
Current age of child and region	10	29.7230	0.0010
Mother's BMI and region	10	25.1905	0.0050
Mother's BMI and weight of child at birth	2	9.2080	0.0100

Amhara region increases by 1-0.862=0.138 as compared to the odds of malnourishment of a child of the same age from Oromiya region. The odds of malnourishment of a child to the effect of current age of a child and Gambela region is 1.262 times higher than the odds of malnourishment of a child to the effect of current age of a child and Oromia region.

The odds of malnourishment of a child to the effect of mother's BMI and Addis Abeba region increases by 1-0.922=0.078 as compared to the odds of malnourishment of a child to the effect of mother's BMI and Oromiya region. The odds of malnourishment of a child from Dire Dawa region increases by 1-0.925=0.075 as compared to the odds of malnourishment of a child from Oromia region who had mothers with the same BMI. The odds of malnourishment of a child from Somali region increases by 1-0.896=0.104 as compared to the odds of malnourishment of a child from Oromia region who had mothers with the same BMI.

The other significant two-way interaction effect was found between mother's BMI and weight of child at birth. Thus, the effect of mother's BMI and child who had large weight at birth compared to small weight at birth was found to be positively associated with malnourishment of under five children (P-value= 0.0016). The corresponding odds ratio was 1.092. The odds of malnourishment of a child to the effect of mother's BMI and a child who had large weight at birth is 1.092 times higher than the odds of malnourishment of a child to the effect of mother's BMI and a child who had small weight at birth.

In addition to the significant interaction effects, Table 4.4 shows that children who had large weight at birth, mother's BMI, mother's work status, educational level (secondary school), region (Affar and Somali), and wealth index (rich) were found to have a significant effect on malnourishment of under five children in Ethiopia.

Table 4.4: Socio-economic, demographic and geographic of effects on response variable for main effects and significant two-way interaction effects.

Main effcts	**Estimate**	**SE**	**OR**	**P-value**
Intercept	-3.1025	0.6849		<0.0001
Current age of child	0.1974	0.1465	1.218	0.1779
Mother's current age	-0.00778	0.00414	0.992	0.0601
Mother's BMI	0.1126	0.0329	1.119	0.0006
Sex of child (ref. = Male)				
Female	-0.00628	0.0523	0.994	0.9044
Weight of child at birth (ref. = Small)				
Average	-0.3588	0.5212	0.698	0.4912
Large	-1.3413	0.5788	0.262	0.0205
Mother work status (ref. = No)				
Yes	-0.1281	0.0607	0.880	0.0348
Educational level (ref. = No education)				
Primary school	-0.6055	0.3329	0.546	0.0689
Secondary school	-1.1423	0.3713	0.319	0.0021
Higher	-0.1525	0.2657	1.165	0.5660
Current marital status (ref. = Married)				
Not married	-0.0320	0.1156	0.969	0.7820
Religion (ref. = Orthodox)				
Catholic	-0.2845	0.3836	0.7524	0.4583
Muslim	0.1362	0.0949	1.1459	0.1511
Other	0.3032	0.2143	1.3542	0.1571
Protestant	0.2140	0.1102	1.2386	0.0521
Region (ref. = Oromia)				
Addis Abeba	1.3826	0.9283	3.985	0.1364
Affar	1.6030	0.7888	4.968	0.0421
Amhara	-0.0621	0.9079	0.939	0.9455
Benishangul	-0.00947	1.0408	0.991	0.9927
Dire Dawa	1.2040	0.8324	3.333	0.1480
Gambela	-0.0578	0.9412	0.944	0.9510
Harari	0.8848	0.7846	2.422	0.2594
SNNP	-0.3989	0.7423	0.671	0.5910
Somali	2.1369	0.7091	8.473	0.0026
Tigray	0.8312	0.8834	2.296	0.3467
Place of residence (ref. = Rural)				
Urban	-0.0178	0.1134	0.982	0.8756
Wealth index (ref. = Poor)				
Middle	0.1176	0.0909	1.125	0.1959
Rich	0.1803	0.0916	1.198	0.0490
Significant interaction effects				
Current age of child and mother's BMI	-0.0247	0.00674	0.9756	0.0002
Current age of child and region (ref. = Oromia)				
Age and Addis Abeba	0.2718	0.1022	1.312	0.0078
Age and Amhara	-0.1482	0.0748	0.862	0.0474
Age and Gambela	0.2328	0.0886	1.262	0.0086
Mother's BMI and region (ref. = Oromia)				
Mother's BMI and Addis Abeba	-0.0808	0.0398	0.922	0.0425
Mother's BMI and Dire Dawa	-0.0782	0.0385	0.925	0.0423
Mother's BMI and Somali	-0.1104	0.033	0.896	0.0008
Mother's BMI and weight of child at birth (ref. Small)				
Mother's BMI and Large	0.0881	0.028	1.092	0.0016

The probability of malnourishment of a child as a function of the independent variables (Table 4.4) is estimated by

$$\hat{P} = \frac{e^{-3.1025} + \sum_{i=1}^{k} \hat{\beta}_i x_i}{1 + e^{-3.1025} + \sum_{i=1}^{k} \hat{\beta}_i x_i} \tag{4.3}$$

where $\hat{\beta}_i' s$ are the estimated coefficients corresponding to $x_i' s$, variables that have significant effect on the response variable.

Since we are modeling the probability of an event in logistic regression the interpretation of parameter estimates is different from the way it is explained or interpreted in multiple regression. The dependent variable in logistic regression is binary. We are not modeling the actual change in the dependent variable. Instead, we are modeling the probability of the event.

The +/- sign of β-Coefficients indicate their respective positive/negative relationship with the probability of malnourishment of a child. Odds ratios can be computed as an exponent to the power of the logistic regression coefficients. We can find the predicted probability of the event from Equation (4.3). When the exponentiated beta value is greater than one, then the probability of higher category increases, and if the exponential beta is less than one, then the probability of higher category decreases.

Based on the result from Table 4.4 the odds ratio for variables, which have significant effect in the probability of malnourishment of a child, are interpreted as follows: the odds of malnourishment of a child who had large weight at birth is 0.262 times the odds of malnourishment a child who had small weight at birth.

The odds ratio of 1.119 for mother's BMI indicates that for a one unit increase in mother's BMI, the odds of malnourishment in under five children increases by 11.9%. Furthermore, mother's work status has a significant effect on the probability of malnourished children (P-value=0.0348). The corresponding odds ratio was 0.880. In addition, mothers who had secondary school education level have a lower odds of having a malnourished child compared to mothers who had no education (OR=0.319). Children who live in Somali region (OR=8.473) were found to be at a higher odds of malnourishment of under five children compared to Oromiya region followed by Affar region (OR=4.968). Moreover, under five children from rich households had a significant effect on the event.

Plots for significant interaction effects

Figures 4.1-4.3 show the predicted probability of malnourishment of a child against significant interaction effects. The interaction effect plot between mother's BMI and regions (these regions were Addis Abeba, Dire Dawa, and Somali) is presented in Figure 4.1. From the figure, it is clearly seen that the predicted probability of malnourishment of a child among under five children increases as mother's BMI increases in Addis Abeba, Dire Dawa and Somali region.

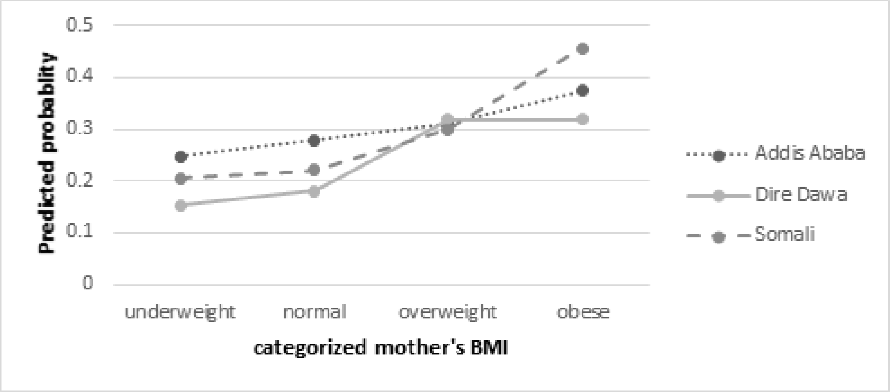

Figure 4.1 – Predicted probability of malnourishment of a child based on the effect of mother's BMI and region.

The relationship between current age of a child and mother's weight status is presented in Figure 4.2. As the children's age increases, the probability of malnourishment of a child decreases across increasing mother's weight status.

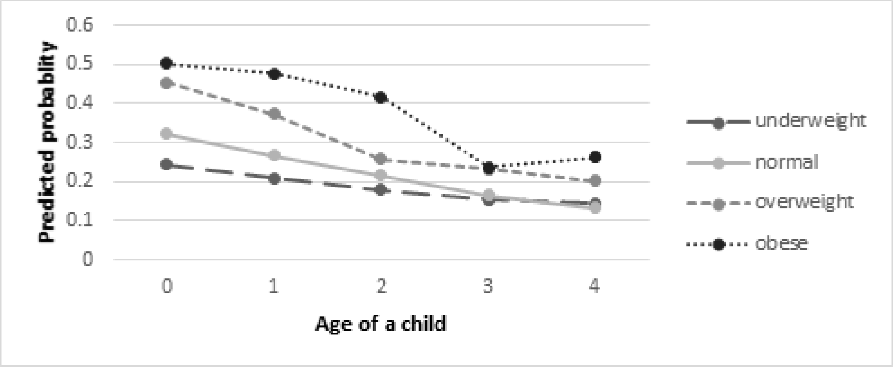

Figure 4.2 – Predicted probability of malnourishment of a child based on the effect of mother's BMI and current age of a child.

The relationship between age of a child and regions (these regions were Addis Abeba, Amhara and Gambela) is presented in Figure 4.3. The figure shows that the probability of malnourishment of a child was almost the same over all ages in Gambela region. As age increases, the probability of malnourishment of a child decreased monotonically in Amhara region followed by Addis Abeba region.

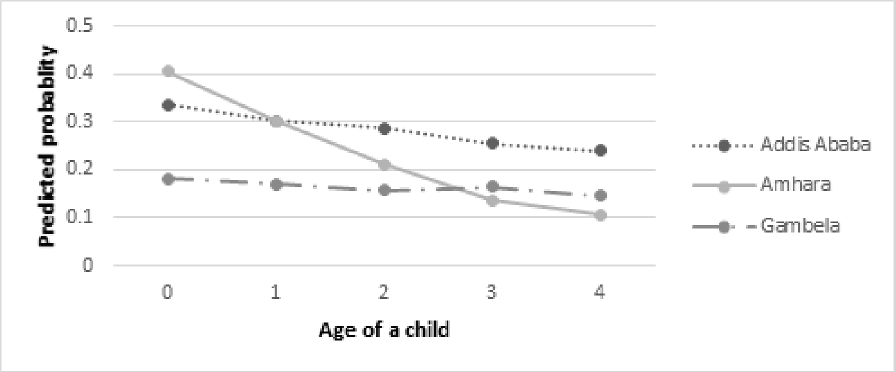

Figure 4.3 – Predicted probability of malnourishment of a child based on the effect of region and age of a child.

Receiver Operating Characteristic Curve (ROC)

A Receiver Operating Characteristic Curve (ROC) is a standard technique for summarizing classifier performance over a range of trade-offs between true positive (TP) and false positive (FP) error rates (Swets, 1988). The true positive rate measures the proportion of observations classified as the true outcome of interest (malnourishment of a child) over all those classified as malnourished children. The false positive rate measures the proportion of observations misclassified as the outcome of interest (malnourishment of a child) over all those classified as malnourished children (Melesse et al., 2016). ROC curve is a plot of sensitivity (the ability of the model to predict an outcome of interest (malnourishment of a child in our case) correctly) versus 1-specificity (the ability of the model to predict an unwanted outcome of interest correctly) for possible cut-off classification probability values π_0 (Swets, 1988). In general, ROC curves are used to check how much the predicted probability agrees with an outcome of interest. The Area Under the Curve should be maximum (close to 1 for a good predictive model).

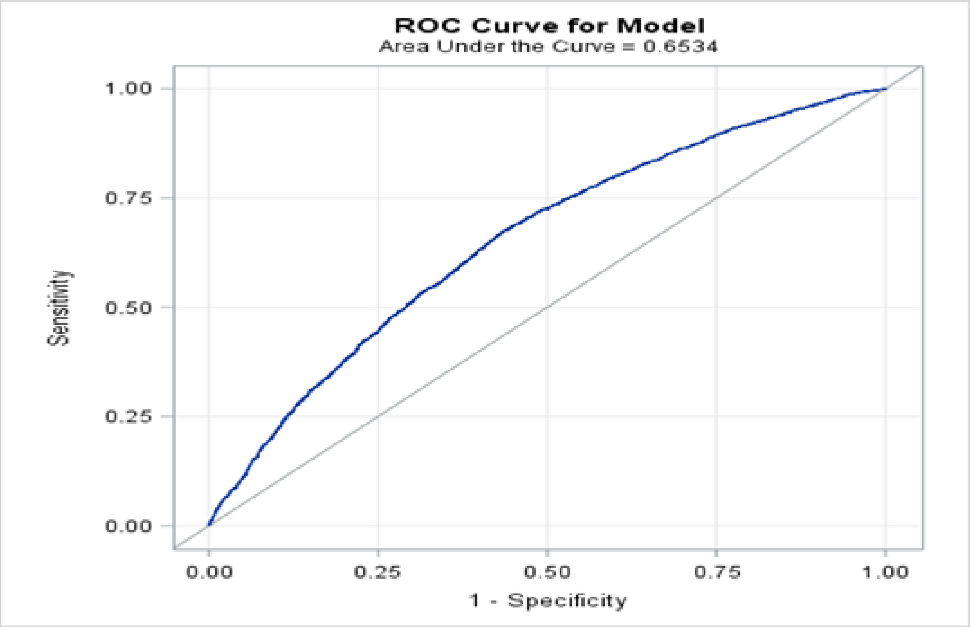

Figure 4.4 – ROC curve for binary logistic regression model.

The Area Under the Curve (AUC) is referred to as the accuracy index, or concordance index, c, in SAS. The more it touches the Y-axis, a value of AUC close to 1, the better prediction power the model has. For better prediction power c must be greater than 0.5 (Hosmer et al., 2013).

In our case the Area Under the Curve=0.6534 (Figure 4.4), which is the same as the c-statistics (Table 4.4) and indicates the moderate predictive power of the model.

4.6 Survey logistic regression

When any sampling method other than simple random sampling is used, survey data analysis software has to be used. The survey analysis method is useful to include the design effect in the estimation of parameters and to adjust the standard errors of the estimates. If the sampling design is not included in the analysis, the standard errors will likely be underestimated, possibly leading to results that seem to be statistically significant, when in fact, they may not be significant. Therefore, this may lead us to biased estimates.

Binary responses can be modeled through binary models that can provide a relationship between the probability of a response and a set of covariates. However, for data which does not come from simple random sampling, the standard logistic regression is not appropriate. According to Rao & Scott (1984), when the data come from a complex survey design with stratification, clustering, and/or unequal weighting, the usual estimates are not appropriate. In these cases, specialized techniques must be applied to produce the appropriate estimates and standard errors. The logistic regression model used to analyze data from complex sampling designs is referred as survey logistic regression models. Survey logistic regression models have the same theory as ordinary logistic regression models (Ayele et al., 2013). The difference between the two is that survey logistic accounts for the complexity of the sampling designs. Simple random sampling designs assume that all units in the population have equal probability of being included in the sample. However, most sample survey data are collected from a finite population with a probability based complex sample design (Rao & Scott, 1981). The main idea here is that simple logistic regression does not account for clustered correlated observations while survey logistic regression does.

4.6.1 Parameter estimation in survey logistic regression

Binary logistic regression with complex survey design uses a modified maximum likelihood estimation called the *pseudo-maximum likelihood estimation* (PMLE). It incorporates element weights in the estimating equation.

Likelihood function in survey logistic regression

Let $U = \{1, 2, ..., N\}$ be a finite population divided into $h = 1, 2, ..., H$ strata, each stratum is further divided into $j = 1, 2, ..., n_h$ primary sampling units (PSU), which is constituted by $i = 1, 2, ..., n_{hj}$ secondary sample units (SSU), comprising of n_{hji} elements. In this case, the first stage primary sampling unit (PSU), was the smallest administrative unit in Ethiopia known as *Kebele*. In the second stage (SSU), households within a *Kebele* were sampled. The response of the i^{th} children in the j^{th} household and h^{th} *Kebele* can be specified as Y_{hji}. Assume, the observed data consists of n'_{hj} SSUs chosen from n'_h PSUs in the h^{th} stratum. Thus, the total number of the observations is given by

$$N = \sum_{h=1}^{H} \sum_{j=1}^{n'_h} \sum_{i=1}^{n'_{hj}} n_{hji} \tag{4.4}$$

Suppose that $\pi_{hji} = P(Y_{hji} = 1|X_{hji})$, is the probability of having a malnourished child in the j^{th} household and h^{th} *Kebele*, and the sampling weight in each sampling unit is denoted by W_{hji}, for unit hji. Thus, the survey logistic regression model is given by

$$logit(\pi_{hji}) = log\left\{\frac{\pi_{hji}}{1-\pi_{hji}}\right\} = \boldsymbol{X}'_{hji}\boldsymbol{\beta} \tag{4.5}$$

where Y_{hji}, X_{hji} and $\boldsymbol{\beta}$ are the categorical response variable, the covariate matrix, and the regression coefficients respectively.

The parameters $\boldsymbol{\beta}$ of the logistic regression model in the complex sampling design are estimated by the Pseudo- maximum likelihood method called weighted maximum likelihood that incorporates the sampling design and the different sampling weights in the estimation of $\boldsymbol{\beta}$ (Hosmer & Lemeshow, 2000; Lumley et al., 2004).

Pseudo-maximum likelihood estimation

The Pseudo-maximum likelihood function for the contribution of a single observation in complex sampling design is given by

$$\pi_{hji}^{W_{hji}Y_{hji}}(1-\pi_{hji})^{(1-W_{hji}Y_{hji})} \tag{4.6}$$

Thus, the Pseudo-maximum likelihood function with weight W_{hji} for a set of n observation is given by

$$L(\boldsymbol{\beta}|W_{hji}, Y_{hji}) = \prod_{h=1}^{H}\prod_{j=1}^{n'_h}\prod_{i=1}^{n'_{hj}}\pi_{hji}^{W_{hji}Y_{hji}}(1-\pi_{hji})^{(1-W_{hji}Y_{hji})} \tag{4.7}$$

$$= \pi_{hji}^{\sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}}W_{hji}Y_{hji}}(1-\pi_{hji})^{1-\sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}}W_{hji}Y_{hji}} \tag{4.8}$$

The main idea of this method is to define a function which approximates the likelihood function of the sampled finite population with a likelihood function formed by the observed sample and the known sampling weights (Hosmer & Lemeshow, 2000; Lumley et al., 2004). In this case, the Pseudo-log-likelihood function is given by

$$\ell(\boldsymbol{\beta}|W_{hji}Y_{hji}) = log\left[\pi_{hji}^{\sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}}W_{hji}Y_{hji}}(1-\pi_{hji})^{1-\sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}}W_{hji}Y_{hji}}\right]$$

$$= log\left(\pi_{hji}^{\sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}}W_{hji}Y_{hji}}\right) + log\left((1-\pi_{hji})^{1-\sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}}W_{hji}Y_{hji}}\right)$$

$$= \sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}} W_{hji}Y_{hji}log(\pi_{hji}) + \left(1 - \sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}} W_{hji}Y_{hji}\right) log(1-\pi_{hji})$$

$$= \sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}} W_{hji}Y_{hji}log(\pi_{hji}) + log(1-\pi_{hji}) - \sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}} W_{hji}Y_{hji}log(1-\pi_{hji})$$

$$= \sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}} W_{hji}Y_{hji}log\left(\frac{\pi_{hji}}{1-\pi_{hji}}\right) + log(1-\pi_{hji}) \qquad (4.9)$$

Substitute equation (4.5) into (4.9)

$$= \sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}} W_{hji}Y_{hji}\boldsymbol{X}'_{hji}\boldsymbol{\beta} + log\left(\frac{1}{1+exp(\boldsymbol{X}'_{hji}\boldsymbol{\beta})}\right)$$

$$= \sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}} W_{hji}Y_{hji}\boldsymbol{X}'_{hji}\boldsymbol{\beta} - log\left(1+exp(\boldsymbol{X}'_{hji}\boldsymbol{\beta})\right)$$

The Pseudo maximum likelihood estimator of $\boldsymbol{\beta}$ is obtained by taking the derivative of the Pseudo-log-likelihood function with respect to beta and equating it to zero.

$$\boldsymbol{\beta} = \frac{\partial \ell(\boldsymbol{\beta}|W_{hji}Y_{hji})}{\partial \boldsymbol{\beta}} = 0$$

$$= \sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}} W_{hji}Y_{hji}\boldsymbol{X}'_{hji} - \left(\frac{1}{1+exp(\boldsymbol{X}'_{hji}\boldsymbol{\beta})}\right) \times exp(\boldsymbol{X}'_{hji}\boldsymbol{\beta}) \times \boldsymbol{X}'_{hji}$$

$$\boldsymbol{\beta} = \sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}} W_{hji}Y_{hji}\boldsymbol{X}'_{hji} - \left(\boldsymbol{X}'_{hji}\left(\frac{exp(\boldsymbol{X}'_{hji}\boldsymbol{\beta})}{1+exp(\boldsymbol{X}'_{hji}\boldsymbol{\beta})}\right)\right) \qquad (4.10)$$

To solve the solutions for Equation (4.10), one can use numerical methods: Newton-Raphson, Fisher scoring or IRLS. SAS PROC SURVEYLOGISTIC for the logistic regression model in the complex sampling design reports three different "Model fit statistics"in the output: Akaike Information Criterion (AIC) introduced by Akaike (Akaike, 1974), Schwarz Criterion (SC) also known as Bayesian Information Criterion (BIC)) introduced by Schwarz (Schwarz et al., 1978), and -2logL. Values of these fit statistics are displayed for two different models: a model with an intercept only, and a model that includes all the specified predictors (a model with an intercept and covariates). The smallest value of AIC is considered the best, and also if the model has the smallest value of SC it is most desirable.

Akaike's Information Criterion (AIC) is given by

$$AIC = -2logL + 2K$$

where k is the number of parameters (including the intercept) in the model. AIC is used for the comparison of models from different samples or non-nested models and the model with the smallest AIC is considered the best (SAS, 2014).

The Schwarz Criterion (SC) also known as Bayesian Information Criterion (BIC) adjusts the -2logL statistics for the number of parameters and is given by

$$-2logL + Klog(n)$$

where n is the overall sample size and k is as explained above for AIC. Like AIC, SC penalizes for the number of predictors in the model and the smallest SC is most desirable.

The most fundamental of the fit statistics, $-2logL$, which is used to compare different models fit to the same data set (nested models), is the maximized value of the logarithm of the likelihood function multiplied by -2. $-2logL$ will always decrease as new explanatory variables (interactions) enter into the model even if they are insignificant. The smaller the deviance, which is distance measure of the log likelihood for the main effect model with that of saturated model, the better the model. -2 *log* likelihood is given by

$$-2log\left(\frac{likelihood \quad for \quad null \quad model}{likelihood \quad for \quad fitted \quad model}\right)$$

4.6.2 Variance estimation in survey logistic regression

There is no direct form to calculate the variance estimators under complex sampling designs. To obtain the variance estimators, a modified maximum pseudo-likelihood is used by some form of replication method, such as Jackknife repeated replication, balanced repeated replication, bootstrap, and Taylor linearization (Hosmer & Lemeshow, 2000; Lumley et al., 2004; Lee & Forthofer, 2006). Many statistical software packages are available for performing the computations (Lepkowski & Bowles, 1996).

In SAS, PROC LOGISTIC does not compute the proper variance estimators for analyzing complex survey data for the categorical outcome. PROC SURVEYLOGISTIC procedure in SAS is designed to perform the necessary and correct computations. PROC SURVEYLOGISTIC procedure fits the linear logistic regression model for a discrete response variable from survey data. The regression parameters and odds ratios are estimated by maximum likelihood method. Standard errors around the estimates are calculated by Taylor expansion approximation (SAS, 2014).

Taylor linearization method for the estimated covariance matrix of $\boldsymbol{\beta}$ is given by

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = (X'DX)^{-1}S(X'DX)^{-1},$$

where X is the design matrix, $D = WV$ is the $n \times n$ diagonal matrix with elements $W_{hji}\pi_{hji}(1-\pi_{hji})$, and S is the pooled estimator within-stratum of the covariance matrix. That is,

$$S = \sum_{j=1}^{H}(1-f_h)\frac{n'_h}{n'_h - 1}\sum_{j=1}^{n'_h}(X_{hj..} - \bar{X}_{h...})(X_{hj..} - \bar{X}_{h...})',$$

where $X_{hji} = W_{hji}\pi_{hji}(1-\pi_{hji})$, $X_{hj..} = \sum_{j=1}^{n'_h} X_{hji}$, and specific mean in the stratum as $\bar{X}_h = \frac{1}{n'_h}\sum_{j=1}^{n'_h} X_{hj..}$. The finite population correction factor is given by $(1-f_h)$, where $f_h = \frac{n'_h}{n_h}$ is the ratio of the number of PSU observed by the total number of the PSU in the stratum h.

The hypothesis test for the significance of the regression coefficients and the test for the goodness of fit of a model also needs to be modified to incorporate the sampling design and the different weights of observation. According to Hosmer & Lemeshow (2000), and Lumley & Scott (2014), the evaluation of the contribution of the covariates is made by the adjusted Wald test (W), with test statistics given by

$$F = \left(\frac{s-p+1}{sp}\right)W$$

$$W = \hat{\boldsymbol{\beta}}'[\widehat{Var}(\hat{\boldsymbol{\beta}})]^{-1}\hat{\boldsymbol{\beta}}$$

where $s = \sum_{h=1}^{H} n'_h - H$ is the total number of the selected PSU (sampled cluster) minus the number of strata, and p is the number of covariates. The P-value can be computed using the above F-distribution with p and $(s-p+1)$ degrees of freedom, that is P-value = $P[F(p, s-p+1) \geq F]$.

4.6.3 Model selection and model checking

Model selection

The option for variable selection procedure, that is forward, backward, and stepwise, in SAS PROC SURVEYLOGISTIC (version 9.4) which is not yet available. Therefore, one should manually select one variable at a time in the model and observe the contribution of each variable effect, then exclude a variable with insignificant effect (one at a time) and observe again the contribution of the remaining variables. This process will continue until the model has only significant effects.

Model checking

Since SAS PROC SURVEYLOGISTIC does not produce plots and Hosmer-Lemeshow statistics, we use the AIC and BIC to compare the models, the likelihood for measuring the goodness of fit considering the complex sampling frame (Lumley & Scott, 2014, 2015). The smaller the AIC and BIC of the full model compared to the corresponding AIC and BIC of the reduced model, the better the full model is. Details of standard criteria for model selection such as AIC and BIC can be found in Akaike (1974); Schwarz et al. (1978); and Lumley & Scott (2014).

4.6.4 Design effect

The sampling variance of a survey statistic is affected by the stratification, clustering, and weighting of selected cases. Stratification may increase the precision of the variance estimate, but clustering and weighting decrease precision (Dowd et al., 2001). Stratified sampling is a process that involves the division or stratification of a population by partitioning population units in the sampling frame into non-overlapping and relatively homogeneous groups called strata. When the strata have been determined, a sample is drawn from each stratum, the drawings being made independently in different strata. If a simple random sample is taken in each stratum, the procedure is called stratified random sampling. One or more of the following reasons are the purpose of stratified sampling:- to reduce sampling error compared to simple random sampling, for administrative convenience, when different parts of population need different sampling procedure, and when separate estimates are required at domain or strata level. Many sampling frames preparation and many selections are the disadvantages of stratified sampling. In cluster sampling the population is first divided into a heterogeneous subset of the population (cluster), then a simple random sampling of clusters is taken. However, its disadvantage is that there is a loss of precision (standard error is high compared with other sampling designs). Detailed discussion of sampling techinques can be found in many literatures (Cochran, 2007; Sharon, 1999).

Because cluster sampling and the cumulative effect of the range of factors affect the precision of a survey statistic and a smaller sample size, an adjustment called the design effect should be used to determine survey sample size. The design effect, deff, is defined as the ratio of the sampling variance of the statistic under the actual sampling design divided by the variance that would be expected for a simple random sample of the same size (Dowd et al., 2001).

$$deff = \frac{Variance(complex \quad design)}{Variance(SRS)}$$

The design effect is used to determine how much larger the sample size or confidence interval needs to be. Usually deff ranges from 1 to 3. It is not uncommon, however, for the design effect to be much larger.

In most cases, the $deft$, which is $\sqrt{deff}$, is preferable to make the determination of design effect, because deft is less variable than deff. The deft shows how much the sample standard error, and consequently the confidence intervals, increase. Thus, for example if the deft is 2, the confidence intervals have to be 2 times as large as they would for a simple random sample. A $deft$=1 indicates no effect of sample design on standard error. The value of $deft$ >1 indicates sample design that inflates the standard error of the estimate. The value of $deft$ <1 indicates sample design that does not inflate the standard error of the estimate.

4.6.5 Application of the binary logistic regression model with complex survey design

The analysis in the subsequent section uses survey logistic regression model, which considers the complexity of the survey design. The probability of malnourishment of under five children was modeled as a function of selected predictor variables described in section 2.2.2.

The AIC, SC, and $-2logL$ of the full model (contains intercept and covariates) is smaller compared to the corresponding criterion of the reduced model (contains intercept only); this indicates that the fitted full model better explains the data (Table

4.5). The P-values corresponding to the likelihood ratio, Score test, and Wald tests

Table 4.5: Survey logistic regression model fit statistics for binary response.

Criterion	Intercept only	Intercept and Covariates
AIC	10094.987	9547.940
SC	10102.164	9921.138
-2 Log L	10092.987	9443.940

are less than 0.05. It means that the overall fitted survey logistic model is significant. There is a significant contribution of independent variables in the prediction of the event (in our case malnurishement of a child). In other words, at least one of the parameters is significantly different from zero. There is a moderate (62.8%) association between the predicted probabilities and the observed responses (actual probabilities). In addition, Table 4.6 shows that the concordant rate was 62.5%.

Table 4.6: Survey logistic regression model evaluation for a binary response.

Model evaluation parameters	F-Value	Num DF	Den DF	P-value
Overall significance				
Likelihood Ratio	11.02	41.1810	25326	<0.0001
Score	10.54	51	565	<0.0001
Wald	9.41	51	565	<0.0001
Association of predicted probabilities and observed response				
Percent Concordant	62.5	Somers'D	0.255	
Percent Discordant	37.0	Gamma	0.257	
Percent Tied	0.6	Tau-a	0.082	
Pairs	12897054	c	0.628	

Type 3 Analysis of Effects, Table 4.7, shows that from four two-way interaction terms, the interactions between mother's BMI and region was found to have significant interaction effect on the response variable (Table 4.7). Moreover, the hypothesis tests for each of the variables in the survey logistic regression model for the binary response individually using multiple degrees of freedom test for the overall effect of the categorical variables. The P-values shown in Table 4.7 indicate that the continuous variables current age of child and mother's current age were found to have a significant effect on the response of interest. Similarly, the overall effect of categorical variables: mother's work status and educational attainment of mother were found to have a statistically significant effect on the probability of malnourishement of under five children.

The output from SAS PROC SURVEYLOGISTIC (version 9.4) is presented in Table 4.8; it shows that the effect of mother's BMI and the child from Addis Abeba region was found to be negatively associated with malnutrition of under five children (P-value=0.0456). The corresponding odds ratio was 0.917. This implies that the odds

Table 4.7: Type 3 analysis of effects for the binary logistic regression with complex survey design

Main effect	F-Value	Num DF	Den DF	P-value
Current age of child	4.80	1	615	0.0289
Sex of child	1.65	1	615	0.1989
Weight of child at birth	0.43	2	614	0.6516
Mother's current age	7.00	1	615	0.0084
Mother's BMI	3.43	1	615	0.0645
Mother's work status	0.48	1	615	0.0491
Educational level	2.33	3	613	0.0132
Current marital status	3.49	1	615	0.0624
Religion	1.67	4	612	0.1555
Region	1.19	10	606	0.2933
Place of residence(rural/urban)	0.95	1	615	0.3299
Wealth index	0.91	2	614	0.4020
Significant interaction effect				
Current age of child and mother's BMI	0.20	1	615	0.6584
Current age of child and region	2.61	10	606	0.1972
Mother's BMI and region	1.36	10	606	0.0041
Mother's BMI and weight of child at birth	0.64	2	614	0.5281

of malnourishment of under five children from Addis Abeba region decrease by (1-0.917=0.083) as compared to the odds of malnourishment of under five children from Oromia region who had mothers with the same BMI. The effect of mother's BMI and the child from Somali region was found to be also negatively associated with malnutrition of under five children (P-value=0.005). The corresponding odds ratio was 0.889. This implies that the odds of malnourishment of under five children from Somali region decrease by 0.111 as compared to the odds of malnourishment of under five children from Oromia region who had mothers with the same BMI.

In addition to the interaction effects, the results from Table 4.8 shows that the probability of malnutrition of under five children has significant association with mother's current age, mother's BMI, mother's working status, educational level of mother (primary school and secondary school) and Harari region when we considered the logistic regression model with complex sampling design.

Based on the result from Table 4.8 the odds ratio for variables, which have significant effect on the probability of malnourishement of under five children are interpreted as follows: the odds of 0.986 for mother's age indicates that for a one year increase in mother's age, the odds of having a malnourished child decrease by 0.014. Furthermore, for one unit increase in mother's BMI, the odds of having a malnourished child will be multiplied by 1.104. The working status of mothers was found to be negatively associated with malnourishment of under five children as compared to mothers who were not working (P-value=0.031). The corresponding odds ratio was

0.835. In addition, the significant effect of covariates in terms of adds ratios in the logistic regression model with complex survey design can be interpreted in the same way as those in the logistic regression model without complex survey design.

Table 4.8: Survey logistic regression model estimation on the binary response for main effects and significant interaction effects.

Main effcts	**Estimate**	**SE**	**OR**	**P-value**
Intercept	-2.4657	0.7390		0.0009
Current age of child	-0.00534	0.2375	0.995	0.9821
Mother's current age	-0.0139	0.00633	0.986	0.0283
Mother's BMI	0.0991	0.0344	1.104	0.0041
Sex of child (ref. = Male)				
Female	-0.0354	0.0766	0.965	0.9044
Weight of child at birth (ref. = Small)				
Average	-0.6758	0.8043	0.509	0.4012
Large	-1.0226	1.0041	0.359	0.3089
Mother's work status (ref. = No)				
Yes	-0.1808	0.0836	0.835	0.0310
Educational level (ref. = No education)				
Primary school	-1.2956	0.4809	0.274	0.0073
Secondary school	-1.3937	0.4892	0.248	0.0045
Higher	0.1519	0.3337	1.164	0.6491
Current marital status (ref. = Married)				
Not married	-0.3395	0.1819	0.7121	0.0624
Religion (ref. = Orthodox)				
Catholic	0.0504	0.3360	1.052	0.8808
Muslim	-0.1289	0.1227	0.879	0.2938
Other	-0.0339	0.3305	0.967	0.9183
Protestant	0.0508	0.1127	1.052	0.6523
Region (ref. = Oromia)				
Addis Abeba	1.4824	1.0436	4.403	0.1560
Affar	-1.0380	1.4610	0.354	0.4777
Amhara	0.3899	1.0214	1.477	0.7028
Benishangul	0.3536	1.2129	1.424	0.7708
Dire Dawa	-2.1058	1.5642	0.122	0.1788
Gambela	-3.2258	3.2966	0.039	0.3282
Harari	-4.1726	1.7892	0.015	0.0200
SNNP	-0.3126	0.6985	0.732	0.6547
Somali	1.7657	0.9463	5.846	0.0626
Tigray	-0.1939	1.0752	0.823	0.8569
Place of residence (ref. = Rural)				
Urban	-0.1564	0.2249	0.855	0.4871
Wealth index (ref. = Poor)				
Middle	0.1782	0.1238	1.195	0.1506
Rich	0.2109	0.1256	1.235	0.0937
Significant interaction effects				
Mother's BMI and region (ref. = Oromia)				
Mother's BMI and Addis Abeba	-0.0862	0.0430	0.917	0.0456
Mother's BMI and Somali	-0.1169	0.0415	0.889	0.005

4.6.6 Comparison of results obtained from binary logistic regression with simple random sampling and with complex survey design

Table 4.9 was constructed to compare standard error, and confidence interval obtained from PROC SURVEYLOGISTIC procedure (Table 4.8) with the ones obtained from PROC LOGISTIC procedure (Table 4.4) based on deff and deft.

Table 4.9 shows the deff and deft value for each significant effect in the study. Thus, the effect of mother's current age has the deff value of 2.3529 and deft value of 1.5339. The deft value equal to 1.5339, indicates that the sample standard error, and consequently the confidence interval are 1.5339 times bigger than they would be if the survey were based on simple random sampling. The effect of mother's BMI has deff=1.0901 and deft=1.0441. The deff value equal to 1.0901 indicates that the sample standard error, and consequently the confidence interval are 1.0901 times bigger than they would be if the survey were based on the same simple random sampling. The effect of currently working mothers has deff=1.8971 and deft=1.3773. The standard errors and confidence interval are 1.3773 times bigger than they would be for simple random sampling. The effect of mothers who had primary education level has deff=15.1469 and deft=3.8919. The deft value equal to 3.8919 indicates that the sample standard error, and consequently the confidence interval, are 3.8919 times bigger than they would be if the survey were based on the same simple random sampling. The effect of mothers who had secondary education level has deff=1.7359 and deft=1.3175. The deft value equal to 1.3175 indicates that the sample standard error, and consequently the confidence interval, are 1.3175 times bigger than they would be if the survey were based on the same simple random sampling. The effect of a child from Harari region has deff=5.2002 and deft=2.2804. A value of deft equal to 2.2804 indicates that the sample standard error, and consequently the confidence interval, are 2.2804 times bigger than they would be if the survey were based on the same simple random sampling. The effect of mother's BMI depending on whether a child is from Addis Abeba region has deff=1.1673 and deft=1.0804. The standard error and the confidence interval are 1.0804 times as bigger than they would be for the same simple random sampling. The effect of mother's BMI depending on whether a child is from Somali region has deff=1.5812 and deft=1.2575. The deft of 1.2575 indicates that the sample standard error, and consequently the confidence interval, are 1.2575 times bigger than they would be if the survey were based on simple random sampling.

Furthermore, information on design effects should also be used when we are planning to determine the sample size of the study. Once we have an estimated design effect, it is straightforward to adjust the required sample size; we need only to multiply the sample size needed under simple random sampling by the estimated design effect. For instance, deff=2.3529 (Table 4.9) indicates that for logistic regression studies with complex survey design, the sample size for mother's current age is 2.3529 times as large as would be needed under simple random sampling (logistic regression without complex survey design).

Table 4.9: Estimated design effects for binary response

			Study result			
Significant main effects	**Estimates (CSD)**	**P-value**	**Var(CSD)**	**Var(SRS)**	**deff**	**deft**
Intercept	-2.4657	0.0009	0.546121	0.4691	1.1642	1.0789
Mother's current age	-0.0139	0.0283	0.00004	0.000017	2.3529	1.5339
Mother's BMI	0.0991	0.0041	0.00118	0.00108241	1.0901	1.0441
Educational level (ref. = No education)						
Primary school	-1.2956	0.0073	1.67858	0.11082	15.1469	3.8919
Secondary school	-1.3937	0.0045	0.23932	0.13786	1.7359	1.3175
Mother's work status(ref.=No)						
Yes	-0.1808	0.0310	0.00699	0.0036845	1.8971	1.3773
Region (ref. = Oromiya)						
Harari	-4.1726	0.0200	3.20124	0.615597	5.2002	2.2804
Significant interaction effects						
Mother's BMI and region (ref. Oromiya)						
Mother's BMI and Addis Abeba	-0.0862	0.0456	0.001849	0.00158404	1.1673	1.0804
Mother's BMI and Somali	-0.1169	0.005	0.001722	0.001089	1.5812	1.2575

CSD(complex survey design), SRS(simple random sampling)

From the above results, we observe that the design effects values are above one. This confirms that there was an under-estimation of variance while using logistic regression, which assumes data was sampled using simple random sampling. Since logistic regression with complex survey design does not assume simple random sampling, the parameter estimates for both models are not the same. However, in some cases, they are closer to one another. One of the assumptions for logistic regression is that the observations are independent, but for a logistic regression with complex survey design this assumption is relaxed thus the model fitted based on logistic regression with complex survey design is better since it accounts for the complexity of the design.

Chapter 5

Ordinal logistic regression

5.1 Introduction

An outcome with more than two categories is known as a polytomous outcome. Let J denote the number of categories for such an outcome. Out of N observations, $Y_1, Y_2, ..., Y_J$ are the frequencies in category $1, 2, ..., J$ with corresponding probabilities, $\pi_1, \pi_2, ..., \pi_J$, respectively. The distribution is the multinomial distribution and can be expressed as follows:

$$P(Y_1, Y_2, ..., Y_J) = \frac{N!}{\prod_{j=1}^{J} y_j} \times \prod_{j=1}^{J} \pi_j^{y_j}$$

The distribution leads to the multinomial (polytomous) logistic regression which is an extension of binary logistic regression. The link function is the multinomial logit model because the probability distribution for the outcome variable is assumed to be a multinomial rather than a binomial distribution. For a polytomous response, it is further important to note whether the response is nominal (consisting of unordered categories) or ordinal (consisting of ordered categories). An outcome variable that has two or more nominal categories can be modeled using multinomial logistic regression. It estimates the odds of being at any category compared to being at the baseline category (comparison or reference category). The model can be treated as a combination of a series of binary logistic regression models.

Suppose Y can take on values coded as $1, 2, ..., J$. Next pick one of the outcome levels, say J, as the reference level. If we assume we have p covariates, then the model is formulated as

$$log\left[\frac{\pi(Y=y_j|x_1,x_2,...,x_J)}{\pi(Y=y_J|x_1,x_2,...,x_J)}\right]=\beta_{j0}+\beta_{j1}x_1+\beta_{j2}x_2+...+\beta_{jp}x_p,$$

where $y_j = 1,2,...,J-1; y_J$ is the base category, which can be any category but is generally the highest one; β_{j0} are the intercepts, and $\beta_{j1},\beta_{j2},...,\beta_{jp}$ are the regression coefficients. Since the model includes $J-1$ comparisons, it estimates $J-1$ logit functions for each predictor (Liu, 2014).

Multinomial logistic regression model uses maximum likelihood procedure to estimate regression coefficients, as it is the case with the binary logistic regression. For nominal categories, one of the categories is designated as a reference or base category and each of the rest of the categories is compared with the reference category (Agresti, 2002a).

Ordinal logistic regression considers any inherent ordering of the levels in the outcome variable and makes full use of the ordinal information (Kleinbaum & Klein, 2010). The incorporation of ordering can result in models that have simpler interpretations. Although ordinal outcomes can be simple and meaningful their optimal statistical treatment remains challenging to many applied researchers (Cliff & Keats, 2003; Clogg & Shihadeh, 1994; Ishii-Kuntz, 1994). Moreover, these models have greater power than the multinomial logit models (Allison, 2012). However, a variable that can be ordered when considered for one purpose could be unordered differently when used for another purpose. Miller & Volker (1985) shows how different assumptions about the ordering of occupations result in different conclusions. Therefore, we need to think carefully before concluding that the outcome is ordinal.

5.2 Ordinal logistic regression model

In the study of the dependence of a response variable on a set of independent variables, the choice of a model is largely determined by the scale of measurement of the response. Epidemiologists are often interested in estimating the risk of adverse events originally measured on an interval scale, but they often choose to divide the outcome into two or more categories in order to compute an estimate of effects (risk or odds ratio). Similarly, response variables originally measured on an ordinal scale (e.g. children's nutritional status (based on weight): underweight, normal/healthy weight, overweight, obese) are often categorized into several binary variables during statistical analysis (Ananth & Kleinbaum, 1997). Although the categories for an ordinal variable can be ordered, the distances between the categories are unknown. Multinomial logistic regression for ordinal responses is normally called ordinal logistic regression. An ordinal logistic regression model is a generalization of a binary logistic regression model, when the outcome variable has more than two ordinal levels. In Stata, the ordinal logistic regression model assumes that the outcome variable is a latent variable, which is expressed in logit form as follows:

$$log\left(\frac{P(Y \leq j|x)}{1-(Y \leq j|x)}\right) = \beta_{j0} + (-\beta_{j1}x_1 - \beta_{j2}x_2 - ... - \beta_{jp}x_p), \tag{5.1}$$

where $P(Y \leq y_j|x_1, x_2, ..., x_p)$ is the probability of being at or below category j, given a set of predictors $v = 1, 2, ..., p$. β_{j0} are the cutoff points (thresholds), and $\beta_{j1}, \beta_{j2}, ..., \beta_{jp}$ are logit coefficients (Liu, 2014).

Ordinal variables are often coded as consecutive integers from 1 to the number of categories. Because of this coding, it is tempting to analyze ordinal outcomes with the linear regression model. However, an ordinal response variable violates the assumptions of linear regression model, which can lead to incorrect conclusions (McKelvey & Zavoina, 1975; Winship & Mare, 1984). With an ordinal response, it is much better to use models that avoid the assumption that the distances between categories are equal. Although many models have been designed for ordinal outcomes, logit and probit models are commonly used as the link function in ordinal regression models (Long & Freese, 2006). Most multinomial regression models for ordinal outcome variables are based on the logit function. The difference between both functions is typically only seen in small samples, because the probit link assumes the normal distribution of the probability of event, whereas the logit link assumes the logistic distribution.

The BMI-for-age is an attempt to quantify the amount of tissue mass (muscle, fat, and bone) in children compared against the percentile for other children of the same sex and age. Based on the amount of tissue mass value, BMI-for-age is then categorized as underweight, normal weight, overweight, or obese. Table 2.1 presents the categorized level of under five children's BMI. This means the creation of a four-category ordinal variable from a continuous variable. This categorized level of BMI is an example of an ordinal categorical variable (Agresti, 2010). We label these four levels of under five children's nutrition status as 1, 2, 3, and 4 where we compare underweight, normal weight, overweight, and obese at the same time. Since this leads to an ordinal variable for nutrition status, an ordinal logistic regression (OLR) is an obvious choice for analysis.

There are many ways of generalizing the logit model to handle ordered categories, such as the partial proportional odds, continuation-ratio, adjacent-category logits, cumulative logits, and stereotype logistic models. Despite this diversity and the vast variety of studies on the subject their use in the public health area is still rare (Ananth & Kleinbaum, 1997; Anderson, 1984; Bender & Benner, 2000; Brant, 1990). This may be attributed not only to their complexity, but especially to the difficulty encountered when it comes to validating their assumptions (Lall et al., 2002). When the dependent variable has only two categories, the usual binary logistic model is appropriate.

5.2.1 Cumulative logits

The cumulative logit model is the most commonly used model for the analysis of ordinal categorical variables and it is widely implemented as the default for ordinal regression analysis in many statistical software packages, such as SAS, SPSS, Stata, S-Plus and R (Liu, 2015). It is used to estimate the cumulative probabilities of being at or below a particular category of the ordinal response variable, conditional on a set of predictor variables. The effects of the independent variables can be interpreted in several ways, including how they contribute to the cumulative odds and their probabilities of being at or beyond a particular category. They can also be interpreted as how these variables contribute to the odds of being at or below a particular category, if the sign is reversed before the estimated logit coefficients and corresponding cumulative odds are computed (Liu, 2009). This model can estimate the cumulative probabilities of being at or beyond a particular value of the ordinal response variable as well (the sign of the cut points needs to be reversed and their magnitude remain unchanged) because below and beyond a particular category are just two complementary directions (Liu, 2009).

Let p_{ij} be the probability that individual i falls into category j of the dependent variable. We assume that the categories are ordered in the sequence $j = 1, ..., J$. Cumulative probabilities are defined as $\pi_{ij} = \sum_{k=1}^{j} p_{ik}$, where π_{ij} is the probability that individual i is in the j^{th} category or lower.

According to Agresti (2002b), one way to use category ordering is to form logits of cumulative probabilities,

$$P(Y \leq j|x) = \frac{exp(\beta_{j0} + X'\beta)}{1 + exp(\beta_{j0} + X'\beta)} = \pi_1(x) + \pi_2(x) + ... + \pi_j(x), j = 1, 2, ..., J \quad (5.2)$$

Equivalently the cumulative logits (logits of cumulative probabilities) can be defined as

$$logit[P(Y \leq j|x)] = log\left[\frac{P(Y \leq j|x)}{P(Y > j|x)}\right] = log\left[\frac{P_1(x) + P_2(x) + ... + P_j(x)}{P_{j+1}(x) + P_{j+2}(x) + ... + P_J(x)}\right],$$

$j = 1, 2, ..., J - 1$. Each cumulative logit uses all J response categories.

In Stata, the logit form of the ordinal logistic regression model that simultaneously uses all cumulative logits can be expressed as follows:

$$logit[P(Y \leq j|x)] = \beta_{j0} + (-\boldsymbol{X'\beta}), \quad j = 1, 2, ..., J, \tag{5.3}$$

where $P(Y \leq j|x)$ is the cumulative probability of the event $(Y \leq j|x)$, β_{j0} are the unknown intercept parameters increasing in j, and $\beta = (\beta_1, \beta_2, ..., \beta_p)'$ is a vector of unknown regression coefficients corresponding to x. Since $P(Y \leq j|x)$ increases in j for fixed x, the logit is an increasing function of this probability.

The cumulative logit model (Equation 5.3) satisfies

$$logit[P(Y \leq j|x_1)] - logit[P(Y \leq j|x_2)] = log\left(\frac{P(Y \leq j|x_1)/P(Y > j|x_1)}{P(Y \leq j|x_2)/P(Y > j|x_2)}\right) = \boldsymbol{\beta}'(x_1 - x_2)$$

An odds ratio of cumulative probabilities is called a cumulative odds ratio. The odds of the event $Y \leq j$ at $x = x_1$ is $exp[\boldsymbol{\beta}'(x_1 - x_2)]$ times the odds of the same event at $x = x_2$. The log cumulative odds ratio is proportional to the distance between x_1 and x_2. The same proportionality constant applies to each logit. Because of this property, cumulative logit model is called the proportional odds model (McCullagh, 1980; Long, 1997; Agresti, 2002b).

Suppose we have J categories that are ordered in the sequence $j = 1, 2, ..., J$. In this logit model we have $J - 1$ cumulative logits.

$$Let \quad \theta_j = log\left(\frac{P_1 + P_2 + ... + P_j}{P_{j+1} + P_{j+2} + ... + P_J}\right), \quad j = 1, 2, ..., J - 1 \quad (cumulative \quad logits)$$

where J is number of classes, thus θ_j have $J - 1$ cumulative logits. Given observed data such that $P_j = \frac{x_j}{N}$, $j = 1, 2, ..., J$, then

$$\hat{\theta}_j = log\left[\frac{P(Y \leq j|x)}{P(Y > j|x)}\right] = log\left(\frac{P_1 + P_2 + ... + P_j}{P_{j+1} + P_{j+2} + ... + P_J}\right), \quad j = 1, 2, ..., J$$

Consider a relatively simple case of a variable with 3 categories ($J = 3$). Thus we have $3 - 1 = 2$ cumulative logits:

$$\hat{\theta}_1 = log\left[\frac{P(Y \leq 1)}{P(Y > 1)}\right] = log\left(\frac{P_1}{P_2 + P_3}\right)$$

$$\hat{\theta}_2 = log\left[\frac{P(Y \leq 2)}{P(Y > 2)}\right] = log\left(\frac{P_1 + P_2}{P_3}\right)$$

$$\underline{\hat{\theta}} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} = \begin{pmatrix} log\left(\frac{P_1}{P_2+P_3}\right) \\ log\left(\frac{P_1+P_2}{P_3}\right) \end{pmatrix} = \begin{pmatrix} log(P_1) - log(P_2 + P_3) \\ log(P_1 + P_2) - log(P_3) \end{pmatrix} = \begin{pmatrix} g_1(\underline{P}) \\ g_2(\underline{P}) \end{pmatrix}$$

To find $var(\hat{\theta}_j)$ we use the delta method for transformation of several variables. The probabilities $p = (p_1, p_2, p_3)'$ has mean $\pi = (\pi_1, \pi_2, \pi_3)'$ and covariance matrix $\sum_{(\pi)}$.

The Variance-Covariance matrix for the estimator $\underline{\hat{\theta}}$ is

$$\sum_{\hat{\theta}} = H' \sum_{(\pi)} H \qquad\qquad (Delta \quad method)$$

$\sum_{(\pi)}$, general form for $J = 3$:

$$\begin{pmatrix} var(P_1) & cov(P_1, P_2) & cov(P_1, P_3) \\ cov(P_2, P_1) & var(P_2) & cov(P_2, P_3) \\ cov(P_3, P_1) & cov(P_3, P_2) & var(P_3) \end{pmatrix} = \frac{1}{N} \begin{pmatrix} \pi_1(1 - \pi_1) & -\pi_1\pi_2 & -\pi_1\pi_3 \\ -\pi_2\pi_1 & \pi_2(1 - \pi_2) & -\pi_2\pi_3 \\ -\pi_3\pi_1 & -\pi_3\pi_2 & \pi_3(1 - \pi_3) \end{pmatrix}$$

H is the matrix of derivatives of $g(\underline{P})$ evaluated at $(\pi_1, \pi_2, \pi_3)'$, which is

$$H = \begin{pmatrix} \frac{\partial g_1(\underline{P})}{\partial P_1} & \frac{\partial g_2(\underline{P})}{\partial P_1} \\ \frac{\partial g_1(\underline{P})}{\partial P_2} & \frac{\partial g_2(\underline{P})}{\partial P_2} \\ \frac{\partial g_1(\underline{P})}{\partial P_3} & \frac{\partial g_3(\underline{P})}{\partial P_3} \end{pmatrix} = \begin{pmatrix} \frac{\partial(log(P_1) - log(P_2+P_3))}{\partial P_1} & \frac{\partial(log(P_1+P_2) - log(P_3))}{\partial P_1} \\ \frac{\partial(log(P_1) - log(P_2+P_3))}{\partial P_2} & \frac{\partial(log(P_1+P_2) - log(P_3))}{\partial P_2} \\ \frac{\partial(log(P_1) - log(P_2+P_3))}{\partial P_3} & \frac{\partial(log(P_1+P_2) - log(P_3))}{\partial P_3} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{P_1} & \frac{1}{P_1+P_2} \\ \frac{-1}{P_2+P_3} & \frac{1}{P_1+P_2} \\ \frac{-1}{P_2+P_3} & \frac{-1}{P_3} \end{pmatrix}$$

Therefore, $\sum_{\hat{\theta}}$ become

$$= \frac{1}{N} \begin{pmatrix} \frac{1}{\pi_1} & \frac{-1}{\pi_2+\pi_3} & \frac{-1}{\pi_2+\pi_3} \\ \frac{1}{\pi_1+\pi_2} & \frac{1}{\pi_1+\pi_2} & \frac{-1}{\pi_3} \end{pmatrix} \begin{pmatrix} \pi_1(1 - \pi_1) & -\pi_1\pi_2 & -\pi_1\pi_3 \\ -\pi_2\pi_1 & \pi_2(1 - \pi_2) & -\pi_2\pi_3 \\ -\pi_3\pi_1 & -\pi_3\pi_2 & \pi_3(1 - \pi_3) \end{pmatrix} \begin{pmatrix} \frac{1}{\pi_1} & \frac{1}{\pi_1+\pi_2} \\ \frac{-1}{\pi_2+\pi_3} & \frac{1}{\pi_1+\pi_2} \\ \frac{-1}{\pi_2+\pi_3} & \frac{-1}{\pi_3} \end{pmatrix}$$

$$= \frac{1}{N}\begin{pmatrix} 1 & \frac{-\pi_2}{\pi_2+\pi_3} & \frac{-\pi_3}{\pi_2+\pi_3} \\ \frac{\pi_1}{\pi_1+\pi_2} & \frac{\pi_2}{\pi_1+\pi_2} & -1 \end{pmatrix}\begin{pmatrix} \frac{1}{\pi_1} & \frac{1}{\pi_1+\pi_2} \\ \frac{-1}{\pi_2+\pi_3} & \frac{1}{\pi_1+\pi_2} \\ \frac{-1}{\pi_2+\pi_3} & \frac{-1}{\pi_3} \end{pmatrix}$$

$$= \frac{1}{N}\begin{pmatrix} \frac{1}{\pi_1}+\frac{1}{\pi_2+\pi_3} & \frac{1}{(\pi_1+\pi_2)(\pi_2+\pi_3)} \\ \frac{1}{(\pi_1+\pi_2)(\pi_2+\pi_3)} & \frac{1}{\pi_1+\pi_2}+\frac{1}{\pi_3} \end{pmatrix}, when \quad J=3, \pi_1+\pi_2+\pi_3=1$$

Thus,

$$var(\hat{\theta}_1) = \frac{1}{N}\left(\frac{1}{\pi_1}+\frac{1}{\pi_2+\pi_3}\right), \implies \widehat{var}(\hat{\theta}_1) = \frac{1}{N}\left(\frac{1}{P_1}+\frac{1}{P_2+P_3}\right)$$

$$var(\hat{\theta}_2) = \frac{1}{N}\left(\frac{1}{\pi_1+\pi_2}+\frac{1}{\pi_3}\right), \implies \widehat{var}(\hat{\theta}_2) = \frac{1}{N}\left(\frac{1}{P_1+P_2}+\frac{1}{P_3}\right)$$

$$cov(\hat{\theta}_1,\hat{\theta}_2) = \frac{1}{N(\pi_1+\pi_2)(\pi_2+\pi_3)}, \implies \widehat{cov}(\hat{\theta}_1,\hat{\theta}_2) = \frac{1}{N(P_1+P_2)(P_2+P_3)}$$

However, SAS uses ordinal logit model that is different from the one used by Stata. For SAS PROC LOGISTIC (the ascending option), the ordinal logit model has the following form:

$$logit\left[P(Y \le y_j|x)\right] = log\left[\frac{P(Y \le y_j|x)}{P(Y > y_j|x)}\right] = \beta_{j0} + \boldsymbol{X'\beta};$$

Using SAS with the descending option, the ordinal logit model can be expressed as:

$$logit\left[P(Y \ge y_j|x)\right] = log\left[\frac{P(Y \ge y_j|x)}{P(Y < y_j|x)}\right] = \beta_{j0} + \boldsymbol{X'\beta},$$

where in both equations β_{j0} are the intercepts, and $\boldsymbol{\beta}$'s are logit coefficients.

The β-coefficients are the ordered log-odds or logit regression coefficients. Besides its positive/negative relationship with the ordinal outcome, interpretation of the ordered logit coefficient is that for a one unit increase in the predictor, the ordinal outcome variable level is expected to change by its respective regression coefficient in the ordered log-odds scale, controlling for all other independent variables in the model. Interpretation of the ordered logit estimates is not dependent on the ancillary (cut points) parameters; the ancillary/thresholds parameters are used to define the changes among category levels of the ordinal response variable. When estimating the odds of being at or below an order response category j the $J-1$ cut points are used to differentiate the adjacent categories of an order response category. In this particular study, the response variable has four ordered categories. β_1 is the cut point for the cumulative logit model for $Y \le 1$ that is level 1 versus levels 2-4; β_2 is the cut point for the cumulative logit model for $Y \leq 2$ that is levels 1 and 2 versus levels 3 and 4; and the final β_3 is used as the cut point for the logit model when $Y \leq 3$ that is levels 1-3 versus level 4.

Maximum likelihood estimation for the cumulative logit models

Maximum likelihood (ML) estimation method is used for cumulative logit. Cumulative logit models assume independent multinomial observations. For subject i, let $y_{ij} = 1$ when the event happens and let $y_{ij} = 0$ otherwise, $i = 1, ..., n$. Then $E(y_{ij}) = \pi_j(x_i)$ the probability that observation i with explanatory variable values x_i falls in category j. For multicategory indicator $(y_{i1}, y_{i2}, ..., y_{iJ})$ of the response for subject i, the cumulative logit model, $logit[P(Y \leq j|x)] = \beta_{j0} + \boldsymbol{X'\beta}$, constrains the $J-1$ response curves that have the same shape. Thus, its fit is not the same as fitting separate logit models for each j. The multinomial likelihood function for the cumulative logit model is based on the product of the multinomial mass functions for the n subjects,

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{J}\pi_j(x_i)^{y_{ij}}\right] = \prod_{i=1}^{n}\left[\prod_{j=1}^{J}\left(P(Y \leq j|x_i) - P(Y \leq j-1|x_i)\right)^{y_{ij}}\right]$$

$$= \prod_{i=1}^{n}\left[\prod_{j=1}^{J}\left(\frac{exp(\beta_{j0} + \boldsymbol{\beta' X_i})}{1 + exp(\beta_{j0} + \boldsymbol{\beta' X_i})} - \frac{exp(\beta_{j-1,0} + \boldsymbol{\beta' X_i})}{1 + exp(\beta_{j-1,0} + \boldsymbol{\beta' X_i})}\right)^{y_{ij}}\right]$$

is viewed as a function of (β_{j0}, β), where $P(Y \leq 0) = 0$. The Log-likelihood function is

$$L(\beta_{j0}, \boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij} log\left[G(\beta_{j0} + \boldsymbol{X_i\beta} - G(\beta_{j-1,0} + \boldsymbol{X_i\beta}))\right],$$

where G denote the inverse link function for the cumulative link model.

The likelihood equations can be solved using Iterative methods (Fisher scoring algorithm or Newton-Raphson method) to obtain the ML estimates of the model parameters (see Agresti, 2010, section 5.1.2).

5.2.2 Continuation-ratio model

It is an alternative method to the proportional odds model for the analysis of categorical data with ordered responses (Fienberg, 1976). When the cumulative probabilities, $P(Y \leq j|x)$, of being in one of the first j categories in the cumulative logit model is replaced by the probability of being in category j that is $P(Y = j|x)$ conditional on being in categories greater than j that is $P(Y > j|x)$, this results in the continuation-ratio model. Therefore the continuation-ratio model can be defined as

$$log\left[\frac{P(Y=j|x)}{P(Y>j|x)}\right] = \beta_{j0} - \boldsymbol{X'\beta}, \qquad j = 1, 2, ..., J$$

Continuation-ratio logit models are useful when a sequential mechanism determines the response outcome, in the sense that an observation must potentially occur in lower category before it can occur in a higher category (Agresti, 2010; Hardin et al., 2007; Long & Freese, 2006). As defined in the cumulative logit model above, for J categories that are ordered in the sequence $j = 1, 2, ..., J$, there are $J-1$ continuation ratio logits. With a three-category outcome, $J = 3$, there are $3 - 1 = 2$ continuation-ratio logits.

$$let \quad \hat{\theta}_j = log\left[\frac{P(Y=j|x)}{P(Y>j|x)}\right] = log\left(\frac{P_j}{P_{j+1}+P_{j+2}+...+P_J}\right), \qquad j = 1, 2, ..., J$$

with $J = 3$ categories, the two set of sequential continuation ratio logits become:

$$\hat{\theta}_1 = log\left[\frac{P(y=1)}{P(y>1)}\right] = log\left(\frac{P_1}{P_2+P_3}\right)$$

$$\hat{\theta}_2 = log\left[\frac{P(y=2)}{P(y>2)}\right] = log\left(\frac{P_2}{P_3}\right)$$

$$\underline{\hat{\theta}} = \begin{pmatrix}\hat{\theta}_1\\ \hat{\theta}_2\end{pmatrix} = \begin{pmatrix} log\left(\frac{P_1}{P_2+P_3}\right)\\ log\left(\frac{P_2}{P_3}\right)\end{pmatrix} = \begin{pmatrix} log(P_1) - log(P_2+P_3)\\ log(P_2) - log(P_3)\end{pmatrix} = \begin{pmatrix} g_1(\underline{P})\\ g_2(\underline{P})\end{pmatrix}$$

$$\sum_{\hat{\theta}} = H' \sum_{(\pi)} H \qquad\qquad (Delta \quad method)$$

Therefore, $\sum_{\hat{\theta}}$ become

$$= \frac{1}{N}\begin{pmatrix} \frac{1}{\pi_1} & \frac{-1}{\pi_2+\pi_3} & \frac{-1}{\pi_2+\pi_3}\\ 0 & \frac{1}{\pi_2} & \frac{-1}{\pi_3}\end{pmatrix} \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & -\pi_1\pi_3\\ -\pi_2\pi_1 & \pi_2(1-\pi_2) & -\pi_2\pi_3\\ -\pi_3\pi_1 & -\pi_3\pi_2 & \pi_3(1-\pi_3)\end{pmatrix} \begin{pmatrix} \frac{1}{\pi_1} & 0\\ \frac{-1}{\pi_2+\pi_3} & \frac{1}{\pi_2}\\ \frac{-1}{\pi_2+\pi_3} & \frac{-1}{\pi_3}\end{pmatrix}$$

$$= \frac{1}{N}\begin{pmatrix} 1 & \frac{-\pi_2}{\pi_2+\pi_3} & \frac{-\pi_3}{\pi_2+\pi_3}\\ 0 & 1 & -1\end{pmatrix} \begin{pmatrix} \frac{1}{\pi_1} & 0\\ \frac{-1}{\pi_2+\pi_3} & \frac{1}{\pi_2}\\ \frac{-1}{\pi_2+\pi_3} & \frac{-1}{\pi_3}\end{pmatrix}$$

$$= \frac{1}{N}\begin{pmatrix} \frac{1}{\pi_1} + \frac{1}{\pi_2+\pi_3} & 0\\ 0 & \frac{1}{\pi_2} + \frac{1}{\pi_3}\end{pmatrix}$$

Thus,

$$var(\hat{\theta}_1) = \frac{1}{N}\left(\frac{1}{\pi_1} + \frac{1}{\pi_2 + \pi_3}\right), \implies \widehat{var}(\hat{\theta}_1) = \frac{1}{N}\left(\frac{1}{P_1} + \frac{1}{P_2 + P_3}\right)$$
$$var(\hat{\theta}_2) = \frac{1}{N}\left(\frac{1}{\pi_2} + \frac{1}{\pi_3}\right), \implies \widehat{var}(\hat{\theta}_2) = \frac{1}{N}\left(\frac{1}{P_2} + \frac{1}{P_3}\right)$$
$$cov(\hat{\theta}_1, \hat{\theta}_2) = 0$$

In general for continuation-ratio logit, $\widehat{cov}(\hat{\theta}_j, \hat{\theta}_J) = 0$.

According to Agresti (2010), an alternative set of continuation-ratio logits, appropriate if the sequential mechanism works in the reverse direction, is

$$\hat{\theta}_j = log\left[\frac{P(Y = j+1)}{P(Y < j+1)}\right] = log\left(\frac{P_{j+1}}{P_1 + P_2 + ... + P_j}\right), \qquad j = 1, 2, ..., J-1$$

Two sets of sequential continuation-ratio logits can be obtained for the above three-category outcome example. These are as follows.

$$\hat{\theta}_1 = log\left[\frac{P(y = 2)}{P(y < 2)}\right] = log\left(\frac{P_2}{P_1}\right)$$

$$\hat{\theta}_2 = log\left[\frac{P(y = 3)}{P(y < 3)}\right] = log\left(\frac{P_3}{P_1 + P_2}\right)$$

This indicates that the two forms of continuation-ratio logits are not equivalent.

5.2.3 Adjacent-category logits

The adjacent-category logit model involves modeling the ratio of the two probabilities, $P(Y = j|x)$ and $P(Y = j+1|x)$ that is, this model considers ratios of probabilities for successive categories $\left(\frac{\pi_1}{\pi_2}, \frac{\pi_2}{\pi_3}, ..., \frac{\pi_{J-1}}{\pi_J}\right)$, where $j = 1, 2, ..., J$. Adjacent-category logit models can be represented as

$$log\left[\frac{P(Y = j|x)}{P(Y = j+1|x)}\right] = \beta_{j0} - \boldsymbol{X'\beta}, \qquad j = 1, 2, ...J$$

For detailed discussions of logit models that are appropriate to handle ordered categories one can refer to Agresti (2010); O'Connell (2006); and Liu (2009, 2014).

5.3 Ordinal logistic regression with complex survey design

The usual proportional odds model assumes that data are collected using simple random sampling by which each sampling unit has an equal probability of being selected from a population. When the data comes from a complex survey design with the use of different strata, clustered sampling techniques, and unequal selection probabilities, it is inappropriate to conduct the proportional odds model analysis for the ordinal response variable without taking the survey sampling design into account. Ignoring these features in data analysis may lead to biased estimates of parameters, incorrect variance estimates and misleading results. The parameters and their variance may be either overestimated or underestimated (Liu, 2015). In such cases, a specialized technique to produce the appropriate estimates and standard errors for ordinal outcome variable should be used. This method takes into account the weight in the survey sampling design.

Features of complex surveys such as sampling weights, strata, and clusters, have been illustrated in literature (Sharon, 1999; Liu, 2015). In Stata, *svy* prefix command for survey data is used to fit the proportional odds model when taking all the elements of survey design features into account. It is necessary to specify strata, cluster and weights before fitting the model. For more details on how to use this command one can use the help *svyset* command in stata software.

5.3.1 Variance estimation in survey ordinal logistic regression

For unbiased variance estimation in complex sampling survey designs that include designs with stratification, clustering, and unequal weighting, the procedure uses the Taylor series (linearization) method or replication (resampling) methods (Binder, 1983; Lee & Forthofer, 2006; Levy & Lemeshow, 2013; Sharon, 1999). The replicated methods estimate variance of a parameter by generating multiple replicates (also called subsamples) from a full sample according to a specific resampling scheme and examining the variability of the subsample estimates. Balanced repeated replication (BRR), jackknife repeated replication (JRR), and bootstrap method are the most commonly used resampling schemes (Lee & Forthofer, 2006; Levy & Lemeshow, 2013).

The Taylor series (Linearization) approximation, also known as the delta method (Kalton, 1983) is the most commonly used method to estimate the covariance-matrix of the regression coefficients for complex survey data. In statistics, the Taylor series linearization is used to obtain a linear approximation to the nonlinear function or statistic and then the variance of the function (Liu, 2015). It is the default variance estimation method used in general purpose software packages, such as Stata, SAS. The Taylor series expansion of the function, f(x) at a point ″a″ is generally expressed as:

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)(x-a)^2}{2!} + \frac{f'''(a)(x-a)^3}{3!} + ...$$

where f', f'' and f''' are the first, second and third derivatives of the function and so on.

The technical details of variance estimation can be found in many literatures (Binder, 1983; Heeringa et al., 2010; Lee & Forthofer, 2006; Levy & Lemeshow, 2013).

5.4 Application of ordinal logistic regression model

In the subsequent section, the proposed model, namely the proportional odds model, was applied to the 2016 DHS data. The response variable used is the ordinal outcome of nutritional status of under five children (see Table 2.1). The same set of explanatory variables used in the previous chapters was used. In addition to the response and explanatory variables, we also assessed two-way interaction effects. The interaction effects were not found significant. Stata *ologit* command was used for model fitting.

Table 5.1 shows the results for the proportional odds model under the simple random sampling assumption. The log likelihood at each iteration shows that ordinal logistic regression, like binary and multinomial logistic regression, uses maximum likelihood estimation, which is an iterative procedure. Iteration 0 is the log likelihood of the ″null″ or ″empty″ model; that is, a model with no predictors. At the next iteration, the predictors are included in the model. At each iteration, the log likelihood increases because the goal is to maximize the log likelihood. When the difference between successive iterations is very small, the model is said to have ″converged″, and the iteration stops (Long, 1997).

The likelihood ratio chi-square $(LR\chi^2)$ tests that at least one of the predictors' regression coefficient is not equal to zero. The number in the parenthesis indicates the degrees of freedom of the chi-square distribution used to test the the null hypothesis using the $LR\chi^2$ statistic and is defined by the number of predictors in the model. The $LR\chi^2$ statistic can be calculated by $-2(L$(null model)-L(fitted model))= -2((-6362.3952)-(-6009.1723))= 708.45, where L(null model) is from the log likelihood of the model with no predictor variable (Iteration 0) and L(fitted model) is the log likelihood from the final iteration (assuming the model converged) with all the parameters.

The P-value of the log likelihood ratio chi-square test with 28 degree of freedom, $LR\chi^2(28) = 708.45$, $Prob > \chi^2 = 0.000$, which indicates that at least one of the logit regression coefficient of the predictors was statistically different from 0, so the full model with all predictors provided a better fit than the null model with no independent variables in predicting cumulative probabilities for under five children's nutrition status (Table 5.1).

Table 5.1 reports three cut-points; $cut1, cut2$, and $cut3$. $cut1$ is the estimated cut point on the latent variable used to differentiate underweight status from normal, overweight, and obese status when values of the predictor variables are evaluated at zero. When the ordinal outcome category is 1 given significant predictor variables (for categorical variables the reference variable evaluated at zero) and had zero value for all other predictor variables, the latent variable falls at or below the first cut point, -2.7741. $cut2$ is the estimated cut point on the latent variable used to differentiate underweight and normal weight status from overweight and obese weight status when values of the predictor variables are evaluated at zero. When the ordinal outcome category is 2 given significant predictor variable and controlling for all other predictor variables in the model, the latent variable falls between the first cut point, -2.7741 and the second cut point, 2.2885. $cut3$ is the estimated cut point on the latent variable used to differentiate underweight, normal, and overweight status from obese status when values of the predictor variables are evaluated at zero. When the ordinal outcome category is 3 given significant predictor variable (reference variable evaluated at zero) and controlling for all other predictor variables in the model, the latent variable falls between $cut2, 2.2885$ and $cut3, 3.5703$ and is classified as overweight. When the ordinal outcome category reaches 4, if the latent variable had a value at or beyond the third cut point, 3.5703, controlling for all other predictor variables in the model would be classified as child with obese nutrition status.

Table 5.1 shows the effect of socio-economic, demographic and geographic factors that have influence on under five children's nutritional status. The cummulative logit model used for the analysis. The estimated logit regression coefficients of current age of child is β=-0.3233 (P-value=0.000). This is the ordered log-odds estimate for a one unit increase in age of a child on the expected nutritional status level given the other variables are held constant in the model. The estimated coefficients of female child is (β=-0.2681, P-value=0.000). The estimated coefficients of weight of child at birth are: large (β=0.5547, P-value=0.000), average (β=0.3776, P-value=0.000). For mother's BMI, (β=0.0604, P-value=0.000), which is the ordered log-odds estimate for one-unit increase in mother's BMI keeping other variables constant. The estimated coefficients for regions are: Affar, β=-0.3951 (P-value=0.001), Dire Dawa, β=-0.6004(P-value=0.000), Gambela, β=-0.4453 (P-value=0.002), Harari, β=-0.3353(P-value=0.016), SNNP, β=0.2134 (P-value=0.047), Somali, β=-0.8988 (P-value=0.000) were found to be significant determinants of under five children's nutritional status.

Substituting the values of the estimated logit coefficients into the Equation (5.3) resulted in $logit[\pi(Y \leq j|x)] = \beta_{j0} + (-\beta_{jp}x)$. By exponentiating the negative logit coefficients ($e^{(-\beta)}$) the odds of being at or below a particular ordinal nutritional status category that is underweight versus being above that category (normal, overweight, and obese) were obtained. Therefore, to estimate the cumulative odds of being at or below a particular under five ordinal nutritional status variable (based on weight) category j for the first predictor, current age of child, the logit form of proportional odds model was used, $logit[\pi(Y \leq j|x_1)] = \beta_{j0} - (-0.3233(age))$. OR= $e^{(0.3233)}$=1.3817, indicating that the odds of being at or below a particular under five ordinal nutritional status variable (based on weight) increased by 38.17% with a one unit increase in the value of current age of a child, holding other variables constant. The estimated cumulative odds of being at or below an ordinal nutritional status (based on weight) category j, for a female child, we calculated $logit[\pi(Y \leq j|x_1)] = \beta_{j0} + (-0.2681(female))$. OR= $e^{(0.2681)}$=1.3075, suggesting that the odds of a female child being at or below a particular under five ordinal nutritional status (based on weight) increased by 30.75%. The estimated cumulative odds of being at or below an ordinal nutritional status (based on weight) category j, for a child who had large weight at birth, we calculated $logit[\pi(Y \leq j|x_1)] = \beta_{j0} + (0.5547(large))$. OR= $e^{(-0.5547)}$=0.5743, suggesting that for a child who had large weight at birth, the odds of being at or below a particular under five ordinal nutritional status variable (based on weight) decreased by $(1 - 0.5743) \times 100\% = 42.57\%$ as compared to small weight of child at birth, controlling for all other independent variables in the model. The estimated cumulative odds of being at or below an ordinal nutritional status (based on weight) category j, for a child who had average weight at birth, we calculated $logit[\pi(Y \leq j|x_1)] = \beta_{j0} + (0.3776(average))$. OR= $e^{(-0.3776)}$=0.6856, suggesting that for a child who had average weight at birth, the odds of being at or below a particular under five ordinal nutritional status variable (based on weight) decreased by $(1 - 0.6856) \times 100\% = 31.44\%$ as compared to small weight of child at birth, controlling for all other independent variables in the model. The odds of being at or below a particular under five ordinal nutritional status for the other significant effects were computed in the same way as above. It was found that for a one-unit increase in the value of mother's BMI, holding other variables constant, the odds of being at or below a particular under five ordinal nutritional status decreased by $(1 - 0.9414) \times 100\% = 5.86\%$ (OR=0.9414). The odds of being at or below a particular under five ordinal nutritional status variable for children from Affar region was 1.4845 (P-value= 0.001) times the odds of children from Oromia region. The odds of being at or below a particular under five ordinal nutritional status variable for children from Dire Dawa region was 1.8228 (P-value= 0.000) times the odds of children from Oromia region. The odds of being at or below a particular ordinal nutritional status category for children from Gambela, Harari and Somali regions were respectively 1.5609 (P-value=0.002), 1.3984 (P-value=0.016), and 2.4567 (P-value=0.000) times the odds of children from Oromia region. However, the odds of being at or below a particular ordinal nutritional status category for children from SNNP was 0.8078 (P-value=0.047) times the odds for children from Oromia region (see Table 5.1).

The odds of being beyond a particular category of ordinal nutritional status are the inverse of those of being at or below a category (Liu & Koirala, 2013). Equation (5.3) can be transformed to $logit[\pi(Y > j|x)] = -\beta_{j0}+\beta_{jp}x$. Odds ratios (Table 5.1) can be used directly for the analysis. In terms of odds ratio (OR), it was found that the odds of being beyond a particular category of ordinal nutritional status was decreased by $(1 - 0.7237) \times 100\% = 27.63\%$ (P-value=0.000) with a one-year increase in current age of child, holding other variables constant. Similarly, the odds of being beyond a particular under five ordinal nutritional status for a female child was 0.7647 times the odds of a male child. The odds of being beyond a particular category of ordinal nutritional status for children who had large weight at birth was 1.7414 times the odds of children who had small weight at birth. The odds of being at or beyond a particular category of ordinal nutritional status for other significant effects can be interpreted in the same way as above.

Table 5.1: Parameter estimates using PO model assuming the observations are independent.

Parameters	**Coeff.(β)**	**St.error**	**OR**	**P-value**
cut1	-2.7741	0.2377		
cut2	2.2885	0.2351		
cut3	3.5703	0.2387		
Current age of child	-0.3233	0.0199	0.7237	0.000
Mother's age	-0.0083	0.0045	0.9916	0.066
Mother's BMI	0.0604	0.0082	1.0622	0.000
Sex of child (ref. = Male)				
Female	-0.2681	0.0537	0.7647	0.000
Weight of child at birth (ref. = Small)				
Large	0.5547	0.0739	1.7414	0.000
Average	0.3776	0.0686	1.4588	0.000
Mother's work status (ref. = No)				
Yes	0.0415	0.0624	1.0424	0.506
Educational level (ref. = No education)				
Primary school	-0.0071	0.0694	0.9928	0.918
Secondary school	0.1200	0.1167	1.1275	0.304
Higher	0.2679	0.1492	1.3072	0.073
Marital status (ref. = Married)				
Not married	-0.1404	0.1120	0.8689	0.210
Religion (ref. = Orthodox)				
Catholic	-0.1449	0.3444	0.8651	0.674
Muslim	0.0377	0.1052	1.0384	0.720
Protestant	0.0083	0.0909	1.0084	0.926
Other	0.0359	0.2206	1.0366	0.870
Region (ref. = Oromia)				
Addis Abeba	0.2281	0.1550	1.2562	0.141
Affar	-0.3951	0.1221	0.6736	0.001
Amhara	-0.1353	0.1235	0.8734	0.273
Benishangul	-0.2240	0.1179	0.7992	0.057
Dire Dawa	-0.6004	0.1507	0.5485	0.000
Gambela	-0.4453	0.1428	0.6406	0.002
Harari	-0.3353	0.1396	0.7150	0.016
SNNP	0.2134	0.1074	1.2379	0.047
Somali	-0.8988	0.1121	0.4070	0.000
Tigray	-0.2392	0.1280	0.7871	0.062
Place of residence (ref. = Rural)				
Urban	0.0801	0.0982	1.0833	0.415
Wealth index (ref. = Poor)				
Middle	-0.0356	0.0833	0.9649	0.669
Rich	0.1277	0.0778	1.1362	0.101

Iteration 0:log likelihood=-6362.3952, LR χ^2 (28)=708.45, Iteration 4:log likelihood=-6009.1723, Prob $> \chi^2$=0.0000

5.5 Application of ordinal logistic regression model with complex survey design

In the subsequent section, the same variables and two-way interaction effects from the previous section are used for data analysis with reference to the Ethiopian DHS data, 2016. Here we investigate the relationship (association) between the response variable, and the explanatory variables by the method of proportional odds (PO) model with complex survey design using the stata *svy: ologit* prefix command. Stata's survey data *svy* prefix command is used to fit the PO model when taking all the elements of survey design features, such as strata, cluster, and weight variables into account (Liu & Koirala, 2013).

The result of the *svy: ologit* is indicated in Table 5.2 below. The *svy: ologit* for PO model that considered sampling design, reports the adjusted Wald test for all parameters rather than the log likelihood ratio chi-square test for the ordinal PO model (Liu, 2015). $F(28, 588) = 13.01, Prob > F = 0.0000$, indicates that the full model with all parameters was significant in fitting the PO model with complex survey design. The logit coefficients and odds ratios in the PO model with complex survey design can be interpreted in the same way as those in the standard PO model.

The three cut off points, when estimating the odds of being at or below a particular ordinal nutritional status category (based on weight), are used to differentiate the adjacent categories of the response variable (ordinal nutritional status). $\alpha_1 = -3.2418$, which is the first cut point for the cumulative logit model for $Y \leq 1$ that is level 1 versus levels 2-4; $\alpha_2 = 1.7371$ is the cut point for the cumulative logit model for $Y \leq 2$ that is levels 1 and 2 versus 3 and 4; $\alpha_3 = 2.9871$ is used as the cut point for the cumulative logit model when $Y \leq 3$ that is levels 1-3 versus level 4.

The results (Table 5.2) revealed that the estimated logit coefficients of current age of child, female children, large and average weight of a child at birth, mother's current age, mother's BMI, mothers who are not married and Affar, Dire Dawa, Gambela, Harari and Somali regions were significant. Therefore, for the predictor, current age of child (β=-0.3186, OR=0.7271) indicates that the odds of being at or beyond a particular ordinal nutritional status category decreased by $(1-0.7271)\times 100\% = 27.29\%$ with a one year increase in current age of child, holding other variables constant; female child (β=-0.2417, OR=0.7852) suggesting that the odds of a female child being at or beyond a particular under five ordinal nutritional status (based on weight) decreased by $(1-0.7852)\times 100\% = 21.48\%$. The odds of being at or beyond a particular ordinal nutritional status for weight of child at birth: large ($\beta = 0.5481$), and average ($\beta = 0.3134$) were 1.7301, and 1.3680 respectively times the odds of small weight of a child at birth; for the predictor mother's age (β=-0.0203, OR=0.9798) indicates that the odds of being at or beyond a particular ordinal nutritional status category decreased by $(1 - 0.9798) \times 100\% = 2.02\%$ with a one year increase in mother's age; for the predictor mother's BMI (β=0.0479, OR=1.0491) indicates that a one-unit increase mother's BMI, holding other variable constant, the odds of being at or beyond a particular under five ordinal nutritional status increased by 4.91% . It was found that the odds of being at or beyond a particular ordinal nutritional status for children born to unmarried mother was 0.7101 (β=-0.3423) times the odds for children born to married mother. The odds of being at or above a particular ordinal nutritional status for children from Affar, Dire Dawa, Gambela, Harari and Somali regions were respectively 0.6411 (β=-0.4445), 0.5554 (β=-0.5879), 0.5422 (β=-0.6120), 0.7243 (β=-0.3224) and 0.4007 (β=-0.9143) times the odds for children from Oromia region (see Table 5.2).

Table 5.2: Parameter estimates using PO model with complex survey design

Parameters	**Coeff.(β)**	**St.error**	**OR**	**P-value**	**[95%C.I for β]**
cut1	-3.2418	0.3566			(-3.9423, -2.5414)
cut2	1.7371	0.3504			(1.0489, 2.4252)
cut3	2.9871	0.3628			(2.2745, 3.6998)
Current age of child	-0.3186	0.0274	0.7271	0.004	(-0.3725, -0.2648)
Mother's age	-0.0203	0.0064	0.9798	0.002	(-0.0331, -0.0075)
Mother's BMI	0.0479	0.0136	1.0491	0.000	(0.0212, 0.0747)
Sex of child (ref. = Male)					
Female	-0.2417	0.0834	0.7852	0.000	(-0.4056, -0.0779)
Weight of child at birth (ref. = Small)					
Large	0.5481	0.1221	1.7301	0.000	(0.3082, 0.7880)
Average	0.3134	0.1004	1.3680	0.002	(0.1161, 0.5106)
Mother's work status (ref. = No)					
Yes	-0.0331	0.0963	0.9673	0.731	(-0.2224, 0.1561)
Educational level (ref. = No education)					
Primary school	-0.0650	0.0966	0.9370	0.501	(-0.2549, 0.1248)
Secondary school	0.0919	0.1645	1.0963	0.577	(-0.2312, 0.4151)
Higher	0.0701	0.2332	1.0726	0.764	(-0.3879, 0.5282)
Marital status (ref. = Married)					
Not married	-0.3423	0.1519	0.7101	0.025	(-0.6406, -0.0440)
Religion (ref. = Orthodox)					
Catholic	-0.2474	0.3597	0.7808	0.492	(-0.9538, 0.4590)
Muslim	0.2161	0.1384	1.2412	0.119	(-0.0557, 0.4879)
Protestant	0.0635	0.1179	1.0655	0.591	(-0.1681, 0.2951)
Other	0.6863	0.6262	1.9864	0.274	(-0.5434, 1.9160)
Region (ref. = Oromia)					
Addis Abeba	0.2395	0.2165	1.2706	0.269	(-0.1858, 0.6648)
Affar	-0.4445	0.1481	0.6411	0.003	(-0.7355, -0.1535)
Amhara	-0.0663	0.1339	0.9357	0.621	(-0.3295, 0.1967)
Benishangul	-0.2702	0.1582	0.7632	0.088	(-0.5809, 0.0404)
Dire Dawa	-0.5879	0.1522	0.5554	0.000	(-0.8868, -0.2889)
Gambela	-0.6120	0.1845	0.5422	0.001	(-0.9745, -0.2496)
Harari	-0.3224	0.1557	0.7243	0.039	(-0.6282, -0.0167)
SNNP	0.1665	0.1381	1.1811	0.228	(-0.1047, 0.4378)
Somali	-0.9143	0.1651	0.4007	0.000	(-1.2387, -0.5899)
Tigray	-0.1783	0.1445	0.8367	0.218	(-0.4622, 0.1056)
Place of residence (ref. = Rural)					
Urban	0.2044	0.2022	1.2268	0.312	(-0.1927, 0.6016)
Wealth index (ref. = Poor)					
Middle	-0.0170	0.1101	0.9830	0.877	(-0.2332, 0.1991)
Rich	0.1908	0.1152	1.2102	0.098	(-0.0355, 0.4171)

F(28,588) = 13.01 , Prob $>$ F = 0.0000

5.6 Comparison of results

Table 5.3 provides the results of the two models, the fitted classical PO model and thereafter PO model with complex sampling design. After complex sampling design was applied to the PO model, the estimated logit coefficients and their standard errors were different from those in the PO model under the simple random sampling assumption. The logit coefficient of the predictors current age of child, female child, Dire Dawa and Harari regions were increased and those of the other significant predictors (large and average weight of child at birth, mother's age, mother's BMI, not married mothers and Affar, Gambela and Somali regions) were decreased.

Compared to the PO model with simple random sampling, the estimated logit coefficient for current age of child in the PO model with complex survey design increased by 1.48%, and its standard error increased by 37.7%; the logit coefficient for female child increased by 10.92%, and its standard error increased by 55.31%; the logit coefficient for Dire Dawa and Harari regions were respectively increased by 2.12% and 4%, with their standard error increased by 0.99% and 11.53%; the logit coefficient for large and average weight of child at birth were respectively decreased by 1.2% and 20.48%, with their standard error increased by 65.2% and 46.35%; the logit coefficient for mother's age, mother's BMI and not married mothers were respectively decreased by 40.9%, 79.3% and 41.01%, with standard error increased by 42.2%, 65.8% and 35.6%; and the logit coefficient for Affar, Gambela and Somali regions were respectively decreased by 11.1%, 27.2% and 1.7%, with their standard error increased by 21.3%, 29.2% and 47.27%.

Further, the standard errors of the significant coefficients in the PO model with complex sampling design were higher as compared to the corresponding standard errors of the significant coefficients in the conventional PO model indicating that standard errors were underestimated when we considered the conventional PO model (Liu & Koirala, 2013; Habyarimana et al., 2014). This is an important distinguishing feature between the models. Analyses ignoring the complex sampling design will lead to a false increased precision and should be avoided.

Models with smaller values of an information criterion are used to considered preferable. However, the number of parameters estimated (k) in the complex model is higher than that of the conventional model. Since we considered the complex nature of the design (weight, cluster and strata) in the PO model with complex survey design, AIC and BIC of the complex model are higher.

Table 5.3: Comparison of results obtained from PO models without and with complex survey design

	PO without CSD				PO with CSD			
Parameters	**Coeff.(β)**	**SE**	**OR**	**P-value**	**Coeff.(β)**	**SE**	**OR**	**P-value**
cut1	-2.7741	0.2377			-3.2418	0.3566		
cut2	2.2885	0.2351			1.7371	0.3504		
cut3	3.5703	0.2387			2.9871	0.3628		
Current age of child	-0.3233	0.0199	0.7237	0.000	-0.3186	0.0274	0.7271	0.004
Mother's age	-0.0083	0.0045	0.9916	0.066	-0.0203	0.0064	0.9798	0.002
Mother's BMI	0.0604	0.0082	1.0622	0.000	0.0479	0.0136	1.0491	0.000
Sex of child (ref. = Male)								
Female	-0.2681	0.0537	0.7647	0.000	-0.2417	0.0834	0.7852	0.000
Weight of child at birth (ref. = Small)								
Large	0.5547	0.0739	1.7414	0.000	0.5481	0.1221	1.7301	0.000
Average	0.3776	0.0686	1.4588	0.000	0.3134	0.1004	1.3680	0.002
Mother's work status (ref. = No)								
Yes	0.0415	0.0624	1.0424	0.506	-0.0331	0.0963	0.9673	0.731
Educational level (ref. = No education)								
Primary school	-0.0071	0.0694	0.9928	0.918	-0.0650	0.0966	0.9370	0.501
Secondary school	0.1200	0.1167	1.1275	0.304	0.0919	0.1645	1.0963	0.577
Higher	0.2679	0.1492	1.3072	0.073	0.0701	0.2332	1.0726	0.764
Marital status (ref. = Married)								
Not married	-0.1404	0.1120	0.8689	0.210	-0.3423	0.1519	0.7101	0.025
Religion (ref. = Orthodox)								
Catholic	-0.1449	0.3444	0.8651	0.674	-0.2474	0.3597	0.7808	0.492
Muslim	0.0377	0.1052	1.0384	0.720	0.2161	0.1384	1.2412	0.119
Protestant	0.0083	0.0909	1.0084	0.926	0.0635	0.1179	1.0655	0.591
Other	0.0359	0.2206	1.0366	0.870	0.6863	0.6262	1.9864	0.274
Region (ref. = Oromia)								
Addis Abeba	0.2281	0.1550	1.2562	0.141	0.2395	0.2165	1.2706	0.269
Affar	-0.3951	0.1221	0.6736	0.001	-0.4445	0.1481	0.6411	0.003
Amhara	-0.1353	0.1235	0.8734	0.273	-0.0663	0.1339	0.9357	0.621
Benishangul	-0.2240	0.1179	0.7992	0.057	-0.2702	0.1582	0.7632	0.088
Dire Dawa	-0.6004	0.1507	0.5485	0.000	-0.5879	0.1522	0.5554	0.000
Gambela	-0.4453	0.1428	0.6406	0.002	-0.6120	0.1845	0.5422	0.001
Harari	-0.3353	0.1396	0.7150	0.016	-0.3224	0.1557	0.7243	0.039
SNNP	0.2134	0.1074	1.2379	0.047	0.1665	0.1381	1.1811	0.228
Somali	-0.8988	0.1121	0.4070	0.000	-0.9143	0.1651	0.4007	0.000
Tigray	-0.2392	0.1280	0.7871	0.062	-0.1783	0.1445	0.8367	0.218
Place of residence (ref. = Rural)								
Urban	0.0801	0.0982	1.0833	0.415	0.2044	0.2022	1.2268	0.312
Wealth index (ref. = Poor)								
Middle	-0.0356	0.0833	0.9649	0.669	-0.0170	0.1101	0.9830	0.877
Rich	0.1277	0.0778	1.1362	0.101	0.1908	0.1152	1.2102	0.098

PO $\Longrightarrow$ Proportional odds model, and CSD $\Longrightarrow$ complex survey design

The AIC and the BIC are two popular measures for comparing maximum likelihood models. AIC and BIC are defined as

$$AIC = -2 \times ln(likelihood) + 2 \times K$$

$$BIC = -2 \times ln(likelihood) + ln(n) \times K$$

where k= number of parameters estimated and n= number of observations.

AIC and BIC can be viewed as measures that combine fit and complexity. The fit is measured negatively by $-2 \times ln(likelihood)$. Complexity is measured positively, either by $2 \times k(AIC)$ or $ln(n) \times k(BIC)$ (see Akaike (1974) and Schwarz et al. (1978)).

Because of the diversity of the models that are available, it is becoming inappropriate to apply just a single criterion such as AIC. There are models that estimate the number of parameters; in some cases, the number of parameters exceeds the number of independent variables. Criteria such as generalized cross-validation (GCV) are used to compare the model quality. However, in practical use, we sometimes need to select models by understanding the character of the models from two points of view. First, how much the model fits the observation. Secondly, how stable the model can make the estimation (Asami, 2016). For standardizing the balance between these two points of view, Asami (2016) suggests a simple way for evaluating the different types of regression models from two points of view: the 'data fitting' and the 'model stability'. Therefore, methods that take into account the complex nature of the design, perform better than those that do not take this into account.

A complete review of types of methods for comparing regression models are beyond the scope of this research, but many of these measures are reviewed in (Asami, 2016; Azen & Budescu, 2006).

Chapter 6

Discussion and Conclusion

The BMI, malnutrition and nutritional status of under five children were studied using different statistical models with reference to the 2016 Ethiopian DHS. The 2016 EDHS sample was designed to provide estimates for the health and demographic variables of interest for Ethiopia as a whole; urban and rural area of Ethiopia and 11 geographical areas. Central Statistical Agency of Ethiopia was the responsible organization for the survey. The 2007 Population and Housing Census results were used as the sampling frame. A nationally representative sample of under five children was used to get information on weight and height measures of children under the age of five years. Overall, for under five children, the number of male and female children were almost equal. The majority of the respondents were residing in rural areas and they were married. More than half of the children were from poor economic class families. Most of the respondents were from Oromia, Somali, SNNP, Tigray and Amhara regions. Many of the respondents were Orthodox Christian and Muslim.

The present work was conducted to determine factors associated with the BMI, malnutrition and nutritional status of under five children in Ethiopia. At the beginning models with two-way interaction effects were considered. The backward elimination technique was used to eliminate non-significant factors for each model considered. The main purpose of the study was to determine factors affecting nutritional status of under five children. This may assist policy makers in their effort to make decision. It will help to prevent children's deaths and to improve children's health, diet and growth status.

In Chapter 3, weighted quantile regression model was fitted by using BMI as response variable. The estimates across different quantile levels allow us to study the impact of predictors on different quantiles of the response variable. This process provides a complete picture of the relationship between the continuous response variable and explanatory variables. Overall, the findings from this model show that age of a child, mother's age, mother's BMI, sex of child, weight of child at birth, regions (Affar, Gambela and Somali) were found to have significant effects on BMI of under five children across different quantile levels. Somewhat surprisingly, mother's work status, educational level of mother, place of residence and wealth index appeared slightly (although not significantly) affecting BMI of under five children when we consider weighted quantile regression model. Research suggests that mother's work status and educational level of mothers may have a significant effect on the BMI of under five children (Taddese et al., 2017; Teller & Yimer, 2000; Engle et al., 1996).

The binary logistic regression model without and with complex survey design were presented in Chapter 4. These models were used to determine factors that affect malnourishment of under five children. The socio-economic, geographic and demographic factors were used as explanatory variables. In addition two-way interaction effects were included in the modeling process. Logistic regression also called a logit model is used to model dichotomous outcome variables. The Hosmer and Lemeshow test was used to test the goodness of fit of the binary logistic regression model. The goodness of fit tests indicate that the logistic model fits the data well. The parameter estimates obtained from both models were compared using design effects. The findings from these models show that the design effect values are above one. This confirmed that there was an underestimation of variance while using logistic regression, which assume data was sampled using simple random sampling. Since survey logistic regression accounts for the complexity of the survey design, it produced parameter estimates that are different from the estimates obtained when simple random sampling was assumed. However, in some cases, they were closer to one another. From the results, we observed that the effect of mother's age, mother's BMI, educational level of mother, mother's work status and region (Harari) were found to have significant effect on the malnutrition of under five children in both models. Malnutrition includes a wide range of nutrient-related deficiencies and disorders, whether it is due to dietary deficiency called under-nutrition, or to excess diet called over-nutrition (Ratzan et al., 2000). The risk of malnourishment of a child decreases for an increase in mother's age. The risk of malnourishment of a child increases for an increase in mother's BMI. The risk of malnourishment of a child born to a mother who was working was lower compared to the risk of malnourishment of a child born to a mother who was not working. The risk of malnourishment of a child born to a mother who had some educational level was lower compared to the risk of malnourishment of a child born to a mother who had no education. The risk of malnourishment of a child in Somali region appeared to be higher compared to the risk of malnourishment of a child in Oromia region.

Based on weight-for-height anthropometric index (Z-score) child nutrition status was categorized into four groups:- underweight, normal, overweight and obese. Since this leads to an ordinal variable for nutritional status, an ordinal logistic regression was presented in Chapter 5. An ordinal logistic regression model is a generalization of a binary logistic regression model, when the outcome variable has more than two ordinal levels. It considers any inherent ordering of the levels in the outcome variable and makes full use of the ordinal information. The most popular way generalizing the binary logit model is to use cumulative logit that handle ordered categories. This model is also known as proportional odds model. The results of the proportional odds model with simple random sampling (conventional proportional odds model) and with complex survey design were presented to determine factors that affect nutritional status of under five children. From the result, it was observed that the current age of child, sex of child, weight of child at birth, mother's BMI and regions (Affar, Dire Dawa, Gambela, Harari and Somali) had significant effects on the nutritional status of under five children in both models. Mother's age, unmarried mother's and SNNP region had no significant effect on nutritional status of children for the conventional proportional odds model. On the other hand, mother's age, unmarried mothers and SNNP region had significant effect on nutritional status of children for proportional odds model with complex survey design. Moreover, the estimated logit coefficients and their standard errors were different for models that assume simple random sampling and models that consider complex survey design. The logit coefficients of the significant predictors current age of child, sex of child and regions (Dire Dawa and Harari) were increased while the logit coefficients of the other significant predictors' weight of child at birth, mother's age, mother's BMI, unmarried mothers and regions (Affar, Gambela and Somali) were decreased in the proportional odds model with complex survey design as compared to the corresponding logit coefficients of the significant predictors in the conventional proportional odds model. Further, the standard errors of the coefficients in the proportional odds model with complex survey design were higher as compared to the corresponding standard errors of the coefficients in the conventional proportional odds model. This confirms that standard errors were underestimated when we considered the conventional proportional odds model. Therefore, the conclusions in this study are based on the proportional odds model with complex survey design. The odds of being at or beyond a specific nutritional status of a child (i.e. the odds of being underweight, normal weight, overweight or obese) decreased for a one unit increase in child's age. The odds of being at or above a specific nutritional status of a female child was lower compared to a male child. The odds of being at or above a specific nutritional status of a child who had large and average weight at birth was higher as compared to a child who had small weight at birth. The odds of being at or above a specific nutritional status of a child decreased for a one unit increase in mother's current age. The odds of being at or above a specific nutritional status of a child increases for a one unit increase in mother's BMI. The odds of being at or above a specific nutritional status of a child born to unmarried mother was lower as compared to a child born to married mother. The odds of being at or above a specific nutritional status of a child from Affar, Dire Dawa, Gambela, Harari and Somali regions was lower as compared to Oromia region.

The findings from this study suggest that studying the determinants of BMI, malnutrition and nutritional status of under five children in Ethiopia is still critical issue that needs to be addressed. Overall, this study shows that mother's current age, mother's BMI, mother's work status, educational level of mother, weight of child at birth and regions (Affar, Dire Dawa, Gambela, Harari and Somali) were found to be dominant determinants of the BMI, malnutrition and nutritional status of under five children in Ethiopia. This confirmed that the BMI, malnutrition and nutritional status of the children were significantly influenced by education, socio-economic and environmental factors. Therefore, policy makers need to focus on the influence of these significant factors to develop strategies that enhance normal or healthy weight status of under five children in Ethiopia. This study also suggests that improving the nutritional status of mothers will consequently improve the nutritional status of their children. Improving the work status of the mothers will enhance the mother's economic status and consequently improve the basic needs of their children. To change weight-related disorders, changes related to children, environmental and social intervention is required to promote and support weight-related change in mothers. The government of Ethiopia needs urgent implementation of programs targeted to the those regions of Affar, Dire Dawa, Gambela, Harari and Somali, which were highly affected by malnutrition of under five children.

It must be borne in mind that this study was conducted based on certain socio-economic and environmental factors. Further research is hence needed to unravel the specific socio-economic and environmental factors and determine whether they serve as influential factors that affect the BMI, malnutrition and nutritional status of under five children and enhance the findings in Chapters 3, 4 and 5. In further study, we will extend this study by considering multilevel modeling, non-parametric and semi-parametric approaches to ordinal logistic regression, Bayesian method, Spatial-temporal analysis and other advanced statistical models. In addition, we will try to identify the trends of malnutrition or nutritional status of the under five children using the available EDHS survey results. For our study, the percentage of missing value was very small (less than 5%). Because of the small percentage, we did not use any missing value techniques. But, for the future direction of this study, we will use missing data analysis technique by exploring the effect of different missing data mechanism such as missing not at random (MNAR), missing at random (MAR) and missing completely at random (MCAR).

References

Achadi, E., Ahuja, A., Bendech, M. A., Bhutta, Z. A., De-Regil, L. M., Fanzo, J., Fracassi, P., Grummer-Strawn, L. M., Haddad, L. J., Hawkes, C., et al. (2016). *Global nutrition report 2016: From promise to impact: Ending malnutrition by 2030*. International Food Policy Research Institute.

Agresti, A. (2002a). Inference for contingency tables. *Categorical Data Analysis, Second Edition*, (pp. 70–114).

Agresti, A. (2002b). Wiley series in probability and statistics. *Analysis of Ordinal Categorical Data, Second Edition*, (pp. 397–405).

Agresti, A. (2010). *Analysis of ordinal categorical data*, vol. 656. John Wiley & Sons.

Agresti, A., & Kateri, M. (2011). Categorical data analysis. In *International Encyclopedia of Statistical Science*, (pp. 206–208). Springer.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Allison, P. D. (1999). *Multiple regression: A primer*. Pine Forge Press.

Allison, P. D. (2012). *Logistic regression using SAS: Theory and application*. SAS Institute.

Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, *26*(6), 1323–1333.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 1–30).

Asami, Y. H. Y. (2016). A method for comparing multiple regression models.

Ayele, D. G., Zewotir, T. T., & Mwambi, H. G. (2013). Spatial distribution of malaria problem in three regions of Ethiopia. *Malaria Journal*, *12*(1), 207.

Azen, R., & Budescu, D. V. (2006). Comparing predictors in multivariate regression models: An extension of dominance analysis. *Journal of Educational and Behavioral Statistics*, *31*(2), 157–180.

Bender, R., & Benner, A. (2000). Calculating ordinal regression models in SAS and S-plus. *Biometrical Journal*, *42*(6), 677–699.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, (pp. 279–292).

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, (pp. 1171–1178).

Buchinsky, M. (1998). Recent advances in quantile regression models: A practical guideline for empirical research. *Journal of human resources*, (pp. 88–126).

Chamberlain, G. (1994). Quantile regression, censoring, and the structure of wages. *Advances in Econometrics*, *1*, 171–209.

CIA (2018). *CIA World Factbook*. Accessed: 2018-02-05.
URL https://www.cia.gov/library/publications/the-world-factbook/geos/et.html

Cliff, N., & Keats, J. A. (2003). *Ordinal measurement in the behavioral sciences*. Psychology Press.

Clogg, C. C., & Shihadeh, E. S. (1994). *Statistical models for ordinal variables*, vol. 4. Sage Publications, Inc.

Cochran, W. G. (2007). *Sampling techniques*. John Wiley & Sons.

CSA, I. (2012). Ethiopia Demographic and Health survey 2011. *Addis Ababa, Ethiopia and Calverton, Maryland, USA: Central Statistical Agency and ICF International*, *430*.

CSA, I. (2016). Ethiopia Demographic and Health Survey 2016. *Addis Abeba, Ethiopia and Calverton, Maryland, USA: Central Statistical Agency and ICF International*.

Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: Theory and implementation. *Available at czep. net/stat/mlelr. pdf*.

Davino, C., Furno, M., & Vistocco, D. (2013). *Quantile regression: Theory and applications*. John Wiley & Sons.

De Onis, M., Brown, D., Blossner, M., & Borghi, E. (2012). Levels and trends in child malnutrition. UNICEF-WHO-The World Bank joint child malnutrition estimates.

Dobson, A. J., & Barnett, A. (2008). *An introduction to generalized linear models*. CRC press.

Dowd, A. C., Duggan, M. B., et al. (2001). Computing variances from data with complex sampling designs: A comparison of stata and spss. *North American Stata Users Group*, *12*.

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, *65*(3), 457–483.

Engle, P., Menon, P., Garrett, J., & Slack, A. (1996). Urbanization and caregiving: Evidence from South and Eastern Africa. *San Luis, California: Department of Psychology and Human Development, California Polytechnic, Stat. University*, (pp. 4–24).

Ferro-Luzzi, A., Scaccini, C., Taffese, S., Aberra, B., & Demeke, T. (1990). Seasonal energy deficiency in Ethiopian rural women. *European Journal of Clinical Nutrition*, *44*, 7–18.

Fienberg, S. E. (1976). *The analysis of cross-classified categorical data*. Springer.

Garcia, M., & Mason, J. (1992). Second report on the world nutrition situation. volume i: Global and regional results. a report compiled from information available to the United Nations agencies of the acc/scn.

Genebo, T., Girma, W., Haider, J., & Demisse, T. (2017). Factors contributing to positive and negative deviances in child nutrition. *The Ethiopian Journal of Health Development (EJHD)*, *12*(2).

Getahun, Z., Urga, K., Ganebo, T., & Nigatu, A. (2017). Review of the status of malnutrition and trends in Ethiopia. *The Ethiopian Journal of Health Development (EJHD)*, *15*(2).

Getaneh, T., Assefa, A., & Tadesse, Z. (1998). Protein-energy malnutrition in urban children: Prevalence and determinants. *Ethiopian Medical Journal*, *36*(3), 153–166.

Gujarati, D. (2014). *Econometrics by example*. Palgrave Macmillan.

Habyarimana, F., Zewotir, T., & Ramroop, S. (2014). A proportional odds model with complex sampling design to identify key determinants of malnutrition of children under five years in Rwanda. *Mediterranean Journal of Social Sciences*, *5*(23), 1642.

Hammer, L. D., Kraemer, H. C., Wilson, D. M., Ritter, P. L., & Dornbusch, S. M. (1991). Standardized percentile curves of body-mass index for children and adolescents. *American Journal of Diseases of Children*, *145*(3), 259–263.

Hao, L., & Naiman, D. Q. (2007). *Quantile regression*. 149. Sage.

Hardin, J. W., Hilbe, J. M., & Hilbe, J. (2007). *Generalized linear models and extensions*. Stata press.

He, X., & Hu, F. (2002). Markov chain marginal bootstrap. *Journal of the American Statistical Association*, *97*(459), 783–795.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis*. CRC Press.

Hien, N. N., & Hoa, N. N. (2009). Nutritional status and determinants of malnutrition in children under three years of age in Nghean, Vietnam. *Pak J Nutr*, *8*(7), 958–964.

Himes, J. H., & Dietz, W. H. (1994). Guidelines for overweight in adolescent preventive services: Recommendations from an expert committee. The Expert Committee on Clinical Guidelines for Overweight in Adolescent Preventive Services. *The American Journal of Clinical Nutrition*, *59*(2), 307–316.

Hosmer, D. W., & Lemeshow, S. (2000). *Special topics*. Wiley Online Library.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2000). Application of logistic regression with different sampling models. *Applied Logistic Regression, Third Edition*, (pp. 227–242).

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*, vol. 398. John Wiley & Sons.

Huang, M. L., Xu, X., & Tashnev, D. (2015). A weighted linear quantile regression. *Journal of Statistical Computation and Simulation*, *85*(13), 2596–2618.

Ishii-Kuntz, M. (1994). Ordinal log-linear models (quantitative applications in the social sciences, no. 97).

Kalhan, S. C., Prentice, A., Yajnik, C. S., et al. (2009). *Emerging Societies: Coexistence of Childhood Malnutrition and Obesity*, vol. 63. Karger Medical and Scientific Publishers.

Kalton, G. (1983). *Introduction to survey sampling*, vol. 35. Sage.

Keys, A., Fidanza, F., Karvonen, M. J., Kimura, N., & Taylor, H. L. (1972). Indices of relative weight and obesity. *Journal of Chronic Diseases*, *25*(6-7), 329–343.

Kleinbaum, D. G., & Klein, M. (2010). Polytomous logistic regression. In *Logistic Regression*, (pp. 429–462). Springer.

Kleinbaum, D. G., Klein, M., & Pryor, E. (2002). Logistic regression: A self-learning text.2002.

Koenker, R. (1994). Confidence intervals for regression quantiles. In *Asymptotic statistics*, (pp. 349–359). Springer.

Koenker, R. (2005). *Quantile regression*. 38. Cambridge university press.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, (pp. 33–50).

Koenker, R., & Zhao, Q. (1996). Conditional quantile estimation and inference for ARCH models. *Econometric Theory*, *12*(5), 793–813.

Krasovec, K., & Anderson, M. A. (1991). Maternal nutrition and pregnancy outcomes: Anthropometric assessment.

Kulasekaran, R. A., et al. (2012). Influence of mothers' chronic energy deficiency on the nutritional status of preschool children in Empowered Action Group states in India. *International Journal of Nutrition, Pharmacology, Neurological Diseases*, *2*(3), 198.

Kutner, M. H., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear statistical models*. McGraw-Hill Irwin.

Lall, R., Campbell, M., Walters, S., Morgan, K., & Co-operative, M. C. (2002). A review of ordinal regression models applied on health-related quality of life assessments. *Statistical Methods in Medical Research*, *11*(1), 49–67.

Lee, E., & Forthofer, R. (2006). Analyzing Complex Survey Data. Quantitative Applications in the Social Sciences series, vol. 71.

Lepkowski, J., & Bowles, J. (1996). Sampling error software for personal computers. *The Survey Statistician*, *35*, 10–17.

Levy, P. S., & Lemeshow, S. (2013). *Sampling of populations: Methods and applications*. John Wiley & Sons.

Liu, X. (2009). Ordinal regression analysis: Fitting the proportional odds model using Stata, SAS and SPSS. *Journal of Modern Applied Statistical Methods*, *8*(2), 30.

Liu, X. (2014). Fitting stereotype logistic regression models for ordinal response variables in educational research (stata). *Journal of Modern Applied Statistical Methods*, *13*(2), 31.

Liu, X. (2015). *Applied Ordinal Logistic Regression Using Stata: From Single-level to Multilevel Modeling*. Sage Publications.

Liu, X., & Koirala, H. (2013). Fitting proportional odds models to educational data with complex sampling designs in ordinal logistic regression. *Journal of Modern Applied Statistical Methods*, *12*(1), 26.

Lives, E., & Nations, R. (2015). Accelerating Inclusive Growth for Sustainable Human Development in Ethiopia.

Long, J. S. (1997). Regression models for limited and categorical dependent variables.

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. Stata press.

Lumley, T., & Scott, A. (2014). Tests for regression models fitted to survey data. *Australian & New Zealand Journal of Statistics*, *56*(1), 1–14.

Lumley, T., & Scott, A. (2015). Aic and bic for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, *3*(1), 1–18.

Lumley, T., et al. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, *9*(1), 1–19.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 109–142).

McCullagh, P., & Nelder, J. (1989). Generalised linear modelling. *Chapman and Hall: New York*.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, *4*(1), 103–120.

McMillen, D. P. (2012). *Quantile regression for spatial data*. Springer Science & Business Media.

Melesse, S., Sobratee, N., & Workneh, T. (2016). Application of logistic regression statistical technique to evaluate tomato quality subjected to different pre-and post-harvest treatments. *Biological Agriculture & Horticulture*, *32*(4), 277–287.

Miller, P. W., & Volker, P. A. (1985). On the determination of occupational attainment and mobility. *Journal of Human Resources*, (pp. 197–213).

Molla, D. A. (2016). *The Ethiopic Calendar*. Accessed: 2018-02-07.
URL http://www.ethiopic.com/calendar/

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models*, vol. 4. Irwin Chicago.

O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. 146. Sage.

Parzen, M., Wei, L., & Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika*, *81*(2), 341–350.

Pietrobelli, A., Faith, M. S., Allison, D. B., Gallagher, D., Chiumello, G., & Heymsfield, S. B. (1998). Body mass index as a measure of adiposity among children and adolescents: A validation study. *The Journal of Pediatrics*, *132*(2), 204–210.

Rao, J. N., & Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, *76*(374), 221–230.

Rao, J. N., & Scott, A. J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, (pp. 46–60).

Ratzan, S. C., Filerman, G. L., & LeSar, J. W. (2000). *Attaining global health: Challenges and opportunities*, vol. 55. Population Reference Bureau Washington, DC.

Rotimi, C., Okosun, I., Johnson, L., Owoaje, E., Lawoyin, T., Asuzu, M., Kaufman, J., Adeyemo, A., & Cooper, R. (1999). The distribution and mortality impact of chronic energy deficiency among adult Nigerian men and women. *European Journal of Clinical Nutrition*, *53*(9), 734–739.

Rutstein, S. (1999). Wealth versus expenditure: Comparison between the dhs wealth index and household expenditures in four departments of Guatemala. *Calverton, Maryland: ORC Macro*.

SAS, I. (2014). Sas/stat r 13.2 users guide. *Cary, North Carolina: SAS Institute Inc*.

Schabenberger, O., & Pierce, F. J. (2001). *Contemporary statistical models for the plant and soil sciences*. CRC press.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Science, P. E. C. (2017). *Introduction to Generalized Linear Models*. Accessed: 2017-10-07.
URL https://onlinecourses.science.psu.edu/stat504/node/216

Sen, A., & Srivastava, M. (2012). *Regression analysis: Theory, methods, and applications*. Springer Science & Business Media.

Sharon, L. L. (1999). Sampling: Design and analysis. *Duxbury Press*, *190*.

Sommerfelt, A. E., & Stewart, M. K. (1994). Children's nutritional status.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Allyn & Bacon/Pearson Education.

Taddese, Z., Larson, C. P., & Hanley, J. A. (2017). Anthropometric status of Oromo women of childbearing age in rural South Western Ethiopia. *The Ethiopian Journal of Health Development (EJHD)*, *12*(1).

Teller, C. H., & Yimer, G. (2000). Levels and determinants of malnutrition in adolescent and adult women in Southern Ethiopia. *Ethiopian Journal of Health Development*, *14*(1), 57–66.

Tseday (2008). *An Ethiopian Journal*. Accessed: 2018-02-07.
URL https://tseday.wordpress.com/2008/09/14/ethiopian-astronomy/

Unicef, et al. (1990). *Strategy for improved nutrition of children and women in developing countries*. Unicef.

USAID (2014). *Ethiopia: Nutrition Profile*. Accessed: 2018-02-05.
URL https://www.usaid.gov/sites/default/files/documents/1864/USAID-Ethiopia-Profile.pdf

Weisberg, S. (2005). *Applied linear regression*, vol. 528. John Wiley & Sons.

Winship, C., & Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review*, (pp. 512–525).

World Health Organization (1996). *Cancer pain relief: with a guide to opioid availability*. World Health Organization.

Worldometers (2018). *Ethiopia Population*. Accessed: 2018-02-12.
URL http://www.worldometers.info/world-population/ethiopia-population/

Yimer, G. (2000). Malnutrition among children in Southern Ethiopia: Levels and risk factors. *Ethiopian Journal of Health Development*, *14*(3).

Yirga, A. A., Ayele, D. G., & Melesse, S. F. (2018a). Application of Quantile Regression: Modeling Body Mass Index in Ethiopia. *The Open Public Health Journal*, *11*(1).

Yirga, A. A., Mwambi, H. G., Ayele, D. G., & Melesse, S. F. (2018b). Factors affecting child malnutrition in Ethiopia based on an ordinal response regression model. *African Health Sciences - African Journals Online*, *18*.

Appendix A

Additional Results

Table 6.1: Parameter estimates at 0.25^{th} quantile

Parameter	**Estimate**	**St.error**	95% **C.I**	**P-value**
Intercept	13.7005	0.2651	(13.1807, 14.2202)	$<$0.0001
Current age of child	-0.1793	0.0167	(-0.2120, -0.1466)	$<$0.0001
Mother's age	-0.0104	0.0040	(-0.0183, -0.0024)	0.0104
Mother's BMI	0.0670	0.0091	(0.0492, 0.0849)	$<$0.0001
Sex of child (ref. = male)				
Female	-0.2722	0.0594	(-0.3887, -0.1557)	$<$0.0001
Weight of child at birth (ref. = Small)				
Large	0.4220	0.0821	(0.2611, 0.5829)	$<$0.0001
Average	0.2413	0.0757	(0.0929, 0.3897)	0.0014
Mother's work status (ref. = Yes)				
No	0.0416	0.0586	(-0.0733, 0.1566)	0.4779
Educational level (ref. = Sec. school)				
No education	-0.2222	0.1516	(-0.5194, 0.0751)	0.1429
Primary school	-0.1572	0.1468	(-0.4449, 0.1305)	0.2843
Higher	-0.1159	0.2313	(-0.5693, 0.3374)	0.6162
Marital status (ref. = Not married)				
Married	0.0093	0.1237	(-0.2332, 0.2517)	0.9403
Religion (ref. = Protestant)				
Orthodox	0.0464	0.0992	(-0.1481, 0.2408)	0.6401
Catholic	-0.3914	0.3589	(-1.0949, 0.3121)	0.2755
Muslim	0.0566	0.0932	(-0.1260, 0.2392)	0.5438
Other	0.2181	0.3113	(-0.3921, 0.8283)	0.4835
Region (ref. = Tigray)				
Addis Abeba	0.2490	0.1512	(-0.0474, 0.5454)	0.0996
Afar	-0.2568	0.1040	(-0.4607, -0.0529)	0.0136
Amhara	0.0184	0.0960	(-0.1698, 0.2066)	0.8477
Benishangul	0.0268	0.0946	(-0.1586, 0.2121)	0.7772
Dire Dawa	-0.2679	0.1475	(-0.5569, 0.0211)	0.0693
Gambela	-0.2627	0.1257	(-0.5092, -0.0162)	0.0367
Harari	-0.0795	0.1570	(-0.3871, 0.2282)	0.6127
Oromia	0.0396	0.0985	(-0.1536, 0.2327)	0.6879
SNNPR	0.4065	0.1084	(0.1941, 0.6190)	0.0002
Somali	-0.8043	0.1036	(-1.0074, -0.6012)	$<$0.0001
Place of residence (ref. = Urban)				
Rural	0.0347	0.1176	(-0.1958, 0.2651)	0.7682
Wealth index (ref. = Rich)				
Middle	-0.0973	0.0747	(-0.2438, 0.0492)	0.1930
Poor	-0.1226	0.0762	(-0.2719, 0.0267)	0.1075

Table 6.2: Parameter estimates at 0.5^{th} quantile

Parameter	**Estimate**	**St.error**	95% **C.I**	**P-value**
Intercept	14.9289	0.2984	(14.3440, 15.5137)	<0.0001
Current age of child	-0.2908	0.0217	(-0.3333, -0.2484)	<0.0001
Mother's age	-0.0149	0.0043	(-0.0233, -0.0065)	0.0005
Mother's BMI	0.0730	0.0100	(0.0533, 0.0926)	<0.0001
Sex of child (ref. = male)				
Female	-0.1699	0.0499	(-0.2676, -0.0722)	0.0007
Weight of child at birth (ref. = Small)				
Large	0.5169	0.0722	(0.3754, 0.6584)	<0.0001
Average	0.2263	0.0678	(0.0934, 0.3593)	0.0008
Mother's work status (ref. = Yes)				
No	0.0589	0.0536	(-0.0461, 0.1640)	0.2715
Educational level (ref. = Sec. school)				
No education	-0.0252	0.1423	(-0.3042, 0.2538)	0.8593
Primary school	-0.0361	0.1363	(-0.3033, 0.2311)	0.7913
Higher	0.1799	0.2749	(-0.3590, 0.7188)	0.5129
Marital status (ref. = Not married)				
Married	0.0416	0.1401	(-0.2330, 0.3162)	0.7664
Religion (ref. = Protestant)				
Orthodox	-0.0413	0.0972	(-0.2318, 0.1492)	0.6707
Catholic	-0.3314	0.3212	(-0.9611, 0.2983)	0.3023
Muslim	-0.1484	0.0967	(-0.3380, 0.0412)	0.1249
Other	0.5450	0.3001	(-0.0434, 1.1333)	0.0695
Region (ref. = Tigray)				
Addis Abeba	0.0533	0.1430	(-0.2270, 0.3336)	0.7093
Afar	-0.2730	0.0931	(-0.4554, -0.0906)	0.0034
Amhara	-0.0096	0.0848	(-0.1758, 0.1566)	0.9097
Benishangul	-0.0851	0.0948	(-0.2710, 0.1007)	0.3693
Dire Dawa	-0.1933	0.1264	(-0.4411, 0.0546)	0.1264
Gambela	-0.4479	0.1316	(-0.7058, -0.1899)	0.0007
Harari	0.0701	0.1373	(-0.1991, 0.3393)	0.6096
Oromia	0.0267	0.0886	(-0.1470, 0.2004)	0.7628
SNNPR	0.2809	0.1028	(0.0793, 0.4824)	0.0063
Somali	-0.8032	0.1066	(-1.0121, -0.5943)	<0.0001
Place of residence (ref. = Urban)				
Rural	-0.1274	0.1180	(-0.3587, 0.1039)	0.2803
Wealth index (ref. = Rich)				
Middle	-0.0936	0.0773	(-0.2452, 0.0580)	0.2264
Poor	-0.1063	0.0662	(-0.2361, 0.0234)	0.1083

Table 6.3: Parameter estimates at 0.75^{th} quantile

Parameter	**Estimate**	**St.error**	95% **C.I**	**P-value**
Intercept	16.2903	0.3318	(15.6398, 16.9408)	<0.0001
Current age of child	-0.4013	0.0239	(-0.4482,-0.3544)	<0.0001
Mother's age	-0.0181	0.0042	(-0.0264, -0.0098)	<0.0001
Mother's BMI	0.0722	0.0146	(0.0436, 0.1008)	<0.0001
Sex of child (ref. = male)				
Female	-0.2169	0.0607	(-0.3359, -0.0979)	0.0004
Weight of child at birth (ref. = Small)				
Large	0.4659	0.0776	(0.3138, 0.6181)	<0.0001
Average	0.2648	0.0705	(0.1265, 0.4030)	0.0002
Mother's work status (ref. = Yes)				
No	0.0363	0.0822	(-0.1248, 0.1974)	0.6589
Educational level (ref. = Sec. school)				
No education	0.1337	0.1565	(-0.1731,0.4404)	0.3931
Primary school	0.0491	0.1608	(-0.2662,0.3644)	0.7601
Higher	-0.0570	0.2495	(-0.5461, 0.4321)	0.8193
Marital status (ref. = Not married)				
Married	0.2700	0.1237	(0.0274, 0.5125)	0.0291
Religion (ref. = Protestant)				
Orthodox	-0.1707	0.1218	(-0.4095, 0.0680)	0.1610
Catholic	-0.5027	0.4790	(-1.4416, 0.4362)	0.2939
Muslim	-0.1297	0.1229	(-0.3707, 0.1112)	0.2913
Other	0.6020	0.4692	(-0.3178, 1.5218)	0.1996
Region (ref. = Tigray)				
Addis Abeba	0.1381	0.1873	(-0.2291, 0.5052)	0.4611
Afar	-0.4732	0.1300	(-0.7281, -0.2184)	0.0003
Amhara	-0.1553	0.0964	(-0.3442, 0.0336)	0.1071
Benishangul	-0.2442	0.1183	(-0.4760, -0.0123)	0.0390
Dire Dawa	-0.5401	0.1272	(-0.7894, -0.2908)	<0.0001
Gambela	-0.5770	0.1584	(-0.8875, -0.2665)	0.0003
Harari	-0.1211	0.1347	(-0.3851, 0.1428)	0.3684
Oromia	0.0672	0.1076	(-0.1437, 0.2781)	0.5324
SNNPR	0.2034	0.1363	(-0.0638, 0.4706)	0.1357
Somali	-0.9451	0.1257	(-1.1915, -0.6988)	<0.0001
Place of residence (ref. = Urban)				
Rural	-0.2769	0.1324	(-0.5364, -0.0173)	0.0366
Wealth index (ref. = Rich)				
Middle	-0.1824	0.1011	(-0.3806, 0.0159)	0.0713
Poor	-0.1665	0.0809	(-0.3251, -0.0079)	0.0396

Appendix B

B.1 Model Fitting using the PROC QUANTREG procedure

The following codes were used to fit the WQR model:

```
PROC IMPORT OUT= WORK.Ashu;
DATAFILE= ”C : \Users\yirga\Desktop\underfiveEDHSdata.sav”
DBMS=SPSS REPLACE;
ods graphics on;
Proc quantreg data=WORK.Ashu;
Class B4 M18 V714 V149 V501 V130 V101 V102 V190;
Model BMICHILD=B8 V012 V439A B4 V101 V102 V130 V190 V714 V149 V501 M18
/quantile=0.05,0.25,0.50,0.75,0.85,0.95;
weight V005A;
Run;
ods graphics off;
```

B.2 Model Fitting using the PROC LOGISTIC procedure

The following codes were used to fit the Binary logistic regression model without and with CSD:

```
PROC IMPORT OUT= WORK.Ashu
DATAFILE= ”C : \Users\yirga\Desktop\project\underfiveEDHSdata.sav”
DBMS=SPSS REPLACE;
RUN;
ods graphics on;
Proc logistic data=WORK.Ashu plots=all;
Class B4(ref=”male”) V101(ref=”Oromia”) V102(ref=”rural”) V130(ref=”Orthodox”)
V190(ref=”poor”) V714(ref=”No”) V149(ref=”No education”) V501(ref=”married”)
M18(ref=”small”) / param=glm;
Model Binary_weightstatus(event=”malnourished”)=B8 V012 V439A B4 V101 V102
V130 V190 V714 V149 V501 M18 B8*V439A B8*V101 V439A*V101 V439A*M18/ctable
```

```
cl lackfit selection=none include=12 scale=none aggregate=(B8 V012 V439A B4 V101
V102 V130 V190 V714 V149 V501 M18);
output out = outdata p = pred-prob;
Run;
ods graphics off;
ods graphics on;
Proc surveylogistic data=WORK.Ashu;
stratum V023/list;
cluster V021;
weight V005;
Class B4(ref=”male”) V101(ref=”Oromia”) V102(ref=”rural”) V130(ref=”Orthodox”)
V190(ref=”poor”) V714(ref=”No”) V149(ref=”No education”) V501(ref=”married”)
M18(ref=”small”) / param=glm;
Model Binary_weightstatus(event=”malnourished”)=B8 V012 V439A B4 V101 V102
V130 V190 V714 V149 V501 M18 B8*V439A B8*V101 V439A*V101 V439A*M18;
Run;
ods graphics off;
```

B.3 Model Fitting using the *Stata:ologit* procedure

The following codes were used to fit the Ordinal logistic regression model without and with CSD:

```
use ”C : \Users\yirga\Desktop\project\underfiveEDHSdata.dta”, clear
global ylist Ordinal nutritional status
global xlist B8 ib1.B4 ib3.M18 V012 V439A ib0.V714 ib0.V149 ib1.V501 ib1.V130 ib4.V101
ib2.V102 ib1.V190 ib1.B4#ib1.V501 ib2.V102#ib0.V149 ib1.V190#ib0.V149
ologit $ylist $xlist, or
svyset V021 [pweight=V005A], strata (V023)
svy:ologit $ylist $xlist
svy:ologit $ylist $xlist, or
```

where, BMICHILD=Body mass index of under five children, B8=current age of child, B4=sex of child, M18=weight of child at birth, V012=mother’s current age, V439A=mother’s BMI, V714=mother’s work status, V149=educational attainment of mother, V501=current marital status, V130=religion, V101=region, V102=place of residence, V190=wealth index, V005A=mother’s individual (weight), V021=primary sampling unit (cluster) and V023=stratum (strata).

Appendix C

EXPONENTIAL FAMILY OF DISTRIBUTION

Exponential family of distribution includes many distributions that are useful for practical modeling such as: Poisson or negative Binomial for count response variable; Binomial, Bernoulli and geometric for the study of discrete responses; Gamma, Normal, Inverse Gaussian, Beta and exponential for the study of continuous responses. A distribution belongs to an exponential family of distributions if its probability density function or probability mass function can be written as:

$$f(Y;\theta,\phi) = exp\left\{\frac{1}{a(\phi)}[\theta Y - b(\theta)] + c(Y,\phi)\right\}$$

where θ is the natural or canonical parameter, $a(\phi)$ is the scale parameter or dispersion and $c(Y_i,\phi)$ is some function of Y_i and ϕ. The mean, $\mu = E(Y) = b'(\theta)$, and the variance, $Var(Y) = \phi b''(\theta)$, can be obtained as follows:
$*$ By definition $\implies \int f(y)dy = 1, \qquad E(y) = \int yf(y)dy,$
$Var(y) = \int (y - E(y))^2 f(y)dy,$ and $\frac{\partial}{\partial y}(g(y)f(y)) = g'(y)f(y) + g(y)f'(y)$
$\therefore \quad \int exp\left\{\frac{1}{a(\phi)}[\theta Y - b(\theta)] + c(Y,\phi)\right\} dy = 1 \qquad \cdots **$
Derivate $(**)$ **wrt** θ **both sides:**
$\frac{\partial}{\partial\theta}\int exp\left\{\frac{1}{a(\phi)}[\theta Y - b(\theta)] + c(Y,\phi)\right\} dy = 0$

$$\int \underbrace{\frac{y - b'(\theta)}{a(\phi)}}_{\text{g(y)}} \times \underbrace{exp\left\{\frac{1}{a(\phi)}[\theta Y - b(\theta)] + c(Y,\phi)\right\}}_{\text{f(y)}} dy = 0 \qquad \cdots * **$$

$\int \frac{y-b'(\theta)}{a(\phi)} \times f(y)dy = 0 \implies \int (y - b'(\theta)) \times f(y)dy = 0$

$$\underbrace{\int yf(y)dy}_{\text{E(y)}} - b'(\theta)\underbrace{\int f(y)dy}_{\text{(density=1)}} = 0 \implies \boldsymbol{E(y) = b'(\theta)}$$

Derivate $(* * *)$ **wrt** θ
$\frac{\partial}{\partial\theta}\int \frac{y-b'(\theta)}{a(\phi)} \times exp\left\{\frac{1}{a(\phi)}[\theta Y - b(\theta)] + c(Y,\phi)\right\} dy = 0$
$\int \frac{-b''(\theta)}{a(\phi)} f(y)dy + \int \left(\frac{y-b'(\theta)}{a(\phi)}\right)\left(\frac{y-b'(\theta)}{a(\phi)}\right) f(y)dy = 0$
$\frac{-b''(\theta)}{a(\phi)} + \frac{1}{(a(\phi))^2}\int (y - b'(\theta))^2 f(y)dy = 0$

$\frac{-b''(\theta)}{a(\phi)} + \frac{1}{(a(\phi))^2} \underbrace{\int (y - E(y))^2 f(y) dy}_{\text{var(y)}} = 0$

$\frac{-b''(\theta)}{a(\phi)} + \frac{var(y)}{(a(\phi))^2} = 0 \implies -a(\phi)b''(\theta) + var(y) = 0$

$\boldsymbol{Var(y) = a(\phi)b''(\theta)}$

Example:- In Chapter 4 our response variable is binary (children catagorized as normal nutrition status or malnourished) which can be assumed to follow the Bernoulli distribution. $Y_i \sim Bernoulli(\pi_i)$, $\pi_i = p(y_i = 1) = \mu_i, 1 - \mu_i = p(y_i = 0)$.

$$f(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}, Y_i = 0, 1$$

$$Implies \quad that, \quad \mu_i = E(Y_i) = (0 \times P(Y_i = 0)) + (1 \times P(Y_i = 1))$$

$$\mu_i = \pi_i$$

We can show that Bernoulli distribution belongs to an exponential family of distribution as follows:

$$f(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i} = exp\{Y_i log\pi_i + (1 - Y_i)log(1 - \pi_i)\}$$

$$= exp\{Y_i[log\pi_i - log(1 - \pi_i)] + log(1 - \pi_i)\}$$

$$= exp\left\{Y_i log\left(\frac{\pi_i}{1 - \pi_i}\right) - (-log(1 - \pi_i))\right\}$$

where $\theta = log\left(\frac{\pi_i}{1-\pi_i}\right)$, implies that $\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$, $1 - \pi_i = \frac{1}{1+e^{\theta_i}}$, $a(\phi) = 1$, $c(y, \phi) = 0$ and $b(\theta_i) = -log(1 - \pi_i) = -log\left(\frac{1}{1+e^{\theta_i}}\right) = -log(1 + e^{\theta_i})^{-1} = log(1 + e^{\theta_i})$

Expressing $f(y_i)$ in terms of θ_i:

$$f(y_i) = exp\{y_i\theta_i - log(1 + e^{\theta_i})\}$$

$$f(y_i) = exp\{y_i\theta_i - b(\theta_i)\}$$

where $\boldsymbol{E(y_i)} = b'(\theta_i) = \frac{1}{1+e^{\theta_i}} \times e^{\theta_i} = \frac{e^{\theta_i}}{1+e^{\theta_i}} = \boldsymbol{\pi_i}$,

$\boldsymbol{Var(y_i)} = \phi b''(\theta_i), b''(\theta_i) = \frac{\partial}{\partial\theta_i}\left(\frac{e^{\theta_i}}{1+e^{\theta_i}}\right) = \frac{e^{\theta_i}}{(1+e^{\theta_i})^2} = \frac{e^{\theta_i}}{1+e^{\theta_i}} \times \frac{1}{1+e^{\theta_i}} = \boldsymbol{\pi_i(1 - \pi_i)}$

N.B: Bernoulli distribution is a special case of the Binomial distribution where n=1 (one trial).

Example:-Suppose $Y_1, Y_2, ..., Y_n$ are independent binomial observations, $Y_i \sim Binomial(n_i, p_i)$:

$$f(Y_i) = \binom{n_i}{Y_i} p_i^{Y_i}(1 - p_i)^{n_i - Y_i}$$

$$= exp\left\{log\binom{n_i}{Y_i} + Y_ilogp_i + (n_i - Y_i)log(1-p_i)\right\}$$

$$= exp\left\{Y_ilogp_i - Y_ilog(1-p_i) + n_ilog(1-p_i) + log\binom{n_i}{Y_i}\right\}$$

$$= exp\left\{Y_ilog\left(\frac{p_i}{1-p_i}\right) - (-n_ilog(1-p_i)) + log\binom{n_i}{Y_i}\right\}$$

where $\theta_i = log\left(\frac{p_i}{1-p_i}\right)$, $a(\phi) = 1$, $c(y,\phi) = log\binom{n_i}{Y_i}$ and $b(\theta_i) = -n_ilog(1-p_i) = n_ilog(1+e^{\theta_i})$.

Expressing $f(y_i)$ in terms of θ_i:

$$f(y_i) = exp\left\{y_i\theta_i - n_ilog(1+e^{\theta_i}) + log\binom{n_i}{Y_i}\right\}$$

where $\boldsymbol{E(y_i)} = b'(\theta_i) = \frac{n_i}{1+e^{\theta_i}} \times e^{\theta_i} = \frac{n_ie^{\theta_i}}{1+e^{\theta_i}} = \boldsymbol{n_ip_i}$,

$\boldsymbol{Var(y_i)} = \phi b''(\theta_i), b''(\theta_i) = \frac{\partial}{\partial\theta_i}\left(\frac{n_ie^{\theta_i}}{1+e^{\theta_i}}\right) = \frac{n_ie^{\theta_i}}{(1+e^{\theta_i})^2} = \frac{n_ie^{\theta_i}}{1+e^{\theta_i}} \times \frac{1}{1+e^{\theta_i}} = \boldsymbol{n_ip_iq_i}, \quad q_i = (1-p_i)$

The above expression indicates that Binomial distribution is a family of exponential family. The dependent variable in logistic regression follows the Bernoulli distribution having an unknown probability, P.

MAXIMUM-LIKLIHOOD ESTIMATION (MLE)

$f(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{n_i - Y_i}$

Liklihood function

$L(\pi) = L(Y_1, Y_2, ..., Y_n|\pi) = \prod_{i=1}^n f(Y_i)$

$= \prod_{i=1}^n(\pi_i^{Y_i}(1-\pi_i)^{n_i-Y_i})$

$L(\pi) = \pi^{\sum Y_i}(1-\pi)^{n-\sum Y_i} \implies Liklihood \quad function$

Log-Liklihood function (the value of π that maximizes $logL$)

$log(L(\pi)) = log[\pi^{\sum Y_i}(1-\pi)^{n-\sum Y_i}]$

$= log\pi^{\sum Y_i} + log(1-\pi)^{n-\sum Y_i}$

$= \sum Y_ilog\pi + (n - \sum Y_i)log(1-\pi) \implies Log-liklihood \quad function$

MLE of π $\quad(\hat{\pi})$

Derivate $logL$ with respect to π

$\frac{\partial log(L(\pi))}{\partial\pi} = \frac{\sum Y_i}{\pi} + \frac{(n-\sum Y_i)}{1-\pi}(-1) = 0$

$= \frac{\sum Y_i}{\pi} - \frac{(n-\sum Y_i)}{1-\pi} = 0$

$= \frac{\sum Y_i(1-\pi) - \pi(n-\sum Y_i)}{\pi(1-\pi)} = 0$

$= \frac{\sum Y_i - \pi\sum Y_i - n\pi + \pi\sum Y_i}{\pi(1-\pi)} = 0$

$= \sum Y_i - n\pi = 0 \implies \boldsymbol{\hat{\pi}} = \frac{\sum Y_i}{n} = \boldsymbol{\bar{Y}} \qquad (MLE \quad of \quad \pi)$

Pseudo-likelihood function with weight W_i (extended, section 4.6.1, Equation 4.10)

$$\implies \pi_i^{W_iY_i}(1-\pi_i)^{(1-W_iY_i)}$$

$$L = Pr(Y_1)Pr(Y_2)...Pr(Y_n)$$
$$= \prod\left(\pi_i^{W_iY_i}(1-\pi_i)^{(1-W_iY_i)}\right)$$
$$= \pi_i^{\sum W_iY_i}(1-\pi_i)^{1-\sum W_iY_i}$$
$$logL = log\left[\pi_i^{\sum W_iY_i}(1-\pi_i)^{1-\sum W_iY_i}\right]$$
$$= log\left(\pi_i^{\sum W_iY_i}\right) + log(1-\pi_i)^{1-\sum W_iY_i}$$
$$= \sum W_iY_ilog\pi_i + \left(1-\sum W_iY_i\right)log(1-\pi_i)$$
$$= \sum W_iY_ilog\pi_i + log(1-\pi_i) - \sum W_iY_ilog(1-\pi_i)$$
$$= \sum W_iY_ilog\left(\frac{\pi_i}{1-\pi_i}\right) + log(1-\pi_i)$$

$$\leftrightarrow \sum W_iY_i\boldsymbol{X}_i'\boldsymbol{\beta} + log\left(\frac{1}{1+e^{\boldsymbol{X}_i'\boldsymbol{\beta}}}\right)$$
$$= \sum W_iY_i\boldsymbol{X}_i'\boldsymbol{\beta} + log(1) - log\left(1+e^{\boldsymbol{X}_i'\boldsymbol{\beta}}\right)$$
$$\frac{\partial}{\partial\boldsymbol{\beta}}\left(\sum W_iY_i\boldsymbol{X}_i'\boldsymbol{\beta} + log(1) - log\left(1+e^{\boldsymbol{X}_i'\boldsymbol{\beta}}\right)\right)$$
$$= \sum W_iY_i\boldsymbol{X}_i' - \left(\frac{1}{1+e^{\boldsymbol{X}_i'\boldsymbol{\beta}}}\right) \times e^{\boldsymbol{X}_i'\boldsymbol{\beta}} \times X_i'$$
$$\implies \boldsymbol{\beta} = \sum W_iY_i\boldsymbol{X}_i' - \left(\left(\frac{e^{\boldsymbol{X}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{X}_i'\boldsymbol{\beta}}}\right) \times X_i'\right)$$