29

Addis Ababa University (Since 1950)

Addis Ababa University

Office of graduate Program

Faculty of science

Department of statistics

Determinants of malaria infection among women in Ethiopia

By

Muluken Derbew

A Thesis submitted to the Office of Graduate Programs of Addis Ababa University in Partial fulfillment of the requirements for the Degree of Master of Science in Applied Statistics

June, 2009

Addis Ababa University
Office of Graduate Program

Faculty of Science
Department of Statistics

Approved by the Board of Examiners:

Sileshi Fanta
.....
Department Head

.....
Signature

Emmanuel Yohannes
.....
Examiner

.....
Signature

Fentaw Abegaz
.....
Examiner

.....
Signature

Tel. 251-1-239465 P.O.Box 1176 Addis Ababa, Ethiopia

TABLE OF CONTENTS

ACKNOWLEDGEMENT

I express my deepest appreciation to my advisor Professor Eshetu Wencheko for his generous contribution to the accomplishment of this research work and for his consistent and stimulating advice during the whole course of the research.

I would like to thank to my teachers Dr. Fantaw Abegaz and Dr. Emanuel G/Yohannes for their sympathetic encouragement during my study.

My striking, deepest and heartfelt thanks are also due to my brothers Tefera Derbew and Tewodros Derbew for their benevolent inputs and all round support during my studies.

My warm thanks and indebtedness are addressed to my friends Zeytu Begashaw, Dejen Tesfaw and Girmachew Kebede, whose compassionate encouragement were the sources of inspiration to me, during my study.

Finally, I would like to thank all my teachers and all the staff members of the Department of Statistics, Addis Ababa University for their kind assistance in many ways.

Abstract

The problem of malaria disease in Ethiopia is compounded by more frequent epidemics, combined P. vivax and P. falciparum infections, and increasing drug and insecticide resistance. This study examines the association between the socio-economic and demographic and health factors and the malaria infection status among women in Ethiopia using logistic regression models. The study also indicates the factors that contribute in explaining the variation of malaria infection status among women across regions. The study is based on 7333 women respondents. The data were obtained from the Central Statistics Agency of the government of Ethiopia in 2005 (EDHS 2005). We use conventional logistic regression modeling to determine the relationship between the explanatory variables and "malaria infection status" and multilevel logistic regression to see whether there exist variations in malaria infection status as relates to women across the regions of Ethiopia. Malaria infection status among women was modeled using socio-economic and demographic and health variables as potential predictors. The results of conventional logistic regression model showed that the explanatory variables "region", "currently pregnant", "wealth index", "type of place of residence", "main floor material" and "age" are found to have a significant association with malaria infection status among women in Ethiopia. Similarly, the interaction of the random parts of "age by main floor material" and "main floor material by wealth index" provided significant effect on "malaria infection status" across regions.

CHAPTER ONE

1. INTRODUCTION

1.1 Background

A vector-borne disease, as the name suggests, is transmitted with the aid of a vector. The term vector refers to a medium, an arthropod or some other agent through which a pathogenic micro-organism is transmitted from an infected individual to another uninfected individual. The mechanism of transmission involves at least three different living organisms, the pathological agents which can either be a bacteria, virus, and protozoa; the vector, generally arthropods such as ticks or mosquitoes; and the human host. Some domesticated or wild animals can act as reservoir for the pathogens till they are exposed to the susceptible human population. Each vector-borne disease is associated with a specific pathogen and a vector which acts as a medium of transmission. The spread of the disease in a particular region depends on the adaptability of the vector to survive and grow in that region.

Malaria remains a major public health problem in the world. It is a major life-threatening vector-borne disease transmitted through mosquitoes. The disease got its name from bad air (mal aria) as it was thought that the disease came from fetid marshes. Later in 1880, it was discovered that the real cause of malaria was Plasmodium, a single cell parasite which can only be transmitted from one person to another by the bite of female Anopheles mosquito of which 3(three) species are of importance, with respect to their breeding, biting and other behavioural characteristics, namely, Anopheles arabiensis, A. funestus and A. gambiae.. The male Anopheles mosquitoes are not involved in disease transmission as they don't require blood to nurture eggs as their female counterparts do. There are 4(four) types of Plasmodium species-Plasmodium falciparum, P. ovale, P. vivax and P. malariae. Of these P. falciparum is by far the most dangerous. Unfortunately, it is also the most common in Africa.

Malaria is the most well known, underestimated disease in the world. More than 2 billion people live in malarious regions of the world. As a result, between 300 to 500 million new infection cases occur every year with over 1.2 to 2.7 million deaths, of which malaria is further responsible for 1(one) out of 4(four) childhood deaths in Africa. About 90% of these occur in Sub-Sahara Africa. A child in Africa dies every 30 seconds because of malaria and those who survive the severe episode of malaria might suffer from learning impairments or brain damage (Anonyms, 21 April 2006). Of the estimated 400 to 900 million episodes of fever occurring yearly in African children, probably about half are due to malaria, resulting in over one million deaths (Breman,J.G.et al., 2001). Although malaria is spread by bites of female Anopheles mosquitoes, it can also spread by other means such as blood transfusions, organ transplants and sharing of needles by intravenous drug (IV drug) users.

In areas of stable endemic malaria transmission in sub-Saharan Africa it has been estimated that in 1995 about 1 million deaths were directly attributable to malaria infection (Snow et al. 1999). Of these deaths, three-quarters were in children below the age of 5 years. In the same population, it is estimated that about 200 million clinical attacks of malaria occurred in the same year. In areas of unstable or epidemic prone malaria in southern Africa (fringe areas), about 2000 deaths and 200,000 clinical episodes occurred and that were not prevented despite malaria control measures in these areas. According to a World Bank report of 1993, malaria accounts for an estimated 35 million disability adjusted life years (DALYs) per year lost in Africa due to ill-health and premature death (World Bank,1993).

The malaria transmission season runs from September to December, following the major rainy season from June to August in Ethiopia. The widespread epidemic has a cyclical pattern of 5 to 8 years that follows major climatic changes. Almost 75 percent of the land is malarious and an estimated 50 million people (68 percent) live in areas at risk of malaria areas at altitude below 2000 meters above sea level are generally considered malarious. However, local transmission has also been detected in areas at altitudes as high as 2500 meters. The transmission pattern is unstable and often characterized by focal and cyclic large scale epidemics. The most recent malaria epidemic, which occurred in

2003, affected 211 districts where more than 2 million clinical cases were recorded (Negash et al., 2005).

Resistance of malaria parasites to drugs is steadily gaining new ground and fresh outbreaks of malaria are being reported from areas which were hardly affected before. According to the statistics of the United Nations Population Division in 1990, malaria is the only disease today (apart from HIV/AIDS) that shows a significant rising tendency. Malaria epidemics are now frequent and spread to areas previously without epidemics.

Pregnant women, young children and elderly individuals are particularly at risk. Malaria in pregnant women increases the risk of maternal death, miscarriage, stillbirth and neonatal death. The greatest challenge faced by communities is to address the issue of malaria. So it is relevant to properly assess the distribution of malaria in order to understand and find an appropriate solution to the socio-economic and demographic problems of the population of Ethiopia by malaria disease.

1.2 Statement of the problem

Malaria is one of the most severe problems faced by the world even today. The annual health and health-related indicators of the Federal Ministry of Health (FMoH) report malaria as the national leading cause of morbidity and mortality. Due to the unstable and seasonal pattern of malaria transmission, the protective immunity of the population is generally low, and all age groups are at risk of infection and disease. Understanding the causative factors and the underlying transmission dynamics of the disease is important for epidemiological research on malaria and its eradication. Only small-scale studies have documented malaria prevalence in Ethiopia. However these studies are not sufficient as compared to the disastrous effect of malaria in the country. Therefore, there is a need for further investigations on the distribution of malaria in Ethiopia in such away that we can clearly indicate the socio-economic and demographic problems in the country. To ensure the development of suitable modeling approach and methodology, based on the available data on the distribution of malaria and other related factors is of utmost importance.

1.3 Research Objectives

General Objective

- ✓ To assess the extent of malaria infection status among women in Ethiopia.

Specific objectives

- ✓ To explore the factors significantly associated with malaria infection status among women in Ethiopia.
- ✓ To estimate the variances and covariances of random effects at regional levels and check their significance.
- ✓ To examine the within and between regional level differences in determining the extent of malaria infection among women in Ethiopia.

1.4 Significance of the study

Ethiopia is one of the most malaria-epidemic prone countries in Africa. Understanding the extent of malaria infection among women is the basic issue in the health sector so that the relevant governmental and non-governmental organizations can appropriately apply all their efforts, knowledge, resources, etc to improve the health status of women in Ethiopia. So this study envisages that it may strengthen the information so far for scaling up and to design effective communication strategy to combat malaria.

1.5 Limitation of the study

1. The target population consists only women of age 15-49 years.
2. The software used in the analysis of multilevel logistic regression (Student version of LISREL software) is restricted to handle few explanatory variables.
3. Women residing in Addis Ababa could not be part of the multivariate analysis. This is because few of them who were malaria positive (malarious) are rejected when data was adjusted for analysis.
4. The time of malaria transmission season runs from September to December 2005. But the data collection was conducted from March 14 to April 20, 2005. This might considerably decrease the number of malaria positive women considered.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview of malaria

Malaria is a common and life-threatening disease in many tropical and subtropical areas. Human malaria is caused by four different species of the protozoan parasite Plasmodium: Plasmodium falciparum, P. vivax, P. ovale and P. malariae.

Transmission

The malaria parasite is transmitted by various species of Anopheles mosquitoes, which bite mainly between sunset and sunrise.

Nature of the disease

Malaria parasites have a complicated life cycle. After injection into the human host from anopheline mosquitoes, the parasites home in the person's liver, where they undergo maturation before being released into the bloodstream, whence they invade red blood cells. They change form and multiply inside the red blood cells, eventually rupturing the cells, releasing still more parasites into the bloodstream. This bloodstream cycle can persist for weeks to years, depending on the species involved. In vivax and *ovale* malaria, some parasites ("hypnozoites") can persist indefinitely in the liver. Meanwhile, back in the bloodstream, some of the parasites differentiate into sexual forms (gametocytes) which, if ingested by another mosquito, can lead to the development of another generation of parasites, ready for transmission to another human host.

Malaria as a disease is therefore closely bounded to conditions which favor the survival of the anopheles mosquito in the form of habitat and breeding sites and which favor the life cycle of the parasite in terms of suitable temperatures. In the absence of any human intervention these conditions are predominantly determined by climatic and environmental factors.

One of the main environmental factors affecting malaria transmission is temperature. The effect of an increase in temperature on the parasite is to shorten the sporogony cycle and hence to accelerate transmission. The duration of sporogony can be calculated by the

formula $n = \frac{T}{t - t_{\min}}$ where n=duration of sporogony in days, t= average temperature in ˇC, and for P.falciparum T =105 ˇ C and t_{min}=16ˇ C. Below 16ˇ C parasite development ceases. Rising temperature also increases transmission by increasing the frequency with which the vector takes blood meals, which increases the growth rate of vector populations through a shortening of the generation time. The optimal range of temperature for most vectors lies between 20ˇ C and 30ˇ C. Higher temperatures reduce the longevity of adult vectors, and hence fewer of them will survive the sporogony cycle to transmit the malaria. There are thus upper and lower thresholds outside which malaria transmission is very inefficient or impossible.

The most severe form is caused by P. falciparum, in which variable clinical features include fever, chills, headache, muscular aching and weakness, vomiting, cough, diarrhea and abdominal pain; other symptoms related to organ failure may supervene, such as: acute renal failure, generalized convulsions, circulatory collapse, followed by coma and death. In endemic areas it is estimated that about 1% of patients with P. falciparum infection die of the disease; the mortality in non-immune individuals with untreated falciparum infection is significantly higher.

The initial symptoms, which may be mild, may not be easy to recognize as being due to malaria. It is important that the possibility of falciparum malaria is considered in all cases of unexplained fever starting at any time between the seventh day of first possible exposure to malaria and three months (or, rarely, later) after the last possible exposure. Any individual who experiences a fever in this interval should immediately seek diagnosis and effective treatment, and inform medical personnel of the possible exposure to malaria infection.

Early diagnosis and appropriate treatment can be life-saving. Falciparum malaria may be fatal if treatment is delayed beyond 24 hours. A blood sample should be examined for malaria parasites. If no parasites are found in the first blood film while there is clinical

suspicion of malaria, a series of blood samples should be taken at 6–12-hour intervals and examined very carefully.

The forms of malaria caused by other Plasmodium species are less severe and rarely life-threatening. Chemoprophylaxis and treatment of falciparum malaria are becoming more difficult because P. falciparum is becoming increasingly resistant to various anti-malarial drugs. Chloroquine resistance of P. vivax is rare and was first reported in the late 1980s in Papua New Guinea and Indonesia. Focal "true" chloroquine resistance (i.e. in patients with adequate blood levels at day of failure) or prophylactic and/or treatment failure have later also been observed in Brazil, Columbia, Ethiopia, Guatemala, Guyana, India, Myanmar, Peru, the Republic of Korea, Solomon Islands, Thailand and Turkey. Resistance of P. malaria to chloroquine has been reported from Indonesia.

Type of parasite

There are over 120 species of the parasite genus Plasmodium. However, only four of these infect humans to cause malaria. These four species of Plasmodium parasites are Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale and Plasmodium malariae. Other species affect other mammals, birds, and reptiles; some of these (rarely) cause human illness. The details of the four main species of Plasmodium that infect humans with malaria are described below:

1. P. vivax – They are found worldwide but most commonly in India, Central and South America. The incubation period in the human body is approximately 8-13 days for the symptoms of the disease to become apparent. Infection by this parasite can sometimes lead to life-threatening rupture of spleen. They hide in liver and can return later once a person is infected .

2. P. ovale –They are found mostly in Africa. This form of malaria has incubation Period of 8-17 days in the infected person and can hide in the liver of partially treated people and return later.

3. P. malariae – They are found in most part of the world but are less frequent than other forms of the malaria parasite. The incubation period for this parasite is 2-4 weeks in the infected person. If the disease is untreated, it can last for many years.

4. P. falciparum–They are responsible for the most life-threatening form of malaria and cause majority of the deaths in the world and are found worldwide. *It* is the only species that appears to directly affect the central nervous system causing neurologic deficits, cognitive sequel and epilepsy. The incubation period for this parasite is 5-12 days. They are resistant to most of the drugs used to treat or prevent malaria.

Distribution of Disease: Malaria is endemic in much of the world, particularly much of the Western Hemisphere between Mexico and Peru-Bolivia-Brazil, sub-Saharan Africa, South and South East Asia, and parts of the Middle East and Turkey. Endemic transmission requires both a pool of infected humans and the presence of competent vector mosquitoes. Although once common in parts of the country, malaria has become an exotic disease in the United States. Nonetheless, one should remain alert to the potential for local transmission.

Period of communicability: As long as infective gametocytes are present in the blood of a patient, that person remains a source of mosquito infection. Untreated or insufficiently treated individuals may be a source of mosquito infection for up to three years, depending on the species as follows:

- P. malaria: up to 3 years.
- P. vivax and ovale: 1-3 years
- P. falciparum: less than 1 year.

Transmission by transfusions may occur as long as asexual forms remain in the circulating blood. Stored blood may remain infective for at least a month. Anopheles Mosquitoes spread Plasmodium for their lifespan.

Incubation period: This is the period from infection with sporozoites by Anopheles mosquito to appearance of first symptoms. Normally from 7-14 days:

- P. falciparum: 12 days.
- P. vivax & ovale: 14 days.

- P. malariae: 30 days.

This is approximately time periods and not rigid periods.

Symptoms:

1. *Uncomplicated Malaria*:
   - Recent history of fever (can be intermittent)
   - Headaches
   - Sweats/chills (cold shaking feeling)
   - Body pains
2. *Moderate Severe Malaria*:
   - Nausea/vomiting
   - Extreme weakness
   - Diarrhoea
3. *Severe Complicated Malaria:*
   - Jaundice
   - Sleepiness (Difficult in rising)
   - Shock (Cold moist skin, low blood pressure, collapse)
   - Convulsions (Fits, uncontrollable body, movements)
   - Respiratory distress (Chest in drawings, rapid breathing)
   - Hypoglycaemia (Low blood sugar)
   - Unconsciousness/Coma
   - Fluid and electrolyte disturbances.
   - Acute renal failure.
   - Acute pulmonary oedema and adult respiratory distress syndrome (ARDS).
   - Circulatory collapse, shock, septicaemia ("algid malaria").
   - Abnormal bleeding.
   - Haemoglobinuria.
   - High fever.
   - Hyperparasitaemia.

Important:

- ✓ These severe manifestations can occur singly or, more commonly, in combination in the same patient.
- ✓ Severe complicated malaria is usually caused by delay in treating an uncomplicated attack of P. falciparum.
- ✓ In high-transmission areas, the risk of severe complicated malaria developing is greatest among young children, and visitors (of any age) from non-endemic areas.

Diagnosis: Diagnosis can be done in two ways, either by visual symptoms or clinical. However, symptomatic diagnosis must always be confirmed clinically by medical personnel to ensure proper and adequate treatment at the end. Prompt diagnosis and effective treatment can prevent complications and death from malaria.

Treatment: The effectiveness of antimalarial drugs differs with different species of the parasite and with different stages of the parasite's life cycle. Therefore Medical treatment should be sought immediately in such a way that the physician will determine the treatment plan most appropriate for individual condition.

High population growth and ecological degradation in the highland and midland areas of Ethiopia have induced population mobility into lowland areas. This has led to increased exposure of people to communicable diseases like malaria.

In epidemic years mortality rates of an extra 40,000 children are not uncommon. In the last major malaria epidemic in December 2003, 3,689 villages in 211 districts were affected, resulting in over 6.1 million cases with an estimated 45,000 to 114,000 estimated deaths attributable to Malaria per annum (Child survival strategy, 2003). About 94,400 children estimated number of lives saved if all Malaria control interventions are fully implemented (Child survival strategy, 2003).

Malaria is still the leading cause of health problem in Ethiopia. In 2004, the disease has been reported as the first cause of illness and death accounting for 15.5% of outpatient visits, 20.4% of admissions and 27% of deaths. The magnitude and periodicity of malaria epidemics in the country has also been on the rise in the past few years.

Out of an estimated 15.3 million malaria cases in Ethiopia, only 4-5 million will be treated in a health facility annually. The remainders often have no medical support. It is estimated that only 20 percent of children under-five population years of age that contract malaria are treated in a facility.

Malaria is prevalent in over 75 percent of the area of the country, putting over 50 million people at risk (out of a countrywide of 77 million). The disease is responsible for largest single cause of morbidity. Large scale epidemics tend to occur every 5-8 years in certain areas due to climatic fluctuations and drought-related nutritional emergencies. Children and pregnant mothers are the most vulnerable. Drought related malnutrition, poor health and poor or no sanitation can leave a weak immune system open to attack from malaria. It can also worsen the effects of malnutrition through malaria-related diarrhea and anemia.

Due to the unstable and seasonal pattern of malaria transmission, the protective immunity of the population is generally low, and all age groups are at risk of infection and disease. Some small-scale studies have documented malaria parasite prevalence between 10.4–13.5% in Gambella (Nigatu, W. et al. 1992) and 7.6–14.1% in Tigray in all age groups [World Health Organization (WHO), 1999]. The malaria indicator survey (MIS) 2007 organized by the ministry of health (MOH) of the government of Ethiopia indicated that the malaria prevalence rate in Ethiopia was 0.7%.

Malaria is also known to speed up the onset of AIDS in anyone who is HIV positive. Those living with HIV in high-risk areas are also amongst the most vulnerable. The situation is exacerbated by the vast distances rural Ethiopians must cover in the countryside to find a clinic or other health facilities with reliable medical supplies. With

day to day survival preoccupying the minds of most parents, walking more than a day for anti-malarial supplies is a daunting task. Ethiopia's Child Survival study showed that an average of 94,400, out of 470,000 child deaths per year, is attributable to malaria. However, if the available malaria control interventions are implemented correctly and effectively, it has been estimated that these malaria related deaths could be reduced by a massive 75%, saving the lives of around 70,000 children every year. (United Nations Children's Fund, 2005).

Home-based management of malaria (HMM) is promoted as a major strategy to improve prompt delivery of effective malaria treatment in Africa. HMM involves presumptively treating febrile children with pre-packaged antimalarial drugs distributed by members of the community. HMM has been implemented in several African countries such as Kenya, Gambia, Zaire, Burkinafaso and Ethiopia. It has the potential to improve treatment delivery and decrease malaria-associated morbidity and mortality. However, the Home-based management of Malaria strategy is a major undertaking, and its implementation should be based on sound evidence of public health benefit.

Malaria is complex but it is a curable and preventable disease. Lives can be saved if the disease is detected early and adequately treated. It is known what action is necessary to prevent the disease and to avoid or contain epidemics and other critical situations. The technology to prevent, monitor, diagnose and treat malaria exists. It needs to be adapted to local conditions and to be applied through local and national malaria control programmes.

2.2 Review of relevant studies

Sudatip (1997) carried out a case-control study of health behavior factors related to malaria infection in Kanchanaburi province. The result showed that the greatest factor of malaria infection was related to staying overnight without mosquito net in the forest. There was no significant association between malaria infection and accepting mosquito nets impregnated with insecticide and taking chemoprophylaxis.

Ittiravivongs et al. (1992) carried out a study in two villages of the same malaria endemicity but use different work pattern (night-and day shift). It was found that uses of mosquito nets and malaria chemoprophylaxis were similar and there was no significant difference in health behavior between the two villages

Rahman et al. (1993) conducted a study on the distribution of malaria in an endemic district in northern Peninsula, Malaysia. The study encompassed the distribution of malaria case according to sex, age and profession. A total of 332 cases were recorded, with 182 cases occurring in males. The highest infection was observed in those who were older than 15 years old. Forest worker (loggers, rattan collectors and forest product gathers) were the groups exposed most to the disease (32.8%), followed by both plantation workers (32.2%) and aboriginal communities (32.2%). Plasmodium falciparum was the most common species of malaria in the area.

Hu et al. (1998) applied disease mapping through GIS with multiple regression analysis to determine the nature and extent of factors influencing malaria transmission in Yunnan province, China. Secondary county-based data during 1990-1996 were collected and analyzed. Maps showed that malaria incidence rates tended to be higher in areas with higher temperature, heavier rainfall, denser forest and lower elevation. Malaria occurred more in border counties along the Yuan River. Malaria was endemic in areas where An. minimus is the major vector. Multiple regression analysis showed that malaria incidence rate increased when temperature, rainfall and forest coverage increased. A border county will have extra higher malaria incidence rate by 6.51/10,000 compared to a county not located along the border. Elevation has a negative correlation with malaria incidence rate in which every 100 meters decrease in elevation will result in 99.8/10,000 increase in malaria incidence rate. Malaria situation in Yunnan is influenced mainly by the combined effects of the physical environment and the presence of efficient vector.

Mauny et al (2004) used an actual data set of 3864 individuals from 38 villages of the Highland Madagascar; a two-level modeling process is presented. Individual **malaria** parasitaemia is modeled step by step according to age (individual factor), altitude, and DDT indoor house-spraying status (village factors). The hierarchical organization of a data set in levels, fixed and random effects and cross-level interactions are considered. Accurate estimations of standard errors, impact of unknown or unmeasured variables quantified and accounted for through random effects, are the highlighted advantages of multilevel modeling. The result also showed that the overall parasite prevalence was 15%, modified by age (<10 years is 23%, ≥10 years is 11%), altitude (below 1300m is 22%, above1300m is 6%), and DDT status (unsprayed 31% and sprayed 8%). While not denying the importance of understanding an aetiological chain, the authors recommend an increased use of multilevel modeling, mainly to identify accurately ecological targets for public health policy.

CHAPTER THREE

Data and Methodology

3.1 The Data

The data on which this thesis research is based on is taken from the Demographic and Health Survey (EDHS 2005), conducted by the Central Statistics Agency of the Government of Ethiopia (CSA). The survey was conducted in the country level. The training of interviewers, editors, laboratory technicians and supervisors for data collection was conducted from March 14 to April 20, 2005. In addition to classroom training, trainees did several days of field practice to gain more experience on interviewing the respondents. A total of 7333 eligible women residing in lowland areas were taken for this particular study. The 2005 Ethiopia Demographic and Health Survey (EDHS) sample is the result of a multi-stage stratified design.

The thesis uses data from this survey for analysis of the demographic and health, and socio-economic factors that influence malaria infection status among women in Ethiopia.

3.2 Variables in the Study

Variables to be included in this study are selected from EDHS 2005. The dependent variable in this study, which is "malaria infection status", is dichotomized as 1 if a respondent is malaria positive (malarious) and 0 if a respondent is malaria negative (non-malarious).

The independent variables/factors included in this study were classified as demographic and health (example age of a respondent) and socio-economic covariates (example wealth index of a respondent).

Table 3.1 Independent variables/factors with their label and category

No	Variable label	Category
1.	Age of the respondent	0=21-29 1=30-49 2=15-20
2.	Region	0=Somali 1=Tigray 2=Afar 3=Amhara 4=Oromia 5=Ben-Gumz 6=SNNP 7=Gambela 8=Harari 9=Addis Ababa 10=Dire Dawa
3.	Type of place of residence	0=Rural 1=Urban
4	De facto place of residence	0=countryside 1=Town 2=city
5.	Educational level	0=No education 1=Primary 2= Secondary and above

6.	**Main floor material**	**0=Natural** **1=Cement/bricks**
7.	**Main roof material**	**0=Natural** **1=Corrugated iron/cement**
8.	**Currently pregnant**	**0=Yes** **1=No**
9.	**Have bednet for sleeping**	**0=No** **1=Yes**
10.	**Wealth index**	**0=Poor** **1=Middle** **2=Rich**
11.	**Has radio**	**0=No** **1=Yes**

3.3 Methodology

3.3.1 Introduction

There are different types of multivariate statistical techniques that can be used to predict a binary dependent variable from a set of independent variables/factors. Among these, multiple linear regression analysis and discriminant analysis are two related techniques that quickly come to mind. However, these techniques bring difficulties when the dependent variable is dichotomous or categorical. Even if linear discriminant analysis does allow direct prediction of group membership, the assumption of multivariate normality of the independent variable as well as equal variance-covariance matrices in

the two groups is required for the prediction rule to be optimal. In this particular study the logistic regression model is used for analysis of data.

3.3.2 Logistic Regression Model

3.3.2.1 Overview

Modeling the relationship between explanatory and response variables is a fundamental issue encountered in statistics. Simple linear regression is often used to investigate the relationship between a single explanatory (predictor) variable and a single response variable. When there are several explanatory variables, multiple regression is used. However, often the response does not assume numerical values. Instead, the response could simply be a designation of one of two possible outcomes (a binary response) e.g. "alive" or "dead", "success" or "failure". Although responses may be accumulated to provide the number of successes and the number of failures, the binary nature of the response still remains.

Logistic regression can be used to predict a dependent variable on the basis of continuous and/or categorical explanatory variables and to determine the percentage of variance in the dependent variable explained by the explanatory variables/factors; to rank the relative importance of explanatory variables/factors; to assess interaction effects; and to understand the impact of covariate control variables. The impact of predictor variables is usually explained in terms of odds ratios.

Logistic regression uses maximum likelihood estimation after transforming the dependent variable into a logit variable (the natural log of the odds of the dependent variable occurring or not). Logistic regression estimates the odds of a certain event occurring. Note that logistic regression calculates changes in the log odds of the dependent variable, not the changes in the dependent variable itself as linear regression does.

Unlike linear regression, logistic regression does not assume a linear relationship between the explanatory variables/factors and the dependent variable, does not require

normality of the dependent variable, does not assume homoscedasticity, and in general has less stringent requirements. It does, however, require that observations be independent and that the explanatory variables be linearly related to the logit of the dependent variable. The predictive success of the logistic regression can be assessed by looking at the classification table, showing correct and incorrect classifications of the dichotomous dependent variable. Goodness-of-fit tests such as the likelihood ratio test are available as indicators of model appropriateness. Data involving the relationship between explanatory variables and a binary response abound in just about every discipline such as engineering, natural sciences, medicine, education, etc. How does one model the relationship between explanatory variables and a binary response variable? This study looks at binary response data and its analysis through logistic regression. In SPSS, binary logistic regression, sometimes called binomial logistic regression, is under Analyze - Regression - Binary Logistic.

3.3.2.2 Background on Odds Ratios

The odds ratio is the natural log of β , where β is the parameter estimate. $\exp(\beta)$ in SPSS output refers to odds ratios. No association between predictor and response means the odds ratio will be about one. If the predictor tends to be positively associated with the response, the odds ratio assumes values greater than one. If the predictor is negatively related to the response, the odds ratio will be less than one. The measure can range from zero to positive infinity. For instance, if $\beta_1 = 2.303$, then the corresponding odds ratio is 10, then we may say that when the independent variable increases by one unit, the odds that the dependent variable = 1 increases by a factor of 10, when other variables are controlled. In SPSS, odds ratios appear as "$\exp(\beta)$" in the "Variables in the Equation" table.

3.3.2.3 Introduction to logistic regression

Logistic Regression model is used when the dependent variable is not continuous but instead has only two possible outcomes, 1 or 0. Possible situations include studies where subjects are "alive" or "dead", "has a disease" or "doesn't have a disease", "purchases product" or "doesn't purchase", "wins race" or "doesn't win", "have" or "do not have" a

particular characteristic and so on. The usual linear regression models cannot be used for such variables because the predicted value needs to be constrained between 0 and 1, which is not possible in linear regression. It also violates the assumption that the variable is normally distributed, since a 1 or 0 variable by definition has a binomial distribution.

Logistic regression model solves this problem by determining the 'odds' of success (1) or failure (0). This is accomplished by estimating something called the Log Odds Ratio, which is just the log of the odds of success (1) divided by the odds of failure (0).

The logistic regression model can have an arbitrary number of parameters and terms in the model representing qualitative variables, quantitative variables and interaction terms. Logistic regression does not rely on distributional assumptions. However, our solution may be more stable if our predictors have a multivariate normal distribution. The procedure is most effective when group membership is a truly categorical variable; if group membership is based on values of a continuous variable, we should consider using linear regression to take advantage of the richer information offered by the continuous variable itself. Coefficients of logistic regression can be used to estimate odds ratios for each of the explanatory variables in the model.
We denote such a response variable by y, and denote the event y = 1 when the subject has the characteristic of interest and y = 0 when the subject does not have the characterstic.

For the purposes of this introduction we will suppose that the subject has a single predictor X that might be related to the response. The logistic regression model defines the probability P(y = 1) as,

$$P(y=1) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \qquad \text{and} \quad P(y=0) = 1 - P(y=1).$$

This model has a convenient representation in terms of the odds of the event y = 1 as,

$$Odds(y=1) = \frac{P(y=1)}{P(y=0)} = \exp(\beta_0 + \beta_1 x)$$

This means the log odds is simply the linear function $\beta_0 + \beta_1 x$ .This model has two parameters, but the parameter β_1 will be of primary interest. It controls the degree of association between the response and predictor variables.

Logistic regression is a powerful technique for fitting models to data with a binary response variable, but the models are difficult to interpret if interactions are present.

3.3.2.3.1 One Qualitative Predictor

To understand the logistic regression model, let us start with the easiest case, a model with one qualitative predictor at two levels. This could correspond to a variable for group membership. Let x = 1 if the subject is in group 1 and x = 0 if the subject belongs to group 2. In this situation the odds that y = 1 for group 1 is $\exp(\beta_0 + \beta_1)$, and the odds that y = 1 for a subject in group 2 is $\exp(\beta_0)$. So the model becomes, $Odds(y = 1/x) = \exp(\beta_0 + \beta_1 x)$. If we construct a ratio of these odds, we can summarize the impact of group membership on the response y.

$$OR = \frac{Odds(x = 1)}{Odds(x = 0)} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

That is, the odds ratio, OR depends only on the model parameter β_1.

Notice that when there is no group effect on the odds, the odds are equal in both the numerator and denominator, and the odds ratio equals 1. This can only happen when $\beta_1 = 0$. The the parameter controls the association between the predictor X and the response y. This implies that we could interpret β_1 as the change in the log odds ratio as x changes from 0 to 1.

3.3.2.3.2 Many Predictors

The logistic regression model can have an arbitrary number of parameters and terms in the model representing qualitative variables, quantitative variables, and interaction terms. The general multivariate logistic regression model with k terms is given by,

$$Odds(Y = 1 / x_1, x_2, ..., x_k) = \exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)$$

$$P(Y = 1 / x_1, x_2, ..., x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}$$

$$\log it \left\{ P(Y = 1 / x_1, x_2, ..., x_k) \right\} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

A model parameter β_i is interpreted as the change in the log odds for a one unit increase in x_i, holding all the other predictors constant, or after adjusting for the other predictors.

3.3.2.3.3 Parameter Estimation and Statistical Inference

3.3.2.3.3.1 Maximum likelihood estimation

The maximum likelihood estimation method is a popular statistical method used for fitting a mathematical model to the data. The modeling of real world data using estimation by maximum likelihood offers a way of tuning the free parameters of the model to provide good fit. Maximum likelihood estimation is used for estimating the parameters in the logistic regression model. The likelihood L (developed below) is maximized in order to achieve better estimates. The higher the maximized value of L, the better the fit of the model. This is assessed on a log scale by computing -2logL, called -2LL. When there are several explanatory variables, different models can be assessed using -2LL as a figure-of-merit. Here is an explanation of maximum likelihood estimation in logistic regression models. We begin with the simpler case of a sample of Bernoulli variables. If we have a sample $y_1, y_2, ..., y_n$ of 0, 1 variables with success probability p, then the log likelihood can be written as:

$$L = y_1 \ln p + (1 - y_1) \ln(1 - p) + y_2 \ln p + (1 - y_2) \ln(1 - p) + ... + (1 - y_n) \ln(1 - p).$$

The maximum likelihood estimator is the value π of which maximizes this; it turns out to be simply the sample proportion of 1's, $\frac{y_1 + y_2 + ... + y_n}{n}$.

In the logistic regression model $p(x) = \frac{1}{1+\exp(-\beta_0 - \beta' x)}$ where β_0 the constant is and β is the vector of logistic regression parameters to be estimated. This is done by maximizing the likelihood by numerical methods.

3.3.2.3.3.2 Statistical Inference

Statistical inference of the model parameter helps us to judge the significance and magnitude of the effects. Statistical inference for one model parameter typically involves either a confidence interval, a hypothesis test, or a confidence interval for the estimated odds ratio. A confidence interval for β_i is:

$$\hat{\beta}_i \pm Z\, se(\hat{\beta}_i)$$

where Z is the appropriate multiplier from the standard normal distribution. Confidence intervals whose endpoints do not contain zero indicate a relationship between the predictor x_i and the response after adjusting for any other predictor variables in the model. Confidence bounds containing zero do not show significant evidence of a relationship between the predictor and response. A hypothesis test of

Ho: $\beta_i = \delta$ vs Ha: $\beta_i \neq \delta$, where δ is some constant, uses the standard normal test statistic:

$$Z = \frac{\hat{\beta}_i - \delta}{se(\hat{\beta}_i)}$$

When $\delta = 0$, this test statistic and p-value are typically given in all statistical package output, and should correspond to the inference that would be made if a confidence interval was computed. When this test gives a small p-value, it will correspond to a confidence interval for β_i that does not contain zero. A confidence interval for the odds ratio exp β_i is obtained by simply exponentiating the confidence limits for the parameter.

Suppose the confidence bounds for β_i are (a, b) then the confidence limits for OR are (exp (a), exp (b)).

3.3.2.3.4 Comparing Models (Model Selection)

Another common inferential task is to examine the usefulness of a set of predictors in explaining the response. The usual approach is to set up a hypothesis test to compare a pair of nested models - one with more terms or predictor variables than the other. The formal set-up is as follows. Null hypothesis $H_0 : Odds = \exp(\beta_0 + \beta_1 x_1 + ... + \beta_m x_m$ versus the alternative hypothesis $H_a : Odds = \exp(\beta_0 + \beta_1 x_1 + ... + \beta_m x_m + ... + \beta_k x_k)$

This test will be examining the benefit of x_{m+1} up to x_k in explaining the response after adjusting for the first m predictor variables. The test statistic is the likelihood ratio statistic;

$$\lambda = \frac{L(H_0)}{L(H_a)}$$

This likelihood ratio statistic can be modified to have an asymptotic chi-squared distribution by taking $-2\log(\lambda) = -2[l(H_0) - l(H_a)] = -2l(H_0) - -2l(H_a)$

where $l(H_0)$ denotes the log likelihood function evaluated under the null hypothesis and $l(H_a)$ denotes the log likelihood function evaluated under the alternative hypothesis. It can be shown that $-2\log(\lambda)$ has an asymptotic chi-squared distribution under the null hypothesis with $k - m$ degrees of freedom.

Model Selection Procedures

In data situations where there are many predictors it is often helpful to use a model selection procedure to obtain a model that uses a subset of the original predictor variables. We may use stepwise model selection procedure. Stepwise selection starts with no predictors in the model and examines each term that could be possibly added and then adds the most significant predictor, or the predictor with the smallest p-value. In the next stage the procedure adds the next most significant term and checks to see if any

previous terms are now non-significant and removes them if they are not significant. This procedure continues until there are no further significant terms to add. So, unlike backward elimination, this procedure builds by adding terms.

3.3.2.3.5 Goodness-of-Fit

It is always important to examine the appropriateness of the fitted models. There are many types of procedures that have been mentioned in the statistical literature. Here we will mention three of the most used methods.

3.3.2.3.5.1 Hosmer and Lemeshow Test

The null hypothesis for this test is that the model fits the data, and the alternative hypothesis is that the model does not fit the data. The test statistic is constructed by first breaking the data set into roughly 10 (g) groups. The groups are formed by ordering the existing data by the level of their predicted probabilities. So the data are first ordered from least likely to have the event to most likely for the event. Then g (often 10) roughly equal sized groups are formed. From each group the observed and expected number of events are computed. The test statistic is,

$$C = \sum_{k=1}^{g} \frac{(O_k - E_k)^2}{v_k}$$

where O_k and E_k are the observed and expected number of events in the k^{th} group, and v_k is a variance correction factor for the k^{th} group. If the observed number of events differs from what is expected by the model, the statistic C will be large and there will be evidence against the null hypothesis. This statistic has an approximate chi-squared distribution with $g-2$ degrees of freedom.

Hosmer and Lemeshow's goodness-of-fit test divides subjects into deciles based on predicted probabilities as illustrated above, and then computes a chi-square from observed and expected frequencies. Then a probability (p) value is computed from the chi-square distribution with g -2 degrees of freedom to test the fit of the logistic model. If

the Hosmer and Lemshow goodness-of-fit test statistic is greater than .05, as we want for well-fitting models, we fail to reject the null hypothesis that there is no difference between observed and model-predicted values, implying that the model's estimates fit the data at an acceptable level. That is, well-fitting models show non-significance on the Hosmer-Lemshow goodness-of-fit test, indicating model prediction is not significantly different from observed values. This does not mean that the model necessarily explains much of the variance in the dependent variable, only that however much or little it does explain is significant.

3.3.2.3.5.2 Likelihood ratio Test

The likelihood ratio test, also called the log-likelihood test, is based on -2LL (deviance). The likelihood ratio test is a test of the significance of the difference between the likelihood ratio (-2LL) for the researcher's model minus the likelihood ratio for a reduced model. This difference is called "model chi-square."

i. Models.

(a) The "Intercept Only" model, also called the null model; it reflects the net effect of all variables not in the model plus error;

(b) The "Final" model, also called the fitted model, which is the researcher's model comprised of the predictor variables; the logistic equation is the linear combination of predictor variables which maximizes the log likelihood that the dependent variable equals the predicted value/class/group. The difference in the -2 log likelihood (-2LL) measures how much the final model improves over the null model.

ii. Test of the overall model.

The likelihood ratio test of the overall model, also called the model chi-square test, is the test shown in the "Final" row in the "Model Fitting Information" in SPSS. When the reduced model is the baseline model with the constant only (i.e, initial model or model at step 0), the likelihood ratio test tests the significance of the researcher's model as a whole. A well-fitting model is significant at the .05 level or better. That is, a finding of

significance (p<=.05 is the usual cutoff) leads to rejection of the null hypothesis that all of the predictor effects are zero. When this likelihood test is significant, at least one of the predictors is significantly related to the dependent variable. When probability (model chi-square) <= .05, we reject the null hypothesis that knowing the independents makes no difference in predicting the dependent in logistic regression.

The likelihood ratio test looks at model chi-square (chi square difference) by subtracting deviance (-2LL) for the final (full) model from deviance for the intercept-only model. Degrees of freedom in this test equal the number of terms in the model minus 1 (for the constant). Model chi-square measures the improvement in fit that the explanatory variables make compared to the null model.

3.3.2.3.5.3 Pearson Test

Another commonly used goodness-of-fit measure is the Pearson hypothesis test. Again the null hypothesis is that the model fits. The test statistic is given by:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(y_i - \hat{p}_i\right)^2}{\hat{p}_i\left(1-\hat{p}_i\right)}$$

where y_i is the observed response and $\hat{p}_i$ is the predicted response or predicted probability for the i^{th} subject. This statistic also has an approximate chi-squared distribution. The degrees of freedom are the number of covariate patterns minus the number of parameters estimated.

3.3.2.3.6 Case Diagnostics

In addition to global examinations of a model, it is also useful to examine the characteristics of individual cases in our data set. We are concerned with potential outliers and also with cases that might unduly (excessively) influence our parameter estimates. There are comparable diagnostics that should be used to identify data problems. The logistic regression provides a variety of such statistics.

3.3.2.3.6.1 Residuals

In logistic regression an outlier would be an observation that had a response y value very different from what the model would predict. There are two main types of residuals: Pearson and Deviance. The Pearson residual is defined as,

$$e_i = \frac{(y_i - \hat{p}_i)}{sd_i}$$

Where sd_i is the estimated standard deviation of the response. This residual will be positive when the event $y_i = 1$ occurs but the predicted probability of this event is lower. Likewise, the residual will be negative if the event did not occur, but the probability was higher than it would occur. These residuals can be viewed on roughly a standard normal scale -3 to +3. Deviance residuals are used in a similar fashion and are defined as

$$\pm 2\sqrt{\ln\left[\frac{y_i}{\hat{p}_i}\right]^{y_i}\left[\frac{1-y_i}{1-\hat{p}_i}\right]^{1-y_i}}$$

3.3.2.3.7 Influence measures

Influence measures that help us analyze the impact of individual cases on our parameter estimates.

3.3.2.3.7.1 Cook's Distance

Cook's distance is a global measure of influence that assesses the impact of case i on the parameter estimate vector β. The definition is,

$$C_i = \frac{e_i h_{ii}}{k(1-h_{ii})}$$

where e_i is the Pearson residual defined earlier and h_{ii} is the leverage of case i and k is the number of parameters in the model. The leverage is a measure of strangeness of the predictor pattern, more atypical (unusual) means higher leverage. Atypically large values of Cook's distance suggest an influential case.

Assumptions of logistic regression

1. Meaningful coding: Logistic coefficients will be difficult to interpret if not coded meaningfully. The convention for binomial logistic regression is to code the dependent class of greatest interest as 1 and the other case as 0. Logistic regression is predicting the log odds of being in the class of greatest interest.

2. Error terms are assumed to be independent (independent sampling). Violations of this assumption can have serious effects. Violations will occur, for instance, in correlated samples and repeated measures designs.

3. Predicts the odds of an event occurring, which is based on the probability of that event occurring. Precisely the Odds of an event occurring is:

$$\text{Odds}=\frac{P}{1-P}$$ [ratio of the probability of an event occurring to the probability of the event not occurring (ranges of 0 to positive infinity)].

4. A "nonlinear" relationship between explanatory variables and the dependent variable; however, this represents a linear relationship between the logit (natural log of the odds of the dependent variable occurring or not) and the set of explanatory variables.

5. Uses a maximum-likelihood rather than least-squares statistical model. In least squares, we select regression coefficients that result in the smallest sum of squared differences between the observed and the predicted values of the dependent variable. In maximum-likelihood, the coefficients that make our observed results "most likely" are selected.

6. Residuals follow a binomial rather than a normal distribution. Normality of variables is not a stringent requirement.

7. Does not assume homoscedasticity.

8. Assumes that there is little or no multicollinearity.

3.3.3 Multilevel Logistic Regression Model

3.3.3.1 Overview

In medical research, the environmental dimension has been neglected in favor of an individual-centered approach. In recent years, the question of whether and how environmental factors could have impact on health has been increasingly explored. For

example, the influence of the social environment on individual health outcome was reported about low birth weight, diastolic blood pressure, or all-cause mortality. Until recently, it was necessary to choose between the individual-centered and the collective-centered (also called ecological) approach for methodological reasons. In the collective approach, therefore, spatial analytical methods and geographical information systems explore diseases at a supra-individual aggregated level. Numbers of cases and prevalence or incidence rates are related to geographical units, and ecological exposure estimations for comparative or predictive purposes are composed on the same scale. In parasitological field research, these methods are increasingly used, notably for malaria.

Although useful, spatial analytical methods could reduce the scope of an investigation since exposure and characteristics of each of the individuals are not taken into account. Indeed, the origin of variation between areas could be explained by a complex combination of factors which are characterizing people (the individual level) or areas (the group level). When an individual factor is a characteristic of subjects who are more likely to be ill, variability of its distribution across areas will influence health outcomes in a given area: this is called a composition effect. So, relationships between individual and 'supra-individual' determinants are of particular interest, especially for investigating the reasons of variation between areas: are the people living in the areas different or are the areas different, i.e. is it a composition or a context effect?

Multilevel logistic regression is used in the analysis of the variability at each level separately. We do not use multilevel analysis, when the levels of the model are crossed instead of nested, when we have to systematically explore a large number of models, and when a less sophisticated technique to solve the research question, is available.

An Example: Why Use a Multilevel Model?

Suppose the data has n clusters with T observations in each cluster. The dependent variable for the i^{th} cluster is y_i and we are interested in knowing what is the mean, μ_i, of the dependent variable in each cluster, $i = 1,2,...,n$. Then $\overline{y}_i = \sum_{t=1}^{T} \frac{y_{it}}{T}$ can be calculated as an

unbiased estimate of μ_i , with standard error $se(\overline{y}_i) = \sqrt{\frac{\text{var}(y_i)}{T}}$ for the i^{th} cluster. In this case the standard error increases to the extent that the clusters are heterogeneous. But there are other estimators available. We could pool units, and estimate μ_i with the grand mean $\overline{y} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it}}{nT}$. This estimator is based on all the data and so has a small standard error. But this grand mean, $\overline{y}$, will be a biased estimator of any of the cluster-specific means. The bias increases to the extent that the clusters are heterogeneous. On the other hand, if the clusters are less heterogeneous or there is little between-cluster variations, then estimating the mean of each cluster by the grand mean is worthwhile.

Between these two estimators one might prefer the pooled estimator on mean square error grounds (trading off some bias for small standard errors), depending on the relative size of the between-cluster variation to the within-cluster variation. We can see, at one extreme that we have no pooling and zero bias in the estimates of μ_i. But these estimates could have large standard errors, for instance, if the within-cluster variation is large (if T is small). At the other extreme, we can see that we have complete pooling and potentially lots of bias if the between-cluster variation is large, but potentially big gain in efficiency (if n is large relative to T).

Thus, choosing any one of the above estimators results in either biased or inefficient estimation. Now, the question is whether there is any estimator that makes $\hat{\mu}_i$ close to the no-pooling estimator if the between-cluster variation is large, but close to the grand-mean if the between-cluster variation is small. The answer to the above question is yes, which is known as a random effect estimator. The estimator dominates the fixed effects (no pooling estimator) in terms of mean-square error, by buying just the right amount of bias relative to the efficiency gain, cluster-by-cluster. The related model of this fixed effect estimator is a multilevel model through which an unbiased and efficient estimate can be obtained.

3.3.3.2 Multilevel Linear Model

The multilevel linear model and its application had been described by various authors [Mason et al (1983), Goldstein (1987, 1995, and 2003), Bryk et al (1992)]. We describe below the multilevel linear model and its basic properties. We first consider a simple linear model for the data with hierarchical structure (with two levels) with a single explanatory variable,

$$y_{ij} = \alpha_j + \beta_1 x_{ij} + e_{ij} \qquad (3.1)$$

where y_{ij} is the outcome variable for the i^{th} unit at level-1 and the j^{th} unit at level-2, α_j is the intercept for the j^{th} unit at level-2 (i.e. it varies across level-2 but has the same value for all the units within each level-2), x_{ij} is the explanatory variable for the i^{th} unit at level-1 and the j^{th} unit at level-2, β_1 is the effect of x_{ij} and e_{ij} is the level-1 random effect. Here, α_j is a random variable rather than a constant and can be written as

$$\alpha_j = \beta_0 + u_j \qquad (3.2)$$

where, β_0 is the intercept (constant across level-2) and u_j is a random effect accounting for the random variation at level-2. Combining both equations (3.1) and (3.2) the two level linear model can be written as

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij} \qquad (3.3)$$

In equation (3.3), u_j and e_{ij} are random quantities which follow normal distributions, $N(0,\sigma_u^2)$ and $N(0,\sigma_e^2)$, respectively, and $Cov(u_j, e_{ij}) = 0$, $Cov(u_j, u_{j'}) = 0$, $j \neq j'$

In this model, β_0 and β_1 are known as fixed parameters. Equation (3.3) is also known as variance component model since the variance of the response is

$$Var(y_{ij} / \beta_0, \beta_1, x_{ij}) = Var(u_j + e_{ij}) = \sigma_u^2 + \sigma_e^2$$

which is the total variation obtained summing level-1 and level-2 variance. The covariance between two units of level-1 (say, i_1, i_2) can be defined as,

$$Cov(u_j + e_{i_1 j}, u_j + e_{i_2 j}) = \sigma_u{}^2$$

The within level-2 or intra-level 2 correlation after controlling the explanatory variable can be obtained from

$$\rho = \frac{\sigma_u{}^2}{\sigma_u{}^2 + \sigma_e{}^2}$$

3.3.3.2.1 The General linear 2-level model

Let there be S level-1 predictors x_{sij} (s=1,2,…, S). Then, the level-1 model is given by:

$$y_{ij} = \beta_{0j} + \sum_{s=1}^{S} \beta_{sj} x_{sij} + \varepsilon_{ij} \qquad (1)$$

Further, assume that there are Q level-2 predictors, z_{qj} (q = 1,2,…, Q). Then, the level-2 model for the intercept is given by:

$$\beta_{0j} = \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} z_{qj} + \delta_{0j}, \qquad (2)$$

and the level-2 model for the slopes is given by:

$$\beta_{sj} = \gamma_{s0} + \sum_{q=1}^{Q} \gamma_{sq} z_{qi} + \delta_{sj} \qquad (3)$$

Substitution of equations (2-3) into equation (1) gives the general linear 2-level model:

$$y_{ij} = \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} z_{qi} + \sum_{s=1}^{S} \gamma_{s0} x_{sij} + \sum_{q=1}^{Q} \sum_{s=1}^{S} \gamma_{sq} z_{qi} x_{sij} + \delta_{0j} + \sum_{s=1}^{S} \delta_{sj} x_{sij} + \varepsilon_{ij}$$

Note that the three models (empty model, random intercept model and random coefficients model) are the special cases of the general linear 2-level model. For instance, the random coefficients model is given by $y_{ij} = \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} z_{qj} + \delta_{0j} + \varepsilon_{ij}$.

The level-2 predictors and cross-level interactions disappear from the general multilevel model, but the disturbances remain heteroscedastic and serially correlated.

3.3.3.3 Multilevel Logistic Regression Model

3.3.3.3.1 Introduction to Multi-level Regression Model

The statistical models used in multilevel analysis are known as hierarchical or multilevel models. These models have previously appeared in different literature under a variety of names including random effect models or random coefficient models [Diggle et al (1994)], covariance components models or variance components models [Searle et al (1992)] and mixed models [Brown et al (2000)].

Multilevel analysis is applicable to a broad range of situations where units at a lower level are nested within units at a higher level. The simplest multilevel model considers only two levels of analysis. The first and most elementary of these levels is usually referred to as level-1 and it is this level that the analysis is focused on. The remaining level is referred to as level-2 and provides the context for the level-1 units. For instance, level-1 units could be voters who are nested in different counties (level-2 units). The dependent variable is always measured for level-1 units, since this is the primary level of analysis. This model can be conceptualized as a two-stage system of equations in which the individual variation within each group of level-2 is explained by an individual-level equation, and the variation across groups in the group-specific regression coefficients is explained by a group-level equation.

There are several straightforward ways to analyze such data where individuals are nested within groups. But these analyses do not take hierarchical effect into consideration. The first is to ignore group identity and focus exclusively on inter-individual variation and on individual level attributes. This approach has the drawback of ignoring the potential importance of group level attributes in influencing individual level outcomes. Besides, if outcomes for individuals within groups are correlated, the assumption of independence of observations is violated, which results in incorrect standard errors and hence inefficient estimates.

A second option is to focus exclusively on inter-group variation and on data aggregated to the group level. This approach considers the correlation between individuals within group but has the drawback of ignoring the role of individual level variables in shaping the outcome.

A third approach is to define separate regressions for each group. This approach allows regression coefficients to differ from group to group, but does not examine how specific group-level properties may affect individual-level outcomes or interact with individual-level variables. In addition, it is not practical when dealing with large numbers of groups or small numbers of observations per group.

A fourth approach is to include group identity in individual-level equations in the form of dummy variables. This approach is analogous to fitting separate regressions for each group and does not allow examination of exactly what group characteristics may be important in explaining the outcome. Besides, this approach treats the groups as unrelated.

Multilevel analysis differs from the approaches mentioned above in the sense that, first: it allows the simultaneous examination of the effects of group level and individual level predictors, second: the non-independence of observations within groups is accounted for, third: groups are not treated as unrelated, but are seen as coming from a larger population of groups, fourth: both inter-individual and inter-group variation can be examined. Thus, multilevel analysis allows dealing with the micro-level of individuals and the macro-level of groups simultaneously.

We shall start considering first a two level logistic regression model with a single explanatory variable.

3.3.3.3.2 Two Level Model with Single Explanatory Variable

Basically, the two level logistic model is equivalent to model (3.3) except for the outcome variable. Let y_{ij} be the binary outcome variable, coded '0' or '1', associated

with level-1 unit i nested within level-2 unit j . Also let p_{ij} be the probability that the response variable equals 1, and $p_{ij} = P_r(y_{ij} = 1)$. Here y_{ij} follows a Bernoulli distribution. Like logistic regression the p_{ij} is modeled using the logit link function. The two level logistic regression model can be written as,

$$\ln(\frac{p_{ij}}{1-p_{ij}}) = \beta_0 + \beta_1 x_{ij} + u_j \qquad (3.3)$$

where u_j is the random effect at level-2. Without u_j, Eq.(3.3) can be considered as a standard logistic model. Therefore, conditional on u_j, the y_{ij}'s can be assumed to be independently distributed. Here u_j is a random quantity and that follows $N(0,\sigma_u^2)$.The model (3.3) can be written as follows splitting up into two models: one for level-1 and the other for level-2.

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_1 x_{ij} \qquad \text{(model: level 1)}$$

$$\beta_{0j} = \beta_0 + u_j \qquad \text{(model: level 2)}$$

3.3.2.3.3 General Logistic two level Model

Recent developments in multilevel analysis now permit for nonlinear model specifications and this opens the door to modeling discrete responses. The simplest multilevel model for discrete responses is that for binary variables. The most common model for such variables is the multilevel logistic model, which modifies the linear multilevel model by specifying the logit link function (Goldstein 1991, 1995). Thus, the outcome of interest is the proportion of cases, P_{ij}, that fall into category 1 of the binary outcome measure and the multilevel model for this proportion can be written in terms of the log-odds ratio:

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right) = \gamma_{00} + \sum_{q=1}^{Q}\gamma_{0q}z_{qj} + \sum_{s=1}^{S}\gamma_{s0}x_{sij} + \sum_{q=1}^{Q}\sum_{s=1}^{S}\gamma_{sq}z_{qj}x_{sij} + \delta_{0j} + \sum_{s=1}^{S}\delta_{sj}x_{sij}$$

Statistical Assumptions

There is no single definitive set of assumptions that apply to all multilevel logistic models. The primary assumptions relevant to multilevel models involving binary outcomes using the logit link function are:

(a) the probability of success Pr $(y_{ij} = 1)$ is identical for individuals within clusters

(b) observations between clusters are independent, whereas pairs of observations within clusters have a common correlation

(c) each random effect is independent and follows a generalized distribution that can be estimated using maximum likelihood

(d) random effects and model predictors at all levels are independent, and

(e) an appropriate model linking y_i and u_i exists with a joint probability density function.

3.3.3.4 Comparison between multilevel and conventional modeling

Conventional logistic regression is performed on a single level of organization (individuals). Neither the region level, nor the correlated structure of the data is considered. Consequently, variability of coefficients across regions (higher level) is not allowed by the modeling process, i.e. the random part defined at the region level (higher level) does not exist. Coefficient values (and odds ratio) are relatively close to those estimated by multilevel modelling. The main difference lies in smaller standard errors in the conventional logistic regression.

CHAPTER FOUR

Statistical Data Analysis

Statistical Data Analysis Using Conventional Logistic Regression

4.1 Introduction

The purpose of this chapter is to measure the effect of the different demographic and health and socio-economic variables on the extent of "malaria infection status" on women at national level. An attempt has been made to identify the most significant variables in determining the extent of "malaria infection status" in the region. The response variable, malaria infection status, is binary assuming only two values 0 and 1. That is, 0 if a respondent is malaria negative (non-malarious) and 1 if a respondent is malaria positive (malarious). In this particular study the logistic regression is used to see the relationship between the proposed independent variables and the response variable. In logistic regression there is no assumption of normality instead we have the assumption of binomial variability. In this study the univariate and multivariate analysis are made. We start our data analysis by giving the summary statistics for the variables considered in the study; we then proceed to the univariate analysis and complete the final model in the multivariate analysis.

4.2 Summary Statistics

This particular study is based on 7333 women of age 15 to 49 years. Of these respondents 6993 (95.4%) are malaria negative (non-malarious) and 340 (4.6%) are malaria positive (malarious). For the sake of simplicity, it is better to classify independent variables with respect to socio-economic, demographic and health factors.

4.2.1 Socio-Economic factors and malaria infection status

From Table 4.1 higher proportion of malaria infection is observed in Somali, Gambela, Tigray and Ben-Gumz regions (8.1%, 7.5%, 6.7% and 6.4%, respectively). But Oromia, Dire Dawa and Harari regions take less contribution to the malaria infection status of women (1.9%, 1.8% and 1.0%, respectively) as compared to other regions of Ethiopia. Although there were few malarious women in Addis Ababa, they were rejected when

data was adjusted for analysis. The rural women residents show higher proportion of being malaria positive (malarious) than urban women residents (5.7% against 1.8%). Those women who are living in the city are at a lower risk of being malarious than town and countryside women residents (1.7%, 2.1% and 5.7% respectively). Of the women who lives in a house with natural floor material (5.2%) are found to be malarious than those women who lives in cement floor house (2.1%). It is apparent from the sample data that women who live in a house made of sheet of corrugated iron roof have relatively less chance to be attacked by malaria (2.7% against 5.8%). Poor women are the most exposed group among the wealth status of women in Ethiopia. It can be seen from the data that a woman who have awareness about malaria through information (radio) have a better chance to prevent malaria than a woman who do not have information (3.2% against 5.7%).

Table 4. 1: Socio-economic factors and distribution of malaria on women in Ethiopia.

Factors	Category	Number Of Malarious women	Percentage of malarious women
Region	Somali	52	8.1
	Tigray	37	6.7
	Afar	43	5.5
	Amhara	25	5.6
	Oromia	18	1.9
	Ben-gumz	48	6.4
	SNNP	42	4.2
	Gambela	55	7.5
	Harari	12	1.8
	Addis Ababa	0	0
	Dire Dawa	8	1.0
Type of place of residence	Rural	304	5.7
	Urban	36	1.8

De facto place of residence	countryside	304	5.7
	town	13	2.1
	City	23	1.7
Main floor material	Natural	312	5.2
	Cement/bricks	28	2.1
Main roof material	Natural	268	5.8
	Corrugated iron	72	2.7
Wealth index	Poor	217	7.1
	Middle	59	3.1
	Rich	64	2.7
Has radio	No	242	5.7
	Yes	98	3.2

4.2.2 Demographic and health factors and malaria infection status

From Table 4.2 we note that malaria infection proportions appear to be almost equal for the youngest and oldest age groups (4.3% and 4.2%, respectively). But it is higher for women in their 20's (5.6%). Malaria infection proportion is lower as a woman's educational level (status) increases (5.4%, 3.7% and 2.5% for women not educated, primary education, secondary and above, respectively). Pregnant women (with malaria infection proportion (10.9%) are remarkably more susceptible to malaria disease than those who are not pregnant (4.1%). About 4.2% of the women, who do not have bednet, were malaria positive (malarious), but this proportion is 4.7% for those who have bednet.

Table 4.2: Demographic and health factors and distribution of malaria on women in Ethiopia.

Factor	category	Number of malaria positive (malarious) women	Percentage of Malaria positive (malarious) women
Age (in years)	21-29	123	5.6
	30-49	124	4.2
	15-20	93	4.3
Educational level	No education	259	5.4
	Primary	54	3.7
	Secondary and above	27	2.5
Currently pregnant	Yes	64	10.9
	No	273	4.1
Have Bednet	Yes	286	4.7
	No	54	4.2

4.3 Univariate analysis

To determine the factors which are significantly correlated with the dependent variable, a preliminary assessment was done using the chi-squared test. Since the Pearson chi-squared test is asymptotically equivalent to the likelihood ratio chi-square test, it can also be used to test the significance of univariate relationships. Independent variables selected for the study were strongly associated with the dependent variable, malaria infection status, as shown in Table 4.3. We note that all independent variables are significant except the variable "have bednet for sleeping". Here the p-value used as a criterion for significance is 0.05. Table 4.3 summarizes the findings of the univariate analysis.

Table 3: Variables in the univariate analysis

Variable	Pearson chi-square	df	p-value (Assymptotic)
Age	6.797	2	.033*
Region	96.290	9	.000*
Type of place of residence	50.907	1	.000*
Educational level	21.452	2	.000*
Main floor material	23.035	1	.000*
Main roof material	36.676	1	.000*
Currently pregnant	59.453	1	.000*
Have bednet for sleeping	.741	1	.389
Wealth index	72.022	2	.000*
De facto place of residence	51.078	2	.000*
Has radio	24.619	1	.000*

*** Significant ($p<0.05$)**

4.4 Multivariate Analysis

The main problem with the univariate approach is that it ignores the possibility that a collection of variables, each of which is weakly associated with the outcome, can become an important predictor of the outcome when taken together (Hosmer and Lemshow, 1989). Using this method, the model that best describes the dependent variable, malaria infection status, is fitted using the explanatory variables. The stepwise forward likelihood method is used to select the best model. The result of logistic regression model is given in table 4.5. The final (optimal) logistic regression model includes only significant variables.

Table 4.4

Categorical Variables Codings

		Frequency	Parameter coding								
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(1)
Region	Somali	643	1.000	.000	.000	.000	.000	.000	.000	.000	.000
	Tigray	553	.000	1.000	.000	.000	.000	.000	.000	.000	.000
	Afar	787	.000	.000	1.000	.000	.000	.000	.000	.000	.000
	Amhara	450	.000	.000	.000	1.000	.000	.000	.000	.000	.000
	Oromiya	936	.000	.000	.000	.000	1.000	.000	.000	.000	.000
	Ben-Gumz	755	.000	.000	.000	.000	.000	1.000	.000	.000	.000
	SNNP	1001	.000	.000	.000	.000	.000	.000	1.000	.000	.000
	Gambela	729	.000	.000	.000	.000	.000	.000	.000	1.000	.000
	Harari	674	.000	.000	.000	.000	.000	.000	.000	.000	1.000
	Dire Dawa	805	.000	.000	.000	.000	.000	.000	.000	.000	.000
De facto place of residence	Countryside	5320	1.000	.000							
	Town	626	.000	1.000							
	City	1387	.000	.000							
Highest educational level	No education	4771	1.000	.000							
	Primary	1468	.000	1.000							
	Secondary and above	1094	.000	.000							
Age of the respondent	21-29	2192	1.000	.000							
	30-49	2987	.000	1.000							
	15-20	2154	.000	.000							
Wealth index	Poor	3061	1.000	.000							
	Middle	1875	.000	1.000							
	Rich	2397	.000	.000							
Main roof material	NATURAL	4647	1.000								
	Corrugated iron	2686	.000								
Type of place of residence	Rural	5320	1.000								
	Urban	2013	.000								
Have bednet for sleeping (household report)	Yes	6041	1.000								
	No	1292	.000								
Has radio	No	4269	1.000								
	Yes	3064	.000								
Currently pregnant	Yes	615	1.000								
	No	6718	.000								
Main floor material	NATURAL	6013	1.000								
	Cement	1320	.000								

Based on these coding schemes given in the above table the logistic regression coefficients can be estimated using the maximum likelihood estimation method. This will be done using SPSS package.

The variables that are found to be significant in the multivariate analysis are: "region", "currently pregnant", "wealth index", "type of place of residence", "main floor material" and "age". The significance of the Wald statistic (under the column with heading Sig) indicates the importance of the predictor variables in the model. A high value of the Wald statistic shows that the corresponding predictor variable is significant.

Table 4.5 Final Logistic Regression Model

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 6(f)	AGE			6.271	2	.043	
	AGE(21-29)	.195	.144	1.821	1	.177	1.215
	AGE(30-49)	-.139	.143	.952	1	.329	.870
	REGION			51.368	9	.000	
	Somali	1.763	.407	18.710	1	.000	5.828
	Tigray	1.602	.418	14.723	1	.000	4.964
	Afar	1.355	.412	10.818	1	.001	3.878
	Amhara	1.536	.434	12.533	1	.000	4.644
	Oromia	.465	.444	1.098	1	.295	1.593
	Ben-Gumz	1.711	.409	17.461	1	.000	5.532
	SNNP	1.272	.411	9.593	1	.002	3.568
	Gambela	1.801	.405	19.788	1	.000	6.054
	Harari	.575	.461	1.556	1	.212	1.777
	TRESDC(Rural)	1.066	.282	14.279	1	.000	2.905
	FLOOR (Natural)	-.626	.309	4.108	1	.043	.535
	PREG(Yes)	.873	.148	34.969	1	.000	2.393
	WLTH			20.242	2	.000	
	WLTH(Poor)	.147	.211	.489	1	.484	1.159
	WLTH(Middle)	-.568	.230	6.091	1	.014	.567
	Constant	-4.746	.378	158.028	1	.000	.009
a Variable(s) entered on step 1: REGION.							
b Variable(s) entered on step 2: PREG.							

c Variable(s) entered on step 3: WLTH.
d Variable(s) entered on step 4: TRESDC.
e Variable(s) entered on step 5: FLOOR.
f Variable(s) entered on step 6: AGE.

4.4.1 Model Checking and Diagnosis

A model should be assessed and diagnosed for model adequacy. Using SPSS package, the output on chi-square goodness of fit statistic and the classification power for the criterion variable (malaria infection status) are shown in the tables given below. A good model is the one that results in the high likelihood of the observed results. This translates to small values of -2LL. If a model fits perfectly, the likelihood is 1 (meaning -2LL is 0).

4.4.1.1 Goodness of fit of the model

i. The columns in the classification table are the two predicted values of the dependent variable, while the rows are the two observed (actual) values of the dependent variable. If the logistic model has homoscedasticity (not a logistic regression assumption), the percent correct will be approximately the same for both rows. The classification table given below shows that, of the 7333 women respondents included in the analysis, 91.1 % were correctly classified. This result shows that the final logistic model performs well with respect to the cut value 0.100.

Table 4.6 Classification Table

Observed		Predicted		
		Malaria Distribution		Percentage Correct
		non-malarious	malarious	
Malaria Distribution	non-malarious	6621	372	94.7
	malarious	281	59	17.4
Overall Percentage				91.1

The cut value is .100

ii. The Omnibus Tests of Model Coefficients gives us a chi-square of 197.199 with p-value 0.000 on 16 df. This is a test of the null hypothesis that adding all the independent variables to the model has not significantly increased our ability to predict the dependent variable (malaria infection status) on women in Ethiopia. So we reject the null hypothesis according to the SPSS output given in Table 4.7; that is our final model fits significantly better than the model with only the intercept (empty model).

Table 4.7 Omnibus Tests of Model Coefficients

	Chi-square	df	Sig.
Model	197.199	16	.000

*Significant (P<0.05)

iii. The -2 Log Likelihood (-2LL) statistic measures the power of the model to predict the dependent variable (malaria infection status). The smaller the value of statistic the better the model is. The -2LL statistic for logistic model that contains only the constant is 2964.447 (see appendix). The goodness of fit statistic for the final logistic regression model given under the model summary as shown in the Table 4.8 indicates that the value of -2 Log Likelihood is 2555.200. It was used to compare the fit of the final model with the empty model. Thus from the above two values we observe that the final model predicts the dependent variable better than the empty model.

Table 4.8 Model Summary

-2 Log likelihood	df	Sig.
2555.200	16	.000*

*Significant (P<0.05)

iv. The Hosmer-Lemshow test is also another way of assessing goodness of fit of the model. Well-fitting models show non-significance on the Hosmer and Lemshow goodness-of-fit test, indicating model prediction is not significantly different from

observed values. As it is indicated from Table 4.9 we do not reject the null hypothesis (the model fits) at α= 0.05 level of significance. This shows that there is no sufficient evidence to reject the null hypothesis and it confirms that the model has a good fit.

Table 4.9: Hosmer-Lemshow Test

Chi-Square	df	Sig.
7.318	8	.503

From the above discussion under (i)-(iv) we can conclude that the fitted model is statistically satisfactory.

Note that: From the analysis of conventional logistic regression the classification table has a power to classify data only at a cut value of .100. But when we observe the data region-wise (for the regions Somali, Tigray, Afar and Ben-Gumz) the classification power increases even at a cut value of .500 as shown in the table given below.

Classification Table					
Region	Observed		Predicted		
			malaria distribution		Percentage Correct
			non-malarious	malarious	
Somali					
	malaria distribution	non-malarious	543	48	91.9
		malarious	39	13	25.0
	Overall Percentage				86.5
Tigray					
	malaria distribution	non-malarious	511	5	99.0
		malarious	30	7	18.9
	Overall Percentage				93.7
Afar					
	malaria distribution	non-malarious	683	61	91.8
		malarious	35	8	18.6
	Overall Percentage				87.8
Ben-Gumz					
	malaria distribution	non-malarious	654	53	92.5
		malarious	35	13	27.1
	Overall Percentage				88.3
The Cut-value is .500					

Statistical Data Analysis Using Multilevel Logistic Regression Model

4.1 Introduction

Malaria is influenced by a web of individual and environmental factors. For a long time analysing these factors concurrently has raised statistical problems. Multilevel modeling provides a new attractive solution, which is still uncommon in tropical medicine. The principle of multilevel modeling is to analyse simultaneously the influence of the factors considered so far. The data set is structured as a succession of nested levels: people live in house, houses are found in villages, villages are found in region etc. Outcomes defined at the lowest level (parasite burden of individuals) are then modeled as a function of variables characterizing the different levels (people, house, village, and region). The aim of this chapter is to demonstrate whether there exists variation in malaria infection among the regions.

A two-level modeling process is presented by using the data set of 7333 women from 10 regions of Ethiopia. The hierarchical organization of a data set in levels, fixed and random effects, are considered.

From a computational point of view, multilevel modeling can be seen as a two-stage process. First, a separate individual level regression is defined for each region. Then, each of the region-specific coefficients is modeled as a function of region variables. So, multilevel analysis allows the partition of the region-specific coefficients: a fixed part that is common across regions and a random part varying among regions.

A chi-square test statistic was applied to assess heterogeneity in the proportion of malaria positive women among the 10 regions. The test yields χ^2=96.290, df = 9, P<0.01. Thus, there is evidence for heterogeneity among the regions with respect to malaria infection status among women in Ethiopia.

4.2 Modeling Process

4.2.1 Intercept-Only Multilevel Logistic Model

We first estimate a model with no predictors (an intercept-only model) that predict the probability of malaria infection status among women. The region level (level 2) variance estimation is 0.4077 as shown in Table 4.11. This variance reflects between-region heterogeneity, regarding malaria infection status of women in Ethiopia. Table 4.11 also provides a deviance-based chi-square test for assessing the goodness of fit of the fitted empty model. The test indicates that the fitted model is good.

Table 4.11 Estimates for empty model

Parameter		**Estimate**	Standard error	**Z-value**	**P-value**
Fixed part	Intercept	3.1230	0.0558	55.9230	0.0000*
		Estimate	Standard error	**Z-value**	**P-value**
Random part	**Level-two variance** ($\delta_{0j}{}^2$)	0.4077	0.1366	2.9837	0.0028*
Deviance-based chi-square		107.3371			0.0000*

*Significance (P<0.01)

4.2.2 **Random Intercept Model and Fixed Explanatory Variables**

In order to identify the effect of some selected explanatory variables a multilevel logistic regression model with random intercept and fixed explanatory variables was estimated using LISREL software and the results are presented in Table 4.12 given below. The deviance based chi-square test for significance of random effects (χ^2 =86.3763 df =1 and P<0.05) indicates that the random intercept model with the fixed explanatory variables is found to be a better fit as compared to the empty model discussed in Section 4.2.1.

From Table 4.12 it is possible to observe that the malaria infection among women varied among the ten regions of Ethiopia. Moreover the explanatory variables, "age", "currently pregnant" and "wealth index" were found to be significant determinants of variation in malaria infection across the regions. As regard to regional difference at level-two variance of the random intercept (σ_0^2) was found to be significant, implying that there exists a remarkable regional difference with respect to the extent of malaria infection status of women.

Table 4.12: Estimates of Random intercept model

Covariates	Estimate	S.E.	Z-value	P-value
Intercept	1.8813	0.1424	13.2117	0.0000*
Age				
15-29	0.2491	0.1171	2.1272	0.0344*
30-49 (Ref.)				
Main floor material				
Natural	-0.1122	0.2191	-0.5121	0.6086
Cement (Ref.)				
Currently pregnant				
Yes (Ref.)				
No	0.9107	0.1461	6.2347	0.0000*
Wealth Index				
Poor (Ref.)				
Rich	0.7219	0.1271	5.6820	0.0000*
Random Part	**Estimate**	**S.E.**	**Z-value**	**P-value**
Random Intercept: $\sigma_0^2 = \text{var}(\delta_{0j})$ intercept variance	0.2611	0.0895	2.9184	0.0035*
Deviance-based chi-square	86.3763			0.0000*

* Significant (P<0.05)

4.2.3 Random Coefficients Model

The random coefficient model is useful because it shows the degree of variability at each level. In Table 4.13 we represent each of the given explanatory variables as age (X_1), main floor material (X_2), currently pregnant (X_3) and wealth index (X_4). The table includes fixed effect coefficients and an overall (level-2) or regional variance constant

term (σ_0^2) together with variance and covariance terms representing the random effects of the respective explanatory variables and their interactions. The significance of these terms is indicated in Table 4.13. According to the overall region variance constant term, the variance of each explanatory variable is found to be significant. Moreover, "age by main floor material" and "main floor material by wealth index" covariance terms are found to be statistically significant. Similarly, the fixed effects currently pregnant and wealth index are found to be significant. Accordingly, the results of the multilevel analysis showed that all explanatory variables shown in the table contribute in explaining the variation of malaria infection across the regions except the factors "main floor material" and "age". The deviance based chi-square test for significance of random effects (χ^2 = 62.4042, df =15) indicates that the random coefficient is statistically significant.

We note that the student version of the software LISREL does not provide model diagnostic tests. We only checked a goodness of fit by using deviance based chi-squared test.

Table 4.13

Results for Fixed and Random Effects of Random Coefficient Model

Fixed Part	**Estimate**	**S.E.**	**Z-value**	**P-value**
Intercept (X_0)	1.8930	0.1427	13.2660	0.0000*
Age (X_1)				
15-29	0.2152	0.1170	1.8388	0.0659
30-49 (Ref.)				
Main floor material (X_2)				
Natural	0.1003	0.2191	0.4575	0.6473
Cement (Ref.)				
Currently pregnant (X_3)				
Yes (Ref.)				
No	0.9228	0.1463	6.3068	0.0000*

$\sigma_4^2 = \text{var}(\delta_{4j})$	0.0227	0.0173	1.3131	0.1891
$\sigma_{01} = \text{cov}(\delta_{0j}, \delta_{1j})$	-0.0532	0.0351	-1.5158	0.1296
$\sigma_{02} = \text{cov}(\delta_{0j}, \delta_{2j})$	-0.2175	0.1493	-1.4572	0.1451
$\sigma_{03} = \text{cov}(\delta_{0j}, \delta_{3j})$	-0.0347	0.0557	-0.6227	0.5335
$\sigma_{04} = \text{cov}(\delta_{0j}, \delta_{4j})$	0.0024	0.0289	0.0822	0.9345
$\sigma_{12} = \text{cov}(\delta_{1j}, \delta_{2j})$	0.1313	0.0616	2.1293	0.0332*
$\sigma_{13} = \text{cov}(\delta_{1j}, \delta_{3j})$	-0.0306	0.0223	-1.3718	0.1701
$\sigma_{14} = \text{cov}(\delta_{1j}, \delta_{4j})$	-0.0191	0.0134	-1.4241	0.1544
$\sigma_{23} = \text{cov}(\delta_{2j}, \delta_{3j})$	-0.0612	0.0804	-0.7614	0.4464
$\sigma_{24} = \text{cov}(\delta_{2j}, \delta_{4j})$	-0.1035	0.0530	-1.9522	0.0430*
$\sigma_{34} = \text{cov}(\delta_{3j}, \delta_{4j})$	0.0178	0.0190	0.9398	0.3473
Deviance-based chi-square	62.4042			0.0000*

Ref. =Reference category

Table 4.14: Level-2 covariance matrix of the random coefficient

	X_0	X_1	X_2	X_3	X_4
X_0	0.353174				
X_1	-0.053231	0.036053			
X_2	-0.217510	0.131268	0.673823		
X_3	-0.034671	0.030588	--0.061189	0.074313	
X_4	0.002379	-0.019134	-0.098096	0.017817	0.022682

Table 4.15: **Level-2 Correlation Matrix of the Random Coefficient**

	X_0	X_1	X_2	X_3	X_4
X_0	1.000000				
X_1	-0.471734	1.000000			
X_2	0.445873	0.842205	1.000000		
X_3	-0.214014	-0.590950	-0.273445	1.000000	
X_4	0.026575	-0.669105	-.793476	0.433962	1.000000

Table 4.16: **Parameter estimates among the Regions**

		Coefficients				
Region		X_0	X_1	X_2	X_3	X_4
Somali	Estimate	-0.5303	0.08565	0.38072	0.05367	-0.0092
	S.E.	0.04514	0.01329	0.31772	0.04421	0.01511
	Z-value	-11.747*	6.44446*	1.19828	1.2139	-0.6112
Tigray	Estimate	-0.0993	0.23989	0.75159	-0.3684	-0.1359
	S.E.	0.03855	0.01431	0.35345	0.05011	0.01592
	Z-value	-2.57588*	16.7673*	2.12644*	-7.3523*	-8.5355*
Afar	Estimate	-0.0585	0.09706	0.50242	-0.0753	-0.1101
	S.E.	0.05254	0.01399	0.33736	0.04799	0.01554
	Z-value	-1.1142	6.93577*	1.48926	-1.5701	-7.0834*
Amhara	Estimate	-0.2575	0.09864	0.46249	-0.0313	-0.0598
	S.E.	0.07657	0.01601	0.37988	0.05666	0.01702
	Z-value	-3.3628*	6.16261*	1.21748	-0.5532	-3.5146*
Oromia	Estimate	0.90627	-0.3305	-1.5448	0.09128	0.18149
	S.E.	0.09429	0.01545	0.22113	0.05994	0.01304
	Z-value	9.61186*	-21.387*	-6.9858*	1.52295	13.9218*
Ben-Gumz	Estimate	-0.5967	-0.0181	-0.0133	0.19635	0.06937
	S.E.	0.05154	0.01259	0.29783	0.04582	0.01409
	Z-value	-11.577*	-1.4416	-0.0446	4.28533*	4.92223*
SNNP	Estimate	-0.2096	-0.0228	0.01165	0.12926	0.03351
	S.E.	0.0557	0.01135	0.23361	0.04578	0.01266
	Z-value	-3.7629*	-2.0112*	0.04986	2.82364*	2.64594*
Gambela	Estimate	-0.5052	0.04667	-0.2338	-0.0585	0.07512
	S.E.	0.05147	0.01134	0.24661	0.04627	0.01277
	Z-value	-9.816*	4.11488*	-0.9481	-1.2649	5.88159*
Harari	Estimate	0.67355	-0.1281	-0.4893	-0.006	0.03031
	S.E.	0.14476	0.0168	0.21549	0.06326	0.01217
	Z-value	4.65276*	-7.6234*	-2.2706*	-0.095	2.48969*
Dire Dawa	Estimate	0.6773	-0.0684	0.1723	0.06907	-0.0748
	S.E.	0.15148	0.01804	0.24858	0.06601	0.01321
	Z-value	4.47126*	-3.7903*	0.69313	1.0463	-5.6611*

*significant ($p<0.05$)

As can be seen from Table 4.16 in Tigray region, all the variables are found to be significant. But it is not true for the other regions.

4.3 Discussion and Interpretation of the results

Although most of the results obtained from the summary statistics matches with the findings of both univariate and multivariate analysis, it has some deviations with respect to certain variables. For example, it showed that women who have bednet for sleeping are at higher risk of malaria than those who do not have bednet (see Table 4.2). But the variable "have bednet for sleeping" is not significant in the multivariate analysis.

From the univariate analysis it can be seen that all explanatory variables under study are significant except the variable "have bednet for sleeping". As we proceed from the univariate analysis to the multivariate analysis 55% of the proposed explanatory variables are found to be significant (see Sections 4.3 and 4.4); of which 33.3% were demographic and health factors and 66.7% were socio-economic factors. These significant variables are factors affecting the malaria infection status among women in Ethiopia.

Let us now interpret the effect of each significant covariate on malaria infection status among women in Ethiopia using the estimated odds ratio given in the final logistic regression model (Table 4.5). Employing a 0.05 criterion of statistical significance, all covariates had significant effect on the malaria infection status on women except the two regions, Oromia and Harari, the two age groups (21-29 and 30-49 years) and one category of wealth index (poor women). The region variable was coded using Dire Dawa as the reference group. So from Table 4.5 it is possible to observe that the odds of women living in Somali and Gambela regions being malarious is almost 6 times higher than Dire Dawa. Also the odds of being malaria positive for Tigray, Amhara and Ben-Gumz women is almost 5 times higher than women living in Dire Dawa. Malaria infection status among women living in SNNP and Afar regions is higher than those women who are residing in Dire Dawa (odds ratio 3.6 and 3.9, respectively). Malaria infection rate for pregnant women is remarkably higher than for non-pregnant women. An amusing result of this study was that the malaria infection rate for rich women is 43% higher than the

middle class women. The study also showed that the odds of malaria infection status for rural women is 3 times higher than urban women.

Another important statistical analysis used in this study was multilevel logistic regression. In the multilevel analysis women are nested within the 10 regions in Ethiopia. Three multilevel models: empty model, random intercept model and random coefficient model were applied in order to explain regional differences in the extent of malaria infection status among women in Ethiopia. The results obtained are discussed as follows.

Before the analysis of data using the multilevel approach, first the heterogeneity of the malaria infection status among women with regard to regions was checked. The fixed part of the effects of explanatory variables included in the models have somewhat similar interpretation as that of the conventional logistic regression discussed above for the national level data. Whereas the random parts of the intercept and explanatory variables provided additional information.

In the three models (empty model, random intercept model and random coefficients model), the overall variance constant term found to be statistically significant which may again imply the differences in the malaria infection status among women. The effect of the random part of the variable "main floor material", on malaria infection status of women differs across regions. Similarly, the interaction of the random parts of "age" by "main floor material" and "wealth index" by "main floor material" provided significant effect on malaria infection status across regions.

CHAPTER FIVE

Conclusions and Recommendations

5.1. Conclusions

- This empirical study based on conventional logistic regression analysis reveals that the factors that affect the malaria infection status among women in Ethiopia are "currently pregnant", "age", "main floor material", "wealth index", "type of place of residence" and "region".

- The result of this study indicated that pregnant women are highly affected by malaria than the non-pregnant women. Also rural women are more exposed to malaria than urban women.

- Multilevel analysis enables the proper investigation of the effects of independent variables measured at different levels on the response variable"malaria infection status". As a result this study showed that there exist variations in malaria infection status among women across regions.

5.2. Recommendation

- Governmental and non-governmental organizations are expected to work hard on creating awareness on how/why to prevent women from being attacked by malaria.

- A special attention should be given to pregnant women in preventing them from being infected by malaria in Ethiopia.

- Great attention should be given to rural women in preventing them from being attacked by malaria.

- Primary health care and malaria prevention programs should be implemented in order to fit to the overall features of the regions to safeguard women from being attacked by malaria.

- Stakeholders should consider the existence of regional variations with respect to the malaria infection status among women in such a way that they might implement their task accordingly.

References

Breman, J.G. et al.,(2001). The Intolerable Burden of Malaria: A New Look at the Numbers, Supplement to the American Journal of Tropical Medicine and Hygiene, Volume 64, Number 1, 2.

Brown, H. and Prescott, R.(2000): Applied mixed models in medicine. Wiley: New York.

Bryk , A. and Raudenbush, S. (1992): Hierarchical Linear Models: Applications and Data Analysis Methods. Newbury Park, CA: Sage.

CARE Nepal, Child Survival Project (2003): Knowledge, Practice and Coverage Final Survey in Kanchanpur district, Nepal.

Diggle, P., Liang, K. and Zeger, S.(1994): Analysis of Longitudinal Data. Oxford University Press, New York.

Goldstein, H. (1987): Multilevel Models in Educational and Social Research. London: Griffin.

Goldstein, H. and Edward, A. (1995): Multilevel Statistical Models. 2nd Edition. London.

Goldstein, H. (2003): Multilevel Statistical Models. London: Arnold; New York: Oxford University Press Inc. 3rd ed.

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

Mason, W., Wong G. and Entwistle, B. (1983): Contextual analysis through the multilevel linear model. Sociol. Methodol., 13:72–103.

Nigatu, W. et al. (1992): Plasmodium vivax and P. falciparum epidemiology in Gambella, south-west Ethiopia

Searle, S., Casella, G. and McCullogh, C. (1992): Variance Components. Wiley: New York.

UN Millennium Project (2005): Coming to grips with Malaria in the New Millennium, London.

WHO (1999). The Community-Based Malaria Control Programme in Tigray, Northern Ethiopia: a Review of Programme Set-up, Activities, Outcomes and Impact, Malaria Control Department, Health Bureau, Tigray, Ethiopia.

LOGISTIC REGRESSION OUTPUTS

Case Processing Summary

Unweighted Cases(a)		N	Percent
Selected Cases	Included in Analysis	7333	100.0
	Missing Cases	0	.0
	Total	7333	100.0
Unselected Cases		0	.0
Total		7333	100.0

a If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
non-malarious	0
malarious	1

Categorical Variables Codings

		Frequency	Parameter coding								
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(1)
Region	Somali	643	1.000	.000	.000	.000	.000	.000	.000	.000	.000
	Tigray	553	.000	1.000	.000	.000	.000	.000	.000	.000	.000
	Afar	787	.000	.000	1.000	.000	.000	.000	.000	.000	.000
	Amhara	450	.000	.000	.000	1.000	.000	.000	.000	.000	.000
	Oromiya	936	.000	.000	.000	.000	1.000	.000	.000	.000	.000
	Ben-Gumz	755	.000	.000	.000	.000	.000	1.000	.000	.000	.000
	SNNP	1001	.000	.000	.000	.000	.000	.000	1.000	.000	.000
	Gambela	729	.000	.000	.000	.000	.000	.000	.000	1.000	.000
	Harari	674	.000	.000	.000	.000	.000	.000	.000	.000	1.000
	Dire Dawa	805	.000	.000	.000	.000	.000	.000	.000	.000	.000
De facto place of residence	Countryside	5320	1.000	.000							
	Town	626	.000	1.000							
	City	1387	.000	.000							
Highest educational level	No education	4771	1.000	.000							
	Primary	1468	.000	1.000							
	Secondary and above	1094	.000	.000							

Age of the respondent	**21-29**	2192	1.000	.000							
	30-49	2987	.000	1.000							
	15-20	2154	.000	.000							
Wealth index	**Poor**	3061	1.000	.000							
	Middle	1875	.000	1.000							
	Rich	2397	.000	.000							
Main roof material	**NATURAL**	4647	1.000								
	Corrugated iron	2686	.000								
Type of place of residence	**Rural**	5320	1.000								
	Urban	2013	.000								
Have bednet for sleeping (household report)	**Yes**	6041	1.000								
	No	1292	.000								
Has radio	**No**	4269	1.000								
	Yes	3064	.000								
Currently pregnant	**Yes**	615	1.000								
	No	6718	.000								
Main floor material	**NATURAL**	6013	1.000								
	Cement	1320	.000								

Block 0: Beginning Block

Classification Table(a,b)

			Predicted		
			malaria infectionstatus		Percentage Correct
	Observed		non-malarious	malarious	non-malarious
Step 0	malaria infectionstatus	non-malarious	6993	0	100.0
		malarious	340	0	.0
	Overall Percentage				95.4

a Constant is included in the model.

b The cut value is .100

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-3.024	.056	2964.447	1	.000	.049

Variables not in the Equation(a)					
			Score	df	Sig.
Step 0	Variables	AGE	6.797	2	.033
		AGE(1)	6.719	1	.010
		AGE(2)	2.684	1	.101
		REGION	96.290	9	.000
		REGION(1)	18.978	1	.000
		REGION(2)	5.708	1	.017
		REGION(3)	1.364	1	.243
		REGION(4)	.916	1	.339
		REGION(5)	17.867	1	.000
		REGION(6)	5.638	1	.018
		REGION(7)	.509	1	.475
		REGION(8)	15.482	1	.000
		REGION(9)	13.694	1	.000
		TRESDC(1)	50.907	1	.000
		EDUC	21.452	2	.000
		EDUC(1)	19.375	1	.000
		EDUC(2)	3.811	1	.051
		HRAD(1)	24.619	1	.000
		FLOOR(1)	23.035	1	.000
		ROOF(1)	36.676	1	.000
		PREG(1)	59.453	1	.000
		BNET(1)	.741	1	.389
		WLTH	72.022	2	.000
		WLTH(1)	71.481	1	.000
		WLTH(2)	12.647	1	.000
		DRESDC	51.078	2	.000
		DRESDC(1)	50.907	1	.000
		DRESDC(2)	10.144	1	.001
a Residual Chi-Squares are not computed because of redundancies.					

Block 1: Method = Forward Stepwise (Likelihood Ratio)

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	108.784	9	.000
	Block	108.784	9	.000
	Model	108.784	9	.000
Step 2	Step	38.017	1	.000
	Block	146.801	10	.000
	Model	146.801	10	.000
Step 3	Step	28.330	2	.000
	Block	175.131	12	.000
	Model	175.131	12	.000
Step 4	Step	11.820	1	.001
	Block	186.951	13	.000
	Model	186.951	13	.000
Step 5	Step	4.030	1	.045
	Block	190.981	14	.000
	Model	190.981	14	.000
Step 6	Step	6.218	2	.045
	Block	197.199	16	.000
	Model	197.199	16	.000

Model Summary			
Step	**-2 Log likelihood**	**Cox & Snell R Square**	**Nagelkerke R Square**
1	2643.615(a)	.015	.047
2	2605.598(a)	.020	.063
3	2577.268(a)	.024	.075
4	2565.448(a)	.025	.080
5	2561.418(a)	.026	.082
6	2555.200(a)	.027	.085
a Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.			

Hosmer and Lemeshow Test			
Step	**Chi-square**	**df**	**Sig.**
1	.000	8	1.000
2	2.371	8	.967
3	7.412	8	.493
4	12.243	7	.093
5	8.178	8	.416
6	7.318	8	.503

Contingency Table for Hosmer and Lemeshow Test						
		malaria infectionstatus = non-malarious		malaria infectionstatus = malarious		Total
		Observed	Expected	Observed	Expected	Observed
Step 1	1	797	797.000	8	8.000	805
	2	662	662.000	12	12.000	674
	3	918	918.000	18	18.000	936
	4	959	959.000	42	42.000	1001
	5	744	744.000	43	43.000	787
	6	425	425.000	25	25.000	450
	7	707	707.000	48	48.000	755
	8	516	516.000	37	37.000	553
	9	674	674.000	55	55.000	729
	10	591	591.000	52	52.000	643
Step 2	1	770	768.709	6	7.291	776
	2	620	620.791	11	10.209	631
	3	842	843.409	16	14.591	858
	4	902	900.991	33	34.009	935
	5	801	798.987	38	40.013	839
	6	394	393.379	20	20.621	414
	7	650	646.513	35	38.487	685
	8	453	459.124	34	27.876	487
	9	622	622.812	46	45.188	668
	10	939	938.283	101	101.717	1040
Step 3	1	666	665.230	5	5.770	671
	2	760	763.761	14	10.239	774
	3	760	759.191	14	14.809	774
	4	725	721.028	17	20.972	742
	5	579	576.240	18	20.760	597
	6	651	650.713	28	28.287	679
	7	820	813.808	40	46.192	860
	8	745	758.483	66	52.517	811
	9	679	679.518	59	58.482	738
	10	608	605.030	79	81.970	687
Step 4	1	642	642.035	5	4.965	647
	2	918	925.570	19	11.430	937
	3	807	801.681	12	17.319	819
	4	703	697.483	14	19.517	717
	5	643	637.433	18	23.567	661

	6	806	809.409	43	39.591	849
	7	643	641.233	37	38.767	680
	8	730	740.513	69	58.487	799
	9	1101	1097.643	123	126.357	1224
Step 5	1	815	814.407	6	6.593	821
	2	690	693.627	13	9.373	703
	3	642	644.467	14	11.533	656
	4	749	740.528	11	19.472	760
	5	696	694.749	22	23.251	718
	6	701	697.595	29	32.405	730
	7	735	735.309	43	42.691	778
	8	699	708.173	62	52.827	761
	9	668	664.733	56	59.267	724
	10	598	599.412	84	82.588	682
Step 6	1	679	678.874	5	5.126	684
	2	727	730.105	12	8.895	739
	3	657	661.599	16	11.401	673
	4	723	719.468	14	17.532	737
	5	733	726.155	17	23.845	750
	6	731	727.638	29	32.362	760
	7	660	656.377	33	36.623	693
	8	710	714.281	55	50.719	765
	9	702	706.925	65	60.075	767
	10	671	671.579	94	93.421	765

Classification Table(a)

	Observed		Predicted		
			malaria infectionstatus		Percentage Correct
			non-malarious	malarious	non-malarious
Step 1	malaria infectionstatus	non-malarious	6993	0	100.0
		malarious	340	0	.0
	Overall Percentage				95.4
Step 2	malaria infectionstatus	non-malarious	6674	319	95.4
		malarious	289	51	15.0
	Overall Percentage				91.7

Step 3	malaria infectionstatus	non-malarious	6709	284	95.9
		malarious	296	44	12.9
	Overall Percentage				92.1
Step 4	malaria infectionstatus	non-malarious	6721	272	96.1
		malarious	293	47	13.8
	Overall Percentage				92.3
Step 5	malaria infectionstatus	non-malarious	6708	285	95.9
		malarious	291	49	14.4
	Overall Percentage				92.1
Step 6	malaria infectionstatus	non-malarious	6621	372	94.7
		malarious	281	59	17.4
	Overall Percentage				91.1
a The cut value is .100					

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	REGION			81.035	9	.000	
	REGION(1)	2.171	.384	32.019	1	.000	8.766
	REGION(2)	1.966	.394	24.907	1	.000	7.144
	REGION(3)	1.751	.388	20.314	1	.000	5.758
	REGION(4)	1.768	.411	18.543	1	.000	5.860
	REGION(5)	.670	.428	2.451	1	.117	1.953
	REGION(6)	1.912	.385	24.607	1	.000	6.764
	REGION(7)	1.473	.389	14.363	1	.000	4.363
	REGION(8)	2.096	.382	30.093	1	.000	8.130
	REGION(9)	.591	.459	1.655	1	.198	1.806
	Constant	-4.601	.355	167.701	1	.000	.010
Step 2(b)	REGION			77.072	9	.000	
	REGION(1)	2.081	.384	29.311	1	.000	8.015
	REGION(2)	1.857	.395	22.083	1	.000	6.401
	REGION(3)	1.681	.389	18.680	1	.000	5.372
	REGION(4)	1.710	.411	17.279	1	.000	5.527
	REGION(5)	.601	.428	1.970	1	.160	1.824
	REGION(6)	1.837	.386	22.642	1	.000	6.276

	REGION(7)	1.392	.389	12.780	1	.000	4.023
	REGION(8)	2.035	.383	28.290	1	.000	7.650
	REGION(9)	.550	.460	1.433	1	.231	1.734
	PREG(1)	.972	.145	44.791	1	.000	2.642
	Constant	-4.658	.356	171.500	1	.000	.009
Step 3(c)	REGION			55.581	9	.000	
	REGION(1)	1.764	.397	19.715	1	.000	5.835
	REGION(2)	1.647	.406	16.431	1	.000	5.190
	REGION(3)	1.372	.401	11.706	1	.001	3.942
	REGION(4)	1.604	.422	14.446	1	.000	4.972
	REGION(5)	.529	.434	1.487	1	.223	1.698
	REGION(6)	1.779	.395	20.302	1	.000	5.926
	REGION(7)	1.403	.396	12.530	1	.000	4.069
	REGION(8)	1.875	.390	23.065	1	.000	6.519
	REGION(9)	.595	.460	1.674	1	.196	1.814
	PREG(1)	.928	.146	40.108	1	.000	2.529
	WLTH			27.010	2	.000	
	WLTH(1)	.504	.159	10.026	1	.002	1.656
	WLTH(2)	-.260	.192	1.841	1	.175	.771
	Constant	-4.728	.358	174.154	1	.000	.009
Step 4(d)	REGION			49.108	9	.000	
	REGION(1)	1.615	.399	16.337	1	.000	5.026
	REGION(2)	1.456	.409	12.654	1	.000	4.291
	REGION(3)	1.192	.404	8.715	1	.003	3.295
	REGION(4)	1.390	.426	10.672	1	.001	4.016
	REGION(5)	.329	.438	.565	1	.452	1.390
	REGION(6)	1.559	.399	15.224	1	.000	4.752
	REGION(7)	1.130	.404	7.840	1	.005	3.096
	REGION(8)	1.657	.395	17.606	1	.000	5.244
	REGION(9)	.554	.461	1.444	1	.230	1.740
	TRESDC(1)	.824	.244	11.397	1	.001	2.280
	PREG(1)	.903	.147	37.959	1	.000	2.466
	WLTH			20.964	2	.000	
	WLTH(1)	.032	.200	.026	1	.871	1.033
	WLTH(2)	-.677	.219	9.545	1	.002	.508
	Constant	-4.875	.364	179.035	1	.000	.008
Step 5(e)	REGION			51.937	9	.000	
	REGION(1)	1.756	.407	18.581	1	.000	5.791

	REGION(2)	1.600	.418	14.684	1	.000	4.953
	REGION(3)	1.337	.412	10.518	1	.001	3.807
	REGION(4)	1.536	.434	12.544	1	.000	4.645
	REGION(5)	.468	.444	1.109	1	.292	1.596
	REGION(6)	1.717	.409	17.593	1	.000	5.569
	REGION(7)	1.267	.411	9.516	1	.002	3.551
	REGION(8)	1.816	.405	20.119	1	.000	6.145
	REGION(9)	.586	.461	1.617	1	.203	1.797
	TRESDC(1)	1.064	.282	14.238	1	.000	2.899
	FLOOR(1)	-.621	.308	4.052	1	.044	.537
	PREG(1)	.906	.147	38.178	1	.000	2.475
	WLTH			19.726	2	.000	
	WLTH(1)	.126	.210	.358	1	.550	1.134
	WLTH(2)	-.579	.230	6.342	1	.012	.561
	Constant	-4.730	.367	166.026	1	.000	.009
Step 6(f)	**AGE**			6.271	2	.043	
	AGE(1)	.195	.144	1.821	1	.177	1.215
	AGE(2)	-.139	.143	.952	1	.329	.870
	REGION			51.368	9	.000	
	REGION(1)	1.763	.407	18.710	1	.000	5.828
	REGION(2)	1.602	.418	14.723	1	.000	4.964
	REGION(3)	1.355	.412	10.818	1	.001	3.878
	REGION(4)	1.536	.434	12.533	1	.000	4.644
	REGION(5)	.465	.444	1.098	1	.295	1.593
	REGION(6)	1.711	.409	17.461	1	.000	5.532
	REGION(7)	1.272	.411	9.593	1	.002	3.568
	REGION(8)	1.801	.405	19.788	1	.000	6.054
	REGION(9)	.575	.461	1.556	1	.212	1.777
	TRESDC(1)	1.066	.282	14.279	1	.000	2.905
	FLOOR(1)	-.626	.309	4.108	1	.043	.535
	PREG(1)	.873	.148	34.969	1	.000	2.393
	WLTH			20.242	2	.000	
	WLTH(1)	.147	.211	.489	1	.484	1.159
	WLTH(2)	-.568	.230	6.091	1	.014	.567
	Constant	-4.746	.378	158.028	1	.000	.009

a Variable(s) entered on step 1: REGION.

b Variable(s) entered on step 2: PREG.

c Variable(s) entered on step 3: WLTH.

d Variable(s) entered on step 4: TRESDC.
e Variable(s) entered on step 5: FLOOR.
f Variable(s) entered on step 6: AGE.

Model if Term Removed					
Variable		**Model Log Likelihood**	**Change in -2 Log Likelihood**	**df**	**Sig. of the Change**
Step 1	**REGION**	-1376.199	108.784	9	.000
Step 2	**REGION**	-1353.560	101.522	9	.000
	PREG	-1321.807	38.017	1	.000
Step 3	**REGION**	-1322.694	68.121	9	.000
	PREG	-1305.873	34.478	1	.000
	WLTH	-1302.799	28.330	2	.000
Step 4	**REGION**	-1311.958	58.468	9	.000
	TRESDC	-1288.634	11.820	1	.001
	PREG	-1299.115	32.782	1	.000
	WLTH	-1294.356	23.265	2	.000
Step 5	**REGION**	-1311.789	62.160	9	.000
	TRESDC	-1288.500	15.582	1	.000
	FLOOR	-1282.724	4.030	1	.045
	PREG	-1297.194	32.970	1	.000
	WLTH	-1291.541	21.664	2	.000
Step 6	**AGE**	-1280.709	6.218	2	.045
	REGION	-1308.484	61.769	9	.000
	TRESDC	-1285.417	15.634	1	.000
	FLOOR	-1279.643	4.087	1	.043
	PREG	-1292.827	30.455	1	.000
	WLTH	-1288.712	22.224	2	.000

Variables not in the Equation(a)					
			Score	df	Sig.
Step 1	Variables	AGE	7.887	2	.019
		AGE(1)	7.411	1	.006
		AGE(2)	4.293	1	.038
		TRESDC(1)	18.885	1	.000
		EDUC	4.930	2	.085
		EDUC(1)	4.884	1	.027
		EDUC(2)	2.299	1	.129
		HRAD(1)	6.257	1	.012
		FLOOR(1)	.801	1	.371
		ROOF(1)	5.103	1	.024
		PREG(1)	47.970	1	.000
		BNET(1)	3.130	1	.077
		WLTH	31.327	2	.000
		WLTH(1)	30.457	1	.000
		WLTH(2)	16.100	1	.000
		DRESDC	21.356	2	.000
		DRESDC(1)	18.885	1	.000
		DRESDC(2)	19.397	1	.000
Step 2	Variables	AGE	5.115	2	.078
		AGE(1)	4.480	1	.034
		AGE(2)	3.396	1	.065
		TRESDC(1)	14.993	1	.000
		EDUC	3.297	2	.192
		EDUC(1)	3.291	1	.070
		EDUC(2)	1.744	1	.187
		HRAD(1)	4.898	1	.027
		FLOOR(1)	.267	1	.605
		ROOF(1)	3.295	1	.069
		BNET(1)	2.589	1	.108
		WLTH	27.678	2	.000

		WLTH(1)	26.239	1	.000
		WLTH(2)	16.441	1	.000
		DRESDC	17.901	2	.000
		DRESDC(1)	14.993	1	.000
		DRESDC(2)	16.876	1	.000
Step 3	Variables	AGE	6.231	2	.044
		AGE(1)	5.248	1	.022
		AGE(2)	4.434	1	.035
		TRESDC(1)	11.549	1	.001
		EDUC	.549	2	.760
		EDUC(1)	.539	1	.463
		EDUC(2)	.408	1	.523
		HRAD(1)	.027	1	.870
		FLOOR(1)	.270	1	.603
		ROOF(1)	.084	1	.772
		BNET(1)	.859	1	.354
		DRESDC	13.719	2	.001
		DRESDC(1)	11.549	1	.001
		DRESDC(2)	11.837	1	.001
Step 4	Variables	AGE	6.243	2	.044
		AGE(1)	5.314	1	.021
		AGE(2)	4.364	1	.037
		EDUC	.845	2	.656
		EDUC(1)	.072	1	.788
		EDUC(2)	.528	1	.467
		HRAD(1)	.019	1	.890
		FLOOR(1)	4.102	1	.043
		ROOF(1)	.364	1	.546
		BNET(1)	.509	1	.475
		DRESDC	3.568	1	.059
		DRESDC(2)	3.568	1	.059
	Overall Statistics		13.924	9	.125
Step 5	Variables	AGE	6.307	2	.043
		AGE(1)	5.424	1	.020
		AGE(2)	4.330	1	.037
		EDUC	.462	2	.794
		EDUC(1)	.181	1	.671
		EDUC(2)	.426	1	.514

		HRAD(1)	.106	1	.744
		ROOF(1)	.145	1	.704
		BNET(1)	.505	1	.477
		DRESDC	2.260	1	.133
		DRESDC(2)	2.260	1	.133
	Overall Statistics		9.921	8	.271
Step 6	Variables	EDUC	.710	2	.701
		EDUC(1)	.514	1	.473
		EDUC(2)	.704	1	.401
		HRAD(1)	.139	1	.709
		ROOF(1)	.155	1	.693
		BNET(1)	.604	1	.437
		DRESDC	2.240	1	.134
		DRESDC(2)	2.240	1	.134
	Overall Statistics		3.597	6	.731
a Residual Chi-Squares are not computed because of redundancies.					

```
Step number: 1

            Observed Groups and Predicted Probabilities

    2000 ô   m                                                          ô
         ó   nm                                                         ó
         ó   nn                                                         ó
F        ó n nn                                                         ó
R   1500 ô n nn                                                         ô
E        ó n nn                                                         ó
Q        ó n nn                                                         ó
U        ó n nn                                                         ó
E   1000 ô nnnn                                                        ô
N        ó nnnn                                                         ó
C        ónnnnn                                                         ó
Y        ónnnnn                                                         ó
     500 ônnnnn                                                         ô
         ónnnnn                                                         ó
         ónnnnn                                                         ó
         ónnnnn                                                         ó
Predicted ôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôô
  Prob:   0             .25             .5              .75             1
  Group:  nnnnnnmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm

          Predicted Probability is of Membership for malarious
          The Cut Value is .10
          Symbols: n - non-malarious
                   m - malarious
          Each Symbol Represents 125 Cases.
```

Step number: 2

Observed Groups and Predicted Probabilities

```
    3200 ô                                                                 ô
         ó                                                                 ó
         ó                                                                 ó
F        ó                                                                 ó
R   2400 ô                                                                 ô
E        ó  m                                                              ó
Q        ó  n                                                              ó
U        ó  n                                                              ó
E   1600 ô  n                                                              ô
N        ón n                                                              ó
C        ón nnn                                                            ó
Y        ón nnn                                                            ó
     800 ônnnnn                                                            ô
         ónnnnn                                                            ó
         ónnnnn                                                            ó
         ónnnnn  mn                                                        ó
Predicted ôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôô
  Prob:   0            .25            .5            .75             1
  Group:  nnnnnnmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm

          Predicted Probability is of Membership for malarious
          The Cut Value is .10
          Symbols: n - non-malarious
                   m - malarious
          Each Symbol Represents 200 Cases.
```

Step number: 3

Observed Groups and Predicted Probabilities

```
    2000 ô                                                                  ô
         ón                                                                 ó
         ón                                                                 ó
F        ón                                                                 ó
R   1500 ôn                                                                 ô
E        ón n                                                               ó
Q        ón n m                                                             ó
U        ón n n                                                             ó
E   1000 ônnnnn                                                             ô
N        ónnnnn                                                             ó
C        ónnnnn                                                             ó
Y        ónnnnn                                                             ó
     500 ônnnnn                                                             ô
         ónnnnnn                                                            ó
         ónnnnnn                                                            ó
         ónnnnnn n mn                                                       ó
Predicted ôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôôô
  Prob:   0             .25             .5             .75             1
  Group:  nnnnnnmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm

          Predicted Probability is of Membership for malarious
          The Cut Value is .10
          Symbols: n - non-malarious
                   m - malarious
          Each Symbol Represents 125 Cases.
```

Step number: 4

Observed Groups and Predicted Probabilities

```
    2000 ô                                                                  ô
         ó                                                                  ó
         óm                                                                 ó
F        ón                                                                 ó
R   1500 ônn                                                                ô
E        ónn                                                                ó
Q        ónn  m                                                             ó
U        ónn mn                                                             ó
E   1000 ônnnnn                                                             ô
N        ónnnnn                                                             ó
C        ónnnnn                                                             ó
Y        ónnnnn                                                             ó
     500 ônnnnnm                                                            ô
         ónnnnnn                                                            ó
         ónnnnnn                                                            ó
         ónnnnnn n mn                                                       ó
Predicted ÒÒÒÒÒÒÒÒÒÒÒÒÒÒÔÒÒÒÒÒÒÒÒÒÒÒÒÒÒÒÒÔÒÒÒÒÒÒÒÒÒÒÒÒÒÒÒÔÒÒÒÒÒÒÒÒÒÒÒÒÒÒÒÒÒ
  Prob:   0            .25            .5            .75            1
  Group:  nnnnnnmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
```

Predicted Probability is of Membership for malarious
The Cut Value is .10
Symbols: n - non-malarious
m - malarious
Each Symbol Represents 125 Cases.

Step number: 5

Observed Groups and Predicted Probabilities

```
     2000 ôm                                                                ô
          ón                                                                ó
          ón                                                                ó
F         ón                                                                ó
R    1500 ôn                                                                ô
E         ónn  m                                                            ó
Q         ónn  n                                                            ó
U         ónn  n                                                            ó
E    1000 ônnnmn                                                            ô
N         ónnnnn                                                            ó
C         ónnnnn                                                            ó
Y         ónnnnn                                                            ó
      500 ônnnnn                                                            ô
          ónnnnnn                                                           ó
          ónnnnnn                                                           ó
          ónnnnnn n nn                                                      ó
Predicted ðòòòòòòòòòòòòòôòòòòòòòòòòòòòôòòòòòòòòòòòòòôòòòòòòòòòòòòòò
  Prob:   0             .25             .5              .75              1
  Group:  nnnnnnmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
```

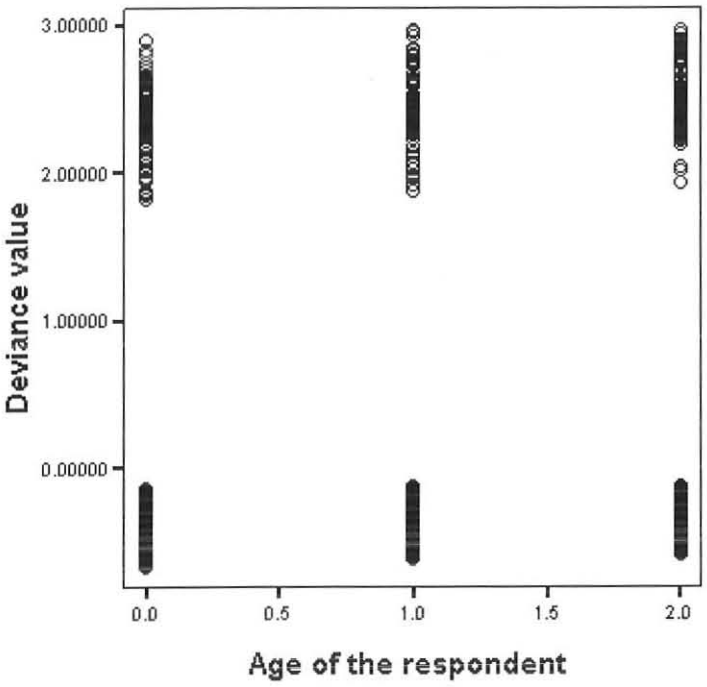
Predicted Probability is of Membership for malarious
The Cut Value is .10
Symbols: n - non-malarious
m - malarious
Each Symbol Represents 125 Cases.

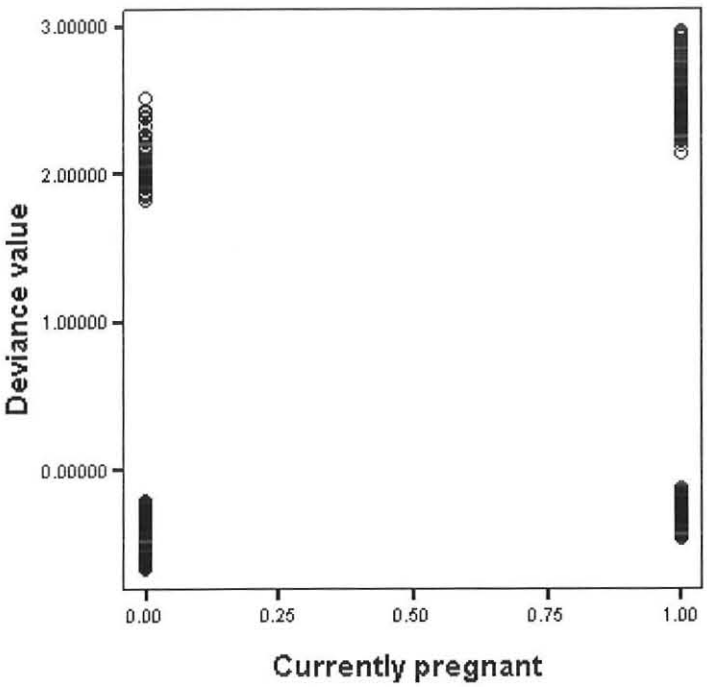
Step number: 6

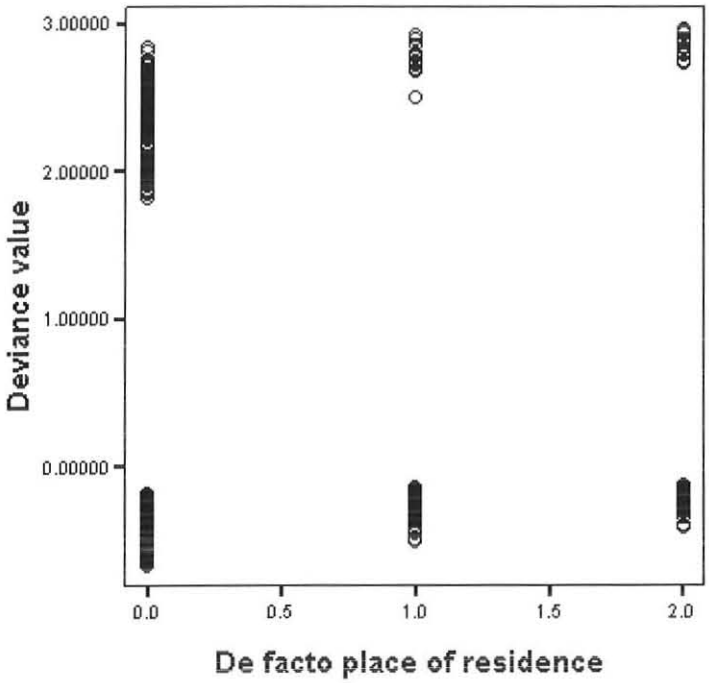
Observed Groups and Predicted Probabilities

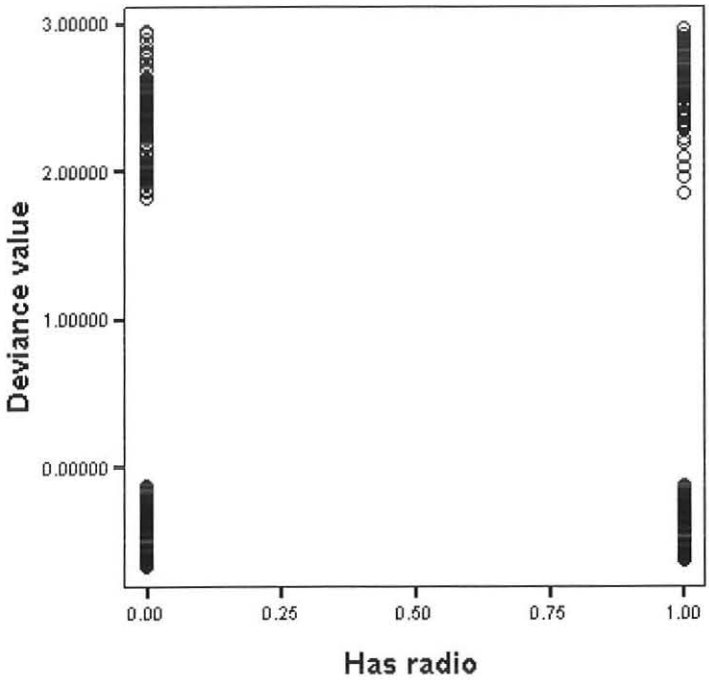
```
    2000 ô                                                                ô
         ó                                                                ó
         ón                                                               ó
F        ónm                                                              ó
R   1500 ônn                                                              ô
E        ónn                                                              ó
Q        ónn                                                              ó
U        ónn                                                              ó
E   1000 ônnnnm                                                           ô
N        ónnnnn                                                           ó
C        ónnnnn                                                           ó
Y        ónnnnn                                                           ó
     500 ônnnnnn                                                          ô
         ónnnnnn                                                          ó
         ónnnnnn                                                          ó
         ónnnnnnn  m                                                      ó
Predicted ðòòòòòòòòòòòòôòòòòòòòòòòòòôòòòòòòòòòòòòôòòòòòòòòòòòòôòòòòòòòòòòòòñ
  Prob:   0             .25             .5              .75             1
  Group:  nnnnnnmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
```

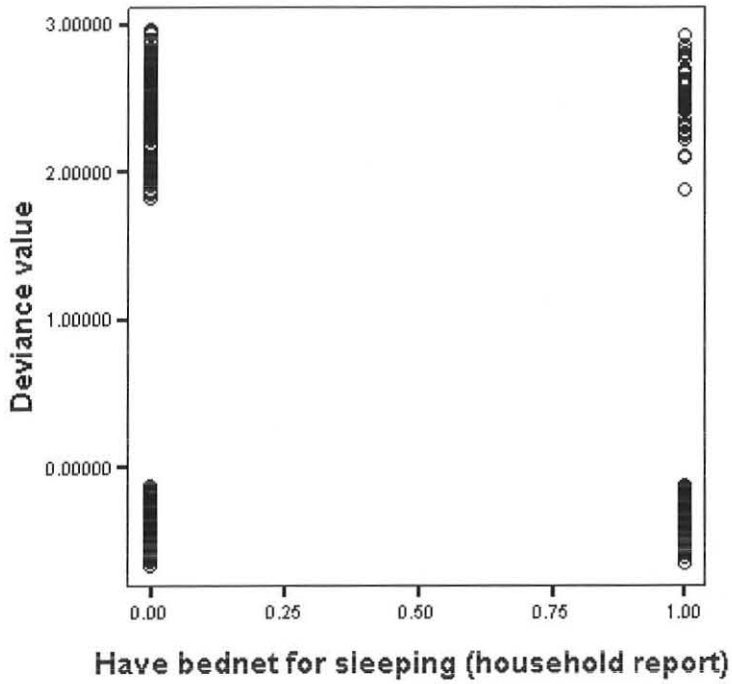
Predicted Probability is of Membership for malarious
The Cut Value is .10
Symbols: n - non-malarious
m - malarious
Each Symbol Represents 125 Cases.

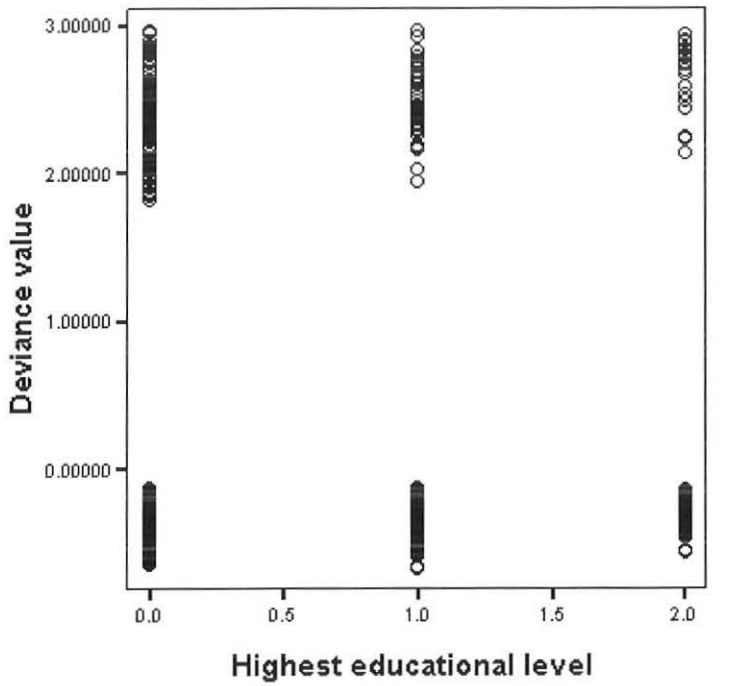

3.00000
2.00000
1.00000
0.00000
Deviance value
0.0
0.5
1.0
1.5
2.0
Age of the respondent

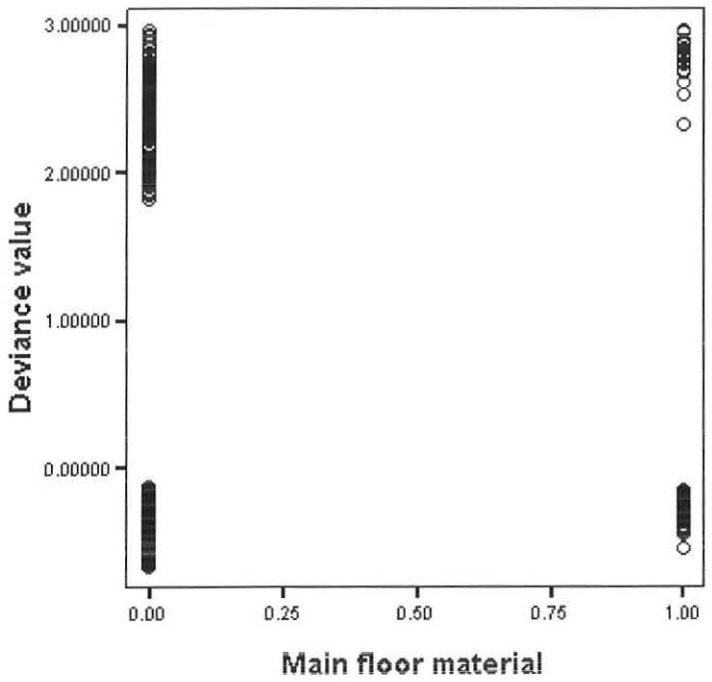

3.00000
2.00000
1.00000
0.00000
Deviance value
0.00
0.25
0.50
0.75
1.00
Currently pregnant

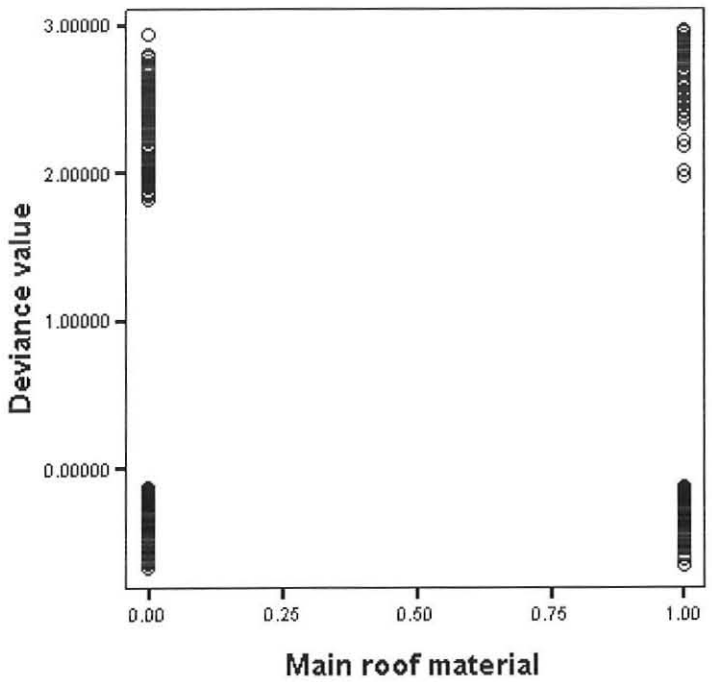

3.00000
2.00000
1.00000
0.00000
Deviance value
0.0
0.5
1.0
1.5
2.0
De facto place of residence

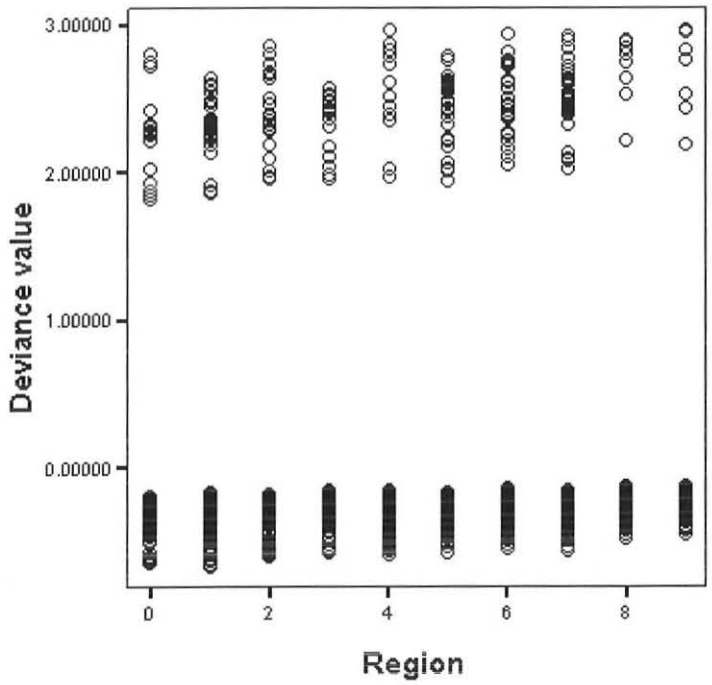

3.00000
2.00000
1.00000
0.00000
Deviance value
0.00
0.25
0.50
0.75
1.00
Has radio

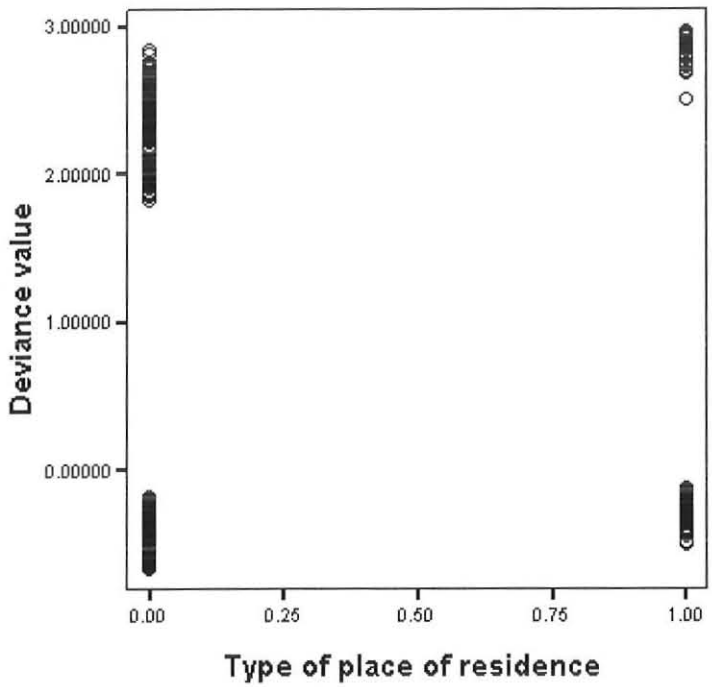

3.00000
2.00000
1.00000
0.00000
Deviance value
0.00
0.25
0.50
0.75
1.00
Have bednet for sleeping (household report)

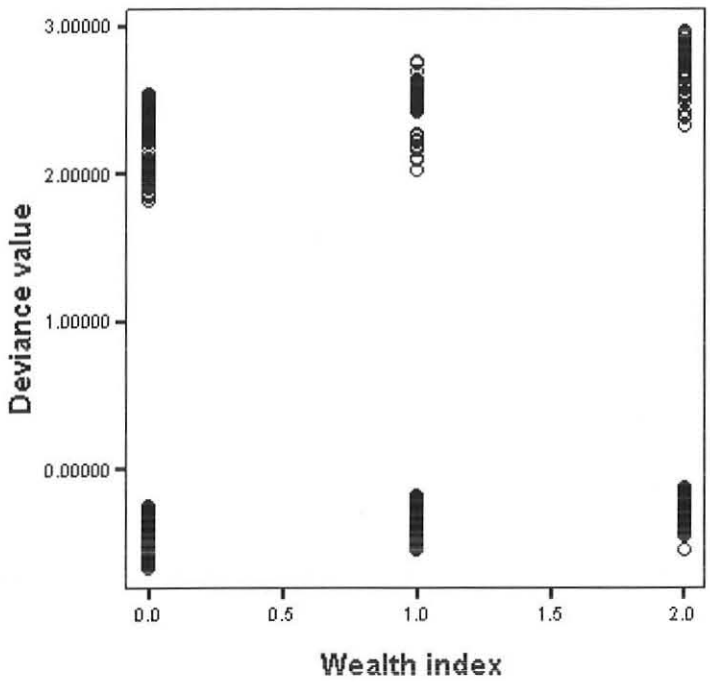

3.00000
2.00000
1.00000
0.00000
Deviance value
0.0
0.5
1.0
1.5
2.0
Highest educational level

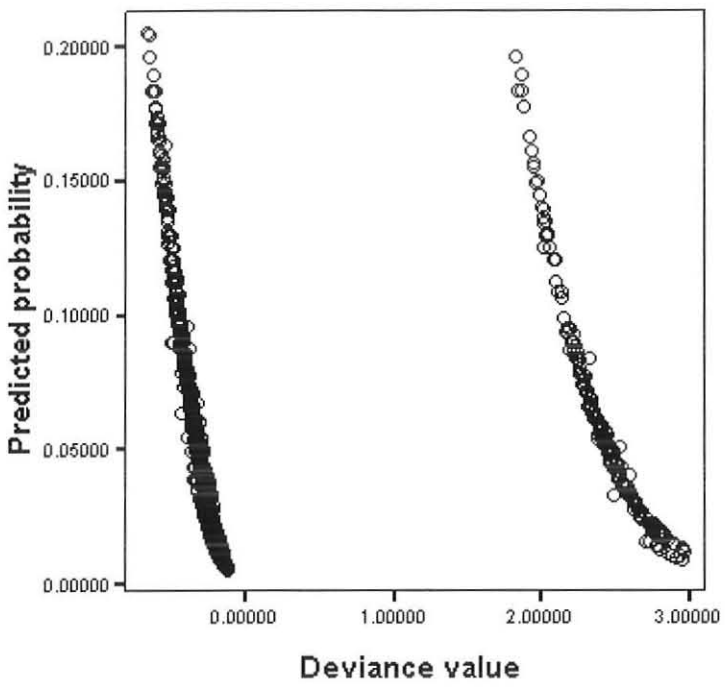

3.00000
2.00000
1.00000
0.00000
Deviance value
0.00
0.25
0.50
0.75
1.00
Main floor material


3.00000
2.00000
1.00000
0.00000
Deviance value
0.00
0.25
0.50
0.75
1.00
Main roof material


3.00000
2.00000
1.00000
0.00000
Deviance value
0
2
4
6
8
Region


3.00000
2.00000
1.00000
0.00000
Deviance value
0.00
0.25
0.50
0.75
1.00
Type of place of residence


3.00000
2.00000
1.00000
0.00000
Deviance value
0.0
0.5
1.0
1.5
2.0
Wealth index


0.20000
0.15000
0.10000
0.05000
0.00000
Predicted probability
0.00000
1.00000
2.00000
3.00000
Deviance value

DECLARATION

I, the undersigned, declare that this thesis is my original work in partial fulfillment for the requirements for the degree of master of statistics. Where other sources of information have been used, they have been acknowledged.

Name: Muluken Derbew

Signature: _____

Place: Faculty of Science, Addis Ababa University

Date: July, 2009

This thesis has been submitted for examination with my approval as a University advisor.

_____

Prof. Eshetu Wencheko