Logistic regression Analysis of exposure to Toxoplasmosis among HIV/AIDS Patients: the case of Mettu Karl Hospital, South west of Ethiopia

Kasim Mohammed Yesuf, Dep't of Statistics, University of Gondar, Gondar, Ethiopia, P.O.Box. 196

Abstract

Background: *Toxoplasma gondii* is food and waterborne opportunistic pathogen that causes severe disease in immunocompromised patients. It is caused by an obligatory intracellular protozoon. It causes a wide range of diseases with toxoplasma encephalitis commonly encountered in HIV/AIDS patients. It has been estimated that one-third of the human population is infected with this parasite. In Addis Ababa, Ethiopia, the prevalence of *toxoplasma gondii* was about 80%.

Methodology: A sample of 120 HIV/AIDS patients has been taken from Mettu Karl Hospital. In this study logistic regression is applied for binary response data. Model comparison is made using the three link functions: logit, probit and complementary log log model. The variables considered in the final model were selected from the candidate using a forward step wises variable selection in R software version 2.13.1. To assess whether the model adequately fit to the data different model checking and model diagnostic have been applied.

Results: The study indicates that the prevalence of toxoplasmosis for HIV\AIDS patients in Mettu Karl town was 60%. Parameter estimations were fairly similar for the three link functions. The logit link function model appears to be the preferred one based on interpretation of the parameters and the possibility to calculate the odds ratio although the three link function model show only a small difference. The analysis shows that some health, economic and risk behavior factors influence exposure of patients to toxoplasmosis. Further, the study shows that the risk

factors of Toxoplasmosis are education, marital status, house toilet, tap water and washing hand and raw meat. On the contrary sex, age, address, occupations, cat presence, fever, fruit, eye disease and compound latrine don't have a significant effect on toxoplasmosis. .

Conclusion: The binary logistic regression model analysis show that; the three link function (logit, probit and c-log-log) as likely as the same. However, due to lose of interpretation of results for such kind of categorical data, the logit link function was considered. The factor that influences toxoplasmosis status can be grouped as risk behavior factors (hand washing and raw meat), demographic and health factors (Education, marital status and house toilet) and economic factors (tap water). Under such grouping we can state that patients who are involved in risky behaviors will have higher risk of contracting toxoplasmosis.

I. Table of Contents

II. List of Tables

III. List of figures

IV. Abbreviations and some operational definition

AIDS - Acquired Immune deficiency syndrome

ANC - Ante Natal Clinic,

ART- Antiretroviral treatment

BSS - Behavior Surveillance Survey

CFT - Complement-fixation test

CNS - Central nervous system

CSA- Central Statistical Agency

CT- Counseling and testing Centers

DHS - Demographic and Health Survey

ECSRNCT-European Collaborative Study and Research Network on Congenital Toxoplasmosis

EDHS - Ethiopia Demographic and Health Survey

ELISA- enzyme-linked immunosorbent assay

HAART- Highly active antiretroviral therapy

HBC - Hepatitis B core

HBV - Hepatitis B virus

HIV- Human Immune-deficiency virus

MDAT- Modified Direct Agglutination Test

MHDPC- Ministry of Health Disease prevention and control

NHANES- National Health and Nutritional Assessment Survey

OLS- Ordinary least square

PLWHA- people living with HIV/AIDS

PMTCT - Prevention of mother-to-child transmission

PMTCT - Prevention of mother-to-child transmission

R-Statistical software

SPSS- Statistical package for social sciences

TE- Toxoplasmic encephalitis

UNAIDS- United Nations AIDS

WIHS- Women's Interagency HIV Study

IgM- Immunoglobulin M

IgG- immunoglobulin G

1. Introduction

1.1. Background for *toxoplasmosis gondii*

Toxoplasma gondii is one of the most important diseases of human and other livestock. It is a protozoan parasite (Techalew, *et al.*, 2009) which is an endemic worldwide organism. It has been estimated that one-third of the world human population is infected with this parasite (Fong, *et al.*, 2010). It has gradually evolved over the past decades to be among the most common opportunistic diseases in HIV/AIDS in developing countries (Monotoya, *et al.*, 2000). The prevalence of seropositivity for antibodies against toxoplasma gondii has been estimated at 3% in Western Europe and 67% to 90% in tropical countries and also up to 14.8% of AIDS patients in Southeast Asia region were reported with toxoplasmosis in the central nervous systems (Nissapatorn, 2008). Transmission is mainly by ingestion of tissue cysts in raw and undercooked meat, exposure to oocysts in contaminated vegetable or contact with faeces of a felid (Monotoya, *et al.*, 2000).

Toxoplasmosis is a sort of serious disease that has the capacity to result in morbidity and mortality in immunocompromised patients. It affects mostly acquired immunodeficiency syndrome (AIDS) patients. *Toxoplasma gondii* can bring systemic disease with multiple sites of infection. It can also contaminate one organ preferentially. Pulmonary toxoplasmosis, estimated to account for 4% of all cases of pneumonia in AIDS patients is the second or third most frequent form of toxoplasmosis, after toxoplasmic encephalitis (Luft, *et al.*, 1988). Extrapolating from the current findings, over a quarter of the HIV positive patients will be at risk of developing cerebral toxoplasmosis. Study in the USA showed that about 30% of AIDS patients previously exposed to toxoplasma and suffered from a cerebral reactivation (Navia, *et al.*, 1986). Consequently, it may be calculated that 8 % of AIDS patients in South East England will experience a life threatening episode of cerebral disease following secondary reactivation of toxoplasmosis. In addition to this 0.5-1 % of these patients may acquire primary toxoplasmosis associated with AIDS each year reflecting the incidence of toxoplasma infection in this group (Holliman, 1990). The study in Iran

showed that seroprevalence of latent Toxoplasma infection in HIV infected individuals to be 18.2% (Davarpanah, *et al.*, 2007). Geographical variation of prevalence of toxoplasmosis has been found: 50% in Mexico (Boto, *et al.*, 1998), 36.7% in Spain (Wanachiwanawinm, *et al.*, 2001), 53.7% in Thailand (Falusi, *et al.*, 2002).

Human toxoplasmosis has been well studied elsewhere in Africa. The estimated seroprevalences in Africa as comparable to many other reports from exposed group was 45.7% (Swai *et al.*, 2009). For instance, seroprevalences of 42.6% and 52.4% were reported from slaughter house workers in Djibouti and Egypt, respectively (Chantal, *et al.*, 1996; Ibrahim, *et al.*, 1997). Occupational contacts including butchers, slaughter house workers, milers, and cow attendants in one state of Joss area, Nigeria revealed 22.6% were infected (Osiyemi, *et al.*, 1985). In contrast, cattle breeders and abattoir personnel from Benin gave 87% positives (Fayomi, *et al.*, 1987) and high-risk groups from city of Pointe-Noire in Congo and Addis Ababa in Ethiopia showed a seroprevalence between 41.9% and 80%, respectively (Candolfi, *et al.*, 1993; Woldemichael, *et al.*, 1998).

The HIV prevalence for the adult of Ethiopia national in 2003 was 4.4%. Out this figure, 12.6% is urban and 2.6% rural and higher prevalence is higher among women (5.0%) than men (3.8%). The 2003 estimate of PLWHA (people living with HIV/AIDS) is 1.5 million, including 96,000 children. There were also 197,000 new infections, 98,000 new AIDS cases, and 90,000 AIDS deaths in the adult population in 2003. A total of 128,000 HIV positive pregnancies and an estimated 35,000 HIV-positive births occurred. Among children aged 0-14 years, there were 35,000 new infections, 25,000 new AIDS cases and 25,000 new AIDS deaths. A total of 4.6 million children fewer than 17 in the country are estimated to be orphans for different reasons. From this figure, 537,000 were due to AIDS reported from MHDPC (EDHS, 2006).

1.2. Statement of the problem

Human infection may be acquired by ingestion of undercooked infected meat containing *Toxoplasma* cysts, ingestion of the oocyst from focally contaminated hands or food, organ transplantation or blood transfusion, transplacental transmission and accidental inoculation of tachyzoites.

However, most of the researches in our country focused on the prevention, on the factors that increase the chance of contracting the disease etc, all dealing on how to prevent it before a person is HIV positive. For example, Alemtsehai, *et al.* (2006) worked on the latter case. It can be said that less attention was given for researches dealing with improving the situation of HIV positives that taking ART. It is this fact that little has been done on the factors that influence the survival/death status of a person given s/he is already HIV positive and is under the follow up of ART that motivated this study.

The previous studies in this area focus only on the prevalence of toxoplasmosis in HIV/AIDS. In addition to this there was no considerable statistical analysis applied to show the risk factors of this parasite.

Thus, this study, tries to apply binary logistic regression to find out the factors that expose HIV /AIDS patients to toxoplasmosis. An outcome like toxoplasmosis (positive or negative), however, have only two values: negative (0) and positive (1). The ordinary linear regression fails for such binary outcome that is why binary logistic regression model is considered.

1.3. Significance of the study

This study will have a great significance and importance in providing relevant information about toxoplasmosis for the proper management of HIV/AIDS patients. On top of this, the result of the study will enable clinicians and policy makers to enhance the awareness of the society about factors which increase the probability of death in HIV patient. The result of this study can also be

used as a source of information to other researchers in the future. In addition to this, the studies compare with the previous studies that already done in the same area; to fill the gaps that missed by the researcher and to clarify the results based on inferential (estimation and hypotheses testing) statistical analysis rather than tables, charts and graphs. In general the study will be able to:

- benefit different stakeholders/activists working in the areas of giving care, support and treatment for HIV positives specifically.
- develop the awareness of the community about the causes or riskfactors of toxoplasmosis.
- Promote effective planning and policy to act on prevention, control and intervention strategies of toxoplasmosis.
- use as a basis for further study in this area.

1.4. Objectives of the Study

1.4.1. General Objective

The general objective of this study is to study the socio-economic, Demographic and health factors that influence toxoplasmosis among HIV-Positive patients by fitting and identifying the best binary logistic regression model using the logit, probit (normit) and complementary log log link functions.

1.4.2. Specific Objectives

- To find out the risk factors associated with *toxoplasmosis gondii* in Mettu Hospital, south west Ethiopia 2010.
- To compare binary logistic regression using logit, probit and c-log-log link functions.
- To show model diagnostic for the different link functions.

2. Literature review

2.1. Toxoplasmosis for the HIV/AIDS patients

Acute cerebral toxoplasmosis is the most common cause of focal neurologic disorder in AIDS patients. If not detected and treated promptly, cerebral toxoplasmosis may cause significant morbidity and mortality. Prophylaxis is key to preventing negative outcomes. All HIV-infected patients should be educated about nonpharmacologic and medical prophylaxis for *T. Gondii* infection, and sero-positive patients should receive either primary or secondary prophylaxis for toxoplasmosis (Jayawardena, *et al.*, 2008).

Toxoplasmosis most of the time occurred in HIV/AIDS patient. The study taken on Mozambique shows that the distribution of toxoplasmosis has high prevalent in HIV-positive patients (31.3%, 18/58), as compared to HIV-negative patients (10.9%, 10/92) ($p < 0.005$), with a 3.6 Odds Ratio for toxoplasma infection in HIV-infected individuals (Sonia, *et al.*, 2010).

The sera of 67 HIV-infected persons without clinical signs of toxoplasma gondii infection and sera of 777 immunocompetent persons from three distinct regions of Czechoslovakia were examined for the presence of toxoplasmic antibodies using the complement-fixation test (CFT). Additionally toxoplasma positive HIV+ individuals were re-examined for the presence of IgG and IgM toxoplasmic antibodies by ELISA methods. Results show that overall prevalence of toxoplasmic antibodies is not significantly greater in HIV-positive subjects (29.8%) than in the general population (26.1%). Similarities between these two tested groups were also documented by a close correlation of their geometrical means of titres (13.9 versus 14.5). All 20 HIV-infected Patients who were positive in CFT were positive in ELISA IgG reaction, and none in ELISA IgM reaction. The detected antibody levels were suggestive of a latent toxoplasma infection only. But because of the risk of the infection reactivation all of these patients should be attended to on a systematic basis (Sýkora, *et al.*, 1992).

The 65 individuals examined for *anti-Toxoplasma gondii* IgG antibodies were also tested for HIV infection by the HIV-Spot Test. Nineteen of the 65 individuals considered in the study were HIV Positive, and of whom 10 were seropositive. Almost all the HIV-Spot Test positive individuals were reactive to the MDAT at dilutions greater than or equal to 1:1024. The majority of the Spot negative individuals were reactive at dilutions of less than or equal to 1:512 (Negash, *et al*., 2008).

A study was done in Addis Ababa, Ethiopia show that from a total of 330 sera (half (50%) of them which were HIV positives) were included for studying sero-prevalence of latent toxoplasma infection among people with and without HIV infection. In bivariate analysis, it is showed that the rate of toxoplasma infection was significantly higher among males than females (93.6% versus 86.6%; p = 0.04). In addition, rate of toxoplasma infection was shown to increase significantly with age (p = 0.003). The prevalence of latent toxoplasma infection was 93.3% (154/165) among HIV positive and 86.7% (143/165) among HIV negative participants (Techalew, *et al.*, 2009).

2.2. Risk factors of toxoplasmosis in HIV/AIDS patients

The sero-prevalence of anti-toxoplasma IgG antibody among HIV-infected and HIV uninfected participants was determined by the study taken in the Addis Ababa, Ethiopia. The overall prevalence rate of latent *T. gondii* infection was found to be 90%. The high sero-prevalence of latent Toxoplasma infection among the study population seems reasonable as raw or insufficiently cooked meat prepared in a various favorite cultural food is consumed. In addition, cats are abundant to cause environmental contamination less and the climate is favorable to favor survival of the parasite (Techalew, *et al.*, 2009).

Wallace, *et al.* (1993) found that among a random survey of 87 HIV infected patients at a particular clinic, 57 (65.5%) from 87 patients had lived with cats in the past 5 years. From this study, 75 of the group were among seo-negative for *Toxoplasmosis gondii* and 12 were however,

is a seropositive. The calculated cat exposure for *toxoplasmosis gondii* which is a seropositive was significant ($P<0.05$).

Most of the time ingestion of tissue cysts occurs when infected meat is eaten raw or undercooked. According to Hoklek (2004), in many parts of the world, 5-35% of pork, 9-60% of lamb, and up to 9% of beef contain *T gondii*. Wild game is often infected with the highest prevalence in black bears and white-tailed deer. Virtually all edible portions of an animal can harbor viable *T gondii* organisms. The parasite can also be transmitted by containers, knives or other utensils, cutting boards or other preparation surfaces contaminated with raw meat. Touching unwashed hands to the face after meat preparation is another source of infection. According to the American Academy of Family Physicians (2005) report oocysts can also be transmitted to food by flies, cockroaches or other insects, or by ingestion of unpasteurized milk and raw eggs.

Moura, *et al.* (2002) reported the world's largest outbreak of water-borne toxoplasmosis. *T. gondii* was isolated from municipal reservoirs and household storage tanks in a small city in Brazil following an outbreak of 600 acute cases of which 8% exhibited ocular toxoplasmosis. Of the total cases, 7 were pregnant with 1 miscarrying and 5 delivering seropositive infants. The studies suggest that suspected source of infection was a cat that lived on and delivered seropositive kittens on the top of the implicated reservoir.

In terms of the clinical scenario, focal neurological defects such as fever, headache, seizure, and hemiparesis were the most common symptoms found in the study sample population. Although these symptoms may mimic those of other neurological diseases, it is nonetheless recommended that the possibility of cerebral toxoplasmosis be considered in every HIV-positive patient with neurological symptoms and that empirical therapy be instituted on wide indications (Nissapatorn *et al*, 2004). Another symptom of Extracerebral toxoplasmosis usually affects the eyes and lungs (Lodenyo, *et al.*, 2007).

A study in Tanzania showed that the seroprevalences of toxoplasmosis is indirectly related with age. The higher seroprevalences rates occurred in age group (≤20 years) as compared to age group (51-60 years), it may be due to an increased exposure of this child age group with pet animals (such as cats), poor sanitary habits and probably low level of hygiene during food preparation as a result of low level of awareness of toxoplasmosis. Surprisingly, and in contrast to other studies, the study found no significant difference in the seroprevalences between males and females suggesting that sex does not influence exposure to and or/ immune response to *T. gondii* or community-specific habits or behaviors that may predispose to exposure (Swai, *et al.*, 2009).

Significant differences were showed between occupation and location of residence terms of marital status. A study indicate that high percentage of patients in both groups were mainly unemployed (75 versus 51%). The majority of patients with TE were single (70%) and resided outside Kuala Lumpur (68%), and that of patients without TE were married (51%) and resided in Kuala Lumpur (57.6%) (Nissapatorn, *et al.*, 2004).

In a study done in China using the logistic regression analysis, ten risk factors (including age, living location, profession, physical condition and so on) were investigated for potential factors associated with T. gondii infection. The results specified that the only factor associated with T. gondii infection was gender (OR = 0.712; 95% CI: 0.564-0.899). The rest of the factors did not show any association with T. gondii infection (Xiao, *et al.*, 2010).

A study taken on south west of Cameroon, as compared to other areas such as Uganda, this area shows a better socio-economic condition and rapid urbanization. Improved access to gate potable water and sanitation facilities though constant breakdowns from time to time continued awareness around good hygiene, education and rare exposure to stray cats are characteristics of the Yaounde region (Assob, *et al.*, 2011).

The study taken in Malaysia show that HIV-positive persons who are Toxoplasma seropositive may also advised about experiencing preventive behavioral practices such as eating only well-

cooked meats, washing their hands after outdoor activities involving soil contact and after contact with cats (Richards, *et al*., 1995).

2.3. Application of logistic Regression model

On the study 'Applying Logistic Regression Model to the Second Primary Cancer Data' (Amr I., 2010), social-demographic risk factors of developing a second primary cancer using logistic regression model were studied. The social-demographic risk factors used are age at first cancer, gender, area the patient live in, marital status, family history, smoking, education and obesity in addition to treatment by radiation. The binary logistic regression model was used to estimate the probability of having second primary cancer. The study showed that smoking, family history, marital status, and education are the significant factors. The 95% CI for the odds ratio for smokers to non-smokers ranges from 3 to 65. The study concludes that: married patients are more susceptible to develop a second cancer; treatment by radiation decreases the susceptibility; a patient with family history is more susceptible to develop second cancer; smokers are more susceptible than non-smokers, and educated patients are less susceptible to develop a second cancer.

A study was done on 'Prevalence of toxoplasmosis in HIV/AIDS patients in Mettu Karl Hospital' in 2010. The major objective of this study was to establish the prevalence of toxoplasmosis in Mettu Karl Hospital. Descriptive statistics like frequency tables, cross tabulations, graphical data presentation techniques like bar graphs, histograms and pie charts were used, and chi-square test was employed. Except chi-square, no other inferential statistics is used. So this study extends to analyze these data using binary logistic regression model. Three link functions are considered. The factors that contribute significantly to toxoplasmosis among HIV/AIDS patients will be identified.

3. Methods

3.1. Study Area and period

The site of the study was conducted in Mettu Karl Hospital that is found in Mettu town; the capital city of Illu Ababora zone. It is located around 600 km south west of Addis Ababa and 265Km away from Jimma city. It has about 1760 m above sea level. The zone has a total population of 1,319,967. It has also owned about 516 kebeles. The study was conducted from November 2010- December 2010 by the principal investigator of Solomon Abera.

3.2. Study Design

A cross sectional study with systematic random sampling technique was used. The study included 120 subjects from all HIV/AIDS patients who came to Mettu Karl Hospital laboratory for CD4 count and ART monitoring during the study period.

3.3. Source Population

All HIV/AIDS patients who came to at Mettu Karl Hospital laboratory for CD4 count and ART monitoring during the study period was the source of population to be determined.

3.4. Study population

Inclusion criteria-HIV positive patients

Exclusion- neonates less than 3 months, they may congenitally acquire IgG antibodies.

3.5. Sample Size Determination and Sampling technique

There are several formulas developed for sample size calculation that conform to different research situations. Using the systematic random sampling, the sample size determination formula adopted for this study was:

$$n = \frac{(z_{\alpha/2})^2 p(1-p)}{d^2}$$

d= margin of error = 0.05

p= prevalence of toxoplasmosis = 0.914

Z= the standard normal value at α = 0.05 level of significance.

Since the sample size of n=120 HIV patients is estimated based on previous studies with the assumed toxoplasmosis prevalence of 91.4% (Myriam, *et al.*, 1993; Negash, *et al.*, 2008).

By taking the prevalence of 91.4% and level of confidence interval at 95%.

So $n = \frac{(1.96)^2.0.914(0.086)}{(0.05)^2} = 120$

3.6. Data collection procedure

The data used in this project were obtained from Jimma University Medical laboratory department. .

3.7. Variables included in the study

3.7.1. The Dependent Variable

The dependant variable is dichotomous random variable of "toxoplasmosis positive/negative Status" (negative=0, positive=1) for an HIV positive person. It is denoted by TOXO.

3.7.2. Independent variables

The variables that are assumed to influence the prediction of positive/negative status are presented in the Table 1 below.

Table-1: List of independent variables, in Mettu Karl Hospital, 2010.

No.	Variable	Code of the Variables	Coding
1	Sex of the patient	SEX	0=Female, 1=Male
2	Age of the patient	AGE1	1=24-29, 0=18-23
		AGE2	1=30-41, 0=18-23
		AGE3	1=42+, 0=18-23
3	Place of Residence	ADR	1=Urban, 0= Rural
4	Education	EDUC1	1=1-4, 0=illiterate
		EDUC2	1=5-8, 0=illiterate
		EDUC3	1=9+, 0=illiterate
5	Occupation	OCC1	1=Merchant, 0=farmer
		OCC2	1=Employee, 0=farmer
		OCC3	1=Unemployed, 0=farmer
6	Marital status	MAR1	1=married, 0=others
		MAR2	1=single, 0=others
7	Presence of tap water	TAPWATER	0=no, 1=yes
8	Compound latrine	CPDLATRINE	0=no, 1=yes
9	House toilet	HOUSETOILET	0=no, 1=Yes
10	Cat presence	CAT	0=No, 1=yes
11	Habit of eating raw meat	RAWMEAT	0=no, 1=yes
12	Habit of washing hand	HAND	0=Always, 1=Sometimes
13	Habit of washing fruit	FRUIT	0=Always, 1= Sometimes
14	Eye disease	EYE	0=no , 1=yes
15	Feeling Fever	FEVER	0=no, 1=yes

3.8. Statistical Analysis

3.8.1. General Introduction

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. It is often the case that the outcome variable is discrete, taking on two or more possible values. Over the last decade the logistic regression model has become, in many fields, the standard method of analysis in this situation. Let us now take a closer look at the statistical modeling of binary response variables, for which the response outcome for each subject is a "success" or "failure." Binary data are the most common form of categorical data, and the methods are of fundamental importance. The most popular model for binary data is logistic regression.

The application of logistic regression model in public health area is very common. Generalized linear models are the most important model for categorical response data. It is used increasingly in a wide variety of applications. Early uses were in biomedical studies but the past 20 years have also seen much use in social science research and marketing. (Agresti, 2002). Logistic regression is an increasingly popular statistical technique used to model the probability of discrete (i.e. binary or multinomial) outcomes. When properly applied, logistic regression analyses yield very powerful insights in to what attributes (i.e., variables) are more or less likely to predict event outcome in a population of interest. These models also show the extent to which changes in the values of the attributes may increase or decrease the predicted probability of event outcome.

Many clinical trials deal with the comparison of populations of subjects with categorical responses. Historically, statistical inference for such studies involved large-sample approximations, and fitting logistic regression models to such data was performed through the unconditional likelihood function. The multivariate normality assumption inherent when categorical predictors are present, logistic regression should be the analysis of choice for most research studies involving a qualitative criterion. With model that posits more realistic nonlinear relationships among variables, Educational researchers can achieve enhanced statistical precision

in estimating prediction equation, which should facilitate improved decision making (King, 2002).

There are many situations in which the response of interest is dichotomous rather than continuous. Examples of variables that assume only two possible values are disease status (the disease is either present or absent) and survival following surgery (a patient is either alive or dead). In general the value 1 is used to represent a "success" or the outcome we are most interested in, and 0 represents a "failure". Just as estimating the mean of a response variable Y when it is continuous, the probability of a dichotomous response (which of course is also its mean) for various values of explanatory variables can be estimated. To do this, Logistic Regression is commonly used.

Logistic regression can be binary or multinomial. The binary or Binomial logistic regression is the type of regression which is used when the dependent variable is a dichotomous and the independent variables are of any type while Multinomial logistic regression is used when the dependent variable has more than two categories. When multiple classes of the dependent variable can be ranked, then ordinal logistic regression is preferred to multinomial logistic regression.

The logistic regression applies maximum likelihood estimation after transforming the dependent into a logit variable (the natural log of the odds of the dependent variable occurring or not). In this way, logistic regression estimates the probability of a certain event occurring. Note that logistic regression calculates changes in the log odds of the dependent variable, not changes in the dependent variable itself as OLS regression does. However, logistic regression has many analogies to OLS regression; logit coefficients correspond to β coefficients in the logistic regression equation, the standardized logit coefficients correspond to beta weights, and a pseudo R^2 statistic is available to summarize the strength of the relationship. Logistic regression does not assume linearity of relationship between the independent variables and the dependent variable,

does not assume homoscedasticity, and in general has less stringent requirements. The logistic regression is also preferred from multiple regression and discriminant analysis as it results a meaningful interpretation, it is mathematically flexible and easily used distribution and it requires fewer assumptions (Hosmer, *et al.*, 2000).

3.8.2. Binary Logistic Regression Model

Binary Logistic Regression is a modeling framework that can be used to describe the relationship of several explanatory variables (in this case the risk factors, X's) to a dichotomous (binary) dependent variable (presence of toxoplasmosis, Y). The outcome variable Y_i (i=1,.....n) follows a Bernoulli probability function that takes value 1 when toxoplasmosis present with probability p_i and takes the value 0 when toxoplasmosis is absent with probability $1-p_i$.

In this study, 15 risk factors are considered as explanatory variables to describe the probability of presence of toxoplasmosis among HIV/AIDS patients. The model is given for the logit, probit and complementary log log link functions as follows:

- **Logit Model**

Using the Logit link function, the model is described in the form of logit (natural logs of the odds):

$$Logit\ (P(Y_i=1)) = Log\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right) = \beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15} = \sum_{j=0}^{16} \beta_j X_{ij}, X_{i0} = 1, \ldots . (1)$$

i=1,2,…,n.

Where

X_j is the j^{th} risk factor

β_0 is the constant (intercept) of the logit model

β_j is the estimated coefficient for each risk factor j for the logit model.

$P(Y_i = 1)$ is the probability that the ith patient will develop toxoplasmosis.

The probability of developing toxoplasimosis for each patient can be expressed interms of the estimated coefficients and given values of the risk factors as below.

$$P(Y_i = 1) = \frac{1}{e^{-(\beta_0 + \beta_1 X_{i1} + ... + \beta_{15} X_{i15})}} \text{..... .... (2)}$$

We might try to solve this problem by fitting the model as $p = e^{\beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15}}$. This equation/model guarantees merely the estimate of p is positive. Otherwise the term $e^{\beta_0 + \beta X}$ although cannot be negative, it can result in a value that is greater than 1.To accommodate this final constraint, we fit a model of the form

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15}}} \text{.....(3)}$$

The expression on the right, called a logistic function, cannot yield a value that is negative or greater than 1; consequently restricting the estimated value of p to the required range (between 0 and 1).

If an event occurs with probability p, then the odds in favor of the event are $\frac{p}{1-p}$ to 1. Thus, if a success occurs with probability $p = \frac{e^{\beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15}}}$

The odds in favor of success are

$$\frac{p}{1-p} = \frac{e^{\beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15}} \Big/ 1 + e^{\beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15}}}{1 \Big/ 1 + e^{\beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15}}}$$

$$= e^{\beta_0 + \beta x}$$

Taking the natural logarithm of each side of this equation,

$$n[\frac{p}{1-p}] = \ln[e^{\beta_0+\beta_1 X_{i1}+..+\beta_{15}X_{i15}}]$$
$$= \beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15} ...$$

Thus, modeling the probability p with logistic function is equivalent to fitting a linear regression model in which the continuous response y has been replaced by the logarithm of the odds of success for a dichotomous random variable. Instead of assuming that the relationship between p and x is linear, we assume that the relationship between $\ln[\frac{p}{1-p}]$ and x is linear. The technique of fitting a model of this form is known as logistic regression using a logit link.

- **Probit (Normit) Model**

Using the probit link function, the probit model is described as the inverse of the normal distribution function.

$$\Pr obit\,(P(Y_i = 1)) = \Phi^{-1}(P(Y_i = 1)) = \beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15} = \sum_{j=0}^{16} \beta_j X_{ij}, X_{i0} = 1, ...(5)$$

i=1,2,…,n.

Where

$\phi^{-1}(p(Y_i = 1))$ is the inverse cumulative Normal probability function.

X_j is the j^{th} risk factor

β_0 is the constant (intercept) of the probit model

β_j is the estimated coefficient for each risk factor j for the probit model.

$P(Y_i = 1)$ is the probability that the i^{th} patient will develop toxoplasmosis.

- **Complementary log log (C-log-log) Model**:

While using the cloglog link function, the model is given as:

$$Log\ (-Log\ (1-P(Y_i=1)) = \beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15} = \sum_{j=0}^{16} \beta_j X_{ij}, X_{i0} = 1, ...(6)$$

i =1,2,…,n.

Where

X_j is the j^{th} risk factor

β_0 is the constant (intercept) of the model

β_j is the estimated coefficient for each risk factor j

$P(Y_i = 1)$ is the probability that the i^{th} patient will develop toxoplasmosis.

The probability of developing toxoplasimosis for each patient can be expressed interms of the estimated coefficients and given values of the risk factors as below.

$$P(Y_i = 1) = 1 - e^{-e^{\beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15}}} \quad .... (7)$$

3.8.2.1. Assumptions regarding Binary Logistic Regression model

As indicated in the previous sections, one advantage of the logistic regression is it gives some relaxation with respect to the usual OLS assumption. Assumptions one should consider for the efficient use of logistic regression is:

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, such as before-after or matched-pairs studies, cluster sampling,

or time-series data. That is, subjects cannot provide multiple observations at different time points.

3.8.3. Model Estimation

3.8.3.1. Maximum likelihood estimation using R software

Suppose we have a sample of n independent observations of the pair (X_i, y_i), i=1,2,……,n, where y_i denotes the value of the dichotomous outcome variable and X_i is a vector of independent variables for the i^{th} subject. Fitting the model in equation (1) requires that we obtain estimates of the values of β_0 and β_i (represented by a vector β').

In linear regression the method used most often for estimating unknown parameters is least squares. In that method we choose those values of β' which minimize the sum of squared deviations of the observed values of Y from the predicted values based upon the model. Under the usual assumptions for linear regressions, the method of least squares yields estimators with a number of desirable statistical properties. Unfortunately when the method of least squares is applied to a model with a dichotomous outcome, the estimators no longer have these same properties (Hosmer, *et al.*, 2000).

The general method of estimation that leads to the least squares function under the linear regression model (when the error terms are normally distributed) is called maximum likelihood. It is this method that provides the foundation for our approach to estimation with the logistic regression model. In a very general sense the method of maximum likelihood yields values for the unknown parameters which maximize the probability of obtaining the observed set of data. In order to apply this method we must first construct a function called likelihood function. The maximum likelihood estimators of these parameters are chosen to be those values which maximize this function. Thus, the resulting estimators are those which agree most closely with the observed data.

If Y is coded as one or zero, then the expression for p(X) in equation (3) above provides for an arbitrary value of β', the vector of parameters) the conditional probability that Y is equal to 1 given X (i.e., p(Y=1|X)). It follows that 1-p(X) gives the conditional probability that Y is equal to 0 given x, p(Y=0|x). Thus, for those pairs (x_i,y_i),where $y_i=1$, the contribution to the likelihood function is $p(x_i)$, and for those pairs where $y_i=0$ the contribution to the likelihood function is $1-p(x_i)$, where the quantity $p(x_i)$ denotes the value of p(x) computed at x_i . A convenient way to express the contribution to the likelihood function for the pair (x_i,y_i) is through the Bernoulli distribution,

$$\theta(x_i) = p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i}, \text{ where } p(x_i) = \frac{e^{\beta_0+\beta_1 X_{i1}+..+\beta_{15}X_{i15}}}{1+e^{\beta_0+\beta_1 X_{i1}+..+\beta_{15}X_{i15}}}, i = 1, 2, ..., n$$

Since the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in the above expression as follows:

$$l(\beta) = \prod_{i=1}^{n} \theta(x_i) \quad \text{.....(8)}$$

The principle of maximum likelihood states that we use as our estimate of β the value which maximizes the expression in equation (8). However, it is easier mathematically to work with the log of equation (8). This expression, the log likelihood is defined as

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \ln[p(x_i)] + (1 - y_i)\ln[1 - p(x_i)]\}$$

To find the value of β that maximizes $L(\beta)$ we differentiate $L(\beta)$ with respect to β_0 and β_i and set the resulting equation to zero. These equations are as follows:

$$\sum_{i=1}^{n} [y_i - p(x_i)] = 0 \text{.....(9)}$$

$$\sum_{i=1}^{n} x_i [y_i - p(x_i)] = 0 \text{.....(10)}$$

And these equations are called the loglikelihood equations. For logistic regression the expressions in (9) and (10) are non linear in β_0 and β_i, and thus require special methods for their solution. These methods are iterative (like Newton Raphson) in nature and have been programmed into available logistic regression packages like R software (version 2.13.1).

3.8.4. Model Building Strategies/Variable Selection

In modeling with many independent variables, one is usually concerned with the goal of selecting those variables that result in the "best" model within the scientific context of the problem. Having a basic plan to follow in selecting the variables for the model, assessing the adequacy of the model both in terms of the individual variables and from the point of view of the overall fit of the model is required for achieving this "best" model. It is also highlighted in Hosmer, *et al*. (2000) that successful modeling of a complex data set is part science, part statistical methods, and part experience and common sense.

The traditional approach to statistical model building involves seeking the most parsimonious model that still explains the data. Recently researchers are shifting to including all scientifically relevant variables in the model, irrespective of their contribution to the model. This is based on the fact that it is possible for individual variables not to exhibit strong association while they do show considerable association when taken collectively. Both approaches have their merits and demerits as indicated in (Hosmer, *et al*., 2000).

The logistic regression modeling approach used in our case, is a forward stepwise likelihood ratio method using R software (version 2.13.1), which begins by selecting the strongest candidate predictor (risk factor), then testing additional candidate predictors (risk factors), one at a time, for inclusion in the model. This stepwise procedure continues until all the candidate risk factors have been thoroughly tested for inclusion and removal. Sometimes the goal of the analysis may be broader, and models containing more variables are sought to provide a more complete picture of possible models. In these cases, use of p-value=0.25 or even larger might be a reasonable choice. Whatever the choice for the variable using p-value judged important enough to include in the model if it is less than 0.25. (Hosmer, *et al*, 2000). So the p-value=0.25 is used in this research to begin the process by selecting candidate variables from bivariate analysis. Other variable selection methods like backward and forward are also possible.

3.8.5. Interpretation of the Parameters

- From $Logit\ (P(Y_i = 1)) = \beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15}$, the odds are an exponential function of X. This provides a basic interpretation of the magnitude of β. Note $\frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = e^{\beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15}}$ is the odds of developing toxoplasmosis for the i^{th} patient.

  For every one unit increase in X the odds increases multiplicatively by e^{β_j}. Thus, e^{β_j} is an odds ratio, the odds at category 1 divided by the odds at the reference category?

- But the model

  $\Pr obit\ (P(Y_i = 1)) = \Phi^{-1}(P(Y_i = 1)) = \beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15}$, do not provide an estimate of the odds ratios.

- Also the model

  $Log\ (-Log\ (1 - P(Y_i = 1)) = \beta_0 + \beta_1 X_{i1} + .. + \beta_{15} X_{i15}$, do not provide an estimate of the odds ratios.

3.8.6. Model Checking and Diagnostic

3.8.6.1. The Likelihood Ratio Test

Once a model has been developed through the various steps as explained in the earlier sections, we now would like to know how, well the model is fitted to the data. This is what is referred to as goodness-of-fit. We say our model fits when summary measures of the distance between the observed response variable $Y' = (y_1, y_2, \ldots\ldots, y_n)$ and its fitted value $\hat{y} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$ are small and the contribution of each pair $(y_i, \hat{y}_i)$, i=1,2,3,……,n to these summary measures is unsystematic and is small.

Taken collectively, does the entire set of the risk factors contribute significantly for the prediction of developing toxoplasmosis? To see whether the variables included in the model together have an improvement than the model containing no variable. A binary logistic model is said to provide a better fit to the data if it demonstrates an improvement over the intercept-only model (also called the null model). An intercept-only model serves as a good baseline because it contains no predictors (risk factor). An improvement over this baseline is examined by using the inferential statistical test called the likelihood ratio test. Let L_1 denote the maximized value of the likelihood function for the fitted model with factors and let L_0 denote the maximized value for the null model. Then L_1 is at least as large as L_0. The likelihood ratio test statistic is

$$\chi^2_{LR} = -2\log\left(\frac{L_0}{L_1}\right) = -2(\log L_0 - \log L_1) = -2(\ell_0 - \ell_1).$$

The log likelihood ratio in logistic regression is analogous to F test in linear regression. The computer output for each model will give the log-likelihood statistic. The difference between log likelihood statistics for two models has an approximate chi-square distribution in large samples. The hypothesis to be tested is: H_0: $\beta_j = 0$ for all j.

$$H_1: \beta_j \neq 0 \text{ for at least one j}$$

So the hypothesis that the full model makes no difference is tested by computing χ^2_{LR}. Having obtained the value of χ^2_{LR} its significance level can be looked up by entering the table of chi-square at q-1 degree of freedom where q is the number of parameters in the full model. A large value of this test statistic would indicate that at least one of the regressed variables in the model is important because it has a nonzero regression coefficient (Montgomery, *et al.*, 2006). Rejection of the null hypothesis in the overall test has an interpretation analogous to that in the multiple linear regressions; we may conclude that at least and the coefficients are different from zero.

3.8.6.2. Wald Test

The Wald test is used to test the statistical significance of each coefficient (β_j) in the model. The Wald statistic is obtained by dividing the value of the estimated β coefficient of the factor of interest by its standard error, i.e. $W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$ for each j.

The hypothesis to be tested is:

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$

For large samples W^2 (Wald squared) will have approximately chi-square distribution with 1 degree of freedom. Thus, Wald statistic provides information as the whether the β coefficient of an explanatory variable (risk factors) is significantly different from zero. It is analogous to the 't' test in multiple linear regression.

3.8.7. Goodness of fit statistics

3.8.7.1. Pearson and Deviance Tests

These are both measures of goodness of fit of the data. The hypothesis sates that:

H_0: the model fits the data well

H_1: the model does not fit the data well

Goodness of fit can be assessed using a Pearson chi-square statistic that compares the observed and expected probabilities of success and failure at each group of observations. The expected number of successes is $n_i \hat{\pi}_i$ and the expected number of failures is $n_i(1 - \hat{\pi}_i)$, i=1,2,…..,n. the Pearson chi-square goodness-of-fit statistic can be compared with the critical values from the chi-squared distribution with n-p degrees of freedom. Small values of the statistic (or a large p-value) imply that the model provides a satisfactory fit of the data (Montgomery, *et al.*, 2006).

When the model is an adequate fit to the data and the sample size is large, the deviance has a chi-square distribution with n-p degrees of freedom, where p is the number of parameters in the model. Small values of the deviance (or a large p-value) imply that the model provides a satisfactory fit to the data (no evidence for a lack of fit of the model), while large values of the deviance imply that the current model is not adequate. This indicates that there is no evidence to suggest that data did not come from population that follows the logistic regression (no evidence for a lack of fit of the model) (Montgomery, *et al*., 2006).[1]

3.8.7.2. Pseudo R^2 statistics

There are also descriptive measures of goodness-of-fit, called Pseudo R^2 statistics, which represent the proportional improvement in the log-likelihood function due to the terms in the model of interest, compared with the minimal model. It is defined by Cox, *et al.* (1989) and (Nagelkerke, 1991), respectively. These indices are variations of the R^2 concept defined for the OLS regression model. In linear regression, R^2 has a clear definition: It is the proportion of the variation in the dependent variable that can be explained by predictors in the model but cannot be computed for logistic regression models because the dependent variable is dichotomous rather than continuous. Attempts have been devised to yield an equivalent of this concept for the logistic model. None, however, renders the meaning of variance explained (Menard, 2000). Furthermore, none corresponds to predictive efficiency or can be tested in an inferential framework (Menard). For these reasons, a researcher can treat these two R^2 indices as supplementary to other, more useful evaluative indices, such as the overall evaluation of the model, tests of individual regression coefficients, and the goodness- of-fit test statistic. The higher the value of pseudo R2, the better the model fits the data well.

3.8.7.3. Information Criteria

The Akaike Information Criterion (AIC) and the Schwartz or Bayesian Information Criteria (BIC) are goodness of fit statistics based on the log-likelihood function with an adjustment for

the number of parameters estimated and for the amount of data. These statistics are usually defined as: $AIC = -2\ell_1 + 2p$ and $BIC = -2\ell_1 + 2p \ln n$ where p is the number of parameters estimated and n is the number of observations. The smaller the AIC and BIC values are the better the model to explain the observed data.

The statistics (except for Pseudo R^2) discussed above summarize how well a particular model fits the data. So a small value of the statistic and, hence a large p-value, indicates that the model fits well. These statistics are not usually appropriate for testing hypothesis about the parameters of nested models, but they can be particularly useful for comparing models that are not nested.

3.8.7.4. Over dispersion

Another issue in assessing the adequacy of the model for binary data is over dispersion. Observations y_i may have observed variances greater than binomial variance $n_i\pi_i(1-\pi_i)$, or equivalently $\text{var}(\hat{\pi}_i)$ may be greater than $\pi_i(1-\pi_i)/n_i$. There is an indicator of this problem if the deviance D is much greater than the expected value of N-p. This could be due to inadequate specification of the model (e.g., relevant explanatory variables have been omitted or the link function is incorrect) or to a more complex structure. One approach is to include an extra parameter ϕ in the model so that $\text{var}(Y_i) = n_i\pi_i(1-\pi_i)\phi$ (Dobson, *et al.*, 2008).

3.8.7.5. Residuals

Pearson and standardized residuals can be used for checking the adequacy of a model. For example, they should be plotted against each continuous variable in the model to check if the assumption of linearity is appropriate and against other possible explanatory variables not included in the model. They should be plotted in the order of the measurements, if applicable, to check for serial correlation. Normal probability plots can also be used because the standardized residuals should have, approximately, the standard normal distribution N(0,1), provided the number of observation for each covariate pattern are not too small.

4. Results and Discussion

4.1. Introduction

The overall proportion of toxoplasmosis in the Mettu Karl Hospital (TOXO=2) is around 60%. The data analysis for this study is started by giving the summary statistics for the variables considered in the study; then proceeds to the bivariate analysis and complete the final model.

4.2. Summary Statistics

4.2.1. Socio-Economic factors in relation to toxoplasmosis

Toxoplasmosis proportions for the HIV/AIDS patient that came to Mettu Karl Hospital for ART treatment shows that the prevalence in urban residence is 58.3%. The proportion of toxoplasmosis in the absence of tap water is 86.1% as compared to the presence of tap water 13.9%. From occupation categories; unemployed has higher percentage 44.4% of toxoplasmosis positive and merchant has less percentage 2.8% as compared to other groups. All the results have been summarized in Table 2 below.

Table-2: Socio-Economic factors by toxoplasmosis status, in Mettu Karl Hospital, 2010.

Socio-Economic factors	Toxoplasmosis status	
	Positive (%)	Negative (%)
Place of residence (ADR)		
Urban	42(58.3%)	28(58.3%)
Rural	30(41.7%)	20(41.7%)
Presence of tap water (TAPWATER)		
Yes	10(13.9%)	24(50%)
No	62(86.1%)	24(50%)
Occupation (OCC)		
Farmer	14(19.4%)	2(4.2%)
Merchant	2(2.8%)	6(12.5%)
Employee	12(16.7%)	6(12.5%)
Unemployed	32(44.4%)	28(58.3%)
Others	12(16.7%)	6(12.5%)

4.2.2. Demographic and Health factors in relation to toxoplasmosis

In this study a total of 120 patients were involved, of which 72 (60%) are toxoplasmosis positive. Male are 26(36.1%) as compared to Females 46(63.9%). Age group between 30-41 has a proportion of 16.7%, but the other groups have the same proportion of toxoplasmosis positive. The illiterate and 9+ groups show a proportion 36.1%, 33.3% respectively. The proportion of toxoplasmosis in the absence of cat is 52.8%. And also the proportion of toxoplasmosis is low in the presence of eye disease 16.7% and fever 36.1%. The results have been summarized in Table 3 below.

Table-3: Demographic and Health factors by toxoplasmosis status, in Mettu Karl Hospital, 2010.

Demographic and health factors	Toxoplasmosis status	
	Positive (%)	Negative (%)
Sex of the patients (SEX)		
Male	26(36.1%)	8(16.7%)
Female	46(63.9%)	40(83.3%)
Age of the patients (AGE)		
18-23	20(27.8%)	8(16.7%)
24-29	20(27.8%)	12(25%)
30-41	12(16.7%)	12(25%)
42+	20(27.8%)	16(33.3%)
Marital status (MAR)		
Married	38(52.8%)	34(70.8%)
Single	22(30.6%)	12(25%)
Others	12(16.7%)	2(4.2%)
Education (EDUC)		
Illiterate	26(36.1%)	12(25%)
1-4	14(19.4%)	4(8.3%)
5-8	8(11.1%)	18(37.5%)
9+	24(33.3%)	14(29.2%)
Compound latrine (CPDLATRINE)		
Yes	42(58.3%)	40(83.3%)
No	30(41.7%)	8(16.7%)
House toilet (HOUSETOILET)		

Yes	6(8.3%)	12(26.1%)
No	66(91.7%)	34(73.9%)
Cat presence (CAT)		
Yes	34(47.2%)	16(33.3%)
No	38(52.8%)	32(66.7%)
Eye disease (EYE)		
Yes	12(16.7%)	14(29.2%)
No	60(83.3%)	34(70.8%)
Fever (FEVER)		
Yes	26(36.1%)	16(33.3%)
No	46(63.9%)	32(66.7%)

4.2.3. Risk Behavior factors in relation to toxoplasmosis

The results of summary statistics for the risk Behavior factors revealed that coincide with the usual instructions of physicians to their patients. The data show that the risk behavior of eating raw meat has higher proportion (70%) of toxoplasmosis positive than not eating raw meat. The proportions of toxoplasmosis among those with habit of washing hand and fruit regularly are 48.8% and 46.9%, respectively.

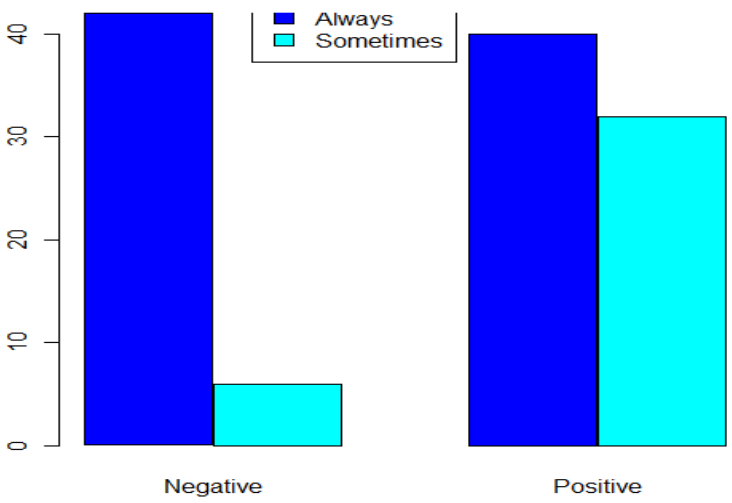


Figure-1: Distribution of Toxoplasmosis on the risk Behavior factors of washing hand, in Mettu Karl Hospital, 2010.

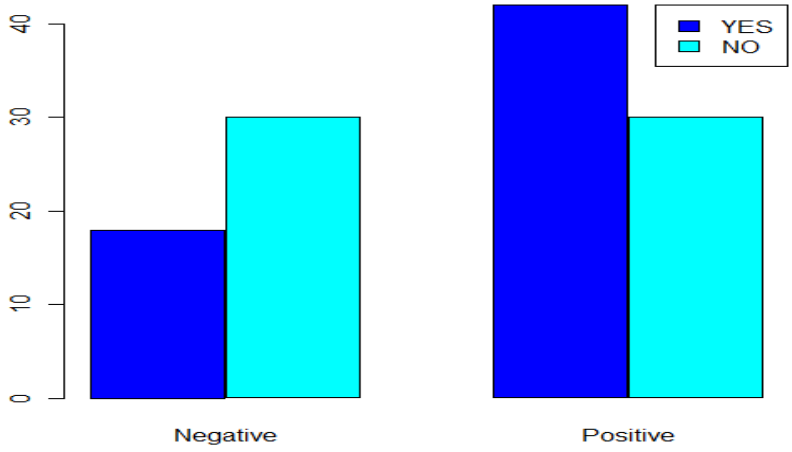


Figure-2: Distribution of toxoplasmosis on the risk factors of raw meat, in Mettu Karl Hospital, 2010.

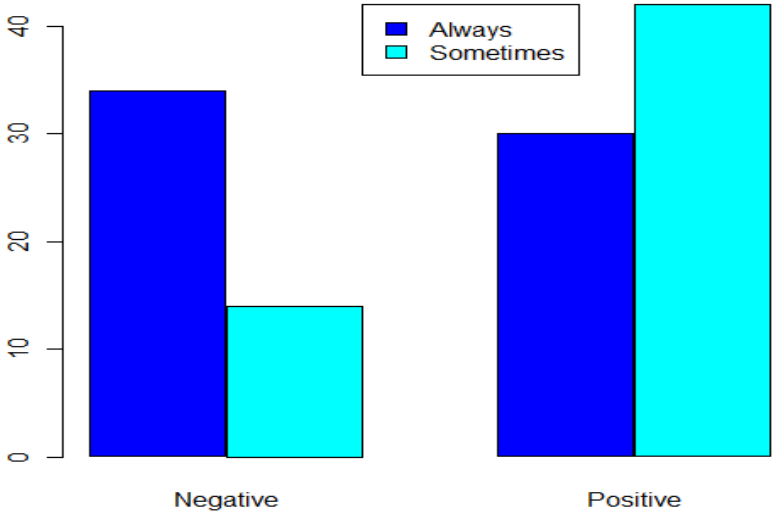


Figure-3: Distribution of toxoplasmosis on the risk factor of washing fruit, in Mettu Karl Hospital, 2010.

4.3. Bivariate Analysis

To see the association between one of the risk factors (independent variables) and the dependent variable (toxoplasmosis) (by taking each of the independent variable at a time), the Pearson Chi-square (χ^2) test is used.

Table- 4: Independent variables used in the Bivariate analysis in Mettu Karl Hospital, 2010.

No	Variable	Pearson Chi-square	
		Value	p-value
1	Sex of the patient (SEX)	5.363	0.021*
2	Age of the patient (AGE)	2.903	0.407
3	Place of Residence (ADR)	2.0736e-30	0.9999
4	Occupation (OCC)	10.903	0.028*
5	Marital status (MAR)	5.736	0.057
6	Education (EDUC)	12.907	0.005*
7	Presence of tap water (TAPWATER)	18.495	0.000*
8	Compound latrine (CPDLATRINE)	8.318	0.004*
9	House toilet (HOUSETOILET)	6.843	0.009*
10	Cat presence (CAT)	2.2857	0.1306
11	Habit of eating raw meat (RAWMEAT)	5	0.025*
12	Habit of washing hand (HAND)	13.582	0.000*
13	Habit of washing fruit (FRUIT)	9.844	0.002*
14	Eye Disease (EYE)	2.651	0.104
15	Feeling Fever (FEVER)	0.098	0.755

*significant (p-value < 0.05)

The variables that are found significant in Table 4 are sex of the patient (SEX), Education (EDC), occupation (OCC), marital status (MAR), Compound latrine (CPDLATRINE), Ownership of toilet (HOUSETOILET), presence of tap water (TAPWATER), eating raw meat (RAWMEAT), washing hand (HAND) and washing fruit (FRUIT).

4.4. Binary Logistic Regression Analysis

One problem with any bivariate 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. Using binary logistic regressions the forward stepwise variable selection criteria carried out using the three link functions.

4.4.1. The Logit Model

In an automated variable selection using the logit link function, when constructing a model, generally we include only risk factors that contribute significantly to the model. The modeling procedure in R offers several methods for stepwise selection of the best risk factors to include in the model. The final model is,

$$\log it(P(Y=1)) = 3.8729 - 3.7290\ TAPWATERye\ s - 3.9797\ EDUC\ 2_{5-8} - 4.0249\ MAR\ 1married + 3.1369\ HANDSometi\ mes + 1.8188\ RAWMEATyes - 2.3263\ HOUSETOILE\ Tyes$$

Table-5: Parameter Estimates of the Binary Logistic Regressions Using Logit Link Function, in Mettu Karl Hospital, 2010.

Variable	Estimate	Std. Error	Z value	p-value
Intercept	3.8729	0.9799	3.952	7.74e-05
TAPWATERyes	-3.7290	0.9589	-3.889	0.000101
EDUC2(5-8)	-3.9797	1.0277	-3.872	0.000108
MAR1married	-4.0249	1.0216	-3.940	8.15e-05
HANDSometimes	3.1369	0.9865	3.180	0.001474
RAWMEATyes	1.8188	0.6845	2.657	0.007879
HOUSETOILETyes	-2.3263	0.8786	-2.648	0.008105

The variables found to be significant in the logit models are Education (EDUC2), marital status (MAR1), tap water (TAPWATER), house toilet (HOUSETOILET) and hand washing (HAND)

and raw meat (RAWMEAT). And this is in effect in line with the results obtained from the univariate analysis. The value of the Wald statistic for individual β coefficients support that the estimated values ($\hat{\beta}$) are significantly different from zero. All the above six categorical variable as compared to the reference groups are significant. The remaining variables which were used in the bivariate analysis are found to be non significant.

4.4.1.1. Model Checking and Diagnostic

The Likelihood Ratio Test: For the above logit model of the toxoplasmosis data, the log likelihood value, ℓ_0, for the model with no explanatory variable (null model) is -79.8388 (df=1) and that of the log likelihood value, ℓ_1, for full model is -38.46087 (df=7) which implies a model chi-square of χ^2_{LR} = -2 (-41.37793) = 82.75586 (df=6) with corresponding p-value 2.647882e-13 which signifies significant increase in the likelihood thereby implying a good fit of the model.

4.4.1.2. Goodness of fit statistics

Pearson and Deviance Tests: These are both measures of goodness of fit of the data. The larger the *P* value the better is the fit of the model to the data. Here for the above model, the Pearson and Deviance test values are 67.354 and 76.922 with their corresponding p-values 0.99965 and 0.9942783 respectively. This indicates that there is no evidence to suggest that the data did not come from population that follows the logistic regression (no evidence for a lack of fit of the model).

Pseudo R^2 statistics:

Since the log likelihood value, ℓ_0, for the model with no explanatory variable (null model) is -79.8388 (df=1) and that of the log likelihood value, ℓ_1, for full model is -38.46087 (df=7), then the Pseudo R^2 value for this model is $(\ell_0 - \ell_1)/\ell_0 = 0.51$ indicating a reasonable fit of the model.

Information Criteria: The Akaike Information Criterion (AIC) and the Schwartz or Bayesian Information Criteria (BIC) is goodness of fit statistics based on the logit model is 94.24 and 169.56 respectively.

Over dispersion: the estimated dispersion parameter, for the logit model is around 0.6929887, which is essentially bellow one and thus there is no evidence of over dispersion.

Residuals:

Plot of Residuals: Of course for binary data, the plot of the raw and Pearson residuals are relatively not informative as there may be few distinct values of the residuals. In general, the plot of the raw, Pearson, standardized residuals and normal probability plot of the standardized residuals is as follows.

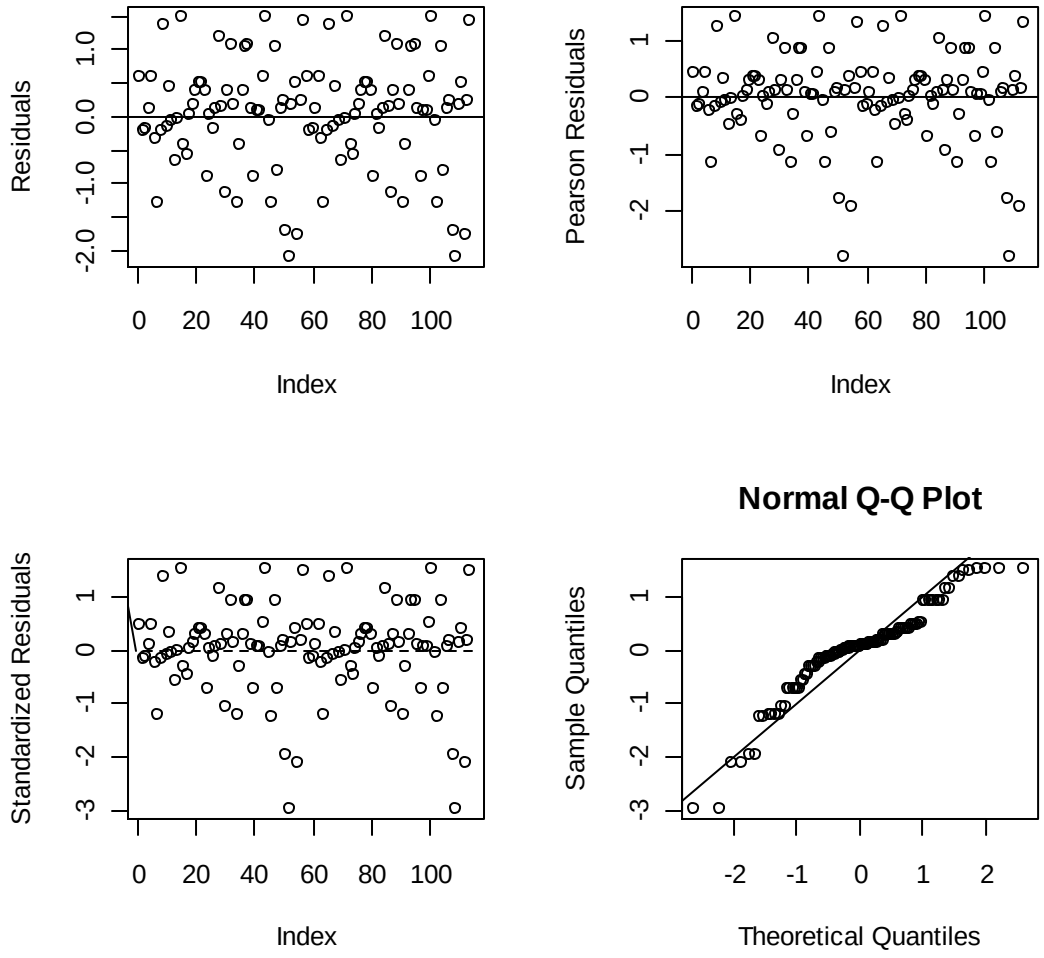


Figure-4: Plot of the raw, Pearson and standardized residuals, and normal probability plot for the logit model, in Mettu Karl Hospital, 2010.

As can be seen from the above plot, the raw, Pearson and standardized residuals are random or no systematic patterns are observed. This indicates that there is no serial correlation between the residuals. Also, in the normal probability plot, the points lie on or near the straight line representing normality and no serious systematic deviations and outlying observations is observed. Therefore the standardized residuals are normally distributed.

4.4.2. The Probit Model

Model formulation for the probit link function is:

$$\log it(P(Y=1) = 2.2999 - 2.1888\ TAPWATERye\ s - 2.3396\ EDUE\ 2_{5-8} - 2.4070\ MAR\ 1married + \\ + 1.8487\ HANDSometi\ mes + 1.0733\ RAWMEATyes - 1.4279\ HOUSETOLET\ yes$$

Table-6: Parameter Estimates of the Binary Logistic Regressions Using Probit Link Function , in Mettu Karl Hospital, 2010.

Variable	Estimate	Std. Error	Z value	p-value
Intercept	2.2999	0.5494	4.186	2.83e-05
TAPWATERno	-2.1888	0.5348	-4.093	4.26e-05
EDUC25-8	-2.3396	0.5710	-4.098	4.17e-05
MAR1married	-2.4070	0.5704	-4.220	2.45e-05
HANDSometimes	1.8487	0.5340	3.462	0.000536
RAWMEATyes	1.0733	0.3906	2.748	0.006004
HOUSETOILETyes	-1.4279	0.5035	-2.836	0.004564

The variables that are found to be significant in the probit link function models as indicated in the Table 6 are Education (EDUC2), marital status (MAR1), tap water (TAPWATER), house toilet (HOUSETOILET) and hand washing (HAND) and raw meat (RAWMEAT). In addition to this, the value of the Wald statistic for individual β coefficients support that the estimated values ($\hat{\beta}$) are significantly different from zero using a p-value < 0.05 for cut-off point. The remaining variables which were used in the bivariate analysis are found to be non significant.

4.4.2.1. Model Checking and Diagnostic for the probit link function

The log likelihood value for null model is -76.88 (df=1) and that of the log likelihood value for full model is -37.8848 (df=7) which implies a model chi-square of χ^2_{LR} = -2 (-41.954) = 83.908 (df=6) with corresponding p-value 1.597e-13 which signifies significant increase in the

likelihood thereby implying a good fit of the model. The Pearson chi-square test and Deviance test have value 65.38502 with the p-value of 0.9998263 indicating that the probit model fits the data. The pseudo R^2 value is around 0.52. This indicates good fit. AIC for the probit model is 93.74 and the BIC is 168.464. The estimated dispersion parameter, for the probit model is around 0.6826091, which is essentially bellow one and thus there is no evidence of over dispersion. Like that of the logit model, as can be seen from the above plot, the raw, Pearson and standardized residuals are random or no systematic patterns are observed which indicates that there is no serial correlation between the residuals. Also, in the normal probability plot, the points lie on or near the straight line representing normality and no serious systematic deviations and outlying observations is observed. Therefore the standardized residuals are normally distributed for the probit model (See the figure in the Appendix page 64).

4.4.3. The Complementary log log (c-log-log) Model

The final model is,

$$\log it(P(Y=1)) = 2.2103 - 2.4605\ TAPWATERno - 2.5648\ EDUC\ 2_{5-8} - 2.7780\ MAR\ 1married - 1.8276\ HOUSETOILE\ Tyes + 1.9251\ HANDSometi\ mes + 1.1356\ RAWMEATyes$$

Table-7: Parameter Estimates of the Binary Logistic Regressions Using C-log-log Link Function, in Mettu Karl Hospital, 2010.

Variable	Estimate	Std. error	Z value	Pr (> | z |)
Intercept	2.2103	0.6061	3.646	0.000266
TAPWATERno	-2.4605	0.5832	-4.219	2.46e-05
EDUC25-8	-2.5648	0.6807	-3.768	0.000165
MAR1married	-2.7780	0.6670	-4.165	3.11e-05
HANDSometimes	1.9251	0.5465	3.523	0.000427
RAWMEATyes	1.1356	0.4094	2.774	0.005534
HOUSETOILETyes	-1.8276	0.6121	-2.986	0.002826

The variables that are found to be significant in the c-log-log link function model as shown in Table 7 are Education (EDUC2), marital status (MAR1), tap water (TAPWATER), house toilet (HOUSETOILET) and hand washing (HAND) and raw meat (RAWMEAT). And this is in effect in line with the results obtained from the bivariate analysis. The value of the Wald statistic for individual β coefficients support that the estimated values ($\hat{\beta}$) are significantly different from zero using a p-value < 0.05 for cut-off point. The remaining variables which were used in the bivariate analysis are found to be non significant.

4.4.3.1. Model Checking and Diagnostic for the c-log-log Model

For the c-log-log model, the log likelihood value null model is -79.8388 (df=1) and that of the log likelihood value for full model is -38.15534 (df=7) which implies a model chi-square of χ^2_{LR} =83.36692 (df=6) with corresponding p-value 1.685e-13 which signifies significant increase in the likelihood. The pearson chi-square test value 65.09523 (df=111) with the p-value of 0.9998439. Also the deviance chi-square test supports that the model fits well with a p-value of 0.995082. The pseudo R^2 value is around 0.51. This indicates good fit of the c-log-log model to the toxoplasmosis data. The AIC value is 93.86 and the BIC is 168.548. The estimated dispersion parameter, for the c-log-log model is around 0.71, which is essentially bellow one and thus there is no evidence of over dispersion. Like that of the logit and probit model, plot of the raw, Pearson and standardized residuals are random or no systematic patterns are observed which indicates that there is no serial correlation between the residuals. Also, in the normal probability plot, the points lie on or near the straight line representing normality and no serious systematic deviations and outlying observations is observed. Therefore the standardized residuals are normally distributed for the c-log-log model (See the figure in the Appendix page 65).

4.4.4. Model Comparison and selection

Binary logistic regressions fitted for the three link function as shown in the above. The risk factors (covariates) that included in the three link function are the same.

.

To select the best model among non nested models is to use the Akaike's information criterion (AIC). The model with the smallest AIC is selected as a best model.

Table-8: Summary of the Likelihood and AIC values for the three models, in Mettu Karl Hospital, 2010.

Model	Log likelihood	No of Parameters	AIC	BIC	Pseudo R^2
Logit	-38.4668	7	94.24	169.56	0.51
Probit	-37.8848	7	93.74	168.46	0.52
C-log-log	-38.1553	7	93.86	168.54	0.51

Therefore from Table 8 as shown above the probit model has the largest log likelihood value than the other two models. Also the AIC and BIC values for the probit model are the smallest values than that of the logit and c-log-log models. And the probit model has also the highest pseudo R^2 which indicates that the model fits the data well than the other two models. Therefore, the probit model seems the best model. However, its difference with the other two is not as such considerable.

One drawback of probit and c-log-log link function is that it is difficult to interpret the coefficients of the risk factors. As a result of this it is impossible to get the odds ratio of the dependant variable. One of the reasons many health researches done using logistic regression model applying logit link function the ease of interpreting to show the relative risk (odds ratio) of the dependant variable due to the risk factors. Based on this, the interpretation of the risk factors to the contribution of toxoplasmosis positive is done using logit link function.

4.4.4.1. Interpretation using the best model

Table-9: Estimates of odds ratio and confidence interval for the logit model, Mettu Karl Hospital, 2010.

Parameter	Odds Ratio	95% of CI
TAPWATERyes	0.024*	0.0037 : 0.1573
EDUC25-8	0.019*	0.0025 : 0.1401
MAR1married	0.018*	0.0024 : 0.1323
HANDSometimes	23.033*	3.3313 : 159.2444
RAWMEATyes	6.165*	1.6115 : 23.5805
HOUSETOILETyes	0.097*	0.0175 : 0.5465

- The odds that patients who have/use tap water are 0.024 times that those patients who do not have/use tap water to develop toxoplasmosis.
- The odds that patients who wash their hand sometimes are 23.033 times that of those patients who wash their hand always to develop toxoplasmosis.
- The odds that married patients will develop toxoplasmosis are 0.018 times that of other (those who are divorced or widowed) groups.
- The odds of patients who were in grade 5-8 in developing toxoplasmosis are 0.019 times that of the illiterate.
- The odds of patients who have the habit of eating raw meat in developing toxoplasmosis are 6.165 times that of those patients who do not have the habit of eating raw meat.
- The odds of patients who have house toilet are 0.097 times that of patients who do not use house toilet to develop toxoplasmosis.

The data shows that the prevalence of toxoplasmosis is 60% (72) of the HIV/AIDS patient. The prevalence of latent toxoplasma infection in Ethiopia was 93.3% (154/165) among HIV positive and 86.7% (143/165) among HIV negative participants (Techalew, *et al.,* 2009).

Although most of the results obtained from the summary statistics matches with the findings of both bivariate and the binary logistic regression analysis; it has some deviations with respect to certain variables. For example, age, location of the patients, eye disease and fever are not significant in the bivariate analysis (and hence are not candidates in the multivariate case).

The study shows that toxoplasmosis positive is not depending on gender. The distribution of toxoplasmosis in HIV/AIDS patient is not different for male and female. The infection rate does not depend on the patient's sex or age (Assob, *et al.,* 2011). This means that HIV-death is equally distributed in all age and gender groups both in rural and urban areas.

As we can see many researches specify that the main factor of toxoplasmosis is cat. Such as Wallace, *et al.* (1993); Moura, *et al.* (2002); Techalew, *et al.* (2009) reported that the exposure of cat for toxoplasmosis seropositive was significant (P-value<0.05). But in this study cat is not show significant in binary logistic regression model

The married group from marital status category shows less likely to develop toxoplasmosis (Odds Ratio=0.018 CI= [0.0024: 0.1323]) than the single groups and other groups. Less toxoplasmosis present in the educational groups of 5-8 (Odds Ratio=0.019, CI= [0.0025: 0.1401] than the illiterates and other educational level groups. Because illiterate persons has less knowledge about HIV/AIDS and the risk factors of toxoplasmosis.

From the study shows that; the person who has house toilet less affected by toxoplasmosis parasite (Odds Ratio= 0.097, CI= [0.0175: 0.5465]) than the person who has not house toilet. It is the main mode of transmission by using cat and soil related practices. The person who use tap water less likely to develop toxoplasmosis (Odds Ratio=0.024, [0.0037: 0.1573]) than the person who do not use tap water. In addition to this the person who wash his hand some time after any

activity without door and during eating food has (Odds Ratio= 23.033, CI= [3.3313: 159.2444]) more likely to develop toxoplasmosis. The result of the analysis specified that raw meat (Odds Ratio=6.165, CI= [1.6115: 23.5805]) is the main risk factors of toxoplasmosis gondii in HIV/AIDS patients.

5. Conclusion and Recommendation

5.1. Conclusion

The binary logistic regression model analysis shows that; the three link function (logit, probit and c-log-log) as likely as the same. Since the aim of the study is to improve the health of HIV-positive individuals by creating awareness for the risk factors of toxoplasmosis, the appropriate model for the health area, especially for categorical data is logistic regression (logit). Because the interpretation of the coefficient of risk factor is suitable for such kind of data

In general the three link function model (logit, probit and c-log-log) based on the data from the Metu Karl Hospital showed that the factors that affect the presence of toxoplasmosis among HIV/AIDS patients include marital status, educational status, tap water, house toilet, hand washing and habit of eating raw meat. And since the rest of the variables do not have significant contribution when included in to the models, they are not included in to the final model and they do not have any influence to have toxoplasmosis.

The results of this study also indicated that toxoplasmosis positive/ negative status doesn't show differences for gender (SEX), age (AGE), and address (ADR) and Occupation (OCC) of the patient. It also doesn't depend on washing fruit (FRUIT), fever of the patient (FEVER) ,compound latrine (CPDLATRIN), cat presence (CAT) and eye disease (EYE).

The factors that influence exposure to toxoplasmosis can be grouped as risk behavior factors (HAND and RAWMEAT), demographic and health risk factors (EDUC2, MAR1 and HOUSETOILET) and an economic factors is TAPWATER. Under such grouping we can state that patients involved in risky behaviors will have higher risk of toxoplasmosis. So the person who eats raw meat and washing hand sometimes after any activity has a high probability to present toxoplasmosis on him/her. The married group in marital status is less likely affected by toxoplasmosis parasite. The measures of economic status is tap water, the absence of tap water

has a chance occurring toxoplasmosis. One of the possible transmissions of toxoplasmosis is due to water born parasites.

5.2. Recommendation

To assess the risk factors of some health related problem the application of logistic regression is most useful in case of categorical data. Rather than using descriptive statistics like frequency, better results could be obtained by considering statistical modeling like logistic regression.

The high prevalence of HIV/AIDS in Ethiopia calls again the high distribution of *T. gondii*. This problem needs serious attention from government and non-government organizations. One of the treatments for HIV patient is ART. As a consequence, ART clinics are now widespread in all corners of the country. And it would be useful to initiate programs that emphasize coping strategies for improving the performance over and above testing and educational campaigns. Patients are being treated in ART clinics to extend their lives.

In addition to this the following recommendation are made for health policy workers and clinicians:

- the results of the study underlined behavioral (risk) factors as important predictors of toxoplasmosis gondii status in HIV/AIDS patients. Stake holders are expected to work hard to bring behavioral change. Currently the emphasis is on bringing behavioral change for prevention purposes. But we should consider it with respect to HIV-positive persons.
- the result of the study indicated that sanitation problem is risk factors of Toxoplasmosis gondii. As a result of this Stakeholders create awareness for the patient about their sanitation during food preparation.

6. References

Agresti, A. (2002, 2nd Ed.) "Categorical Data Analysis" Gainesville, Florida.

Alemtsehai, A. and Eshetu, W. (2006). Association of HIV Infection with some Selected Factors and Modeling the Chance of Contracting HIV: the Case of Hawassa and Its Surrounding. *Journal of Ethiopian Statistical Association*, **15**: 27-48.

Amr, I. A. (2010). Applying Logistic Regression Model to the Second Primary Cancer Data. Ain Shams University, Egypt: Pp-15-17.

Assob, J.C.N., Njunda, A.L., Nsagha, D.S., Kamga, H.L., Weledji, P.E. and Che, V.B. (2011). Toxoplasma Antibodies amongst HIV/AIDS Patients attending the University teaching Hospital Yaounde, in Cameroon. *AFR. J. CLN. EXPER. MICROBIOL,* **12**: 1119-123.

Boto, d.e., Los, B.A., Vega-Aleman, D., Galvez-Acebal, J., et al. (1998). Prevalence of latent toxoplasma infection in HIV infection patients. *Ann Med Intern,* **15**: 298-300.

Buve, A., Carael, M. and Hayes, R.J. (2001). Multi-center Study on Factors Determining Differences in rate of Spread of HIV in sub-Saharan Africa: Methods and Prevalence of HIV Infection. *AIDS*, **15**: 5-14.

Candolfi, E., Berg, M. & Kien, T. (1993). Prevalence of Toxoplasmosis in Pointe –Noire in Congo: study of the sampling of 310 subjects. *Bulletin de la Société de pathologie exotique et de ses filiales,* **86**: 358-62.

Chantal, J., Bessiere, M.H., Guenno, Ble., Magnava, I.J.F. and Dorchies, P. (1996). A serological survey of some zoonotic diseases among abattoir personnel in Djibouti. (French) *Bulletin de la Societe de Pathologie Exotique,* **89**: 353-357.

Cox, D. R., and Snell, E. J. (1989, 2nd Ed.). "The Analysis of Binary Data". London :Chapman and Hall.

Davarpanah,M.A., Mehrabani, D., Neirami, R., Ghahremanpoori, M. and Darvishi, M. (2007).Toxoplasmosis in HIV/AIDS patients in Shiraz, southern Iran. *Iranian Red Crescent Medical Journal*, **9**: 22-27.

Dobson, A.J., Barnett, A.G. (2008, 3rd Ed.). "An Introduction to Generalized Linear Models". Herston :Kelvin grove..

Ethiopia Demographic and Health Survey. Central Statistical Agency Addis Ababa, Ethiopia,September 2006 (USAID). Ethiopia. AIDS in Ethiopia. Fifth edition. June 2004.Ethiopia.

Falusi, O., Fench, A.A., Seaberg, E.C., et al. (2002). Prevalence and predictors of toxoplasma seropositivity in women with and at risk for human immunodeficiency infection. *Clin Infect Dis*, **35**: 1414-7.

Fayomi, B., Duong, T.H., Karagirwa, A., Marquet, N., Djivoh, C. & Combescot, C. (1987). Serological study of Toxoplasmosis and echinococcosis in cattle breeders and abattoir personnel in Benin. *Médecine Tropicale : Revue du Corps de santé colonial*, **47:** 149-51.

Fong, M.Y., Wong, K.T., Rohela, M., Tan, L.H., Adeeba, K., Lee, Y.Y. and Lau, Y.L. (2010). Unusual manifestation of cutaneous toxoplasmosis in a HIV-positive Patient. *Tropical Biomedicine,* **27**: 447-450.

Hoklek, M. and Safdar, A., (2004). *Toxoplasmosis*. Retrieved November 15, 2005 from http://www.emedicine.com/med/topic2294.htm

Holliman, R.E. (1990). Serological study of the prevalence of toxoplasmosis in asymptomatic patients infected with human immunodeficiency virus. *Epidemiol.Infect.*, **105**: 415-418.

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

http://www.cdc.gov/eis/conference/archives/2002ProgramFinal.pdf

Ibrahim, B.B., Salaam, M.M.I, Garish, N.I. and Hardy, F.M. (1997). Serological and histopathological studies on *Toxoplasma gondii* among the workers and the slaughtered animals in Tanta Abattoir, Gharbia Governate. *Journal of the Egyptian Society of Parasitology*, **27**: 273-278.

King, J.E. (2002). Logistic regression: Going beyond point-and-click. Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA), **141**: 13-15.

Jayawardena, S., Singh, S., Burzyantseya, O. and Clarke, H. (2008). Cerebral Toxoplasmosis in Adult Patients with HIV Infection. *J a y a w a r d e n a e t a l : C e r e b r a l To x o p l a s m o s i s*, pp. 1 7 – 2 4

Johnson S.A. (2004). HIV/AIDS Mortality Differential Across Provinces in Kenya and Through Time . Texas Tech University.

Lodenyo, H., Sitati, S.M. and Rogena, E. (2007). Reactivated toxoplasmosis presenting with non-tender hepatomegaly in a patient with HIV infection. *African Journal of Health Sciences,* **14**: 97-98.

Luft, B.J. and Remington, J.S. (1988). Toxoplasmic encephalitis. *J Infect.,* 157: pp.1-6

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, **54**: 17–24.

Montgomery, D.C., Peck, E.A. and Vining, G. (2006, 4th Ed). "Introduction to Linear Regression Analysis ". Tempe, Atlanta and Blacksburg.

Montoya, J. G. and Rosso, F. (2000). Diagnosis and management of toxoplasmosis. Clin Perinatol, **32**: 705-726.

Moura, L, Wada, M., Carmo, E, Dusi, R., Tuboi, S. H., Camargo, N. et al. (2002, April 23). *Epidemic of toxoplasmosis associated with ingestion of contaminated municipal water – Brazil.* Paper presented at the Epidemic Intelligence Service 51st Annual Conference. Retrieved November 15, 2005 from

Myriam, G., Ayenew, N., Amha, G. (1993). Sero-epidemiological survey of *T.gondi* infection in Ethiopia. *Ethiop Med J*, **31**: 201-204.

Nagelkerke, N. J. D. (1991). A note on the general definition of the coefficient of determination. *Biometrika*, **78**: 691-692.

Navia, B.A., Petito, C.K., Gold, J.W., Cho, E.S., Jordan, B.D. and Price, R.W. (1986). Cerebral toxoplasmosis complicating the acquired immune deficiency syndrome: clinical and neuropathological findings in 27 patients. *Annals of Neurology*, **19**: 224-238.

Negash,T., Tilahun, G., and Medhin, G. (2008). Seroprevalence of *Toxoplasma gondii* in Nazaret town, Ethiopia. *East African Journal of Public Health.*, **5**: 54-65.

Nissapatorn, V. (2008). Lessons learned about opportunistic infections in Southeast Asia. *Southeast Asian J Trop Med Public Health,* **39**: 625-641.

Nissapatorn, V., Lee, C., Quek, K.F., Leong, C.L., Mahmud, R. and Abdullah, k.A. (2004). Toxoplasmosis in HIV/AIDS patients: A Current Situation. *Jpn. J. Infect.,* **57**(160): pp.160-165.

Osiyemi, T.I., Synge, E.M., Agbonlahor, D.E. & Agbavwe, R., (1985). The prevalence of *Toxoplasma gondii* antibodies in man in Plateau State and meat animals in Nigeria. *Transactions of the Royal Society of Tropical Medicine & Hygiene,*79: 21-30.

Richards, F.O.,Kovacs,J.A. and Luft, B.J. (1995). Preventing toxoplasmic encephalitis in persons infected with human immunodeficiency virus. *Clin.Infect.Dis.*, **21**: pp.49-56.

cytomegalovirus, HIV and treponemal infections among pregnant women in Cotonou. *Repub. Benin. Acta Trop.*, **59**: 271-277.

Sonia, P.B., Bernardete, R., Luciana, R., Heitor, F.D. and Ricardo,T. (2010). Preliminary Report of HIV and Toxoplasmosis gondii occurrence in Pregnant Women from Mozambique. *Rev. Inst. Med. Trop. Sao Paulo,* **52**: 291-295.

Swai, E.S. and Schoonman, L. (2009). Seroprivalence of Toxoplasmosis gongii infection amongst residents of Tanga District in north-east Tanzania.*Tanzania Journal of Health Research,* **4**: 205-209.

Sýkora, J., Zástěra, M., Stanková, M. (1992). Toxoplasmic antibodies in sera of HIV-infected persons**.** *Folia Parasitol (Praha).*, **39**: 177-80.

Techalew, S., Mekashaw, T., Endale, T., Belete, T. and Ashenafi, T. (2009). Sero-prevalence of latent *Toxoplasma gondii* infection among HIV-infected and HIV-uninfected people in Addis Ababa, Ethiopia: A comparative cross-sectional study. *BMC Research Notes*, 2:213. In: Kloos, H. and Berhane.Y. (2006): Pp. 692-700.

Thompson, D., Health, A. and Milwaukee, W.I. (2009). Ranking Predictors in Logistic Regression. West Michigan.Pp. 23-35.

Verheij, R.A. (1996). Explaining urban-rural variations in health: A review of interactions between individual and environment. *Social Science and Medicine*, **42**: 923-935.

Wallace, R.M., Rossetti, J.R., L.V.N. and Olson, E.P. (1993). Cats and Toxoplasmosis Risk in HIV- infected adults.*JAMA*, **269**: 76-77.

Wanachiwanawinm, D., Sutthent, R., Chokephaibulkit, K., Mahakittikun, V., Ongrotchanakun, J. and Monkong, N., (2001). Toxoplasma gondii antibodies in HIV and non-HIV infected Thai pregnant women. *Asian Pac J Aller Immunol*, **19**: 291-293.

Williams, C.C. and Windebank, J. (2000). Helping people to help themselves: Policy lessons from a study of deprived urban neighborhoods in Southampton. *Journal of Social Policy*, **29**: 355-373.

Woldemicheal, T., Fontanet, A.L., Sahlu, T., Gilis, H., Messele, T., Rinke de Wit, T.F., Yeneneh, H., Coutinho, R.A. and Gool, T.V. (1998). Evaluation of the Eiken latex agglutination test for anti-*Toxoplasma* antibodies and seroprevalence of *Toxoplasma* infection among factory workers in Addis Ababa, Ethiopia. *Trans R Soc Trop Med Hyg*, **92**: 401-403.

Xiao, Y., Yin, J., Jiang, N., Xiang, M., Hao, L., Lu, H., Sang, H., Liu, X., Xu, H., Ankarklev, ., J., Lindh, J. and Chen, Q. (2010). Seroepidemiology of human Toxoplasma gondii infection in China. *Xiao et al. BMC Infectious Diseases*, **10**: 3-5.

7. Appendix

```
########## LOGIT LINK FUNCTION ########################

> ##### the final logit model

> LL<-glm(TOXO ~ TAPWATER + EDUC2 + MAR1 + HAND + RAWMEAT  +
HOUSETOILET ,data=toxodata,family=binomial)

> summary(LL)

Call:

glm(formula = TOXO ~ TAPWATER + EDUC2 + MAR1 + HAND + RAWMEAT +

    HOUSETOILET, family = binomial, data = toxodata)

Deviance Residuals:

     Min        1Q    Median        3Q       Max

-1.91822  -0.35270   0.08208   0.56025   1.50525

Coefficients:

               Estimate Std. Error z value Pr(>|z|)

(Intercept)      3.8729     0.9799   3.952 7.74e-05 ***

TAPWATERyes     -3.7290     0.9589  -3.889 0.000101 ***

EDUC25-8        -3.9797     1.0277  -3.872 0.000108 ***

MAR1married     -4.0249     1.0216  -3.940 8.15e-05 ***

HANDSometimes    3.1369     0.9865   3.180 0.001474 **

RAWMEATyes       1.8188     0.6845   2.657 0.007879 **

HOUSETOILETyes  -2.3263     0.8786  -2.648 0.008105 **
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 157.807  on 117  degrees of freedom

Residual deviance:  76.922  on 111  degrees of freedom

  (1 observation deleted due to missingness)

AIC: 90.922

Number of Fisher Scoring iterations: 7

> L<-glm(TOXO ~ TAPWATER + EDUC2 + MAR1 + HAND + HOUSETOILET,data=toxodata,family=binomial("logit"))

> summary(L)

Call:

glm(formula = TOXO ~ TAPWATER + EDUC2 + MAR1 + HAND + HOUSETOILET,

family = binomial("logit"), data = toxodata)

Deviance Residuals:

    Min       1Q   Median       3Q      Max

-2.1868  -0.3192   0.1834   0.7560   1.6515

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept)      4.0767     0.9197   4.433 9.31e-06 ***

TAPWATERyes     -3.3634     0.8689  -3.871 0.000109 ***

EDUC25-8        -3.2779     0.9384  -3.493 0.000477 ***
```

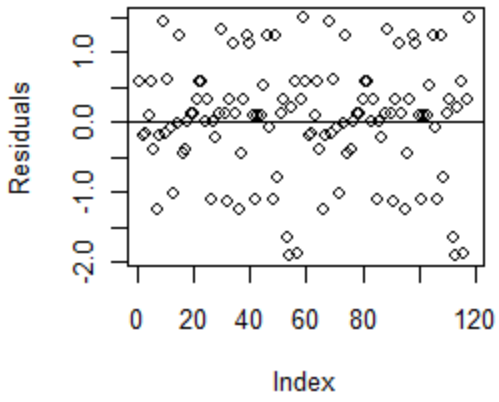
```
MAR1married     -3.7499     0.9337  -4.016 5.92e-05 ***
HANDSometimes    2.9237     0.8869   3.296 0.000979 ***
HOUSETOILETyes  -1.7817     0.7863  -2.266 0.023454 *
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 157.81  on 117  degrees of freedom
Residual deviance:  85.50  on 112  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 97.5
Number of Fisher Scoring iterations: 6
> ## odds ratio of logit model
> exp(LL$coefficients)
   (Intercept)     TAPWATERyes          EDUC25-8     MAR1married
HANDSometimes
   48.08074917       0.02401660       0.01869148       0.01786507
23.03268945
    RAWMEATyes HOUSETOILETyes
    6.16456875     0.09765894
> LL<-glm(TOXO ~ TAPWATER + EDUE2 + MAR1 + HAND + RAWMEAT  +
HOUSETOILET,data=toxodata,family=binomial("logit"))
>  L<-glm(TOXO ~ TAPWATER + EDUC2 + MAR1 + HAND +
HOUSETOILET,data=toxodata,family=binomial("logit"))
```

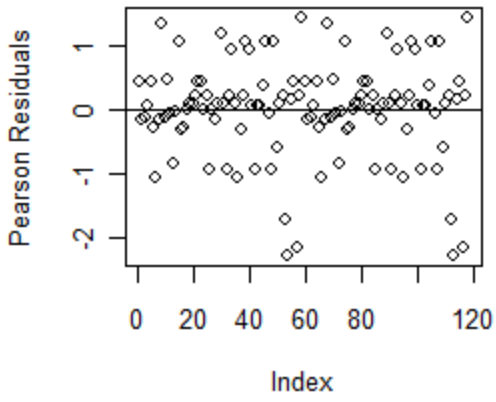
```
> L0<-glm(TOXO~1,data=toxodata,family=binomial)

> anova(L0,LL,test="Chisq")#likelihood Ratio Test

Analysis of Deviance Table

Model 1: TOXO ~ 1

Model 2: TOXO ~ TAPWATER + EDUC2 + MAR1 + HAND + RAWMEAT +
HOUSETOILET

  Resid. Df Resid. Dev Df Deviance P(>|Chi|)

1       117    157.807

2       111     76.922   6 80.885 2.638e-13 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(L,LL,test="Chisq")#likelihood Ratio Test

Analysis of Deviance Table

Model 1: TOXO ~ TAPWATER + EDUC2 + MAR1 + HAND + HOUSETOILET

Model 2: TOXO ~ TAPWATER + EDUC2 + MAR1 + HAND + RAWMEAT +
HOUSETOILET

  Resid. Df Resid. Dev Df Deviance P(>|Chi|)

1       112     85.5

2       111     76.922  1   8.578   0.01493 *

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> ###### confidence interval of the parameter

> confint(LL,level=0.95)

Waiting for profiling to be done...
```

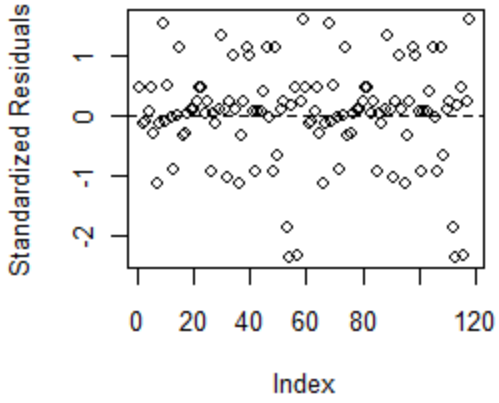
```
                    2.5 %     97.5 %
(Intercept)     2.1554493  6.0850578
TAPWATERyes    -5.9163528 -2.0662773
EDUC25-8       -6.3164770 -2.1934404
MAR1married    -6.3400792 -2.2602595
HANDSometimes   1.4292803  5.4094062
RAWMEATyes      0.5707465  3.3076036
HOUSETOILETyes -4.2174018 -0.7313676
> #Pearson Chisquare Test
> p.chi<- sum(resid(LL, type = "pearson")^2)
> p.chi
[1] 67.3538
> df<-LL$df.resid
> df
[1] 111
> pp<-1-pchisq(p.chi,df )
> pp
[1] 0.99965
> #Deviance (G2)Test
> dp<-1-pchisq(LL$deviance,df)
> dp
```

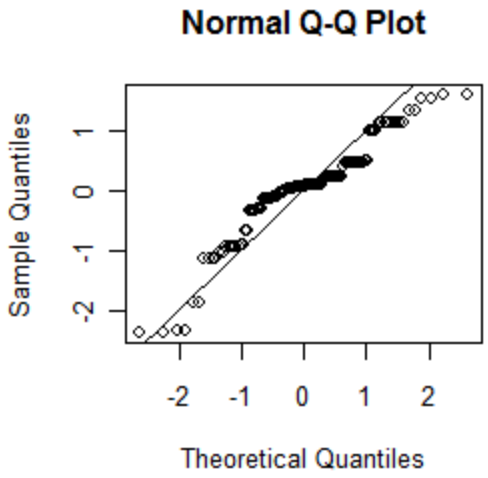
```
[1] 0.9942783

> #Pseudo R2

> r0<-logLik(L0)

Error in logLik(L0) : object 'L0' not found

> r0

Error: object 'r0' not found

> r1<-logLik(LL)

> r1

'log Lik.' -38.46087 (df=7)

> R<-(-76.88+37.42)/(-76.88)

> R

[1] 0.5132674

> aic<- -2*-37.42+2*10

> aic

[1] 94.24

> bic<- -2*-37.42+2*10*log(114)

> bic

[1] 169.564

> ########over dispersion

> p.chi/df

[1] 0.606791
```

```
> LL$deviance/df
[1] 0.6929887
> summary(LL)$dispersion
[1] 1
> #Residuals
> ###### Raw residuals #######
> r.res<-resid(LL)
> par(mfrow=c(2,2))
> plot(r.res,ylab="Residuals")
> abline(0,0)
>  ########Pearson Residuals #######
> p.res<-resid(LL,type="pearson")
> plot(p.res,ylab="Pearson Residuals")
> abline(0,0)
> ### Adjusted/standardized Residuals #######
>  hii<-hatvalues(LL)
>  adj.res<-resid(LL,type="pearson")/sqrt(1-hii)
>  plot(adj.res,ylab="Standardized Residuals")
>  abline(h=0,lty=2)
> qqnorm(adj.res)
> abline(0,1)
```



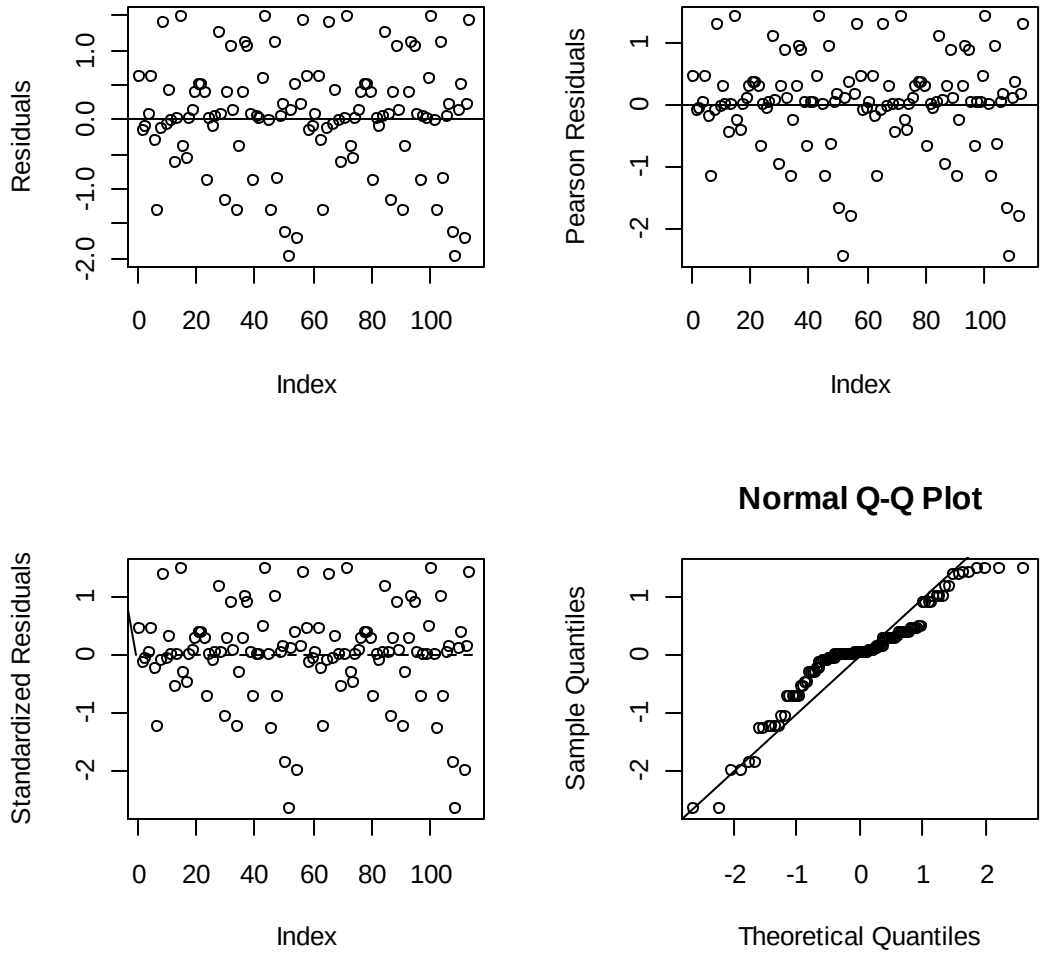


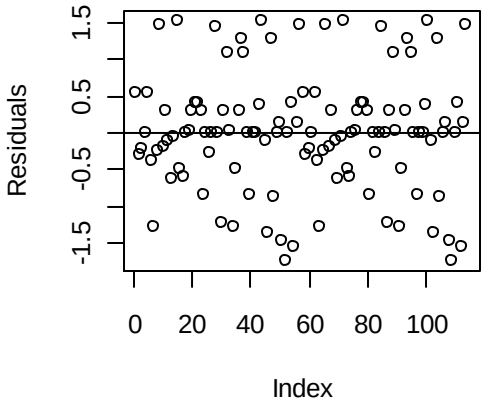


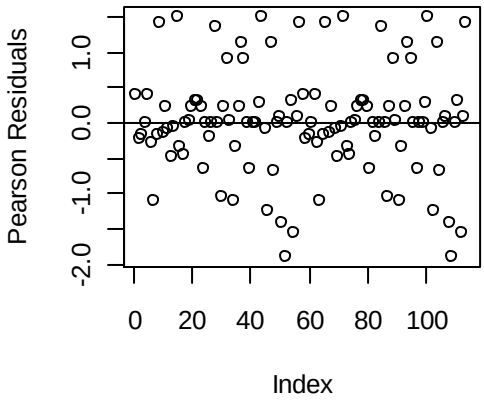


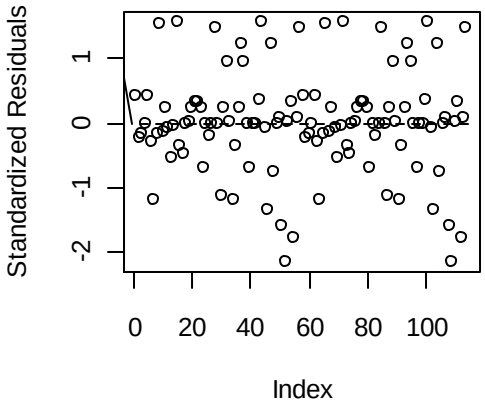

```
#######   Leverages #########

p<-sum(hii)

>  n<-nrow(toxodata)

>  plot(hii,ylim=c(-1,1),ylab="Leverages")

>  c.value<-2*p/n
```

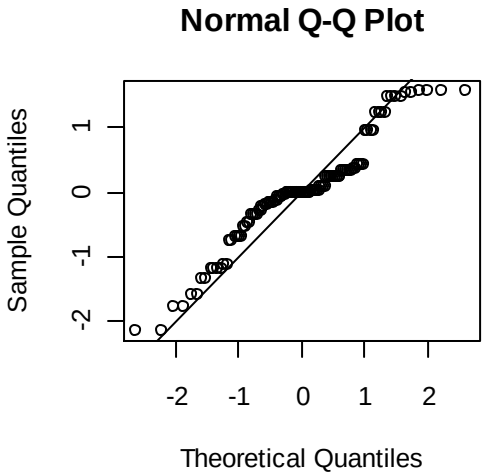
```
> c.value

[1] 0.1754386

>  abline(c.value,0)

>  abline(-c.value,0)

>  identify(1:n,hii)
```




Residuals
1.5
0.5
-0.5
-1.5
0
20
40
60
80
100
Index


Pearson Residuals
1.0
0.0
-1.0
-2.0
0
20
40
60
80
100
Index


Standardized Residuals
1
0
-1
-2
0
20
40
60
80
100
Index


Normal Q-Q Plot
Sample Quantiles
1
0
-1
-2
-2
-1
0
1
2
Theoretical Quantiles