Estimation of Disease Prevalence in the Presence of Missing Data

Elhadji Moustapha SECK

MS300-0003/15

A Thesis submitted to Pan African University Institute for Basic Sciences, Technology and Innovation in partial fulfillment of the requirement for the award of the degree of Master of Science in Mathematics (Statistics Option)

2017

DECLARATION

I do hereby declare that this is my original work and has not been presented in partial fulfillment of any other degree award in any other university.

Signature: ..... Date .....

Elhadji Moustapha SECK

This research thesis has been submitted for examination with our approval as University Supervisors.

Signature ..... Date .....

Dr. Ngesa Owino Oscar

Taita Taveta University, Taita Taveta, Kenya

Signature ..... Date .....

Prof. Abdou Ka Diongue

Universite Gaston Berger de Saint Louis, Saint Louis, Senegal

DEDICATION

This thesis is dedicated to my mother, who taught me that the best kind of knowledge to have is that which is learned for its own sake, she also taught me that even the largest task can be accomplished if it is done one step at a time. I also dedicate this work to my uncle Said Diaw who has encouraged me all the way and whose encouragement has made sure that I give it all it takes to finish that which I have started. To all my friends especially Abdoulaye Ndiaye, my brothers Baba Ngom and Cheikh Mbacke Faye who have been affected in every way possible by this quest. This thesis is also dedicated to all of my sisters for their endless love, support and encouragement. Thank you. My love for you all can never be quantified. God bless you.

ACKNOWLEDGMENT

First of all, I give glory to God for bringing me this far and giving me good health and knowledge while writing this research. I would like to thank Dr. Ngesa Owino Oscar and Prof. Abdou Ka Diongue, my supervisors for supporting my work. I would also like to thank Prof. Mwita and my coordinator Dr. Orwa for their constant encouragement. I would like to thank the administration staff and Prof. Gabriel Magoma, director PAUISTI, for their sincere and valuable guidance and encouragement extended to me. I am grateful for the support provided by Cheikh Loucoubar, Mareme Diarra and Mamadou Tamsir Diop at Institut Pasteur de Dakar. I wish to also thank my parents, and numerous friends who endured this long process with me, always offering support and love. Finally, I would like to extend my sincere appreciation and thanks to African Union for giving me this opportunity.

TABLE OF CONTENTS

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

AIDS	Acquired Immune Deficiency Syndrome
ART	Antiretroviral Treatment.
CC	Complete Case.
CI	Confidence Interval.
DHS	Demographic and Health Surveys.
HIV	Human Immunodeficiency Virus.
MAR	Missing At Random.
MCAR	Missing Completely At Random.
MNAR	Missing Not At Random.
TasP	Treatment as Prevention.
UNAIDS	United Nations Programme on Human Immunodeficiency Virus and Acquired Immune Deficiency Syndrome.
WHO	World Health Organisation.

ABSTRACT

Missing data are commonly encountered in most medical research. Unfortunately, they are often neglected or not properly handled during analytic procedures, and this may substantially bias the results of the study, reduce the study power, and lead to invalid conclusions. In this study, we introduce key concepts regarding missing data in survey data analysis, provide a conceptual framework on how to approach missing data in this setting, describe typical mechanisms of missing data, and use a theoretical model for handling such data. We consider a case where the variable of interest (response variable) is binary and some of the observations are missing and assume that all the covariates are fully observed. In most cases, the statistic of interest, when faced with binary data is the prevalence. We develop a two stage approach to improve the prevalence estimates: in the first stage, we use a logistic regression model to predict the missing binary observations and then in the second stage we recalculate the prevalence using the observed binary data and the imputed missing data. Finally we study the asymptotic properties of the prevalence estimator. Such a model would be of great interest in research studies involving HIV in which people usually refuse to donate blood for testing yet they are willing to provide other covariates. The prevalence estimation method is illustrated using simulated data and applied to HIV/AIDS data from the Kenya AIDS Indicator Survey, 2007.

CHAPTER 1

INTRODUCTION

1.1 Background of study

Prevalence in epidemiology is the proportion of a population found to have a condition (typically a disease or a risk factor such as smoking or seat-belt use). It is arrived at by comparing the number of people found to have the condition with the total number of people studied, and is usually expressed as a fraction, as a percentage or as the number of cases per people. It is difficult to overestimate the importance of obtaining accurate information on the prevalence. Accurate estimates of disease prevalence are critical for tracking the epidemic, designing and evaluating prevention and treatment programs, and estimating resource needs. A potential threat to the validity of survey-based prevalence estimates is that not all individuals eligible to participate in a survey can be contacted, and some who are contacted do not consent to be tested (Hogan et al., 2012).

If any data on any variable from any participant is not present, then the researcher is dealing with missing or incomplete data. The problem of missing data is a common occurrence in most medical research (Horton and Laird, 2001). In clinical trials and observational studies, complete data are often not available for every subject. Missing data may arise because of many circumstances: the unavailability of converting measurements, survey nonresponse, study subjects failing to report to a clinic for monthly evaluations, respondents refusing to answer certain items on a questionnaire (Ibrahim et al., 2005). Respondents may refuse to answer a question because of privacy issues or the person taking the survey does not understand the question. Perhaps, the respondent would have answered,

but the answer, he or she might have given was not one of the options presented. Perhaps there wasn't enough time to complete the questionnaire or the respondent just lost interest. Every survey question without an answer is a missing data point.

It is rare, even under the strictest protocols, to complete a biological or medical study with absolutely no missing values. While many investigators consider missing data a minor nuisance, ignoring them is potentially very problematic (Haukoos and Newgard, 2007). In fact, investigators should attempt to use all available data to perform the most efficient study possible, to reduce bias, and to provide the most valid estimates of risk and benefit. A bias which is known as systematic error, may result directly from the inappropriate handling of missing values. A primary goal in the analysis of a medical study is to minimize bias so that valid results are presented and appropriate conclusions are drawn. While bias may be introduced into research through several other mechanisms (e.g., study design, patient sampling, data collection, and or other aspects of data analyses), native methods of handling missing data may substantially bias estimates while reducing their precision and overall study power, any of which may lead to invalid study conclusions. When a large proportion of missing data exist or when there are missing data for multiple variables, these effects may be dramatic. Despite these concerns and the development of sophisticated methods for handling missing data that allow for valid estimates with preservation of study power, many studies continue to ignore the potential influence of missing data, even in the setting of clinical trials (Haukoos and Newgard, 2007).

Previous authors have suggested that non-participation may lead to bias in human immunodeficiency virus (HIV) prevalence estimates, but official estimates of HIV prevalence in sub-Saharan Africa relies heavily on population-based surveys, which often have low participation rates (Hogan et al., 2012). An analysis

of the Demographic and Health Surveys (DHS), which are the most common nationally representative surveys for HIV prevalence in sub-Saharan Africa, reveals average rates of non-participation in HIV testing of 23% for adult men and 16% for adult women in the region, with a high of 37% for men in Zimbabwe 2005–2006 and a low of 3% for women in Rwanda 2005 (Mishra et al., 2008), and the most recent national population-based survey in South Africa reported an overall non-participation rate of 32% for HIV testing among adults (Hogan et al., 2012). Analyses of the DHS have adjusted HIV prevalence estimates for testing non-participation by imputing missing HIV test results with probit regressions, controlling for differences in observed characteristics between testing participants and non-participants, such as gender, urban residence, wealth and indicators of sexual behaviour (Mishra et al., 2008; Hogan et al., 2012). Based on this conventional imputation approach, non-participants were estimated to have higher HIV prevalence than participants in about half of the DHS examined, but this did not result in substantially different estimates of overall HIV prevalence when compared with the complete-case estimates that ignored missing observations (Mishra et al., 2008). These results have been interpreted to mean that non-participation in HIV testing surveys is likely to have minimal impact on prevalence estimates (Mishra et al., 2008; Hogan et al., 2012). However, the conventional imputation approach has two important limitations. First, it assumes that no unobserved variables associated with HIV status influence participation in HIV testing. Second, it ignores regression parameter uncertainty in the imputation model, resulting in confidence intervals (CI) that are too narrow.

1.2 Statement of the problem

Accurate estimates of disease prevalence are critical for tracking the epidemic, designing and evaluating prevention and treatment programs, and estimating

resource needs. A potential threat to the validity of survey-based prevalence estimates is that not all individuals eligible to participate in a survey can be contacted, and some who are contacted do not consent to be tested. Incomplete participation in testing can lead to selection bias, and a recent paper found evidence for substantial downward bias in existing national HIV prevalence estimates for Zambian men due to selective survey non-participation (Hogan et al., 2012). For example in low and middle income countries data are often derived from HIV testing conducted as part of household surveys, where participation rates in testing can be very low. A low participation rates may be attributed to HIV positive individuals being less likely to participate because they fear disclosure, in which case, estimates obtained using conventional approaches to deal with non-participation, such as imputation-based methods and complete-case (CC), will be biased (Ibrahim et al., 2005). However, the conventional imputation approach has two important limitations. First, it assumes that no unobserved variables associated with HIV status influence participation in HIV testing. Second, it ignores regression parameter uncertainty in the imputation model, resulting in confidence intervals (CI) that are too small. The evaluation of possible bias in HIV prevalence estimates for other African countries is thus important for HIV research and policy (Hogan et al., 2012)

1.3 Justification of the study

Policy interventions targeted to control the diseases epidemic, improve population health, and reduce diseases-related health disparities, are often motivated by prevalence data obtained from testing as part of national or regional surveillance (Beyrer et al., 2011; De Cock et al., 2006). Particularly in low and middle income countries without developed health systems infrastructure, data obtained from nationally representative samples of the population of interest are a power-

ful source of information for establishing the current numbers of being positive for a test, as well as the change in diseases prevalence over time (Boerma et al., 2003). This information is important for governments to be able to cost policy interventions, to implement these interventions, and to plan and forecast future demands on the health care system and public finances. The development of new antiretroviral treatment (ART) for reducing viral load and stabilizing the health status of HIV positive individuals, and subsequent initiatives using treatment-as-prevention (TasP), which aims to reduce the transmission of HIV by placing infected individuals on treatment as soon as possible, is a very promising development for combating the HIV epidemic (Marra et al., 2015).

However, to be most effective, these programs will require accurate prevalence data on hard to reach and at risk populations (Kranzer et al., 2012). The recent success of ART means that improving treatment access in sub-populations with high HIV prevalence or which have seen increases in HIV prevalence will have potentially large payoffs (Tanser et al., 2013; Bor et al., 2013). In addition to identifying the most suitable groups for these policy interventions, prevalence data are important for evaluating the effectiveness of large-scale programs. Establishing whether a population-based policy or intervention acted to reduce disease prevalence will require population-based prevalence data. In low and middle income countries, estimates of HIV prevalence obtained from nationally representative household surveys are now considered the gold standard (Burma et al., 2003). These data are generally obtained from home-based testing which takes place after survey respondents complete a standard interview (Marra et al., 2015). After the interview, the surveyor conducting the interview will ask the respondent to participate in a blood test for HIV, generally to be collected by finger prick, following the recommended guidelines specified by the World Health Organization (WHO) and the Joint United Nations Program on HIV and AIDS (UNAIDS).

Similar data collection procedures take place as part of the demographic surveillance site, which track the residents of specific geographic areas, and which are another important source of data in HIV prevalence (Tanser et al., 2008). For HIV surveys which are designed to be nationally representative, a random sample of the population is approached with an offer for HIV testing.

However, these HIV survey data can be affected by non-participation, because some of those who are eligible for testing opt out. This non-participation can occur through a variety of mechanisms, including directly declining to test for HIV when a respondent is approached to test after an interview, or being an eligible respondent for HIV testing, but not being present when the interviewers seek to contact the person to interview (Marra et al., 2015). Even if, ex ante, the eligible population for the survey is either the complete population of interest (as at surveillance sites), or a random sample (in household surveys), ex post the surveyed group who consent to HIV testing may not be representative of the population of interest due to this non-participation. Selection bias can occur if HIV prevalence among those who participate in testing differs from those who do not participate in testing. In many contexts, the extent of non-participation is substantial. For example, at some demographic surveillance sites, less than half of eligible respondents participate in testing (Tanser et al., 2008). In the nationally representative DHS, non-participation can also be common, for example, 37% of eligible male respondents failed to participate in testing in Malawi in 2004. In general, the treatment of missing information in survey data has the potential to have a substantial impact on both the parameter estimates and the policy recommendations derived from these surveys (Marra et al., 2015). In the worst case scenario, where missing information caused by non-participation are a symptom of selection bias, conventional estimates can be substantially biased. Therefore, modeling this non-participation in HIV testing is important from a

policy perspective.

1.4 Objectives of the study

1.4.1 General Objective

Our general objective in this study is to improve prevalence estimates using missing data approach.

1.4.2 Specific Objectives

(i) To estimate the prevalence when there are some missing cases

(ii) To study the asymptotic properties of the estimator

(iii) To use simulated data to ullistrate the method used and then applied it to real data

CHAPTER 2

LITERATURE REVIEW

2.1 Missing Values Mechanisms

We present three classifications of missing data, as discussed by Rubin (1976). We denote the data we intended to collect, by Y, and we partition this into $Y = \{Y^0, Y^m\}$, where Y^0 is observed and Y^m is missing. Note that some variables in Y may be outcomes/responses, some may be explanatory variables/covariates. Depending on the context, these may all refer to one unit, or to an entire dataset. Corresponding to every observation Y, there is a missing value indicator R, defined as:

$$R = \begin{cases} 1 & if\ y\ is\ observed \\ 0 & if\ y\ is\ missing \end{cases}$$

with R corresponding to Y.

The key question for analyses with missing data is, under what circumstances, if any, do the analyses we would perform if the data set were fully observed lead to valid answers? As before, "valid" means that effects and their standard errors are consistently estimated, tests have the correct size, and so on, so inferences are correct. The answer depends on the missing value mechanism. This is the probability that a set of values is missing given the values taken from the observed and missing observations, which we denote by:

$$Pr(R|Y^0, Y^m).$$

2.1.1 Missing Completely At Random (MCAR)

Suppose the probability of an observation being missing does not depend on observed or unobserved measurements.

In mathematical terms, we write this as:

$$Pr(r|y^0, y^m) = Pr(r).$$

Then we say that the observation is Missing Completely At Random, (MCAR). Note that in a sample survey setting MCAR is sometimes called uniform non-response. If data are MCAR, then consistent results with missing data can be obtained by performing the analyses we would have used had their been no missing data, although there will generally be some loss of information. In practice, this means that, under MCAR, the analysis of only those units with complete data gives valid inferences (Roth, 1994; Schlomer et al., 2010).

2.1.2 Missing At Random (MAR)

After considering MCAR, a second question naturally arises. That is, what are the most general conditions under which a valid analysis can be done using only the observed data, and no information about the missing value mechanism, $Pr(r|y^0, y^m)$? The answer to this is when, given the observed data, the missing-ness mechanism does not depend on the unobserved data. Mathematically, this is written as:

$$Pr(r|y^0, y^m) = Pr(r|y^0).$$

.

This is termed Missing At Random (MAR). This is equivalent to saying that the behaviour of two units who share observed values have the same statistical behaviour on the other observations, whether observed or not Schafer and Graham (2002); Schlomer et al. (2010).

2.1.3 Missing Not At Random (MNAR)

When neither MCAR nor MAR hold, we say the data are Missing Not At Random, abbreviated MNAR. In the likelihood setting the missingness mechanism is termed non-ignorable.

What this means is, even accounting for all the available observed information, the reason for observations being missing still depends on the unseen observations themselves. To obtain valid inference, a joint model of both Y and R is required (that is a joint model of the data and the missingness mechanism).

Unfortunately, we cannot tell from the data at hand whether the missing observations are MCAR, NMAR or MAR (although we can distinguish between MCAR and MAR). In the MNAR setting it is very rare to know the appropriate model for the missingness mechanism.

Hence the central role of sensitivity analysis; we must explore how our inferences vary under assumptions of MAR, MNAR, and under various models. Unfortunately, this is often easier said than done, especially under the time and budgetary constraints of many applied projects.

2.2 Survey of missing data techniques

Missing data are ubiquitous throughout the social, behavioral, and medical sciences. For nearly a century, methodologists have been studying missing data problems. Unfortunately, most of these techniques require a relatively strict assumption about the cause of missing data and are prone to substantial bias. These methods have increasingly fallen out of favor in the methodological literature (Enders, 2010; Wilkinson and Task Force, 1999), but they continue to enjoy widespread use in published research articles (Enders, 2010; Peugh and Enders, 2004). Much has been written about statistical methods in order to handle

incomplete data (Little and B, 1987) for a comprehensive review.

Many of these approaches have focused on missing outcomes. But covariates in regression models are often missing, either by design or circumstance. Little (1992) reviewed a number of approaches to estimation of regression models with missing covariates, including complete case estimation, likelihood-based methods and ad hoc methods. Robins, Zhao and Rotnitzky, (1994) suggested a class of semiparametric estimators based on inverse probability weighted estimating equations similar to a method proposed by Zhao and Lipsitz (1992). Ibrahim (1990) described a maximum likelihood method using the EM algorit (Dempster et al., 1977) for generalized linear regression models with missing categorical covariates. The major breakthroughs came in the 1970s with the advent of multiple imputation and maximum likelihood estimation routines (Dempster et al., 1977; Horton and Laird, 2001). At the same time, Rubin (1976) outlined a theoretical framework for missing data problems that remains in widespread use today. Multiple imputation and maximum likelihood have received considerable attention in the methodological literature during the past 30 years. When a non - response is unrelated to the missing values of the variables, then the non-response is said to be incurable (Little and B, 1987).

The literature for generalized linear model with incomplete observations, however, is sparse. Ibrahim et al. (2005) discussed incomplete data in the generalized linear models. have proposed a method for estimating the parameters in binomial regression models when the response variable is missing and the missing data mechanism is non-ignorable. Ibrahim and Lipsitz (1996) proposed a conditional model for incomplete covariates in parametric regression models. (Ibrahim et al., 2005) proposed a method for estimating the parameters in generalized linear models with missing covariates and a non-ignorable missing data mechanism.

Intuitively, when the subjects with missing covariate values differ from those

with complete data with respect to the outcome of interest, then the results of a traditional data analysis omitting the missing cases may no longer be valid. Because standard techniques for regression models require full corporate information, then one simple way to avoid the problem of missing data is to analyze only those subjects who are completely observed. This method, known as a complete case analysis, is the technique most commonly used with missing values in the covariates and/or response. The complete case analysis is still the default method in most software packages, despite the development of statistical methods that handle missing data more appropriately. It is known that when the data are not missing completely at random (MCAR), the complete case analysis can be biased. Further, when the data are MCAR so that the complete case analysis is unbiased, as the fraction of missing data increases, the deletion of all subjects with missing data is unnecessarily wasteful and quite inefficient. Another ad hoc method for handling with missing covariate data is to exclude those covariates subject to missingness from the analysis. Because this procedure can lead to model misspecification then is not recommended (Ibrahim et al., 2005).

Researchers have been slow to adopt maximum likelihood and multiple imputation and still rely heavily on traditional missing data handling techniques (Enders, 2010; Peugh and Enders, 2004). In part, this may be due to lack of the software options, as maximum likelihood and multiple imputation did not become widely available in statistical package until the late 1990s. However, the technical nature of the missing data literature probably represents another significant barrier to the widespread adoption of these techniques. Unless missing data are a deliberate feature of the study design, then it is important to try to limit them during data collection, since any method for compensating for missing data requires unverifiable assumptions that may or may not be justified. Since data are still likely to be missing despite these efforts, it is important to try to collect covariates

that are predictive of the missing values, so that an adequate adjustment can be made. In addition, the process that leads to missing values should be determined during the collection of data if possible, since this information helps to model the missing-data mechanism when the incomplete data are analyzed. Then it is reasonable to seek ways for incorporating incomplete cases into the analysis. In this study, we purpose a two stage procedures for inferring missing data then improving the estimated prevalence with the imputed values based on the logistic model (Enders, 2010).

2.3 Review of the Generalized Linear Models (GLMs)

Introduced in a 1972 by Nelder and Wedderburn, Generalized Linear Models (GLMs) analysis comes into play when the error distribution is not normal and/or when a vector of non linear function of the response $\eta(Y) = (\eta(Y_1), \eta(Y_2), \cdots, \eta(Y_n))'$, has Y itself as expectation the vector $X\beta$.

It involves three components:

- An exponential family model for the response.
- A systematic component via a linear predictor.
- A link function that connects the means of the response to the linear predictor.

In GLM, the response variable distribution must be a member of the exponential family of distribution.

A random variable that belongs to the exponential family with a single parameter θ has a pdf :

$$f(u, \theta) = s(u)t(\theta) \exp \{a(u)b(\theta)\} \tag{2.1}$$

Where s,t,a,b are all known functions.

So Equation (2.1) can be written as:

$$f(u,\theta) = \exp\{a(u)b(\theta) + d(u) + c(\theta)\} \tag{2.2}$$

Where $a(u) = \ln(s(u))$, $c(\theta) = \ln(t(\theta))$.

When $a(u) = u$, the distribution is said to be in canonical form.

$b(\theta)$ is called the natural parameter.

Parameter other than the the parameter of interest θ are called nuisance parameters

2.3.1 Fitting the Generalized Linear Models (GLMs)

Suppose we have a set of independent observations $\left(Y_i, \widetilde{x_i}'\right)$, $i = 1, 2, \cdots, n$, $\widetilde{x_i}' = (x_{i1}, x_{i2}, \cdots, x_{ip})$ from some exponential type distribution of canonical form.

Then the joint pdf is

$$f(Y_1, \cdots, Y_n, \theta, \phi) = \prod_{i=1}^{n} \exp\{y_i b(\theta_i) + d(y_i) + c(\theta_i)\} \tag{2.3}$$

$$= \exp\left\{\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right\} \tag{2.4}$$

θ is the vector of interest $= (\theta_1, \cdots\cdots, \theta_n)$, and ϕ is the vector of nuisance parameters.

We would hope that the variation in Y_i or $E(Y_i) = \theta_i$ values could be explained in terms of the $\widetilde{x_i}'$ values.

We would hope that we could find a suitable Link Function $g(\theta_i)$ such that the model:

$$g(\theta_i) = \widetilde{x_i}'\,\beta \tag{2.5}$$

held where $\beta = (\beta_1, \cdots, \beta_p)'$ is a vector of regression coefficients. This link function is often the natural parameter.

CHAPTER 3

METHODOLOGY

3.1 Introduction

We restrict the models to a case where the response variable is binary and make an assumption that all the covariates are available. The study introduces key concepts regarding missing data in survey data analysis, provide a conceptual framework for how to approach missing data in this setting. The theoretical background and the method behind the analysis of the data are presented. We develop a theoretical model for handling missing data by using logistic regression technique, demonstrate how the prevalence can be estimated in the presence of missing data by using the logistic regression model and develop a two stage approaches to improve the prevalence estimates. In the first stage, we come up with a model to predict the missing binary observations and then in the second stage we recalculate the prevalence using the observed binary data and the imputed binary data. The model is tested using simulated data.

3.2 Statistical Model

To predict the values for the missing data and to identify the underlying determinants which have significant effect on the prevalence, a statistical model will be employed. Therefore, due to the binary nature of the outcome variable in this study, being positive or negative, a binary logistic regression model will be used for the given data.

3.2.1 The logistic Regression

Logistic regression was first proposed in the 1970s as alternative techniques to overcome limitations of ordinary least square (OLS) regression in handling dichotomous outcome. Logistic regression has been used in epidemiological research, where often the outcome variable is the presence or absence of some disease. In the logistic regression analysis the aim is to find the best fitting and most parsimonious, yet biologically reasonable, a model for describing the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables. The key quantity in any regression problem is the mean value of the outcome variable, given the value of the independent variable. This quantity is called the conditional mean and will be expressed as $E(Y/x)$, where Y denotes the binary or dichotomous outcome variable and x denotes a value of the independent variable.

Many distribution functions have been proposed for use in the analysis of a dichotomous outcome variable (the response variable).

There are two primary reasons for choosing the logistic distribution.

(i) from a mathematical point of view it is an extremely flexible and easily used function, and

(ii) it lends itself to a biologically meaningful interpretation.

Furthermore, assume that the outcome variable has been coded as 0 or 1 representing the absence or presence of the characteristic, respectively.

To fit the logistic regression model to a set of data requires that we estimate the values of $\beta_i^{'}s$, the unknown parameters. What distinguishes the logistic regression model from the linear regression model is that the outcome variable in logistic regression is categorical and most usually binary or dichotomous. Consider a binary random variable Y which defines the absence or presence of characteristic

of a disease. Suppose we have a sample of size n independent observations of the pair (x_i, y_i), where y_i represents the observed values for the i^{th} individual and let y be the column vector containing the elements y_i. Y can be considered as a column vector of n Bernoulli random variables Y_i. Let π be a column vector also of length n with elements $\pi_i = P(Y_i = 1/x_i)$, this means that π_i is the probability of success for any given observation for the i^{th} observation. In the linear component of the model, we have the design matrix and the vector of parameters to be estimated. The design matrix of the independent variables which are categorial predictors, X, is composed of n rows and $k+1$ columns, where k is the number of independent variables which are specified in the model.

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \tag{3.1}$$

In each row of the design matrix, the first element $x_{i0} = 1$ .This value x_{i0} is called the intercept. The parameter vector, is a column vector of length $k+1$.

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \tag{3.2}$$

In each of the k columns of independent variable settings in X, there is one corresponding parameter, and plus β_0, for the intercept.

The logistic regression model equates the logit transform, the log-odds of the probability of a success, to the linear component:

$$g(x) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=0}^{k} x_{ij}\beta_j \tag{3.3}$$

$$i = 1, \cdots, n$$

In this transformation, the importance is that $g(x)$ has many of the desirable properties of a linear regression model. The logit, $g(x)$, is linear in its parameters, may be continuous, and may range from $-\infty$ to ∞ depending on the range of x. By taking the exponential of Eq(3.3) , we find that the odds for the i^{th} unit are given by:

.

$$\left(\frac{\pi_i}{1-\pi_i}\right) = \exp\left\{\sum_{j=0}^{k} x_{ij}\beta_j\right\} \tag{3.4}$$

$$i = 1, \cdots\cdots, n$$

Solving for the probability π_i in the logit model in Equation (3.4) gives the following model

$$\pi_i = E\left(Y_i = 1/x_i\right) = \frac{\exp\left\{\sum_{j=0}^{k} x_{ij}\beta_j\right\}}{1+\exp\left\{\sum_{j=0}^{k} x_{ij}\beta_j\right\}} \tag{3.5}$$

3.3 Prevalence Estimation

Consider a population that, consists of n living individuals who can be infected or not by a disease. The disease status of individual i is represented by the binary indicator y_i, which is equal to 1 if individual i is positive and is equal to 0 otherwise.

$$y_i = \begin{cases} 1 & if\ the\ i^{th}\ individual\ has\ the\ desease \\ 0 & if\ the\ i^{th}\ individual\ doesn't\ have\ the\ desease \end{cases}$$

$$\theta = Prevalence = Pr(T = 1) = \frac{\sum_{i=1}^{n} y_i}{n} \tag{3.6}$$

where T is a random variable that represents the variability of the indicator of disease status in the population.

Thus, disease prevalence is just the proportion of infected people. Our aim is to estimate $Pr(T = 1)$ from sample surveys when the disease status may be missing for some cases.

By the law of total probability, we can write the disease prevalence as:

$$Pr(T = 1) = Pr(T = 1|R = 1)Pr(R = 1) + Pr(T = 1|R = 0)Pr(R = 0), \tag{3.7}$$

where R is a binary indicator equal to 1 if disease status is known and to 0 otherwise.

$$R = \begin{cases} 1 & if\ y_i\ is\ observed \\ 0 & if\ y_i\ is\ missing \end{cases}$$

The missing data problem arises because the data tell us nothing about $Pr(T = 1|R = 0)$.

Let I be the set of indices for the observed values and J be set of indices for the missing values.

From 3.7 and the fact that $Pr(A, B) = Pr(A/B) \times Pr(B)$, we have:

$$Pr(T = 1) = Pr(T = 1, R = 1) + Pr(T = 1, R = 0)$$

Since

$$Pr(T=1,R=1)=\frac{\sum_{i\epsilon I} y_i^0}{n}$$

,

and

$$Pr(T=1,R=0)=\frac{\sum_{i\epsilon J} y_i^m}{n}$$

,

$$Pr(T=1)=\frac{\sum_{i\epsilon I} y_i^0}{n}+\frac{\sum_{i\epsilon J} y_i^m}{n} \tag{3.8}$$

From $Eq(3.8)$, to estimate the prevalence $Pr(T=1)$, we will find the estimated missing values. And for that we will use the logistic regression model to estimate the probability of success for those missing values.

3.3.1 Fitting the Logistic Regression using the observed Outcomes

Denote Y^0 the matrix of the observed data and y^0 the matrix of the corresponding value i.e $R=1$.

Let π_i^0 be the probability of success for the i^{th} individual such that R=1.

Note that $Y_i^0 \sim Bern\,(\pi_i^0)$

Denote by X^0 be the matrix that contains all the explanatory variables corresponding to Y^0.

To illustrate this, let us consider the following table:

$$y_i^0=\begin{cases}1 & if\ the\ i^{th}\ individual\ has\ the\ desease\\ 0 & if\ the\ i^{th}\ individual\ doesn't\ have\ the\ desease\end{cases}$$

By making an assumption that all the covariates are available, we can fit a logistic regression model considering only the observed data and the corresponding

variables.

$$logit\left(\pi_i^0\right) = \left(X_i^0\right)^T \beta \tag{3.9}$$

Where X_i^0 is the vector of the explanatory variables such that $R = 1$.

From model (3.9), we can now estimate the parameters.

3.3.2 Parameter Estimation

Estimating the $k + 1$ unknown parameters in the logistic regression is one of our goals in this study. To achieving this, we use the maximum likelihood estimation which entails finding the set of parameters for which the probability of the observed data is greatest. In linear regression, the method used most often to estimate unknown parameters is least squares. In this method, we choose those values of β_i that 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 regression, the least squares method yields estimators with a number of desirable statistical properties. Unfortunately, when the least squares method is applied to a model with a dichotomous outcome the estimators no longer have these desirable properties. The maximum likelihood equation is derived from the probability distribution of the dependent variable.

To be able to use this method, the likelihood function which expresses the probability of the observed data as a function of the unknown parameters must be first constructed. The maximum likelihood estimators of these parameters are chosen to be those values that maximize this function. Thus, the resulting estimators are those that agree most closely with the observed data no longer have these same properties. 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 maximum likelihood. This is the method used to estimate the

logistic regression parameters. In a very general sense, the maximum likelihood method yields values for the unknown parameters that maximize the probability of obtaining the observed set of data.

This will be denoted by $Pr\left(Y=1|x\right)$. It follows that the quantity $1-\pi_i$ gives the conditional probability that Y is equal to zero given x, $Pr\left(Y=0|x\right)$.

Thus, for those pairs (x_i, y_i), where $y_i = 1$, the contribution to the likelihood function is π_i, and for those pairs where $y_i = 0$, the contribution to the likelihood function is $1-\pi_i$, where the quantity π_i denotes the value π computed at x_i

A convenient way to express the contribution to the likelihood function for the pair (x_i^0, y_i^0) is through the term

$$\xi\left(x_i^0\right) = \left(\pi_i^0\right)^{y_i^0}\left[1-\pi_i^0\right]^{1-y_i^0} \tag{3.10}$$

Because of the independence of the observation, the likelihood function is obtained by taking the product of the terms given in (3.10)

$$l\left(y^0/\beta\right) = \prod_{i=1}^{n}\xi\left(x_i^0\right) \tag{3.11}$$

To derive estimates of the unknown β parameters, as in the univariate case, we need to maximize this likelihood function. We follow the usual steps, including taking the logarithm of the likelihood function, taking $(k+1)$ partial derivatives with respect to each β parameter and setting these $(k+1)$ equations equal to zero, to form a set of $(k+1)$ equations in $(k+1)$ unknowns. Solving this system of equations gives the maximum likelihood equations. The maximum likelihood equations in the logistic regression are nonlinear $\beta_j's$, and thus require special methods for finding their solution. These methods are iterative in nature and have been programmed into available logistic regression software

Let us consider the following two first likelihood equations

$$\sum_{i\epsilon I} \left(y_i^0 - \pi_i^0\right) = 0 \tag{3.12}$$

and

$$\sum_{i\epsilon I} x_i^0 \left(y_i^0 - \pi_i^0\right) = 0 \tag{3.13}$$

Each such solution, if any exists, specifies a critical point. The critical point will be a maximum if the matrix of second partial derivatives is negative definite; that is, if every element on the diagonal of the matrix is less than zero. Another useful property of this matrix is that it forms the variance-covariance matrix of the parameter estimates.

3.3.3 Testing for the significance of the model

After having estimated the coefficients in the regression, it is standard practice to assess the significance of the variables in the model. This usually involves testing a statistical hypothesis in order to determine whether the independent variables in the model are "significantly" related to the outcome variable. One approach to testing for the significance of the coefficient of a variable in any model is to see whether the model that includes the variable in question tells us more about the outcome (or response) variable than a model that does not include that variable. This can be done by doing a comparison between the observed values of the response variable with those predicted by each of the two models; the first with and the second without the variable in question. The mathematical function used in comparing the observed and predicted values depends on the particular problem. If the predicted values with the variable in the model are better, or more accurate in some sense, than when the variable is not in the model, then we can say that the variable in question is significant. It is important to note that we are not considering the question of whether the predicted values are an

accurate representation of the observed values in an absolute sense (this would be called goodness of fit). Instead, our question is posed in a relative sense.

For the purposes of assessing the significance of an independent variable we compute the value of the following statistic:

$$G = -2\ln\left(\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}}\right) \tag{3.14}$$

The first step in this process is usually to assess the significance of the variables in the model. The likelihood ratio test for overall significance of the k coefficients for the independent variables in the model is performed based on the statistic G given in (3.14). Under the null hypothesis that the k "slope" coefficients for the covariates in the model are equal to zero, the distribution of G is chi-square with k degrees of freedom.

Rejection of the null hypothesis (that all of the coefficients are simultaneously equal to zero) has an interpretation analogous to that in multiple linear regression; we may conclude that at least one, and perhaps all k coefficients are different from zero.

Before concluding that any or all of the coefficients are nonzero, we may wish to look at the univariate Wald test statistics which is obtained by comparing the maximum likelihood estimate of the slope parameter, β_j , with an estimate of its standard error,

$$W_j = \frac{\hat{\beta}_j}{\hat{se}\left(\hat{\beta}_j\right)} \tag{3.15}$$

Under the hypothesis that an individual coefficient is zero, these statistics will follow the standard normal distribution. Thus, the value of these statistics may give us an indication of which of the variables in the model may or may not be significant.

Considering that the overall goal is to obtain the best fitting model while minimizing the number of parameters, the next logical step is to fit a reduced model, containing only those variables thought to be significant, and compare it with the full model containing all the variables. The likelihood ratio test comparing these two models is obtained using the definition of G given in Equation (3.14).

It has a distribution that is chi-square with k degrees of freedom under the hypothesis that the coefficients for the variables excluded are equal to zero and has a P value of $P\left[\chi^2\left(k\right) > G\right]$.

If the P value is large, we conclude that the reduced model is as good as the full model.

3.3.4 Interpretation of the Coefficients in the Logistic Regression

After fitting a model the emphasis shifts from the computation and assessment of significance of the estimated coefficients to interpretation of their values. The interpretation of any fitted model requires that we can draw practical inferences from the estimated coefficients in the model. The question addressed is:

What do the estimated coefficients in the model tell us about the research questions that motivated the study?

For most models, this involves the estimated coefficients for the independent variables in the model. The estimated coefficients for the independent variables represent the slope or rate of change of a function of the dependent variable per unit of change in the independent variable. Thus, interpretation involves two issues:

(i) determining the functional relationship between the dependent variable and the independent variable, and

(ii) appropriately defining the unit of change for the independent variable.

For a linear regression model we recall that the slope coefficient, β_1 , is equal to the difference between the value of the dependent variable at $x + 1$ and the value of the dependent variable at x, for any value of x. In the logistic regression model $\beta_1 = g(x + 1) - g(x)$. That is, the slope coefficient represents the change in the logit for a change of one unit in the independent variable x. Proper interpretation of the coefficient in a logistic regression model depends on being able to place meaning on the difference between two logits.

3.3.5 Missing Values Estimation

Now From the selected model, we estimate the probability $\hat{\pi_i}^m$, i.e the probability of success for each missing outcome as follow:

$$\hat{\pi}_i^m = \frac{\exp\left\{(X_i^m)^T \hat{\beta}\right\}}{1 + \exp\left\{(X_i^m)^T \hat{\beta}\right\}} \tag{3.16}$$

Where $\hat{\beta}$ is the vector of the coefficients estimated from model (3.9) and X_i^m is the vector of the explanatory variables such that R is equal to 0.

$\hat{\pi_i}^m$ is called the maximum likelihood estimate of π_i^m. This quantity provides an estimate of the conditional probability that Y_i^m is equal to 1, given that x is equal to x_i^m . As such, it represents the fitted or predicted value for the logistic regression model.

Once we have the estimated probabilities $\hat{\pi_i}^m$, we define $\hat{y}_i^m$

$$\hat{y}_i^m = \begin{cases} 1 & if\ \hat{\pi}_i^m \geq \alpha \\ 0 & if\ \hat{\pi}_i^m < \alpha \end{cases} \tag{3.17}$$

Where α is a value that depends on the data.

This means that from those who were missing, if the predicted probability $\hat{\pi}_i^m$ is greater than or equal to α, then one can conclude that individual i is positive.

If $\hat{\pi}_i^m$ is strictly less than α, then that individual i is negative.

We define$\hat{Y^m}$ as the dataset containing all the imputed missing values. We can now calculate the estimated prevalence denoted by $\{Prevalence\}_{est}$ using the full data set containing $\left\{Y^0,\, \hat{Y}^m\right\}$.

We define:

$$z = \sum_{i\epsilon I} y_i^0 + \sum_{i\epsilon J} \hat{y}_i^m \tag{3.18}$$

z is now the number of individuals who have the disease in the full dataset $\left\{Y^0,\, \hat{Y}^m\right\}$.

Now the estimated prevalence from the full data set $\left\{Y^0,\, \hat{Y}^m\right\}$ is given by:

$$\tilde{\theta} = Prevalence_{est} = \frac{z}{n} = \frac{\sum_{i\epsilon I} y_i^0}{n} + \frac{\sum_{i\epsilon J} \hat{y}_i^m}{n}$$

And $\tilde{\theta}$ is the observed value of a random variable $\hat{\Theta}$, the estimator of θ in *eq.* (3.8), which is:

$$\hat{\Theta} = \frac{Z}{n} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{\sum_{i\epsilon I} Y_i^0 + \sum_{i\epsilon J} \hat{Y}_i^m}{n}$$

Where

$$Z = \sum_{i\epsilon I} Y_i^0 \; + \; \sum_{i\epsilon J} \hat{Y}_i^m$$

$\hat{\Theta}$ is the estimator of the estimate prevalence $\tilde{\theta}$.

3.4 Asymptotic Properties of the Estimator $\hat{\Theta}$

3.4.1 Consistency

In statistics, a consistent estimator or asymptotically consistent estimator is an estimator a rule for computing estimates of a parameter θ_0 having the property that as the number of data points used increases indefinitely, the resulting sequence of estimates converges in probability to θ_0. This means that the distri-

butions of the estimates become more and more concentrated near the true value of the parameter being estimated, so that the probability of the estimator being arbitrarily close toθ_0 converges to one. In other words, a consistent sequence of estimators is a sequence of estimators that converge in probability to the quantity being estimated as the index (usually the sample size) grows without bound. Mathematically, a sequence of estimators $\{t_n; n \geq 0\}$ is a consistent estimator for parameter θ if and only if, for all $\varepsilon > 0$, no matter how small, we have

$$\lim_{n\to\infty} Pr\left(|t_n - \theta| < \varepsilon\right) = 1$$

In practice, however, we often make use of more convenient conditions that are easier to verify and that guarantee the above definition of consistency is met.

Theorem. *3.1*

Assume that $\{X_1, X_2, ...\}$ is a sequence of independent random variables, each with finite expected value μ_i and variance σ_i^2 .

If T_n is a sequence of estimators of $\tau(\theta)$ satisfying

$$\lim_{n\to\infty} Bias_\theta\left(T_n\right) = 0 \tag{3.19}$$

$$\lim_{n\to\infty} Var_\theta\left(T_n\right) = 0 \tag{3.20}$$

for all θ, then T_n is consistent for $\tau(\theta)$. Here $\text{Bias}_\theta(T_n) = E(T_n) - \theta$.

We Have

$$E\left(\hat{\Theta}\right) = \frac{\sum_{i\epsilon I} E\left(Y_i^0\right) + \sum_{i\epsilon J} E\left(\hat{Y}_i^m\right)}{n} \tag{3.21}$$

$$= \frac{\sum_{i\epsilon I} \pi_i^0 + \sum_{i\epsilon J} \hat{\pi}_i^m}{n}$$

$$= \frac{\sum_{i=1}^{n} \pi_i}{n}$$

Where π_i now represents the probability of success in the full data set $\left\{Y^0,\ \hat{Y}^m\right\}$

From the likelihood equations, an interesting consequence of (3.12) by fitting the model in the full data set with the imputed missing values is that

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \pi_i$$

This implies that:

$$\frac{\sum_{i=1}^{n} \pi_i}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

This implies that:

$$E\left(\hat{\Theta}\right) = \frac{\sum_{i=1}^{n} y_i}{n} = \theta$$

This implies that $\hat{\Theta}$ is an unbiased estimator of θ.

Furthermore,

$$Var\left(\hat{\Theta}\right) = \frac{\sum_{i\epsilon I} Var\left(Y_i^0\right) + \sum_{i\epsilon J} Var\left(\hat{Y}_i^m\right)}{n^2}$$

$$= \frac{\sum_{i\epsilon I} \pi_i^0\left(1-\pi_i^0\right) + \sum_{i\epsilon J} \hat{\pi}_i^m\left(1-\hat{\pi}_i^m\right)}{n^2}$$

This implies that:

$$Var\left(\hat{\Theta}\right) = \frac{\sum_{i=1}^{n} \pi_i\left(1-\pi_i\right)}{n^2}$$

since $\quad \pi_i\left(1-\pi_i\right) \leq 1$

This implies that:

$$Var\left(\hat{\Theta}\right) = \frac{\sum_{i=1}^{n} \pi_i \left(1 - \pi_i\right)}{n^2} \leq \frac{n}{n^2} = \frac{1}{n} \tag{3.22}$$

This implies that:

$$Var\left(\hat{\Theta}\right) \longrightarrow 0,$$

as

$$n \longrightarrow \infty$$

From the above theorem, $\hat{\Theta}$ is consistent for θ

3.4.2 Normality

An asymptotically normal estimator is a consistent estimator whose distribution around the true parameter θ approaches a normal distribution as the sample size n grows.

Mathematically, a consistent estimator $\hat{\Theta}$ of a parameter θ, is defined to have an asymptotic normal distribution if

$$\frac{\hat{\Theta} - \theta}{\sqrt{\frac{V\left(\theta\right)}{n}}} \xrightarrow{d} N\left(0, 1\right) \qquad as\ n \longrightarrow \infty$$

for some quantity $V\left(\theta\right)$

Theorem. *3.2 (Lyapunov CLT)*

Suppose $\{X_1, X_2, ...\}$ is a sequence of independent random variables, each with finite expected value μ_i and variance σ_i^2. Define

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2 \tag{3.23}$$

If for some $\delta > 0$, the Lyapunov's condition

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} E\left(|X_i - \mu_i|\right)^{2+\delta} = 0 \tag{3.24}$$

is satisfied, then a sum of $\frac{X_i - \mu_i}{s_n}$ converges in distribution to a standard normal random variable, as n goes to infinity:

$$\frac{1}{s_n} \sum_{i=1}^{n} \left(X_i - \mu_i\right) \xrightarrow{d} N\left(0, 1\right).$$

Equivalent to:

$$\sqrt{n} \left(\frac{\bar{X} - \bar{\mu}_n}{\frac{s_n}{\sqrt{n}}} \right) \xrightarrow{d} N\left(0, 1\right)$$

Where:

$$\bar{X} = \sum_{i=1}^{n} X_i \tag{3.25}$$

and

$$\bar{\mu}_n = \sum_{i=1}^{n} \mu_i \tag{3.26}$$

Since our $Y_i's$ are following $bernoulli\left(\pi_i\right)$ and they are independent from each other.

$$E\left(Y_i = 1 | x_i\right) = \pi_i \tag{3.27}$$

and

$$Var\left(Y_i = 1 | x_i\right) = \pi_i \left(1 - \pi_i\right) \tag{3.28}$$

$$s_n^2 = \sum_{i=1}^{n} \pi_i \left(1 - \pi_i\right) \tag{3.29}$$

Let

$$X_{ni} = X_i - \pi_i \tag{3.30}$$

for any $\delta > 0$,

$$E\left|X_{ni}\right|^{2+\delta} \leq E\left(X_{ni}^2\right) = \pi_i \left(1 - \pi_i\right) \leq 1 \tag{3.31}$$

This implies that

$$\frac{1}{s_n^{2+\delta}}\sum_{i=1}^{n} E\left|X_{ni}\right|^{2+\delta} \leq \frac{1}{s_n^{2+\delta}}\sum_{i=1}^{n} Var\left(X_{ni}\right) = \frac{1}{s_n^{\delta}} \tag{3.32}$$

Therefore,

if $s_n \longrightarrow \infty$(which is clearly true because π_i is bounded away from 0 and 1), the Lyapunov condition is satisfied and so

$$\frac{\sum_{i=1}^{n} X_{ni}}{s_n} \xrightarrow{d} N\left(0,1\right) \tag{3.33}$$

This implies that

$$\sqrt{n}\left(\frac{\left(\bar{Y}-\bar{\pi}_n\right)}{\frac{s_n}{\sqrt{n}}}\right) \xrightarrow{d} N\left(0,1\right) \tag{3.34}$$

Where:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}, \tag{3.35}$$

$$\bar{\pi}_n = \frac{\sum_{i=1}^{n} \pi_i}{n}, \tag{3.36}$$

$$s_n = \sqrt{\sum_{i=1}^{n} \pi_i\left(1-\pi_i\right)}, \tag{3.37}$$

From the likehood equations:

$$\sum_{i=1}^{n}\left(y_i - \pi_i\right) = 0 \tag{3.38}$$

This implies that the true prevalence is:

$$\theta = \frac{\sum_{i=1}^{n} \pi_i}{n} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{3.39}$$

Since

$$\hat{\Theta} = \frac{\sum_{i=1}^{n} Y_i}{n} \tag{3.40}$$

and

$$Var\left(\hat{\Theta}\right) = \frac{\sum_{i=1}^{n} \pi_i \left(1 - \pi_i\right)}{n^2} \tag{3.41}$$

$$\left(\frac{\left(\hat{\Theta} - \theta\right)}{\frac{V_n\left(\theta\right)}{n}} \right) \xrightarrow{d} N(0,1)$$

Where:

$$V_n\left(\theta\right) = \frac{s_n}{n} \tag{3.42}$$

Thus, we conclude that the estimator $\hat{\Theta}$ has an asymptotic normal distribution.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Simulated Data

We simulate 3000 binary observations from a logistic regression model where the outcome variable is called Disease and covariates Age, Sex, Ever married, Urban, Educational level, Condom use and other. For the first time, we assume that both the outcome variable and the covariates are fully observed, then we compute the true prevalence. After that, we consider a case where the variable of interest (response variable or outcome variable) is binary and some of the observations are missing and assume that all the covariates are fully observed. In this simulation study, we consider two steps. Firstly, we create randomly 10%, 20%, 30% and 50% of missing data along the outcome variable over 1000 simulation runs using Monte Carlo simulation. Now after creating these missing values, we use the method described above to estimate the prevalence over the 1000 simulation and then take the average estimates of the prevalence of each of these four scenarios. Secondly we only create missing values among those whose disease status is positive to examine the sensitivity to a non random missing data. We also use our method to estimate the prevalence. It is well known that it is possible to estimate the probability of occurrence of disease status from a logistic model. The estimates prevalence without the missing values and the estimates prevalence from our method are further compared with the true prevalence. A Wald test statistic based on the parameter estimate divided by its standard error estimate was used to calculate the proportion of rejections for a Wald test of the null hypothesis that the true parameter is equal to the chosen parameter. When

the null hypothesis was that the true parameter value was zero, a likelihood ratio test for the significance of the variable was computed.

These values of W_j in *eq.*(3.13) are given in the fourth column in Table 4.1 Under the hypothesis that an individual coefficient is zero, these statistics will follow the standard normal distribution. The $p - values$ are given in the fifth column of Table 4.1

Table 4.1: Estimated Coefficients of all variables from the from the fitted model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.259	0.151	-1.717	0.086
condom use (yes)	0.226	0.075	3.008	0.002
Sex (Male)	-0.198	0.075	-2.644	0.008
Ever married (yes)	0.043	0.075	0.586	0.558
Urban (yes)	0.065	0.075	0.873	0.382
Age	0.012	0.002	6.032	1.62e-09

If we use a level of significance of 0.05, then we can conclude that only the variables condom use, Sex and Age are significant and the others are not significant.

If our goal is to obtain the best fitting model while minimizing the number of parameters, the next logical step is the reduced model containing only those variables thought to be significant, and compare it to the full model containing all the variables. The results of fitting the reduced model are given in Table 4.2

Table 4.2: Estimated Coefficients for the variables age, condom use and sex from the reduced model

	Estimate	Std. Erro	z value	Pr(>|z|)
(Intercept)	-0.262	0.120	-2.184	0.028
Age	0.012	0.002	6.051	1.44e-09
Condom use Yes	0.229	0.075	3.053	0.002
Sex Male	-0.202	0.075	-2.697	0.007

The value of the statistic comparing the models in Table 4.1 and in Table 4.2 is $G = -3.7454$

Which has a $p - value\ Pr(\chi^2(k) > -3.7454) = 0.8086$, where k is the number of degrees of freedom in this case. Since the $p - value$ is large, exceeding 0.05, we conclude that the reduced model is as good as the full model. Thus there is no advantage to include the other variables in the model.

4.1.1 Simulation Results

Table 4.3: Average estimates of the prevalence, their average bias and their 95% confidence intervals over 1000 simulation runs for 10%, 20%, 30%, 40% and 50% of missng values. The true prevalence is 0.603

% of missing values	Average estimates of the Prevalence withoutmissing values	Average Estimates of the Prevalence	Bias	95 % CI
10%	0.601	0.596	-0.007	0.578-0.613
20%	0.596	0.611	0.008	0.593-0.628
30%	0.594	0.591	-0.012	0.574-0.609
40%	0.594	0.609	0.006	0.592-0.627
50%	0.591	0.590	-0.005	0.572-0.607
True Prevalence		0.603		

Table 4.3 displays the average estimates of the prevalence based on our method, their bias and their 95% confidence intervals and the average estimates of the prevalence without the missing values over 1000 simulation runs using Monte Carlo simulation. These are compared to the true prevalence (0.603) shown in the last line of this table, which uses the full database before creating the missing values from the disease status. We find that these average estimates of the prevalence based on our method described above are almost identical to those in column 2 which were based only on observations without missing data. These averages estimate of prevalences are both close to the true prevalence. We note also that the prevalence obtained by ignoring the missing values and the estimates prevalence obtained from our method are very similar. The estimated prevalences based on our approach presented similar estimates of prevalence that are close to the true prevalence. From these results, we can see that if the missingness

is created randomly or involves only those which are negative, the prevalence without the missing values is close to the true prevalence, meaning that the true prevalence might not be affected. However, our method can still be used to estimate the prevalence for some missing cases as shown in the table. There are two other important features of these results. First, we find that the estimates prevalence based on this approach and the one obtained by ignoring the missing cases are almost identical to the true prevalence. Second, the confidence intervals obtained from our method contains always the true prevalence. These confidence intervals are not so wide, and they include the true prevalence, indicating that the uncertainty to rule out selection bias is not higher. When the amount of missing observations increased, we realise that our method still continues to produce almost unbiased estimates. However, our approach is easy to implement, it does not require any assumptions about the nature of the missing data, and it allows us to obtain reliable intervals from a statistical point of view. Therefore, we conclude that even if the prevalence without the missing data is close to the true prevalence, our method can still be used to find the estimated prevalence that will be closed to the true prevalence.

Now let us consider the case where we assume that all the missing observations are positive.

Table 4.4: Summary of the disease status when there are some missing cases among those whose disease status is positive

Sample size N= 3000			
% of missing values	Positive disease status	Negative disease status	Number of missingness
0 %	1800	1200	0
10 %	1620	1200	180
20 %	1440	1200	360
30 %	1260	1200	540
40 %	1080	1200	720
50 %	900	1200	900

Table 4.5: Estimated Prevalence when there are only missing values among those whose disease status is positive

% of missing values	Prevalence without the missing values	Estimated Prevalence	Bias	95 % CI
10%	0.574	0.597	-0.006	0.575-0.608
20%	0.545	0.587	-0.016	0.568-0.603
30%	0.512	0.618	0.015	0.600-0.636
40%	0.473	0.612	0.012	0.594-0.630
50%	0.428	0.598	-0.002	0.581-0.615
True Prevalence		0.60		

Table 4.5 shows the estimate prevalence when there are only some missing cases among those whose disease status is positive. To examine the sensitivity to a non random missing data, missing values were created among those whose disease status is positive. Even if it is rare, it is possible because individuals might know

they are positive because they have been tested before or fear they are positive because of private information on own sexual behavior. Those who refuse to take the test may simply not believe that the results cannot be traced back to the individual, and they may fear for exposure of being found out to be infected with the disease. This fear is likely to be higher among those with high-risk behavior, which in turn is an unobserved determinant of the disease-status. For the first time, we use the full database without any missing values and calculate the true prevalence. For the second time, we create 10%, 20%, 30%, 40% and 50% of missing values, then we compute the prevalence without the missing values for each of these four scenarios. Finally, we use the method described above to estimate the prevalence using both the observed values and the imputed missing values. Using simulated data, we find that when the missing cases are among those whose disease status is positive, the true disease prevalence can be affected by the presence of missing values. Our results show that the estimated prevalence from the method described above is better than the prevalence calculated by ignoring the missing values. As the number of missing values increases, the prevalence without those missing values decreases. According to our results, the prevalence could be much lower, as a larger part of the non respondents could be infected. This can be seen from the table by comparing the prevalence calculated by ignoring the missing values from the true prevalence. If we ignore the missing data and compute directly the prevalence from the observed data we realise that the prevalence can be different from the true prevalence because of the missing data. When the number of missing values is higher, the estimates prevalence from our method are significantly higher than the prevalence without the missing values. As we can see from the table, an important finding is that when the number of missing values is higher, the estimates prevalence without the missing values substantially underestimate the true prevalence. But from the table, when using our method to estimate the prevalence by using the full

dataset containing both the observed and the predicted missing values, we obtain a prevalence that is very close to the true prevalence.We can see also from the table that the true prevalence is always lying inside the confidence interval. Thus the method described in this study can still be used to estimate the disease prevalence when there are some missing cases.

4.2 HIV Data from Kenya

The 2007 KAIS was conducted among a representative sample of households selected from all eight provinces in the country, covering both rural and urban areas. A household was defined as a person or group of people related or unrelated to each other who live together in the same dwelling unit or compound, share similar cooking arrangements, and identify the same person as the head of household. The household questionnaire was administered to consenting heads of sampled, occupied households. All women and men aged 15-64 years in selected households who were either usual residents or visitors present during the night before the survey were eligible to participate in the individual interview and blood drawn, provided they gave informed consent. For minors aged 15-17 years, parental consent and minor assent were both required for participation. Participants could consent to the interview and blood draw or to the interview alone. The inclusion criteria may have captured non-Kenyans living as usual residents or visitors in a sampled household. Military personnel and the institutionalized population (e.g. imprisoned) are typically not captured in similar household-based surveys may have been included in the 2007 KAIS if at home during the survey.

Administratively, Kenya is divided into eight provinces. Each province is divided into districts, each district into divisions, each division into locations, each location into sub-locations, and each sub-location into villages. For the 1999 Pop-

ulation and Household Census, the Kenya National Bureau of Statistics (KNBS) delineated sub-locations into small units called Enumeration Areas (EAs) that constituted a village, a part of a village, or a combination of villages. The primary sampling unit for Kenya's master sampling frame, and for the 2007 KAIS, is a cluster, which is constituted as one or more EAs, with an average of 100 households per cluster.

The master sampling frame for the 2007 KAIS was the National Sample Survey and Evaluation Programme IV (NASSEP IV) created and maintained by KNBS. The NASSEP IV frame was developed in 2002 based on the 1999 Census. The frame has 1800 clusters, comprised of 1,260 rural and 540 urban clusters. Of these, 294 (23%) rural and 121 (22%) urban clusters were selected for KAIS. The overall design for the 2007 KAIS was a stratified, two-stage cluster sample for comparability to the 2003 KDHS. The first stage involved selecting 415 clusters from NASSEP IV and the second stage involved the selection of households per cluster with equal probability of selection in the rural-urban strata within each district. The target of the 2007 KAIS sample was to obtain approximately 9,000 completed household interviews. Based on the level of household non- response reported in the 2003 KDHS (13.2% of selected households), 10,375 households in 415 clusters were selected for potential participation in the 2007 KAIS.

Table 4.6: Summary of the HIV status

Sample size N = 11338

Number of missingness	Positive disease status	Negative disease status	% of missing values
3401	641	7296	30 %

Table 4.7: Estimated Coefficients from the fitted model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.303	0.340	-9.710	< 2e-16
herpes (Yes)	2.148	0.115	18.682	< 2e-16
Age 20-24	0.495	0.273	1.813	0.069
Age 25-29	0.561	0.273	2.051	0.040
Age 30-34	0.694	0.275	2.525	0.011
Age 35-39	0.490	0.279	1.753	0.079
Age 40-44	0.351	0.286	1.228	0.219
Age 45-49	0.175	0.291	0.603	0.546
Age 50-54	0.118	0.308	0.384	0.701
Age 55-59	-0.604	0.353	-1.711	0.087
Age 60-64	-0.774	0.407	-1.899	0.057
Final Marital status Married,+2 partner	0.248	0.142	1.742	0.081
Final Marital status Divorced / Separated / Widowed	0.909	0.110	8.257	< 2e-16
Final Marital status Never Married	0.096	0.158	0.611	0.541
Ever used condom No	-0.537	0.092	-5.777	7.62e-09
Education level Primary	0.059	0.103	0.576	0.564
Education level Secondary	-0.204	0.141	-1.442	0.149
Education level Higher	-0.664	0.210	-3.157	0.001
STI No	-0.660	0.204	-3.231	0.001

Now we can analyze the model fitting and interpret what the model is telling us. As for the statistically significant variables, all the variables have a small p-value suggesting a strong association of these variables with the probability of being positive.

Table 4.8: Table of deviance

	Df	Deviance Resid.	Df	Resid. Dev	Pr(>Chi)
NULL			7936	4507.9	
herpes	1	569.69	7935	3938.2	< 2.2e-16
Age	9	63.09	7926	3875.1	3.382e-10
FinalMaritalstatus	3	69.15	7923	3806.0	6.494e-15
Ever_used_condom	1	31.08	7922	3774.9	2.481e-08
educationlevel	3	16.18	7919	3758.7	0.001
STI	1	9.60	7918	3749.1	0.001

The difference between the null deviance and the residual deviance shows how our model is doing against the null model (a model with only the intercept). The wider this gap, the better. Analyzing the table, we can see the drop in deviance when adding each variable one at a time. A large p-value here indicates that the model without the variable explains more or less the same amount of variation. From the Table 4.8, we can see that these variables are significant according to their p-value.

4.2.1 Results from the HIV Data

Table 4.9: Estimated HIV Prevalence and its confidence interval

Prevalence without the missing values	Estimated Prevalence	95 % CI
0.080	0.095	0.090 - 0.101

This Table 4.9 shows the HIV estimated prevalence from Kenya when there are some missing cases by using our method described above. Using HIV/AIDS data from the Kenya AIDS Indicator Survey 2007 HIV, where there are some missing cases along the outcome variable, we find that the true HIV prevalence might be affected by the presence of missing as shown in our simulation studies. Our results show that the estimated prevalence from our method is higher than the prevalence calculated by ignoring those missing values. According to the simulation studies, the estimates prevalence from our method are always close to the true prevalence and the confidence intervals contain also true prevalence, thus we can conclude that this estimated prevalence (0.095) from our method would be close the true prevalence which could be contained in the confidence interval (0.090 - 0.101) from the table

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

Incomplete data are a pervasive problem in medical research, and ignoring them or handling them inappropriately may bias study results, reduce power and efficiency. Appropriate handling of censored values in medical research especially when dealing with prevalence should be a substantial concern of investigators, and planning for the integration of valid incomplete data methods into the analysis is important. This study shows that non-participation in disease testing may be an important source of bias in disease prevalence estimates. However, our approach is easy to implement. It does not require many assumptions, and it allows to obtain the estimated prevalence and reliable confidence intervals from a statistical point of view. This method allows to have disease estimated prevalence that can be close to the true prevalence.

Moreover, we stress the fact that it is important to design well surveys to reduce non response, either unit and item non response. It is also critical to include in the data information, such as interviewer's characteristics, fieldwork procedures etc, as they can be used as instrumental variables. Two approaches to the problem are to reduce the frequency of missing data in the first place and to use appropriate statistical techniques that account for the missing data.

REFERENCES

Beyrer, C., Baral, S., Kerrigan, D., El-Bassel, N., Bekker, L., and Celentano, D. (2011). Expanding the space: inclusion of most-at-risk populations in HIV prevention, treatment, and care services. *Journal of Acquired Immune Deficiency Syndromes*, 57(Suppl I):S96–S99.

Boerma, J., Ghys, P., and Walker, N. (2003). Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. *Lancet*, 362(9399):1929–1931.

Bor, J., Herbst, A. J., Newell, M.-L., and Bärnighausen, T. (2013). Increases in adult life expectancy in rural South Africa: valuing the scale-up of HIV treatment. *Science*, 339:961–965.

De Cock, K. M., Bunnell, R., and Mermin, J. (2006). Unfinished Business-Expanding HIV Testing in Developing Countries. *The New England journal of medicine*, 354(5):440–442.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

Enders, C. K. (2010). *Applied Missing Data Analysis*. The Guilford Press, New York London.

Haukoos, J. S. and Newgard, C. D. (2007). Advanced Statistics: Missing Data in Clinical Research-Part1: An Introduction and Conceptual Framework. *Society for Academic Emergency Medicine*, 14:662–668.

Hogan, D. R., Salomon, J. A., Canning, D., Hammitt, J. K., Zaslavsky, A. M., and Barnighausen, T. (2012). National HIV prevalence estimates for sub-Saharan

Africa: controlling selection bias with Heckman-type selection models. *Sex Transm Infect*, 88:i17–i23.

Horton, N. J. and Laird, N. M. (2001). Maximum Likelihood Analysis of Logistic Regression Models with Incomplete Covariate Data and Auxiliary Information. *Biometrics*, 57:34–42.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-Data Methods for Generalized Linear Models: A Comparative Review. *American Statistical Association Journal of the American Statistical Association*, 100(469):332–347.

Kranzer, K., Govindasamy, D., Ford, N., Johnston, V., and Lawn, S. (2012). Quantifying and addressing losses along the continuum of care for people living with HIV infection in sub-Saharan Africa: a systematic review. *Journal of International AIDS Society*, 15(2):1738–17533.

Little, R. J. and B, R. D. (1987). *Statistical Analysis with Missing Data*. John Wiley and Sons, New York.

Marra, G., Radice, R., Barnighausen, T., Wood, S. N., and McGovern, M. E. (2015). A Unified Modeling Approach to Estimating HIV Prevalence in Sub-Saharan African Countries. Technical report, University College London, London.

Mishra, V., Barrere, B., Hong, R., and Khan, S. (2008). Evaluation of bias in HIV seroprevalence estimates from national household surveys. *Sex Transm Infect*, 84(Suppl I):63–70.

Peugh, J. L. and Enders, C. K. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research*, 74(4):525–556.

Roth, P. L. (1994). Missing data: A Conceptuel Rewiew for Applied Psychologists. *Personnel Psychology*, 47(3):537–560.

Schafer, J. L. and Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, 7(2):147–177.

Schlomer, G. L., Bauman, S., and Card, N. A. (2010). Best Practices for Missing Data Management in Counseling Psychology. *Journal of Counseling Psychology*, 57(1):1–10.

Tanser, F., Barnighausen, T., Grapsa, E., Zaidi, J., and Newell, M.-L. (2013). High Coverage of ART Associated with Decline in Risk of HIV Acquisition in Rural KwaZulu-Natal, South Africa. *Science*, 339(6122):966–971.

Tanser, F., Hosegood, V., Barnighausen, T., Herbst, K., Nyirenda, M., Muhwava, W., Newell, C., Viljoen, J., Mutevedzi, T., and Newell, M.-L. (2008). Cohort Profile: Africa Centre Demographic Information System (ACDIS) and population-based HIV survey. *International Journal of Epidemiology*, 37:956–962.

Wilkinson, L. and Task Force, o. S. I. (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist*, 54(8):594–604.