Statistical approaches for handling longitudinal and cross sectional discrete data with missing values focusing on multiple imputation and probability weighting

A thesis submitted to the University of KwaZulu-Natal
for the degree of Doctor of Philosophy of Science
in the College of Agriculture, Engineering & Science

By
Aluko Omololu Stephen
School of Mathematics, Statistics & Computer Science
April 2018

Contents

List of Tables

Abstract

In the case of longitudinal studies, measurements are taken repeatedly over a sequence of time points on the same experimental unit. The key property intrinsic to repeated measures data is that observations within an individual or unit of measurement are correlated. This correlation structure must be taken into account when modeling such data. But in cross sectional survey studies, measurements are recorded once for each unit, but we can still have correlated data for example measurements on individuals from the same household, students within a class in a school, patients within a hospital and many more. In both cross-sectional and longitudinal studies, missingness may occur in either covariates or outcome variable or both. The use of standard regression, analysis of variance and multivariate analysis of variance techniques may produce biased results because these methods entail some mathematical assumptions that do not hold for repeated measurements data. However, some appropriate methods that are capable of dealing with the incomplete nature of the data with the possibility of identifying the reasons for incomplete data in the analysis are used. This thesis studies and compares the performance of several methods for missing data that can be used under specific underlying assumptions and data types both discrete and continuous. These assumptions covers ignorability and non-ignorability but we focus more on former which comprises missing completely at random (MCAR)and missing at random (MAR).

Actually the objective is to handle data with either missing covariate or response or both. The thesis begins with the case of data with incomplete outcome. In Chapter 2, the thesis compares the performance of three different enhancements of the generalized estimating equations (GEE); the weighted GEE, multiple imputation (MI) based GEE and doubly

robust (DR) based GEE. Chapter 3 compares and reveals the strengths of MI (fully conditional specification) against the inverse probability weighting method with a real data and simulation studies. In Chapter 4, the thesis presents the use of MI to handle ordinal longitudinal data with missing observations in both the predictor and response variables. The MI strategies are multivariate normal imputation (MVNI), fully conditional specification (FCS) and Expectation maximization (EM) which are compared in both application and simulation studies. We used ordinal negative binomial count model in the analysis and applied the model in the ordinal simulation study. Chapter 5 presents the conclusion derived from each chapter of this thesis and the possible areas of further research.

Declaration

The work described in this thesis was carried out under the supervision and direction of Prof. H. Mwambi in the School of Mathematics, Statistics, & Computer Science, University of KwaZulu-Natal (PMB), from April 2015 to April 2018.

No portion of the work referred to in this dissertation has been submitted in any form to any university or institution of learning for any degree or qualification. The thesis represents my original work except where due reference and credit are given.

Sign: ....

Aluko Omololu Stephen Date

Sign: ....

Prof. H. Mwambi Date

Publications

Aluko Omololu S and Mwambi H. A comparison of three different enhancements of the generalized estimating equations methods in handling incomplete longitudinal binary outcome. *Global Journal of Pure and Applied Mathematics*, 13(10):7669–7688, 2017.

Aluko Omololu S and Mwambi H. The use of fully conditional specification of multiple imputation and inverse probability weighting to model the pulmonary disease occurrence in survey data with non-response. *ARPN Journal of Engineering and Applied Sciences*, Accepted.

Aluko Omololu S and Mwambi H. Statistical methodology for handling ordinal longitudinal responses with monotone dropout patterns using multiple imputation. *International Journal of Pure and Applied Mathematics*, In review.

Acknowledgments

My profound appreciation goes to my supervisor, Prof. Mwambi, who has mentored me in research, critical thinking and scientific writing. I am highly indebted to his invaluable insight, suggestions and motivation during the programme. The research insight and expertise provided by him has further improved the application of this research.

I would like to thank the School of Mathematics, Statistics, & Computer Science, University of KwaZulu-Natal, Pietermaritzburg Campus. I am deeply thankful to my big family for all what they have done to me to get over the recent hard time. Last but not least I would like to acknowledge and thank my friends in SMSC for helping me with ideas and suggestions for my research and who generously supported me.

Chapter 1

Introduction

1.1 Introduction

The word termed missing or incomplete data are indications that certain types of information are missing concerning a particular interest in the analysis, this shows that not all the needed information were recorded. This occurs in studies on individuals which is the case of longitudinal studies, but not limited to animals, plants or groups of individuals. In addition, the occurrence of missing data is very common in statistical and design analysis, but not of interest in any fields of research as frequently encountered by the statisticians. Missing data may be seen as a nuisance, create problems when it has to do with the analyses of incomplete data because most softwares and standard statistical techniques are designed to assume complete data for the variable(s) included in the analysis. The importance of the statistical analysis is to produce valid and efficient results about the population of interest. It is achievable with or without incomplete data, but the process may be complicated in the presence of missing data. Adequate planning should be given to the study in order to reduce or avoid missing data and make the study scientifically sound. However, irrespective of the efforts involved in the planning design and collection of data; missing data is unavoidable in data-based research (Allison [1], Carter [2], Regoeczi and Ridel [3], Rudas [4], Stumpf [5]). As detailed in (McKnight et al., [6]), the problems of incomplete data are

classified under three categories which affects either complete individuals or specific items, namely cases, variables and occasions. The first category happens when the individuals in the planned study fail to give data for a study. The second, missing variables, the missing data happen when individuals provide part information but not all variables. In the third category, missing occasions (i.e., follow-up data), this is when participants are available for some but not all of the data collection periods in a study. Other contributory factors to incomplete data apart from human and study design errors are: (1) when the questions raised poses no significant importance to such individual; (2) when the questionnaire is lengthy an individual may not attempt all the sections; and (3) illness or other uncomfortable conditions of the participants. In addition, missingness may be experienced regardless of the study field due to data entry errors or omission of data. This may occur when the interviewer fails to ask a particular question in a way that the respondent would have understood. This happens especially when the interviewer have inadequate training.

Explaining the effects of missing data can be in terms of the amounts, the patterns of missing data and also the techniques applied to handle it which also have implication on the interpretation of the statistical study of the analysis. Almost all the reasons of missingness can be identified with three obvious challenges. The first is associated with loss of information, efficiency or power. Second is handling of data problems, computation and analysis as a result of discrepancies in the data patterns and non-applicability of standard software. According to (Barnard and Meng, [7]), the third is marked biased when there are differences between the observed and unobserved data. Alternatively, it may mean that the data is insufficient to obtain any meaningful inference from the study. Furthermore, it is also difficult to specify the impact missing data might have displayed in the study analysis statistically. The extent of the impact of missing data on study inferences depend on:

- The amount of incomplete data: The impact is in relation to the study inferences. When there is greater amount of missing data; it has a great impact on the statistical results. In other words, the power of the statistical tests can be compromised (De

Leeuw et al., [8]).

- The mechanisms of missing data: The process that causes missing data may affect the validity of the statistical inferences. If the process depends on causal effects factors, the missing data can have dramatic impact on the validity of the results.

- The methods or techniques the statistician or data analyst will apply to handle these incomplete data (Musil et al., [9], Streiner [10]).

Missing data are an important issue to address in many disciplines of science including medical and epidemiological studies, psychometry, econometrics and surveys (Friedman et al., [11], Green et al., [12], Piantadosi [13]) and epidemiological studies (Kahn and Sempos [14], Clayton and Hills [15], Lilienfeld and Stolley [16], Selvin [17]). As detailed in studies conducted by (Schafer et al., [18], Rubin [19], Rubin et al., [20]) to mention but few. Apart from these, there are other examples in the context of observational and experimental data in non-human life settings such as environmental, agricultural and biological studies. The focus in the current thesis is on both longitudinal and cross-sectional data. There are several earlier studies on the problem of missing data largely concerned with algorithmic and computational solutions to the induced lack of balance or deviations from the intended study design (Molenberghs and Verbeke, [21]). Foremost research on missing data included the work by (Affifi and Elashoff, [22]) and (Hartley and Hocking, [23]). Thereafter, we have in literature other applications such as Expectation Maximization (EM) introduced by (Dempster et al., [24]), data imputation and augmentation (see, Rubin [19]; Tanner and Wong [25]) which has strong combination of computing resources of solving the difficulties associated with computation. These and other studies ushered in a revolution to the idea of handling missing data in statistical analysis.

1.1.1 Patterns of missing data

The pattern of missing data means the description and explanation of the totality of the dataset; where the values are observed and where values are missing respectively. The techniques to handle missing data are varied. There are those that are robust in handling any missing data pattern and some which are limited to specific missing data patterns. In addition, having identified the variables that define the pattern, a suitable analysis procedure may be proposed. In order to investigate the missing data pattern, it is of necessity to identify the cases and variable that contribute to the missing data (Schafer and Graham, [26], Allison, [27]). In a standard approach, let $Y = (y_{ij})$ be an $(n \times K)$ rectangular dataset containing incomplete data, with ith row $y_i = (y_{i1}, ..., y_{ik})$, where y_{ij} denotes a value of variable Y_j for individual i. Furthermore, when missing data are present, we define the missing data indicator matrix, R. In addition, R is defined as follows: R equal 1 if, y_{ij} is observed, and 0 otherwise. The missing data pattern may simply be defined by the matrix R whose (i, j)th element is R_{ij}. The matrix R is two-dimensional by individual and variable but for repeated measurements we have a third dimension denoting time. Thus typically the notation would be y_{ijt} the value if Y_j for individual i at time t.

1.1.2 Missing data mechanisms

Reasons while data are missing are unknown and also out of control of the analysts/statisticians, but the assumptions about the approach that generates the data and its implications for statistical inferences are necessary. In the analysis of incomplete data, missing data mechanisms which inform the methodologies to be employed are necessary to explain the dependencies between unobserved data and the missingness process. (De Leeuw [28], Little and Rubin, [29]). *Dropout mechanism* is used in relation to when participants drop out of a clinical trial study prematurely, especially when it is longitudinal studies. Many researchers misuse the term "dropout" in trials even when missingness occurs not because

a subject chooses to dropout, but because the protocol is not written to follow-up participants after treatment discontinuation. Some of the reasons that may warrant discontinuation can range from adverse effects to lack of efficiency, or both. As described in (Rubin [30], Little and Rubin, [31]), missing data mechanisms can be broadly classified into three categories. First, data is *Missing Completely at Random* (MCAR) when the mechanism that generates the missing observations is a truly random process unrelated to any measured or unmeasured features of the study participants. *Missing at Random* (MAR) is the second category, when the missingness mechanism is random meaning conditional on the observed measured characteristics, the missingness mechanism is independent of the unobserved measurements. The final category, *Missing Not at Random* (MNAR) is when the missingness process depends on unobserved measurements and possibly on the observed measurement features of the study trial. Each mechanisms refer to the probability of missingness, given information about the variable(s) with the missing data, associated variables and a hypothetical mechanism underlying the missing data. The classifications described by (Rubin, [30]) are in relation to the bias level that missingness may have on statistical analysis where it is stated that the potential impact of MCAR is negligible and MNAR potential is of greatest impact. Moreover, it is a bit difficult to differentiate which categories of missingness are in focus, unless one understands the rationale for participant's dropping out. This complexity was addressed in (Molenberghs et al., [32]) where it is reported that there is no formal-based distinction between MAR and MNAR. This is because for any MNAR model there exists a MAR counterpart that fits the data rightly, but their difference is in the prediction of what is unobserved. Furthermore, the role of the MNAR model is in sensitivity analysis which suggests that there are changes in assumptions, the conclusions from the primary (typically MAR) analysis are also changed. Accordingly (Molenberghs and Verbeke, [33] and Molenberghs and Kenward, [34]), defined sensitivity analysis as an analysis in which several statistical models are simultaneously considered under different missing data scenarios.

1.1.3 Ignorable approach

In addition to fully understanding the different missing data mechanisms broadly speaking there are two classes of incomplete data as described by (Rubin, [30]): these are *ignorable* and *non-ignorable* missing data. Under the assumption that missing data happen under either MCAR or MAR mechanism, the problem is termed ignorable, and the missingness process need no explicit modeling. When data are ignorable, the likelihood-based approach and Bayesian frameworks allow to ignore the missingness process since they use only observed data conditional on the model being correctly specified (Little and Rubin, [31]) On the other hand, if the data are MNAR; the missingness process needs to be modeled (Little and Rubin, [35]). When applying incomplete data classifications, it is important to point out that ignorability applies to the missingness mechanism and should be construed to mean that the analyst should ignore the missing values. This implies that the factors that cause missingness are unrelated or weakly related to the estimated intervention effect. In a restricted sense, the term refers to whether missingness mechanisms must be modeled as part of the parameter estimation process or not (Allison, [27]). Moreover, ignorability becomes useful when the need arises to evaluate the impact of incomplete data in the analysis and its inferences. Because the way they are missing is random, MCAR data would not have had systematic effect between the complete and incomplete observations on results. However, a systematic process underlying the missingness is experienced in the case of MAR data, but this effect can be modeled using the observed data (McKnight et al., [6]). For explanation, assume that x is the auxiliary variable observed for the sample, y is the study variable subject to missingness and r is the response missingness of the status variable, and that interest is to find the best prediction model for y in terms of x. In predicting the incomplete data a prediction model can be used especially when the response missing data mechanism is ignorable, which means the relationship between y and x in the respondents hold for the non-responding part of the sample. Intuitively, an incomplete data is ignorable when the study variable y is independent of the value of the status variable r

given the auxiliary variable x. But when the missing mechanism is non-ignorable it means the probability of y being missing depends on y itself, despite controlling of x thereafter. Thus, it should be noted that MNAR mechanism violates the principle of ignorability and requires adequate measures to account for the effects of data that is MNAR which is known as non-ignorable case. Furthermore, because the effect of non-ignorable mechanism is unknown this implies inadequate information from the dataset used in the analysis to allow the statistician to model and study the approach in which the data are missing. In fact, handling of non-ignorable missingness must be with caution because it is a bit difficult than ignorable missing data. As detailed in Thijs et al., [36], making a satisfactory analysis of data under the non-ignorable is impossible. In research, many authors have studied the role of non-ignorable missingness in different applications, see (Belin et al., [37], Wachtar [38], Little and Rubin, [35], Demirtas and Schafer [39]). As described in (Verbeke and Molenberghs, [40]), to investigate the impact of deviations from the ignorable missingness on the inferences, there is need to apply sensitivity analysis in which models for the non-ignorable mechanism process play an important role. This thesis deals with models under the ignorability mechanism assumption.

1.1.4 Methods of handling missing data

Having to look at the challenges that can come up when there are incomplete data, the following question is forced upon analysts. What techniques can be used to solve these potential problems? The goal is to use techniques that generate unbiased inferences.
Several techniques have been raised in handling the incomplete data. A holistic review of some of these techniques is given by (Schafer and Graham, [26]) which range from *ad hoc* methods to the *model-based* methods which include multiple imputation and augmentation techniques. Thus, an important discussion regarding the technical details of handling incomplete data is provided by (Schafer and Olsen, [41]). In addition, (Rubin, ([19], Van der Laan and Robins, [42], Molenberghs and Verbeke, [21], Tsiatis, [43] and Molenberghs

and Kenward, [34]) have all conducted thorough studies on methods that can be used to handle incomplete data. Some techniques adopt different approaches to addressing missing covariates and outcomes. Although the incomplete data problem is ubiquitous, there is still no firm decision on what statistical processes should be applied for the analysis or the circumstances under which they should be used. In literature, we have several methods that can be applied to handle incomplete data, as stated before ranging from simple *ad hoc* to model-based methods. There is need to understand these methods and appropriately applied in relation to missing data and should be proved theoretically before use in practical terms. Moreover, the use of each method depends on the specific missingness mechanism, but it should be noted that it is difficult to identify the missingness data mechanism. Before 1970's, incomplete data were handled through editing (Schafer and Graham, [26]). Until the mid 1970's, the major techniques that were applied to deal with missing data included case deletion, single imputation and maximum likelihood estimation. Right from the mid 70's to the 1980's, other techniques like expectation maximization (EM) algorithm and multiple imputation using maximum likelihood estimation were developed to solve wide range of problems. During this time, a framework of inference from incomplete data was developed by (Rubin, [30]). Thus, the expectation maximization (EM) procedure was propounded by (Dempster et al., [24]) as a famous instrument for making full use of maximum likelihood (ML) for handling the missing data analysis. A breakthrough initiative of multiple imputation was formulated by (Rubin, [44] and Little and Rubin, [19]), but many challenges were identified in relation to creation of multiple imputations in terms of computational facilities and capabilities available (Schafer and Olsen, [41]). In 1980's, a lot of developmental resources were available in solving this difficulty such as improved computer technology and new methods for Bayesian simulation (Schafer, [45]). The setbacks associated with single imputation and case deletion were documented by (Little and Rubin, [19]). Solutions to problems present in using maximum likelihood estimation were advanced in the 1990's. At that time the EM algorithm was extended to different forms, such stochastic EM algorithm (SEM), stochastic expectation conditional

maximization algorithm (SECM). Also at the same time, Bayes simulation methods such as Markov Chain Monte Carlo (MCMC) and data augmentation were developed. In recent times, analysts concentrate on more modern methods that avoid the specification of a full parametric models (Robins et al., [46]). Right from 1995, several approaches have been discussed and developed for dealing with incomplete data with diverse applications. Recently, other methods have been proposed to assess the sensitivity of the inferences to the distribution of missing data mechanisms (Verbeke and Molenberghs, [47]). In view of non-ignorability mechanism setting, the main interest has been dropout in longitudinal clinical trials data where subjects may drop out of the study for reasons closely related to the responses being measured (Diggle and Kenward, [48], Little, [49]; Verbeke and Molenberghs, [47]; Molenberghs and Verbeke, [33]; Molenberghs and Kenward, [34]). Next we briefly discuss method commonly used to deal with missing data and review the existing literature in which the effectiveness of these techniques are examined in the analysis of missing data.

1.1.5 Deletion methods

It is obvious that there are many ways to deal with missing data. Without loss of generality we refer to the case of longitudinal data. One of these ways is to discard individuals with incomplete sequences, and then only the complete data are analyzed. (Nie et al., [50]).This technique is what is termed as the *deletion methods*. Under this technique, there is no replacement or imputation of missing observations and adjustments to account for the incomplete values. The method however suffers from some of the obvious inefficiencies such as losing data for statistical power, although of different degrees and failure to give the missing data mechanism proper attention. The advantage of the method is that it is simple and easy to use with many of the standard statistical softwares. Thus, (Brown [51]) states that some the of deletion methods are good options, but only when used under specific circumstances (i.e., when the amount of missing data is small and the data are MCAR, for

example, the complete case discussed below and the available case which uses all available case and discards data only at the level of the variable, not at observation level). In other words, these circumstances are not common. (McKnight et al., [6]) suggest that deletion methods should be avoided when possible. In addition, no recommendation for the use of any deletion methods except in rare cases where the amount of missing data is small (Little and Rubin, [35]). Hence, we briefly describe the complete case as a deletion method, giving explanation of its usefulness, strengths and weaknesses.

Complete case analysis

Complete case or list-wise deletion analysis is a simple deletion procedure in which the analysis uses only those individuals with full recorded information. Concerning the variables in context, complete case relies on the available observations. In fact, this method has several advantages. To start with, it is simple to use because the method is effective and may be satisfactory when the amount of missing data is limited. However, in such situation it becomes pertinent that the deleted cases are not unduly influential (Schafer and Graham, [26]). Another advantage is that its conduct is easy. In most statistical softwares, it is used as default but details of implementations vary.

The disadvantages of this method are as follows: (1) it is capable of producing biased estimates because of the loss of statistical power especially when drawing conclusions for sub-populations; and (2) when data are not MCAR, technique can lead to serious biased inferences. However, the validity of the method relies on MCAR data (Little and Rubin, [35]), but even when MCAR holds, it can still be inefficient (Schafer and Graham, [26]). Thus, (McKnight et al., [6]) state that one should give careful consideration before the use of this method regardless of its ease of use. In addition, one can easily conclude that complete case can be misleading. Thus, examples were presented by (Kenward et al., [52]) and (Wang-Clow et al., [53]) where complete case led to misleading inferences.

1.1.6 Imputation-based techniques

In contrast to methods highlighted above, the focus is on the techniques that produce possible values for the incomplete data. These options fall under *imputation* methods, where one "fills-in" (imputes) the missing data to obtain a full dataset, which is then analyzed by standard statistical techniques without concern as if the set represented the true and complete dataset (Rubin ,[19], (Little and Rubin, [35]). This gives the main idea behind the frequently used approaches for imputation which includes: *single* and *multiple imputation* (Little and Rubin, [35]). For the purpose of the appropriate evaluation of imputation uncertainty; multiple imputation fills in more than one value for each missing observation (Rubin, [19], Little and Rubin, [35]). The difference between multiple and single imputation is that; single imputation procedure substitutes only one value for every missing observation in the dataset (Little and Rubin [35], [29]). In this section, we outline a number of single imputation techniques whose validity lies under the assumption of ignorable missing data mechanism (Rubin, [19], Allison, [27], Schafer and Graham, [26]).

There are several single imputation methods which include: (1) mean imputation; where the missing items are replaced with the estimated mean of the dataset for that variable; (2) last observation carried forward (LOCF) is where the last observed value in longitudinal or repeated measures context is used to replace or impute the missing observation; (3) regression imputation, where the missing data are imputed using the prediction taken from a multiple regression analysis; (4) hot Deck imputation, in which the missing data can be replaced with the observed data taken from a matched data from the variables that contain non-missing data; and (5) stochastic regression imputation, where the missing observations are replaced by a value that is predicted using regression imputation plus a residual that is drawn to reflect uncertainty in the predicted value. The procedure of single imputation techniques are flexible in dealing with incomplete data and its implementation is fast in several statistical software packages (such as SAS, R, SPSS, Winbugs and others). On the other hand, to reproduce accurately known population inferences (parameter estimates

and standard errors), each of these methods have been found to be marred by a number of deficiencies (Schafer and Graham, [26]). The problems associated with these methods are as follows: (1) these techniques perform poorly even when the ignorable missing data mechanism (MCAR or MAR) holds, a situation where the usability is limited to restricted set of assumptions (Allison, [27]); (2) capable of producing biased results that may or may not be predictable; (3) When these methods are used, the standard errors and standard deviations are capable of being underestimated, and the possibility committing type-I error is higher (See, Allison, [27]).The variability of the estimators is also underestimated since imputed data are treated as observed data; and (4) these methods have high chances of producing inconsistent point estimates even when data are MCAR.

1.1.7 Data augmentation methods

In data augmentation methods the shortcoming associated with the deletions techniques are avoided. These methods obtain parameter estimates from the available data as well as from either the probability model or an underlying distribution. As opposed to the single imputation techniques, no replacement of missing observations in data augmentation methods is done. During parameter estimation, missing values are taking into account in this algorithm as follows, observed data and the relationships between observed data and several underlying ensure that parameter estimates from the observed data are augmented with the additional information provided by the proposed probability model or underlying distribution. In view of incomplete data, Maximum Likelihood (ML), Expectation Maximization (EM), Markov Chain Monte Carlo (MCMC) and weighting methods are considered to be augmentation methods. In contrast to this, (McKnight et al., [6]) argued that the classification of many of the augmentation methods are not clear-cut especially the MCMC, ML and EM methods. In addition, (Allison, [27]) describes the MCMC method as an augmentation method within the use of multiple imputation. According to (Little and Rubin, [35]), the ML and EM techniques are described as model-based techniques, while (Schafer, [45])

also describes these approaches as data augmentation methods. Thus, we briefly discuss a few of these techniques which include ML, EM and weighting methods.

Maximum likelihood (ML)

The use of or by definition the ML is not designed to deal with missing data as for example, the LOCF or multiple imputation. ML is an estimation algorithm that estimates parameters under different models or distributional assumptions such as structural equation models (SEM) and ordinary least squares (OLS) regression. Generalized linear models (GLMs) assuming exponential family of distributions also fall under ML estimation. Thus we give a brief explanation on the ML as a technique for handling incomplete data. First, (Little and Rubin, [29]) are among the early researchers to give examples where ML is applied to missing data problems. In addition, in many diverse ways ML is seen to be an excellent procedure for dealing missing data. The only caveat is the likelihood formulation should be capable of accommodating both complete and incomplete data sequences or structures. ML performs well and gives unbiased estimates especially when missing data are ignorable (MCAR or MAR) (Arbuckle, [54]). Therefore, the description of ML under this assumption is fairly easy. When this assumption is met, the ML estimators for incomplete data give estimates that have the following properties: unbiased estimates even when the samples are large, asymptotically efficient (small standard errors) and asymptotic normality is satisfied which mean the estimates approximate a normal distribution which may be applied to exploit a normal approximation for statistical results, such as obtaining confidence interval and p-values (Mcknight et al., [6]). The implementation of ML in most statistical software packages is possible.

Expectation maximization (EM)

The EM algorithm was developed by (Dempster et al., [24]). This process calculates and imputes a value for each missing observation based on the best prediction model. The

EM algorithm is a very general iterative algorithm for ML estimation in incomplete data problems (Little and Rubin, [31]). This algorithm also falls under the less restrictive MAR assumption. The rationale behind the use of ML is to deal with missing variable issue in a more inclusive manner, and the complications associated with ML estimation when trying to solve smaller complete data problems. The EM algorithm deals with missing items using the following procedures: (1) fill-in the observations for missing variables by using the estimated values generated by ML; (2) estimate parameters based on data in step 1; (3) re-estimate parameters based on the parameter estimates from step 2; and (4) re-estimate parameters based on the re-estimated data from step 3, and so on, continue the iterative process until the last procedure converges on a result that negligibly differs from the previous result. There are only two steps on each iteration of the EM algorithm, namely: the expectation and maximization steps (Little and Rubin, [31]). Each step is completed in one iteration within each algorithm cycle; and this process is repeated until a convergence criterion is satisfied. These steps were justified in theory; as shown by (Dempster et al., [24] and Little and Rubin, [31]). In accordance to (Dempster et al., [24]), the inferences obtain from the fitted parameters (on convergence) are the same with the local maximum of a likelihood function which is the maximum likelihood estimate in the case of a unique maximum. The disadvantages of the EM algorithm are two: first, it takes time to converge, and second, it does not provide direct measure of precision for the maximum likelihood estimates. Many researchers have proposed alternatives to overcome these setbacks, and we make reference to the methods as provided by (Louis, [55], McLachlan and Krishnan, [56], Rubin, [57] and Baker, [58]).

Weighting methods

Weighting methods are based on the observed values as developed by (Flanders and Greenland, [59], and Zhao and Lipsitz, [60]). In view of this, when all the missing items in the analysis are ignored, the remaining observed values are weighted with respect to how their distribution approximates the full sample or population. The use of these methods

help correct for either the standard errors present in the parameters or the population variable. In order to obtain an appropriate weights, the predicted probability of each response is estimated from the data for the variable with incomplete values. Furthermore, weighting methods stand as a better alternative under certain circumstances, especially when the pattern of missing data is monotone or is under univariate analysis.

In the view of survey data, (Rubin, [19]) describes many techniques for applying and estimating weights. Under an appropriate joint model of outcome and covariates, weighting methods are in some situations expected to have inferences that are the same with those of multiple imputation (Schafer and Graham, [26]). In biostatistics study, a weighting regression model was developed by (Rubin et al., [20]) which needs an explicit model for missingness but relaxes parts of the parametric assumptions in the data model. The extension of the generalized estimating equations (GEE) developed by (Liang and Zeger, [61]) was the new weighting technique. This new technique is termed weighted generalized estimating equations (WGEE). The validity of the classical GEE method is only when data are MCAR, but the WGEE method can accommodate missing data if they are MAR, as long as the model for the missing data with respect to the observed outcomes or covariates is correctly specified (Rubin et al., [20]). In addition, (Rubin et al., [20]) discusses and extends this method and also (Robins et al., [62]). We give elaborate detail of this in Chapter 1. In fact, several statistical software packages can perform weighting methods. A wide range of researchers have conducted studies on the weighting methods, such as (Schluchter and Jackson, [63], Ibrahim [64], (Lipsitz and Ibrahim, [65] [66], (Horton and Laird [67], Carpenter et al., [68] and (Seaman and White, [69]).

1.1.8 Non-ignorability models in longitudinal data

All of the methods fail to give an optimal solution to the challenge of non-ignorable missing mechanism. This type of missingness creates a challenge especially in the area of longitudinal data setting. In longitudinal studies, there is correlation in the observations

that are repeatedly measured over time and part may be lost due to dropout from the trial. Missingness happens due to dropout (premature withdrawal) or attrition, which indicates a unique situation coming up from the longitudinal studies where subjects do not complete the trial for whatever reasons. Dropout is a special case of monotone missing data pattern (Diggle and Kenward, [48], Little, [49], Molenberghs et al., [70], Michiels et al., [71]). In literature, many applications indicate that it is pertinent to accommodate dropouts in the modeling process especially (Diggle and Kenward, [48], Little [72], [73], [49], Verbeke and Molenberghs [47] and Molenberghs et al., [74]). On the other hand, it is suggested that modeling of measurement process jointly with dropout may be considered more appropriate. However, in the area of non-ignorable dropout, a totally satisfactory statistical analysis of the data used in the study is not feasible, and one needs to be careful when handling the non-ignorable case.

In the situation of non-ignorable missingness, advanced modeling strategies have been developed to model the joint distribution of the dropout indicators pattern and the measurements process (observed and missing measurements included). Summaries as provided by (Little, [49], Verbeke and Molenberghs, [47] and Molenberghs and Kenward, [34]), the possibility of modeling the joint distribution of the measurements and dropout indicators rely on at least three factorizations. First, there is *outcome-dependence factorization*, where the dropout indicators are conditioned on the measurements. In the case of continuous longitudinal data, the adoption of the technique was given by (Diggle and Kenward, [48]). The second is *pattern-dependence factorization*, where the distribution of the measurements is a mixture of the distribution for individuals of distinct sub-groups determined by their dropout patterns. Lastly, is the *parameter-dependence factorization* which is conditional on the group of parameters shared by the two components so that the measurements process and dropout indicators are conditional independent. In relation and based on the above-stated factorizations, there are three types of modeling strategies: selection models, pattern-mixture models and shared pattern models. In accordance to (Vach and Blettber, [75], Molenberghs et al., [76], and Verbeke et al., [77]), the limitation to any of these model

factorizations is that they are sensitive to the assumptions made on the measurements model and the dropout mechanisms. Thus, the different analysis models may have a unique effect on the inferences obtained from the same study.

1.1.9 Research objectives

The aim of this thesis is to investigate the different methods of handling longitudinal and cross sectional missing data, with major focus on the non-Gaussian setting. In clinical trials, categorical (binary and ordinal) and counts data with missing covariates, responses or both are very frequent but the methods of dealing such data are less standard, due to non-availability fo a simple analogue to the normal distribution. However, we begin with the non-Gaussian case. The specific objectives are:

1. To compare three different enhancements of the generalized estimating equations, namely the weighted generalized estimating equations, multiple imputation based generalized estimating equations and doubly robust based generalized estimating equations in dealing with incomplete binary responses subject to MAR dropouts.

2. To investigate the difference between the multiple imputation and inverse probability weighting methods when applied on incomplete outcomes subject to MAR dropouts.

3. To investigate various multiple imputation strategies to handle ordinal responses with monotone dropout patterns.

4. To demonstrate and contrast the application of two families of MNAR-based models, namely selection and pattern mixture models for investigating the potential influence that dropout might exert on the dependent measurement of the considered data as modeling frameworks that could be used for sensitivity analysis.

1.1.10 Thesis outline

This thesis has a collection of 3 research papers which have been submitted for publication in international accredited journals. Out of these papers, one has been published and one has been accepted for publication while the remain are still in review. The papers start from Chapters 2 to 5 in which each chapter presented as a stand alone and not as a continuation of the existing one. On the other hand, the general concepts in these chapters are somehow interconnected as a way to achieve the total goal of the thesis. In Chapter 1, the introduction and general ideas to the thesis is provided. The remaining parts of the thesis is outlined as follows. As earlier stated, the objective was to handle discrete data, therefore the thesis began with the case of binary outcome. In Chapter 2, the thesis presents a comparative study of three different enhancements of the generalized estimating equations methods in handling incomplete longitudinal binary outcome. A simulation study was conducted to study the performance of these techniques. In Chapter 3, the thesis compared two methods of handling incomplete data. These are multiple imputation and inverse probability weighting to handle a survey data with non-response. A simulation study is conducted to study the efficiency of the methods. Chapter 4 deals with multiple imputation and likelihood-based approach in handling ordinal longitudinal responses with monotone dropouts. The multiple imputation and likelihood-based strategies involved are multivariate normal imputation, fully conditional specification and expectation maximization which are compared in both simulation studies and real data application. In the application, the dataset is on patients with HIV-Lung health concerns (HIV-lung data). In chapter 5, the focus is on the conclusion and areas of further research.

Chapter 2

A comparison of three different enhancements of the generalized estimating equations method in handling incomplete longitudinal binary outcome

Abstract

This paper compares the performance of three techniques of analyzing incomplete longitudinal binary outcome data when the missingness is due to dropout. It is assumed the response data are missing at random. We consider three modifications of the generalized estimating equations (GEE) based on inverse probability weighting (IPW) and multiple imputation (MI). In the weighted GEE (WGEE), observations are weighted by the inverse of the probability of being observed. The multiple imputation (MI) combined with GEE analysis is commonly known as MI-GEE. In this approach, the missing observations are filled multiple times with the predicted values from the imputation model followed by a GEE analysis. The so-called doubly-robust (DR) technique combines the multiply imputed binary responses with IPW and then applying GEE to the completed data sets. A simulation study is first used to compare the performance of the methods followed by an application to a clinical trial data on Amenorrhea. The simulation and empirical example results revealed better performance for DR-GEE compared to WGEE and MI-GEE, but MI-GEE

was evidently superior than WGEE and quite close to DR-GEE.

2.1 Introduction

In a longitudinal study, each individual or unit is measured at several time points which provides the opportunity to study changes over time for the variable(s) and effects of interest. In several areas of biomedical research the response variable is binary or in general non-Gaussian, for instance, the presence or absence of an ailment in an individual included in a clinical trial to compare two or more treatments or interventions. The most widely used approach for handling binary longitudinal responses is the generalized linear mixed model (GLMM) (Fitzmaurice et al., [78] and Molenberghs and Verbeke, [33]). In the presence of missing data this model imply complex and hard to manipulate likelihoods for moderate and large sequences of repeated measurements. Generalized estimating equations (GEE) are on alternative modeling technique (Liang and Zeger, [61]), but needs some enhancements to deal with incomplete longitudinal data where missing completely at random (MCAR) is ruled out. The essence of this technique is that, it allows confining attention to the mean structure given that there is a willingness to adopt a working assumption about the association structure for the repeated measurements.

Thus, the main difference is that the GLMM is a conditional or random effects likelihood based model while the former namely the GEE is a marginal model. The strength of the GLMM is that it is likelihood based and hence in the presence of missing data the less stricter missing at random (MAR) assumption can be used instead of the restrictive MCAR assumption.

In the presence of incomplete data, GEE suffers from its frequentist nature where its validity is restricted to the MCAR assumption, meaning the missingness is independent of both unobserved and observed responses (Rubin, [30]). To overcome this defficiency, another member of the GEE technique was introduced by (Robins et al., [79]). This is the WGEE because it allows for the weaker MAR assumption, where the missingness is independent

of the unobserved data given the observed data (Bang and Robins [80], Rubin [30]). Under WGEE, the technique uses the inverse of the individual's probability of being observed as a weight to the estimating equations in order to reduce bias in the regression parameter estimates. In addition, weights can be at the subject level or observation level. In our study, we adopt the weight at subject level.

The GEE method is one of the most common techniques for the analysis of non-Gaussian correlated data. It is advantageous when one specifies the mean structure correctly for the parameter estimates to be consistent and asymptotically normal. In deriving the method, the association parameter(s) among the repeated measures are taken as nuisance parameters. The GEE technique is attractive because it helps avoid dealing with complex and sometimes intractable likelihoods and it naturally gives rise to population-averaged parameters that are of interests in most studies. When longitudinal data is incomplete, the non-response can occur at any time from the beginning of the study. There are three possible patterns of missing data that can be observed for the response: first, dropout is when an individual leaves the study prematurely for reasons known or not known to the investigator and does not return. This generally falls under the monotone pattern of non-response. Second, intermittent non-response occurs when an individual leaves and returns to the study after some period of non-response, and possibly a repeat of the same once, twice or more times and lastly a pattern that may be monotone at earlier times and a dropout later. Missing data is also possible in the covariates, but our focus in this paper is on missing data in the outcome variable. In the presence of incomplete data, three challenges are of concern. First, between the observed and missing data there can be potential serious bias due to systematic differences. Secondly, handling incomplete data and statistical inference can be complex and lastly, the loss of efficiency can be substantial. Furthermore, in order to have inferences that are valid; it is pertinent to have a good knowledge of the missing data mechanism that could have generated the non-response and properly accounting for it in the analysis.

Doubly robust estimators (DR-GEE) are seen as an appealing modification, or extension

of the ordinary GEE to handle data that are subject to MAR mechanism. The doubly robust (DR) estimating equations method has been developed as an extension of the WGEE method, where the idea is to integrate the weights with the use of a predictive imputation model for the missing data given the observed data. In effect, the DR estimation method produces parameter estimates that are consistently correct given correct specification of either the weights or the predictive imputation model, but not necessarily both.

DR techniques have widely received attention in the literature in the last decade (Bang and Robins [80], Carpenter et al., [81], Jolani et al., [82], Jose et al., [83], Tsiatis and Davidian [84]). More importantly, the inclusion of the inverse of the propensity score into the imputation model gives an increasing robustness to the imputations against misspecification of the imputation model. The uniqueness of this technique is that it gives analysts two routes to validate inferences, instead of only one. However, the method lacks generalization to intermittent missing observations, where the individuals return to the study after skipping one or more visits. Multiple imputation (MI) is one of the alternative approaches (Mehrotra et al., [85]) which relies on the MAR assumption. Under this approach missing values are imputed several times, and the resulting completed datasets are analyzed using a standard technique like GEE. Several authors like (Beunckens et al., [86]) have worked on the combination of MI and GEE, such that missing data are multiply imputed, and then inferences are obtained based on GEE. These inferences are combined into a single summary using Rubin's pooling rules and hence the method has become commonly known as MI-GEE. Moreover, the important requirement in this method are just like any other technique for imputation, namely the imputation model needs to be specified correctly. That the model should include all important covariates; including auxiliary ones to make it rich in informing the missing values predictive distribution. By its very nature MI-GEE does not suffer from the intermittent missing data problem.

A study conducted by (Satty et al., [87]), compared the two types of GEE (WGEE and MI-GEE) to the likelihood-based GLMM for analyzing longitudinal binary outcomes with dropout. The use of extended or enhanced GEEs to other categorical outcomes has and

is also gaining popularity. For example, authors in (Toledano and Gatsonis, [88]) used a WGEE method to accommodate arbitrary patterns of a missing responses and missingness in key covariates. A recent paper from Donneau et al., [89], compared through a simulation study two multiple imputation methods (multivariate normal imputation and ordinal imputation regression) for longitudinal ordinal data subject to dropout. In another paper, the same authors compared joint modeling and fully conditional specification approaches for non-monotone missingness patterns (Donneau et al., [89]). Single robust versions of GEE are used in the two papers mentioned above and they treated only missing covariates and response respectively. In a recent paper, Jose et al., [83] proposed a doubly robust approach for the analysis of longitudinal ordinal data with intermittently missing response and covariates under MAR.

In this paper, we re-visit the incomplete binary data problem. Our interest is thus on the combination of the DR and GEE for incomplete longitudinal binary data when the missing data pattern is monotone. This method involves multiply imputing binary responses using the DR approach and then applying GEE to the complete data sets. However, we also introduce a novel idea to find out whether using different working correlation structures, namely; compound symmetry (CS), first order autoregressive AR(1) and TOEP for estimation would affect the parameter estimates and standard errors under the specified methods. This paper is organized as follows. Section 2.2 defines the GEE notations and an overview of the GEE method. Section 2.3 outlines the WGEE, MI-GEE and DR-GEE approaches. A simulation study is presented in section 2.4. Data on Amenorrhea from clinical trial contracepting women is openly available online https://content.sph.harvard.edu is analyzed in section 2.4.4. The paper ends with a discussion and conclusion in section 2.5.

2.2 The generalized estimating equation (GEE)

In the situation where population-averaged effects are of interest, the most widely used model to analyzing discrete longitudinal data are the GEE which falls among the popular marginal models. The GEE technique was proposed by (Liang and Zeger, [61]), as an extension of the generalized linear model (GLM) to the case of correlated data in the context of longitudinal studies and correlated data in general.

Suppose $y_{ij}, j = 1,...,n_i,\ i = 1,...,N$, represents the jth response at time t_{ij} for the ith individual with a vector of covariates x_{ij}. Thus n_i are the measurements on individual i, and let n be the maximum number of measurements per individual, i.e. $n = max_i\{n_i\}$ if all the planned repeated measurements were obtained. This is the most general setting but if $t_{ij} = t_j$ and $n_i = n$ for all i, then we have the most balanced design with no missing values. We assume that the responses on the ith individual are held in the vector $Y_i = [y_{i1},...,y_{in_i}]'$ and the corresponding vector of means is $\mu_i = [\mu_{i1},...,\mu_{in_i}]'$. Under the generalized linear model formulation, the marginal mean μ_{ij} of the response y_{ij} is related to a linear predictor through a link function $g(\mu_{ij}) = x_{ij}^{'}\beta$ or as others may prefer $\mu_{ij} = h(x_{ij}^{'}\beta)$ where $h = g^{-1}$, and the variance of y_{ij} depends on the mean through a variance function $v(\mu_{ij})$, since $\text{var}(y_{ij}) = a(\phi)v(\mu_{ij})$ given $a(\phi)$ is the additional overdispersion parameter in the exponential family formulation. However, since we have repeated measurements the GLM has to be modified to account for the correlation of observations within an individual. This leads to a modified estimating equation for the model parameters. The generalized estimating equation, used to estimate the parameters of interest in the vector β in the marginal model is given by

$$S(\beta) = \sum_{i=1}^{N} \frac{\partial \mu_i'}{\partial \beta} \mathbf{V}_i^{-1}(Y_i - \mu_i(\beta)) = 0 \tag{2.1}$$

where $\mathbf{V}_i$ is the covariance matrix of Y_i which is specified through the working correlation matrix $R_i(\alpha)$ as

$$\mathbf{V}_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2} \tag{2.2}$$

Here, A_i is an $n_i \times n_i$ diagonal matrix whose jth diagonal element is $v(\mu_{ij})$, the variance function at μ_{ij}, from the assumed linear exponential family distribution. If $R_i(\alpha)$ is the true correlation matrix of Y_i, then V_i will be the true covariance matrix of Y_i and in this case the resulting standard errors of parameters are referred to as model based standard errors. Otherwise, it is suffices to use the empirical robust standard errors because in reality it is hard to discern the true covariance structure.

In the GEE estimation technique, the only requirements are the specification of the mean model and correlation structure of the vector Y_i; such that the specification of the full joint distribution of the correlated responses is not needed. The joint marginal distribution is complex to specify because often for non-Gaussian longitudinal responses, the joint distribution involves high-order associations and integration. However, the regression parameter estimates from the GEE are consistent even when the working covariance specification through $R_i(\alpha)$ is incorrect. When the marginal effects are of interest and the responses are not continuous, the GEE is a very common choice. Nevertheless, the GEE approach can lead to biased estimates when the underlying missingness mechanism is not MCAR. One of the methods that can produce unbiased estimates is the WGEE and is briefly described in the following section.

2.3 Methods for handling incomplete data

The weighted generalized estimating equations and multiple imputation are the two commonly used methods for missing data under the MAR mechanism. The missingness mechanism can be described via a statistical model for the probability of observing a missing value. A reasonable assumption about the mechanism is important for methods that are used to handle missing data. In general, missingness mechanisms are classified into three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Rubin, [30]). The three mechanisms are briefly discussed below in the

contest of longitudinal data. We confine attention to dropout as the missing data pattern. First for each potential outcome Y_{ij} define a binary indicator variable R_{ij} which takes the value $r_{ij} = 0$ if Y_{ij} is missing and $r_{ij} = 1$ if Y_{ij} is observed.

- A missingness mechanism is said to be MCAR if the probability of a missing response is independent of its past, current and future responses conditional on the covariates. That is $P(r_{ij} = 0|Y_i, X_i) = P(r_{ij} = 0|X_i)$.

- A missing mechanism is said to be MAR if the probability of a missing response is independent of its current and future responses conditional on the observed past responses and the covariates. That is
  $P(r_{ij} = 0|r_{ij-1} = 1, X_i, Y_i) = P(r_{ij} = 0|r_{ij-1} = 1, X_i, y_{i1}, ..., y_{ij-1})$
  MAR is a weaker assumption than MCAR. In fact, MCAR is a special case of MAR, thus an analyst is better of working with the superior MAR than the MCAR assumption.

- A missing mechanism is said to be MNAR if the probability of a missing response depends on the unobserved responses. MNAR is the most general and the most complex missing data mechanism to deal with. Thus, there is no further reduction to $P(r_{ij} = 0|r_{ij-1} = 1, X_i, Y_i)$.

Sections 3.1 to 3.3 present a brief discussion of the missing data methods considered in the current paper.

2.3.1 The weighted generalized estimating equation (WGEE)

Currently, there are two weighting methods that can be used to construct the WGEE for estimating the regression parameters β, when dropout is the missing data pattern. There are two possible weights; observation-specific weights and subject-specific weights versions as outlined in (Lin and Rodriguez, [90]). The weighting methods produce parameter

estimates that are consistent provided the data are MAR.

WGEE based on observation-specific weights

The weight ω_{ij} for y_{ij} is defined as the inverse probability of observing y_{ij}. In other words, $\omega_{ij} = P(r_{ij} = 1|X_i, h_{ij})^{-1}$ where $h_{ij} = (y_{i1}, ..., y_{ij-1})$ denotes the observed response history. Let W_i be a $n_i \times n_i$ diagonal matrix whose jth diagonal is $r_{ij}\omega_{ij}$. Then the weighted generalized estimating equation (Preisser et al., [91], Robins and Rotnitzky, [92]) is given by

$$S_{ow}(\beta) = \sum_{i=1}^{N} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} W_i (Y_i - \mu_i(\beta)) = 0 \tag{2.3}$$

Unlike the standard GEE, the weighted estimating equation is unbiased when observations are appropriately weighted, leading to consistent parameter estimates of β. Practically, the weights ω_{ij} are unknown and they have to be estimated using an appropriate model for r_{ij}, such as the logistic regression model under the MAR assumption. Specifically, suppose $\lambda_{ij} = P(r_{ij} = 1|r_{ij-1}, X_i, h_{ij})$ denote the probability of observing the response y_{ij} given previous observed responses, under the MAR assumption. Using the observed data, λ_{ij} can be predicted from the logistic regression model, logit $(\lambda_{ij}) = z_{ij}'\alpha$, where z_{ij} are predictors that usually include the covariates x_{ij}, the past responses and indicators for visit times and α is a vector of parameters. The dropout process implies that the estimated probability of observing y_{ij} can be expressed as a cumulative product of conditional probabilities given by

$\hat{P}(r_{ij} = 1|X_i, h_{ij}) = \lambda_{i1}(\hat{\alpha}) \times ... \times \lambda_{ij}(\hat{\alpha})$

With the estimated weights $\hat{\omega}_{ij} = \hat{P}(r_{ij} = 1|X_i, h_{ij})^{-1}$, we solve the estimating equation $S_{ow}(\beta)$, from which the regression parameters β are estimated. There is a similarity in the algorithm for solving the WGEE and standard GEE.

When the dropout process is MAR; the following algorithm fits marginal models by using

the observation-specific WGEE method. The steps are:

S1. Fit a logistic regression with data (r_{ij}, z_{ij}) to obtain an estimate of α and estimate the weights, $\hat{\omega}_{ij} = \hat{P}(r_{ij} = 1|X_i, h_{ij})^{-1} = [\lambda_{i1}(\hat{\alpha}) \times ... \times \lambda_{ij}(\hat{\alpha})]^{-1}$, where $\lambda_{ij}(\hat{\alpha})$ is the predicted probability obtained from the logistic regression.

S2. Compute an initial estimate of β by using an ordinary generalized linear model, assuming independence of the responses.

S3. Compute the working correlation matrix $\mathbf{R}$ based on the standardized residuals, the current estimate of β and the specified structure of $\mathbf{R}$.

S4. Compute the $n_i \times n_i$ estimated covariance matrix: $\mathbf{V}_i = \phi A_i^{1/2} \hat{\mathbf{R}}_i(\alpha) A_i^{1/2}$

S5. Update $\hat{\beta}$:

$$\hat{\beta}_{r+1} = \hat{\beta}_r + \left[\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right) \mathbf{V}_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right)'\right]^{-1} \left[\sum_{i=1}^{N} \frac{\partial \mu_i'}{\partial \beta} \mathbf{V}_i^{-1} W_i (Y_i - \mu_i)\right]$$

S6. Steps S3-S5 are repeated until convergence.

In SAS, to estimate the probabilities for dropout as well as to pass the weights (predicted probabilities) to be used for WGEE, the “dropout" and “dropwgt" macros introduced by (Molenberghs and Verbeke, [33]) are used for this purpose and the macros need no modification. The variables “dropout” and “previous” are constructed through the use of the dropout macro. The dropout outcome variable is discrete indicating an individual drops

out of the study before the end of follow-up, whereas, the previous variable refers to the outcome at previous occasions. Second, the dropwgt macro is used to pass the weights to the individual observations in the WGEE. Such weights are calculated as the inverse of the cumulative product of conditional probabilities, estimated as $\hat{\omega}_{ij} = 1/(\hat{\lambda}_{i1} \times ... \times \hat{\lambda}_{ij})$. This simply means the use of the predicted probabilities from the fitted missingness model to calculate the weights. Lastly, once the earlier two steps are executed appropriately, the last step is implemented by specifying the weights, by means of the weight statement in SAS procedure GENMOD. In specifying the working correlation matrix, there are a number of choices but in our case the first order autoregressive AR(1), TOEP and compound symmetry (CS) are chosen. However, there are two features to note about procedure GENMOD:

- This procedure is more appropriate for an independence working correlation matrix structure. In the GENMOD procedure, the weight statement procedure does not properly include weights when other correlation structures are used.

- The GENMOD procedure regards the weights as fixed. Consequently, the standard errors of the regression parameters from the two-step approach are conservative, which leads to narrower confidence intervals and conservative inference (Fitzmaurice et al., [78].

In SAS/STAT 9.4, the above deficiencies are better handled with the new GEE procedure which also provides appropriate standard errors. Furthermore, PROC GEE also handles a variety of working correlation structures. Thus in our case, we use PROC GEE in order to exploit this flexibility

WGEE based on subject-specific weights

The subject-specific weighted method is quite different from the observation-specific weighted method because it assigns a single weight for all observations within an individual. This means all the observations from an individual receive the same weight. Using this technique, one obtains the regression parameter estimates from the subject-specific weighted generalized estimating equation given by

$$\mathbf{S}_{sw}(\beta) = \sum_{i=1}^{N} \frac{\partial \mu_i'}{\partial \beta} \mathbf{V}_i^{-1} w_i (Y_i - \mu_i(\beta)) = 0 \tag{2.4}$$

where the weight w_i for individual i happens to be the inverse probability of an individual i dropping at the observed time (Fitzmaurice et al., [93], Preisser et al., [91]). Remember w_i is a scalar, as opposed to the weight matrix W_i in the observation-specific WGEE. The estimating equation from the subject-specific weighted method is unbiased after the observations have been weighted properly, and this produces consistent estimates for the regression parameters β.

However, the weight w_i needs to be estimated because they are unknown. Assume that m_i is a dropout indicator for the individual i, where $m_i = \sum_{j=1}^{n_i} r_{ij} + 1$. The first visit observation y_{i1} is assumed to be always observed with $r_{i1} = 1$. Thus, the values of m_i are 2,...,n_i. Note that $m_i = n+1$ indicates that individual i completes all the n visits, which were set aprior by design.

The definition of the weight w_i is as follows: if an individual i drops out before completing the last visit (i.e. $m_i < n+1$), then $w_i = P(r_{im_i} = 0, r_{im_i-1} = 1 | X_i, h_{ij})^{-1}$. Otherwise, the individual completes all the n visits (i.e. $m_i = n+1$), and $w_i = P(r_{in_i} = 1 | X_i, h_{ij})^{-1}$

As with observation-specific weights, the dropout process implies that subject-specific weights can be estimated as a cumulative product of conditional probabilities:

- $\hat{w}_i = P(r_{im_i} = 0, r_{im_i-1} = 1 | X_i, Y_i)^{-1} = [\lambda_{i1}(\hat{\alpha}) \times ... \times \lambda_{im_i-1} \times (1 - \lambda_{im_i-1}(\hat{\alpha}))]^{-1}$, if $m_i < n+1$

- $\hat{w}_i = P(r_{in_i} = 1|X_i, Y_i)^{-1} = [\lambda_{i1}(\hat{\alpha}) \times ... \times \lambda_{in}(\hat{\alpha}))]^{-1}$, if $m_i = n+1$

After the estimation of λ_{ij} by using the appropriate logistic regression for the dropout process, the subject-specific weights $\hat{w}_i$ can be obtained. There is a clear similarity in the algorithm that fits the marginal models for subject-specific WGEE technique and observation-specific WGEE technique, thus the fitting algorithm is not repeated here. Thus, the same SAS macro can be adapted to fit the subject-specific WGEE model.

2.3.2 Multiple imputation based GEE (MI-GEE) approach

Multiple imputation is a simulation-based approach for filling in the missing values multiple times which leads to multiple complete data sets. It is assumed that the model for the vector of repeated measurements Y_i is described by the parameter vector β. In the imputation stage, the objective is to impute the missing values with draws from the conditional distribution $f(y_i^m|y_i^0, \beta)$. However, β is not known hence an estimate for it denoted by $\hat{\beta}$, has to be obtained from the data, after which $f(y_i^m|y_i^0, \hat{\beta})$ is used to fill the missing observations. In the process, it means that we generate draws from the distribution of $\hat{\beta}$, which requires that we take into account the sampling uncertainty of estimating β. Another alternative is the Bayesian method, where the uncertainty about β is incorporated by means of using some prior distribution for β. However, after the formulation of the posterior distribution of β, the following imputation algorithm can be adopted: a random $\hat{\beta}$ is drawn first from the posterior distribution of β. The posterior distribution is approximated by the normal distribution. Then a random $\tilde{Y}_i^m$ is selected from $f(y_i^m|y_i^0, \hat{\beta})$. The so-imputed missing values are next augmented to the observed data, producing complete data, $Y_i = (Y_i^0, \tilde{Y_i^m})$. These are used to obtain $\widehat{\beta}$ and its variance, $V = \hat{Var}(\hat{\beta})$. The steps mentioned above are independent and repeated a number of times, say M times, to generate $\hat{\beta}^m$ and $\hat{V}^m$, for $m = 1, ..., M$. Moreover, the last step as stated above is when the results of the analysis

from the M completed (imputed) data are combined into a single inference. The overall estimated parameter for β and V are as follows:

$$\bar{\hat{\beta}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{m}, \tag{2.5}$$

and

$$V = W + \left(\frac{M+1}{M}\right)B \tag{2.6}$$

where

$$W = \sum_{m=1}^{M}\frac{\hat{V}^{m}}{M} \tag{2.7}$$

and

$$B = \sum_{m=1}^{M}\frac{(\hat{\beta}^{m}-\bar{\hat{\beta}})(\hat{\beta}^{m}-\bar{\hat{\beta}})'}{M-1} \tag{2.8}$$

The within-imputation variance and between-imputation variance are denoted by W and B respectively (Rubin and Little, [94]). With the description of MI above, this gives an insight to another method of handling missingness when the MAR-based MI is combined with a GEE analysis as the substantive analysis model. MAR-based MI hinges on the flexibility of the MI procedure hence the need to understand the idea on uncongenial imputation. Uncongeniality was introduced by (Meng, [95]) for an inconsistent imputation model in relation to the substantive analysis model. As stated by (Meng, [95]), one of the greatest strength of MI is that these two models (substantive and imputation) can be inconsistent in the sense that the two models need not be derived from the same overall model for the complete data. This method has become known as the MI-GEE technique, where M multiple data sets are subjected to the GEE analysis before the combination or pooling step. This serves as an alternative to likelihood and WGEE inference.

2.3.3 Doubly robust based GEE (DR-GEE)

The doubly robust DR method is an alternative approach that uses the inverse probability weights (IPW) to refine estimates of the model parameters (Bang and Robin, [80]), within a GEE analysis. In this technique, there is a requirement for the specification of two models: the first model is on the distribution of the complete data which include the outcome and covariates, and secondly a model for the missingness mechanism. The parameter estimates would be asymptotically unbiased when one of the models is correctly specified. On the other hand, the methods can be unstable in practice, especially when both models are misspecified (Tsiatis and Davidian, [84]), and can be disastrous when the propensity score (i.e. the probability of being observed) are close to zero (Tsiatis and Davidian, [84], Vansteelandt et al., [96]). In the current application, we combine IPW with MI and the GEE as the analysis model to construct DR-GEE. The robustness of the imputation model is enhanced by ensuring adequate information is included in the model, while avoidance of bias from the final inference is the target.

The main idea of the DR-GEE estimation; is to estimate the propensities for each incomplete variable conditional on the other variables, and impute the missing values on that variable by the inclusion of the propensity functions (i.e. IPW) into the imputation model. The results of the analysis from M completed (imputed) data are combined into a single inference with the GEE. The expectation of this method is to be readily robust, and by design it is aimed at handling incomplete data with any pattern of missingness.

Doubly robust estimation

As a caricature of the analyses, it helps to consider the estimation of a population mean outcome in the presence of incomplete data. This problem shows fundamental challenges involving inverse probability weighting. Consider a study design that aims to obtain independent and identical distributed data. Let $\{(Y_i, X_i), i = 1, ..., n\}$, with Y_i as the outcome

and X_i a set of auxiliary covariates for individual i. In the presence of missing data, the estimation of the mean $E(Y)$ is complicated by the fact that Y_i is not available for all individuals. Let R_i denote the missingness indicator, coded as $R_i = 1$ when Y_i is observed and $R_i = 0$ if Y_i is missing. The observed data can then be described as the random sample $\{Z_i(R_iY_i, R_i, X_i), i = 1, ..., n\}$ as illustrated in (Vermeulen and Vansteelandt, [97]). Assume that the covariates X_i contain sufficient information to explain missingness so that the missing at random assumption, $Y_i \perp R_i | X_i$ (Tsiatis, [43]), holds. Let μ represent its (unknown) population value; in particular, $E(Y) = \mu$. When the outcome data are missing, consistent estimation of μ requires specification of at least one of the two following working models as stated by (Robins et al., [46]). The probability of observing the data which is referred to as the propensity score (PS), is the first working model. This is taken as $P(R = 1|X) = \pi(X;\gamma)$, for which we assume $\pi(X;\gamma) > 0$ with probability one, where $\pi(X;\gamma)$ is a known function, smooth in γ which is an unknown $p-$dimensional parameter; for example, a logistic regression model $\pi(X;\gamma)$=expit$(\gamma_1 + \gamma_2^T X)$. This model is denoted as $M(\gamma) = \{\pi(X;\gamma) : \gamma \in \Re^P\}$. The second working model is for the conditional mean outcome $E(Y|X) = m(X;\beta)$, where $m(X;\beta)$ is a known function, smooth in β which is an unknown q-dimensional parameter; for example, a linear model $m(X;\beta) = \beta_1 + \beta_2^T X$ for a continuous outcome Y. This model is denoted as $M(\beta) = \{m(X;\beta) : \beta \in \Re^q\}$. As outlined in (Scharfstein et al., [98]), the DR estimator of μ, is $\tilde{E}_n(U) = n^{-1}\sum_{i=1}^n U_i$, can be obtained as

$$\hat{\mu}_{DR}(\hat{\gamma}, \hat{\beta}) = \tilde{E}_n \left[\frac{RY}{\pi(X, \hat{\gamma})} - \frac{R - \pi(X, \hat{\gamma})}{\pi(X, \hat{\gamma})} m(X, \hat{\beta}) \right] \tag{2.9}$$

for root$-n$ consistent and asymptotically normal estimators $\hat{\gamma}$ and $\hat{\beta}$ for the parameters γ and β (Tsiatis, [43]). This estimator is consistent for μ under the union model $M(\gamma) \cup M(\beta)$ as long as one but not necessarily both working models are correctly specified. If the intersection model $M(\gamma) \cap M(\beta)$ holds, that is, both working models are correctly specified, the DR estimator in equation (9), is locally efficient (Tsiatis, [43]) under model $M(\gamma)$. It then has the smallest asymptotic variance within the class of all estimators that are consistent and asymptotically normal under $M(\gamma)$, provided that also $M(\beta)$ is correctly specified,

with more explanation in (Vermeulen and Vansteelandt, [97]). Note that if $R_i = 0$ in equation (9) then the contribution in the summation is $m(X_i, \hat{\beta})$. If on the other hand $R_i = 1$ and $0 < \pi(X_i, \hat{\gamma}) = u_i < 1$ such that $u_i^{-1} = K$ then the contribution in the summation is $KY_i - (K-1)m(X_i, \hat{\beta})$. However, as a caution it is important that $\pi(X_i; \gamma)$ is bounded away from zero in the sense that $\pi(X_i; \gamma) \geq \delta_0 > 0$, otherwise one may be faced with undefined terms in the summation.

2.4 Simulation study

2.4.1 Data generation

We simulated data in order to mimic the non-Gaussian longitudinal clinical trial data. In the simulation, 1000 random samples of sizes $N = 100$, 250 and 500 individuals were drawn. The individuals were assumed to have been assigned to two treatment arms (Higher dose=1 and Mild dose=0). The measurements were taken at four time points $(j = 1, 2, 3, 4)$. Y_{ij} is the response variable measurement from individual i, at time j. The two levels of the response take the values 1 or 0 representing the event or non-event respectively. We modeled the event probability as a function of the explanatory variables. A marginal model for each binary response variable Y_{ij} is the focus and we assumed the logistic regression model. We thus generated the longitudinal binary outcome according to the following marginal model

$$logit\ P(y_{ij} = 1) = \beta_0 + \beta_1 dose_i + \beta_2 time_i + \beta_3 dose_i \times time_{ij} \tag{2.10}$$

where the model parameters are β, where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$. In the model, the fixed categorical effects are treatment $(dose)$, time (t) and the treatment-by-time interaction $(dose \times t)$. However, time was taken as a continuous variable. We fixed $\beta_0 = 0.50, \beta_1 = 1.00, \beta_2 = 0.70, \beta_3 = -1.25$. We used AR(1) as the working correlation matrix, with common correlation ρ=0.70. Dropouts were created on the complete simulated datasets using

different settings of missingness rate on the response. The dropouts were imposed on the response variable Y_{ij}. For the MAR mechanism to be achieved, after simulating a data set without missing data, we adopted the following strategy. We assume that dropout can occur after the first time point. Thus in this study, four dropout patterns are possible. That is, dropout at second, third, fourth time points or no dropout. According to Satty et al., [87], the data generated at time j and the subsequent times were assumed to be dependent on the values of outcome measured at time $j-1$. In our study, we retained the criterion that if the dependent variable (Y_{ij}) was positive $(i.e. Y_{ij}=1)$, then the individual dropped out at the next time point, is j+1. We generated dropouts of approximately 10%, 20% and 30% on each sample size respectively. We considered a monotone missing data pattern in our simulation where the only source of dropout was an individual's withdrawal.

2.4.2 Measures of performance of the techniques

The performance of the different techniques were assessed using two criteria criteria: the relative bias (RB) and root mean square error (RMSE). SAS/STAT 9.4 was used to perform the statistical analyses and to produce the results. In each case, the covariance structure used in the enhanced GEE is the compound symmetry to account for correlation in the data. The performance criteria used are briefly discussed below.

Relative bias

The relative bias (RB) is defined as the fractional difference between the averaged estimate and the true value. It is expressed as $RB = \frac{\bar{\hat{\beta}}-\beta}{\beta}$, where β is the true parameter value of interest. If number of simulations performed is represented as S then $\bar{\hat{\beta}} = \sum_{i=1}^{s} \frac{\hat{\beta}_i}{S}$. The estimate of interest within each of the $i = 1,...,S$ simulation is $\hat{\beta}_i$. In addition, $\bar{\hat{\beta}}$ is simply the estimate averaged over all simulations.

Root mean squared error

The mean squared error (MSE) is defined as the averaged squared difference between the parameter estimates and its corresponding true value. MSE is equal to the sum of the variance and squared bias of the parameter estimates. The RMSE is defined as the square root of MSE. This is calculated as
$RMSE = \left[(\bar{\hat{\beta}} - \beta)^2 + Var(\bar{\hat{\beta}})\right]^{1/2}$, where $Var(\bar{\hat{\beta}}) = \sum_{s=1}^{S} \frac{(\hat{\beta}_i - \bar{\hat{\beta}})^2}{(S-1)}$, where S is the number of replications. The importance of RMSE is that it measures the overall precision or accuracy, therefore it is used to evaluate the performance of estimation methods. In general, the more effective technique would have a smaller RMSE (Huang and Carriere, [99]).

2.4.3 The analysis

In this section, we discuss the results of the simulation study that compares the three techniques, namely the WGEE, MI-GEE and DR-GEE under different dropout settings. The imputation model for the MI-GEE and DR-GEE methods are specified accordingly while WGEE requires no imputation. The simulation study also considers the correct specified model for the imputation model for both the MI-GEE and DR-GEE. The measurement at first time point is assumed to be observed for each individual. The incomplete data set were multiply imputed and analyzed by MI-GEE and DR-GEE techniques respectively. We incorporate weights to analyze the WGEE. In the case of MI-GEE, dose and response status at other time points were included as covariates in the imputation model. The logistic regression was used to estimate the propensity scores for the DR-GEE approach, which in turn were used in the imputation model. We set the number of imputations to 50.

From Table 2.1, under the sample size of 100; it can be observed that the relative bias was smaller under the DR-GEE method showing better asymptotically unbiased parameter estimates, except for β_3 under 20% dropout setting when compare with WGEE. In addition, MI-GEE performs better than WGEE in terms of relative bias, except for (β_0 and β_2) under

30% dropout. It is also observed that the RMSE based on the DR-GEE was marginally smaller than the MI-GEE except for β_3 under 20%. The RMSE are slightly smaller in MI-GEE, except for (β_0 and β_2) under 30% dropout. However, the results obtain for the MI-GEE under the sample size of 250 performs closely to DR-GEE in terms of RB than WGEE, except for (β_2) and (β_3) under 10%, (β_0, β_2 and β_3) under 20% and β_3 under 30% dropout settings respectively. It is observed that the RMSE based on the MI-GEE is close to DR-GEE than WGEE,except for (β_2) and (β_3) under 10%, (β_0, β_2 and β_3) under 20% and β_3 under 30% dropout settings. In addition, the results obtain from the sample size of 500 clearly shows that the performance of DR-GEE and MI-GEE in terms of relative bias and RMSE are better than WGEE, except for β_1 and β_3 under 10%, β_0 and β_3 under 20% and 30% dropout respectively. But small and high sample sizes produce efficient results under the DR-GEE which is better in performance than WGEE. This points to the greater efficiency of the estimators of the DR-GEE method.

2.4.4 The application

The data used in this paper show the application of the three modifications to the GEE procedure when dealing with longitudinal data with missing observations. The data set used is from a longitudinal clinical trial study of women that used contraception during the four consecutive months Out of the 1,151 women available for the study each of them were randomly assigned to one of two treatments available: 100 mg or 150 mg of depot-medroxyprogesterone acetate (DPMA) representing the low and high dose of the drug respectively. The Amenorrhea status in each of the four months was measured as the response variable. The research question was on the effect of treatment on the rate of the Amenorrhea over time.

Let y_{ij} denote the Amenorrhea status of the *i*th woman at the *j*th visit, *j=1,...,4*, and suppose $\mu_{ij} = P(y_{ij} = 1|x_{ij})$ denote the probability of a positive Amenorrhea status at visit j to

Table 2.1: Simulation study: relative bias (RB) and root mean squared error (RMSE) values for the different parameters under the three models; WGEE, MI-GEE and DR-GEE under MAR mechanism over 1000 samples: N=100, 250 and 500 individuals, for monotone dropout.

Sample	Drp	Par	WGEE		MI-GEE		DR-GEE	
			RB	RMSE	RB	RMSE	RB	RMSE
100	10%	β_0	**0.1984**	**0.0909**	0.0722	0.0888	0.0722	0.0888
		β_1	**0.0844**	**0.2617**	0.0644	0.0699	0.0244	0.0365
		β_2	**0.0933**	**0.0849**	0.0864	0.0613	0.0866	0.0614
		β_3	**-0.4444**	**0.4253**	-0.2995	0.3746	-0.3114	0.3895
	20%	β_0	**0.1908**	**0.2046**	0.1610	0.1715	0.1542	0.981
		β_1	**0.2312**	**0.3427**	0.0135	0.0301	0.0250	0.0367
		β_2	**0.1247**	**0.1212**	0.0837	0.0593	0.0806	0.0572
		β_3	-0.1217	0.1959	-0.5338	0.6674	**-0.5363**	**0.6705**
	30%	β_0	0.4098	0.2827	**0.5682**	**0.2858**	0.5372	0.2705
		β_1	**0.2535**	**0.3019**	0.2441	0.2458	0.2443	0.3455
		β_2	0.1927	0.1652	**0.2917**	**0.2045**	0.2757	0.1934
		β_3	**-0.1294**	**0.2330**	-0.0673	0.0852	-0.0757	0.0956
250	10%	β_0	**0.6270**	**0.3663**	0.5576	0.2797	0.5532	0.2775
		β_1	**0.2044**	**0.3288**	0.1760	0.1782	0.1114	0.1148
		β_2	0.1886	0.1569	**0.1993**	**0.1398**	0.1973	0.1384
		β_3	-0.0885	0.1584	-0.2609	0.3263	**-0.2639**	**0.3300**
	20%	β_0	0.6878	0.3890	0.8144	0.4078	**0.8176**	**0.4094**
		β_1	**0.2382**	**0.3556**	0.1753	0.1774	0.1015	0.0111
		β_2	0.2441	0.1897	0.3196	0.2239	**0.3211**	**0.2250**
		β_3	-0.1394	0.2152	-0.5231	0.6440	**-0.5369**	**0.6712**
	30%	β_0	**0.8926**	**0.4851**	0.8876	0.4443	0.8590	0.4300
		β_1	**0.1953**	**0.3716**	0.1765	0.1787	0.1011	0.1049
		β_2	**0.3339**	**0.2500**	0.3150	0.2207	0.3010	0.2109
		β_3	-0.1555	0.2591	-0.7172	0.8966	**-0.7184**	**0.8981**
500	10%	β_0	**0.1674**	**0.2111**	0.0524	0.0347	0.0554	0.0359
		β_1	0.0212	0.2543	0.4239	0.4248	**0.4300**	**0.4309**
		β_2	**0.0767**	**0.1075**	0.0674	0.0484	0.0021	0.0107
		β_3	-0.1768	0.1191	**-0.2158**	**0.2576**	-0.2094	0.2521
	20%	β_0	0.2564	0.2271	0.3484	0.6963	**0.3704**	**0.1866**
		β_1	**0.1150**	**0.2863**	0.1009	0.1046	0.0178	0.0328
		β_2	**0.1854**	**0.1084**	0.1419	0.0998	0.1527	0.1073
		β_3	-0.1039	0.1837	-0.4410	0.5514	-0.4524	0.5656
	30%	β_0	**0.6442**	**0.3748**	0.4500	0.2201	0.4262	0.2143
		β_1	**0.0658**	**0.3085**	0.0545	0.0611	0.0325	0.0428
		β_2	**0.2201**	**0.1818**	0.1967	0.1380	0.1859	0.1305
		β_3	-0.0159	0.1692	**-0.6423**	**0.8030**	-0.6392	0.7991

individual i given covariates information x_{ij}. In order to determine the effect of treatment on the rate of Amenorrhea over time, we consider the following marginal model:
$\text{logit}\ (\mu_{ij}) = \beta_0 + \beta_1 time_{ij} + \beta_2 dose_i + \beta_3 time_{ij}^2 + \beta_4 dose_i \times time_{ij} + \beta_5 dose_i \times time_{ij}^2$

Of the 1,151 women in this study, 576 are from the low-dose group, and 575 are from the high-dose group. For the low-dose group, 62.67% of the women completed the trial; for the high-dose group, 61.39% of the women completed this trial. Thus, both groups have substantial dropouts.

We considered the following logistic regression model for the missingness mechanism to obtain the weights for the wGEE:

$$\begin{aligned} logit\ p(r_{ij} = 1 | r_{ij-1} = 1, dose_i, time_{ij}, y_{ij-1}) =& \alpha_0 + \alpha_1 I(time_{ij} = 2) + \alpha_2 I(time_{ij} = 3) \\ &+ \alpha_3 dose_i + \alpha_4 y_{ij-1} + \alpha_5 dose_i \times y_{ij-1} \end{aligned} \tag{2.11}$$

Equation (2.11), is the logistic regression for the missingness model where the second and third terms are the copy of time used as a class or factor variable. The fifth term is to relate the probability that a participant will dropout to previous Amenorrhea status. The last term relates the probability that a participant will dropout to the interaction of dose and previous Amenorrhea status.

The large fraction of missing observations pose a challenge in this trial study, where the pattern of missingness is a monotone dropout. Using standard GEE may produce biased estimates because it is near impossible to justify an MCAR assumption. Furthermore, complete case analysis would result to a heavy loss of data due to a large fraction of missing values. WGEE is possible with a monotone missingness pattern and difficult when the pattern of missingness is intermittent. Imputation strategy also gives consistent parameter estimates of interest.

The results from the three modifications of the GEE procedure are shown in Table 2.2. The first one is the weighted method (WGEE) using observation-specific weights model; the

Table 2.2: Parameter estimates (Est), standard errors (SE), p-value obtained from the Amenorrhea data under the methods of (WGEE), MI-GEE and DR-GEE under MAR mechanism using different working correlation structure.

		WGEE			MI-GEE			DR-GEE		
Cor str	Par	Est	SE	$Pr > \|t\|$	Est	SE	$Pr > \|t\|$	Est	SE	$Pr > \|t\|$
	β_0	-2.2057	0.1391	<.0001	-1.9573	0.2340	<.0001	-1.7523	0.2411	<.0001
	β_1	0.3672	0.1691	0.0298	0.4578	0.1825	0.0121	0.2496	0.1971	0.2054
CS	β_2	-0.4233	0.2068	0.0407	-0.3923	0.3340	0.2401	-0.2958	0.3417	0.3867
	β_3	0.0857	0.0500	0.0868	0.0097	0.0332	0.7712	0.0125	0.0363	0.7299
	β_4	0.5850	0.2536	0.0211	0.5726	0.2607	0.0280	0.5071	0.2793	0.0189
	β_5	-0.1530	0.0743	0.0395	-0.1095	0.0473	0.0207	-0.1042	0.0512	0.0418
	β_0	-2.2039	0.1392	<.0001	-1.9502	0.2347	<.0001	-1.7435	0.2417	<.0001
	β_1	0.3659	0.1689	0.0302	0.4555	0.1827	0.0127	0.2461	0.1975	0.2126
	β_2	-0.4156	0.2064	0.0440	-0.3340	0.3329	0.3158	-0.2760	0.3427	0.4207
AR(1)	β_3	0.0860	0.0501	0.0858	0.0097	0.0332	0.7746	0.0123	0.0363	0.7340
	β_4	0.5851	0.2527	0.0206	0.5596	0.2600	0.0314	0.5043	0.2795	0.0712
	β_5	-0.1547	0.0743	0.0374	-0.1078	0.0471	0.0220	-0.1036	0.0510	0.0422
	β_0	-2.2012	0.1418	<.0001	-1.9425	0.2362	<.0001	-1.7270	0.2439	<.0001
	β_1	0.3644	0.1687	0.0308	0.4532	0.1830	0.0133	0.2404	0.1979	0.2243
TOEP	β_2	-0.4004	0.2087	0.0551	-0.3279	0.3351	0.3278	-0.2790	0.3469	0.4212
	β_3	0.0861	0.0501	0.0855	0.0093	0.0331	0.7791	0.0121	0.0362	0.7378
	β_4	0.5809	0.2512	0.0207	0.5571	0.2605	0.0325	0.5051	0.2805	0.0718
	β_5	-0.1565	0.0743	0.0391	-0.1078	0.0469	0.0217	-0.1035	0.0509	0.0419

Notes: The missing value is on the response variable and are approximately 62.67% and 61.39% on low-dose and high-dose groups respectively.

second is the multiple imputation using multiple imputation in SAS/STAT 9.4 before GEE; and the third is the doubly robust technique using inverse probability weighting and imputation models, respectively. Three different working correlation structures were adopted. However, we briefly explain the result obtain using compound symmetry (CS) because the result is similar to other results. Thereafter, we compare the results of the three methods used. Furthermore, it is also noted that the p-value for β_3 i.e quadratic time effect is not significant under all the three techniques. All the techniques provided the same conclusion for the effect of dose (β_2). The negative effect of dose indicates that the rate of change of log odds probability of Amenorrhea over time is lower in the group receiving 150 mg compared to the reference group receiving 100 mg of depot-medroxyprogesterone acetate (DMPA). Then β_5 i.e. dose and quadratic time effect shows negative effect which indicates that the rate of change of log odds Amenorrhea over quadratic time depends on the dose, and is non-linear.

For purpose of comparison, under the WGEE TOEP produces the lowest parameter estimates except for β_3, β_4 and β_5 under the CS. In the case of MI-GEE, TOEP records the best values without any exceptions, but under the DR-GEE the situation is different as TOEP gives parameter estimates that are smaller than other methods, except for AR(1) where β_2 and β_4 produce the lowest values. Furthermore, the standard errors produced are small and closer to one another. This study has given an insight that the use of an appropriate correlation structure could produce better parameter estimates. Furthermore, different p-values were observed between WGEE and (MI-GEE and DR-GEE). Somehow, there is close similarity between WGEE, MI-GEE and DR-GEE, but the common feature between MI-GEE and DR-GEE is the imputation component and seems to give them an edge over WGEE despite all being valid under MAR.

These results show that the two techniques perform better than the WGEE. Combining these results and the relative performance for the simulation study suggests that both MI-GEE and DR-GEE which are imputation based are quite strong methods. The WGEE had

smaller standard errors but this did not change the overall inference and conclusion. The appealing feature in using the DR-GEE method is its doubly robust property. Quite different p-values especially between WGEE and (MI-GEE and DR-GEE). There is close similarity between (MI-GEE and DR-GEE) and WGEE. The common feature between MI-GEE and DR-GEE is the imputation component and seems to give them an edge over WGEE despite all being valid under MAR.

2.5 Discussion and conclusion

In this paper, the focus was on the performance of three different techniques for handling longitudinal binary data, under the MAR assumption with monotone dropout as the pattern of missingness. In addition, we prioritize the use of different working correlation structures to find out whether it would affect the parameter estimates and standard errors substantially. Therefore, we presented three stand alone enhancements to the generalized estimating equations for incomplete binary longitudinal data under MAR. Three methodologies were used namely multiple imputation, inverse probability weighting and its doubly robustness counterpart. The main focus was on DR-GEE technique for handling incomplete binary measurement because it combines both the weighting and imputation remedies to handle incompleteness. Furthermore, another attraction to this method is that it needs only the correct specification of at least one of the models, but not necessarily the two. However, in the simulation results when one of the missingness or outcome models is correct; the doubly robust estimators are consistent and present small-sample bias when compare with the single robust alternatives WGEE and MI-GEE. But DR-GEE has smallest standard errors than WGEE and MI-GEE especially when the sample size is small. In real application, the predictive model was not misspecified and this made the doubly estimators had a great potential of reducing the bias when the MAR assumption is correct.

In our study, we adopted different working correlation structures and observed differentials in the parameter estimates under the different methods used. We observed smaller

estimates under TOEP than we have under AR(1) and CS which is an indication that parameter estimates under TOEP consistent and better that AR(1) ans CS. On the comparison between WGEE and MI-GEE (Beunckens et al., [86] and Clayton et al., [100]) among others provided evidence of preference of MI-GEE over WGEE in longitudinal binary data. In addition, in the study conducted (Molenberghs and Verbeke, [33]) assumed independent working correlation space, but in our study different correlation structures were used which serve as an extensions.

Chapter 3

The use of fully conditional specification of multiple imputation and inverse probability weighting to model the pulmonary disease occurrence in survey data with non-response

Abstract

Incomplete data is a frequent occurrence in many research areas especially cross sectional survey data in epidemiology, health and social sciences research. In this paper, the effect of missing observations were accounted for by using multiple imputation (MI) and inverse probability weighting (IPW) methods. Generally, multiple imputation has the ability to draw multiple values from plausible predictive distrirbution for the missing values. However, under the inverse probability weieghting procedure the weights are the inverse of the predicted probabilities of response estimated from the missingness models of incomplete variables. A simulation study is conducted to compare methods and the use of the methods to mitigate bias induced by missing data in a cross-sectional survey data. The application and simulation results show the benefit of the IPW compared with the MI. The former performs well but not as the latter.

3.1 Introduction

Non-response in population surveys are becoming a great challenge especially to the health sector. However, the common and earliest technique to handle such problem is the use of complete case analysis which most statistical software applications have capabilities to do. This technique is called listwise deletion and many statisticians and clinicians adopt the approach simply because it is easy to use and readily available in almost all analysis software. This approach is valid when the missingness assumption is that of missing completely at random (MCAR) and the details of this method and others are detailed in (Rubin, [19]). However, justifying this approach in real applications may not be attainable and is almost impossible. In fact, there are many ad hoc methods that handle the missing observations especially where missing values are substituted plausibly, such as last observation carried forward (LOCF), the mean and regression predictions (single imputation) that can be resorted to rather than complete case. However, even these seemingly better methods than the CC method have their own shortcomings. One of the major difficulties of these methods is when there is high percentage of incomplete observations as explained by (Rubin, [19] and Sterne et al., [101]). Biased parameter estimates and loss of relationship among variable may be possible especially when the complete data does not give a true representative of the target population, and in such situation the MCAR assumption is violated. According to (Sterne et al., [101]), the single imputation method and other related methods may produce unrealistic small standard errors because uncertainty about the imputation values are not emphasized. The purpose why survey data have incomplete data are many, (Rubin, [19], Sterne et al., [101], Baraldi and Enders, [102], Kalton and Brick, [103], Rubin and Little, [94]). Incomplete information arise when an element in the planned population is mistakenly excluded on the survey's sampling frame, and this results to what is called non-coverage. Many researchers among them (Rubin, [19], Rubin and Little, [94], Lohr [104]) state that some elements may have zero probabilities of being included in the sample population. Thus, total/unit non-response is defined as when a sampled person fails to

take part in the survey. However, the occurrence of total non-response may occur when a participant refuses to be recorded or when an individual fails to actively participate in the survey due to some reasons ranging from the sensitivity of the questions, language of communication to non-availability on the day of the interview as stated by (Chinomona and Mwambi, [105]). For household surveys the availability of a person to be interviewed on the scheduled day of the interview is essential but other forms other than physical presence can be done such telephonic interviews, on line questionnaires among others. The failure of the selected individual to provide an accurate response(s) to one or many question(s) is called item non-response. Thus partial non-response is simply defined as when a non-response falls between unit and item non-response. This can occur as for example when a respondent cuts off the phone conversation in the middle of the interview or in a multiphase survey, the respondent provides data for some but not all phases of data collection (Rubin [19], Kalton and Brick, [103], Lohr [104]).

Before one starts to deal or handle missing data in a statistical analysis it is important to understand the different missing data mechanisms. Missing data mechanisms are broadly classified into three namely: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Different techniques have been developed that handle non-response in a survey data. But the form of remedial measures depends on the mechanism that generated the missing data. Many of these techniques that handle non-response range from the traditional methods like deletion, weighting adjustments to current novel methods that use multiple imputation combined with powerful analysis methods which may also integrate weighting at the observation or unit of measurement level such as the doubly robust GEE analysis. For total non-response and non-coverage; weight adjustments are mostly appropriate. Individuals with complete data receives greater weights, so as to compensate for inadequacies coming from non-respondents. In the situation of non-coverage, the use of weighting adjustments become imperative at is handles external data sources especially when there is no complete information from the sampled

individual. In item non-response, one of the appropriate ways to compensate for missing data is through multiple imputation. Incomplete observations are filled plausibly with imputed values. The two techniques just stated are used to compensate for partial non-response. In this research study, we embraced the use of multiple imputation and inverse probability weighting methods to handle the incomplete observations, for the purpose of obtaining estimates that are unbiased in modeling the prevalence of chronic obstructive pulmonary disease in three countries from Southern America, namely Argentina, Chile and Uruguay. We used socio-demographic risk factors and illnesses variables as covariates in the analysis.

Multiple imputation is a Monte Carlo technique that utilizes a Bayesian inference paradigm as indicated by (Rubin and Little, [94]). The complete datasets are then analyzed providing combined estimates of effects that are consistent with true values. Furthermore, unbiased parameter estimates and confidence interval are obtained from m complete datasets that incorporate missing data uncertainty. The parameter estimates and their variances that account for both the within-imputation and across-imputation variability; are derived from the mean of multiple imputed estimates obtained from the multiple analyses (Baraldi and Enders, [102], Heeringa et al., [106], Pigott [107], Schafer [45], Schafer and Olsen [41]). However, correct specification of the imputation model based on the MAR assumption is is important and necessary. In addition, to account for variability due to the missing observations multiple imputation accounts for this in the variance formula. One version of the Bayesian approach is that which employs the fully conditional specification (FCS) method to impute observations; from the posterior distribution of the missing data given the observed data as utilized by (Berglund, [108]).

Inverse probability weighting (IPW) is one of the methods that can also reduce the estimation bias. An additional advantage of inverse probability weighting is in its ability to correct for unequal sampling fractions. When a survey is conducted, the sample is expected to be representative, that is, everyone is equally likely to be sampled, but practically few or no individuals with rare or unusual characteristics will be chosen (Seaman and White,

[109]). However, interest may be on such individuals. In order to ensure that adequate number of individuals are sampled, sampling weights are employed. In the population, every individual is given a sampling weight and the probability that such individual is chosen is proportional to this weight. As outlined in (Seaman and White, [109]), in such an approach the sample estimates of population parameters may be biased, because the sample is slightly different from the population.

In our study, we check that underestimation of the occurrence of chronic obstructive pulmonary disease may likely occur due to dependent-item non-response. Therefore, the incomplete data were adjusted and the occurrence of the disease in the survey data was also re-estimated. Using IPW method adjusts for non-response by specifying a regression model for the missingness mechanism given fully-observed covariates. To obtain valid inferences, this method relies on two assumptions: it assumes that the missingness process is independent of the fully-observed covariates; and relies on a correct specified regression model for the missingness process. In order to harness the strength of the second assumption; we compare it with the correct-specification of MI (FCS). When both models are correct, the advantage of MI (FCS) is that it has better efficiency than IPW because the former uses the entire sample while the latter uses the complete cases only.

The paper is organized as follows. Section 3.2 discusses the material used for the analysis. Section 3.3 gives a brief discussion of the missing data mechanisms and methods. In section 3.4, we presented the results of the simulation study. The analysis results of the chronic obstructive pulmonary disease data as reported in (Bardach et al., [110]) are presented in section 3.5. The paper ends with a discussion and conclusion in section 3.6.

3.2 Material

3.2.1 Data

We used the research data obtained from “ de Excelencia en Salud Cardiovascular para el Cono Sur" center for excellence in cardiovascular health for the southern cone (CESCAS). It is a country-level population-based household survey. The study was designed with the goal of examining the occurrence, as well as the prevalence of cardiovascular and chronic obstructive pulmonary diseases in the general population including determination of risk factors for the diease.

The study was based on a sample of 8,000 non-institutionalized adult men and women between the ages of 35 and 74 years old (2000 per site) coming from Bariloche and Marcos Paz (Argentina), Temuco (Chile) and Canelones (Uruguay). In the study, specially trained interviewers conducted a household survey to uncover information about lifestyle (diet, physical activity, quality of life, smoking, alcohol consumption), socio-demographic data (age, sex, occupation, conditions of life), access to and utilization of health services (consultations, laboratory analysis, hospitalizations, etc.), risk factors and illnesses (high blood pressure, diabetes mellitus, cardiovascular and pulmonary problems, among others). Once the questionnaire was finished, the interviewers invited the participants to visit the assigned health centers to complete baseline evaluations (physical examination, blood test, electrocardiogram and spirometry). Two years from the initial visit participants were required to visit the clinic in the assigned health centers. During the clinical visit, a number of measurements were taken including; taking laboratory blood sample measurements (lipids total cholesterol, HDL-cholesterol, LDL-cholesterol and triglycerides, glucose and plasma creatinine), physical measurements (arterial tension, height, weight, waist and hip circumference) and electrocardiogram (ECG) readings. Between 5% and 10% of the samples selected at random were repeated with the purpose quantifying the variability among the samples. The study employed an electrocardiogram with 12 derivations that is standardized

at 25 mm/sec and at 1 mV of amplitude.

However, spirometry was repeated to assess changes in lung function over time and the development of pulmonary disease on the basis of the established criteria. Spirometry is the most frequently used pulmonary function test and enables health professionals to make an objective measurement of airflow obstruction and assess the degree to which it is reversible. As a diagnostic test for chronic obstructive pulmonary disease, it is a reliable, simple, non-invasive, safe and inexpensive procedure. Participants did up to 8 forced expiratory maneuvers to obtain 3 American Thoracic Society (ATS) acceptable maneuvers, with forced vital capacity (FVC) and forced expiratory volume (FEV1) in the first second reproducible within 150 mL, according to the ATS recommendations. Then, albuterol 200 μg was administered by inhalation and the test was repeated 15 minutes later. Participants with any of the following conditions were excluded from the performance of spirometry. These included pregnant women, participants with active tuberculosis and all the individuals that underwent Eye surgery or retinal detachment, thoracic or abdominal surgery and Myocardial infarction within the last three months.

Under the 2010 CESCAS data, a stratified three-stage sampling design was used to obtain the data using 2010 population census figures as the sampling frame. The first stage consisted of randomly sampling the census radii of each location, which are stratified by socio-economic level. The second stage was conducted at the level of households contained in each radius. In the third stage, individuals between the ages of 35 and 74 were selected. In this study, the disease status is the outcome variable which ia a binary response with indication either a respondent's status of chronic bronchitis is positive, negative or missing. Two classes of variables namely - demographic and lifestyle (were used as covariates with missing observations) were included as potential factors that can cause chronic obstructive pulmonary disease status. These risk factors included gender, marital status, age group, religion, blood cholesterol and the status (presence) of asthma, chronic bronchitis, pneumonia and severe wheezing. In Table 1, we display the response and covariates used and their percentages of missing observations. The range of the incomplete observations are

quite variable from 0% to 87.80%.

Table 3.1: Frequencies and percentages of missing values in each variable

variable	frequency of missing values	% of missing values
COPD	40	0.53
Gender	0	0.00
Marital Status	5	0.07
Agegroup	0	0.00
Religion	4	0.05
Blood Cholesterol	5	0.07
Asthma	36	0.48
Chronic Bronchitis	51	0.68
Pneumonia	63	0.84
Severe Wheezing	6612	87.8

3.3 Missing data mechanisms and methods

3.3.1 Types of missing data mechanisms

Based on the theory of missing values as discussed in (Rubin, [19]), we introduce briefly the concept of missing data mechanisms. Suppose $Y = \{Y_{obs}, Y_{mis}\}$, is the complete data where we let the observed data be Y_{obs} and unobserved data be Y_{mis} respectively. Let M represents the missing data indicator matrix of the same dimension as Y such that value in row i and column j $M(i, j)$ is equal to 0 if the value in Y is missing and 1 if it is observed. Data are MCAR if $P(M|Y) = P(M)$ for all Y that is, the fact that data are missing is not dependent on any observed or unobserved values for any of the variables (Chinomona and Mwambi, [105]). That is the probability that a respondent does not report an item value is completely independent of the true underlying values of all the observed and unobserved variables (Rubin and Little, [94]). Missing observations are not systemic and the observed values can be expressed as a random sub-sample of the complete data (Chinomona and Mwambi, [105]). Under MCAR it is assumed that the observed data is a true representation of the complete sample and possibly the target population, thus inference on parameters of interest can be made, based on the complete case.

MAR mechanism operates when missingness is related to other measured or observed variables in the data, but not to the underlying unobserved values of the incomplete variable, that is the hypothetical values that would have resulted had the data been all observed (Baraldi Enders, [102]). Then under MAR $P(M|Y) = P(M|Y_{obs})$ for all Y. In the context of estimation of parameter in the measurement and missingness model both MCAR and MAR fall under the ignorable missingness. The MNAR holds if missing data is neither MCAR nor MAR. The MNAR operates when the missingness depends on both the unobserved and possibly observed values of Y, that is $P(M|Y) = P(M|Y_{obs}, Y_{mis})$ which has no further simplification. The MNAR mechanism in contrast to MCAR and MAR mechanisms fall under non-ignorable missingness mechanism.

In this research, we focus on the MAR ignorability assumption. Missing data were recorded in some variables which may possibly result from inconsistencies in the responses given for the measured variables or during data computation.

3.3.2 Methods

Multiple imputation

The population quantity to be estimated is represented by θ. The statistic that would be used to estimate θ if complete data were available is denoted by $\hat{\theta} = \hat{\theta}(Y_{obs}, Y_{mis})$ and the variance is represented by $U = U(Y_{obs}, Y_{mis}^{(l)})$. When Y_{mis} is accounted for we suppose to have $m \geq 2$ independent imputations, $Y_{mis}^{(l)}, ..., Y_{mis}^{(m)}$ and the estimates from the imputed datasets are calculated as $\hat{\theta}^{(l)} = \hat{\theta}(Y_{obs}, Y_{mis}^{(l)})$ and the estimated variances $U^{(l)} = U(Y_{obs}, Y_{mis}^{(l)})$, $l = 1, ..., m$.

The overall estimate of θ as an average is computed as follows

$$\bar{\theta} = \frac{1}{m} \sum_{l=1}^{m} \hat{\theta}^{(l)}. \tag{3.1}$$

In addition, we obtained the standard error of $\bar{\theta}$ as the square root of the total estimated variance given by

$$T = (1 + m^{-1})B + \bar{U}, \tag{3.2}$$

where B is the between-imputation variance given by

$$B = \frac{\sum_{l=1}^{m} (\hat{\theta^{(l)}} - \bar{\theta})^2}{m-1},$$

and the within-imputation variance $(\bar{U})$ is given by

$$\bar{U} = \frac{\sum_{l=1}^{m} U^{(l)}}{m}.$$

The confidence interval (CI) for the population quantity, θ from the combined multiple imputed estimate is computed using its standard error of $\bar{\theta}$ and critical value from the student's t-distribution as $CI(\theta) = \bar{\theta} \pm t_{\tilde{v}_{mi}, \frac{\alpha}{2}}$ where $\tilde{v}_{mi}$, are the required degrees of freedom as detailed in (Rubin,[19]).

Inverse probability weighting

The IPW also adjusts for item non-response by creating from the complete cases the so called pseudo-population. In this case individual weights are obtained by the inverse of the conditional probability of being observed given fully observed predictors. In the resulting pseudo-population, the participants' responses with complete data represent themselves and those with the same features who had incomplete information on the variable of choice, (Wirth et al.,[111]). In other words, under correct model specification and under the missing at random assumption, missing data information in the pseudo-population is a chance mechanism unrelated to the observed or unobserved information, as stated by (Hernan et al.,[112]). In the complex sampling framework, the inverse probability weights are modified to adjust simultaneously for item non-response and the probability of being chosen into the study population respectively (Wirth et al.,[111]). As earlier stated by (Moore et al., [113] in the missing covariate context, the final weight W_i^* for each individual i is constructed by multiplying the inverse probability weight $\bar{W}_i = \frac{1}{\pi_i(\tilde{M}_i, \hat{\alpha})}$ by the survey weight $W_{i,s}$. The maximum likelihood estimate predicting the probability that the outcome is observed is represented by $\pi_i(\tilde{M}_i, \hat{\alpha})$. That is $W_{i,s}^* = \bar{W}_i^* W_{i,s}$. According to (Wirth et al.,[111]), the resulting inverse probability weighted regression estimator is given by the weighted sample average.

$$\hat{\mu}_{IPW} = \frac{\sum_{i=1}^{N} (W_{i,s}^*, Y_i)}{\sum_{i=1}^{N} W_{i,s}^*}.$$

When longitudinal data is considered, the corresponding IPW for the longitudinal data model is outlined in (Seaman and White, [109]). The model is basically a generalized

linear model for regression of a scalar response Y on covariates X as the analysis model. In longitudinal observations, Y_i and X_i represent the values of Y and X for individual $i(i = 1,...,n)$ and θ be the model parameters. Let $R_i = 1$ if Y_i and X_i are observed and $R_i = 0$ otherwise.

Using the IPW procedure discussed in (Seaman and White, [109]), the analysis model is also fitted only to the complete cases, but some complete cases receive more weight than the other. The solution of the IPW score equation of the estimator $\hat{\theta}$ is given as:

$$\sum_{i=1}^{n} R_i w_i U_i(\theta) = 0. \tag{3.3}$$

where w_i is the weight given to individual i, R_i is the missing value indicator, U_i is the first derivative of the log likelihood function with respect to θ for the ith observation. The weight w_i is derived as the inverse of the probability that individual i is a complete case which equals $P(R = 1|X,Y,H)$ or an estimate thereof. The weight w_i is unknown and has to be estimated. A logistic regression model also known as the missingness model is fitted to the missing data indicator variable R where predictors are drawn from the set X,Y,H. Here, H denotes additionally known predictors that are informative about the missingness process. Thus the missingness model may include more predictors than there are the analysis model. The w_i are then the inverse of the fitted probabilities of being complete. The missingness model is a generalized linear model for

$$P(R = 1|X,Y,H). \tag{3.4}$$

which may reduce $P(R = 1|H)$. The idea is to include sufficient variables that are credible. Two things are necessary, first the predictors H need to be informative about the missingness process and secondly, we should correctly model the relation between these predictors and the probability of being a complete case. The initial step is to consider the predictors to include, assuming (unrealistically) that we have an idea of the model relation between the predictors and the probability of being a complete case.

3.3.3 The analysis model

Since we are dealing with complex survey data the survey logistic regression model, which is a generalized linear model (GLM) as the analysis model for both the IPW and m multiple imputed data sets. The GLM was introduced for the first time by (McCullagh and Nelder, [114]) and expanded later by (McCullagh and Nelder, [115]) as a unified regression technique that explains the variations in both normal and non-normal (such as binary) response variables using a set of covariates. For example, to formulate a GLM for a binary response variable Y_i one can assume it satisfies the binomial model properties, meaning that $Y_i \sim Bin(n_i, \pi_i)$ and the predictor vector variables x_i relates to Y_i through a link function $g(\mu_i)$ where $\mu_i = E(Y_i)$ for $i = 1, ..., n$. Based on the concept of GLM, the ensuing regression model gives information about the variation in the probabilities π_i using the set of predictors based on the equation given by

$$\pi_i(x_i) = g^{-1}(x_i'\beta), \tag{3.5}$$

where the parameters to be estimated from the data are contained in the vector β which is $(p+1)$-dimensional where p is the number of covariates in the model. Specifically for a logistic regression GLM, the logit link model is given by

$$logit(\pi(x_i)) = log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = x_i'\beta. \tag{3.6}$$

Other link functions such as the probit link can be used for a binary response. In the complex sampling design, the parameters are estimated via a pseudo-likelihood estimation method rather than the maximum likelihood which is applicable under the classical GLM as outlined in (McCullagh and Nelder, [114]). In our case we use the survey logistic regression. The null hypothesis about β_j, where $j = 1, ..., p$ can be preferred using the Wald design-based test statistics that $\beta_j = 0$ and design-based confidence intervals that give information on the likelihood and uncertainty associated with the estimates of each β_j.

3.3.4 Statistical analyses

To estimate the effects of risk factors for the prevalence of pulmonary disease adjusted for the item non-response in the survey data, we use two approaches to correct for non-response namely multiple imputation and inverse probability weighting methods. The multiple imputation method described in sub-section 3.2.1 was used to produce multiple complete data sets for analysis that accounts for the variability about the missing observations. In SAS/STAT 9.4, we used “proc mi" to carry out the imputation. This approach specifies the conditional distributions of variables with missing observations conditioned on the other variables in the survey data and the algorithm for imputation iterates in sequential order through the variables to impute the missing observations using the specified models. This approach is termed the fully conditional specification (FCS) method. As stated by (Schafer and Olsen, [41]), the procedure for multiple imputation is performed using the markov chain monte carlo (MCMC) technique making use of an iterative data augmentation approach. Furthermore, (Berglund and Heering, [116]) described the implementation of MI following a framework for estimation and inference based upon a three step process: the first step is the formulation of the imputation model and imputation of missing data using PROC MI with FCS as the selected method. The analysis of complete data sets using standard SAS procedures such as the simple logistic regression (that assume the data are identically and independently distributed or from a simple random sample) or SURVEY procedures for analysis of data from a complex sample design is the second step. In this step, we performed design-based logistic regression using PROC SURVEYLOGISTIC. The PROC SURVEYMEANS and PROC SURVEYFREQ allow to give correct means and produce predicted frequencies of the datasets. Lastly, the results of the output from the previous 2nd step are combined using the PROC MIANALYZE. A key assumption made in the MI and MIANALYZE procedures is that the missing data are missing at random (MAR) or in other words, the probability that an observation is missing depends on observed data Y_{obs} but not missing Y_{mis} (Rubin, [19]).

It is pertinent to guarantee correct incorporation of the complex sample design features and weights into the MI framework. For these to be captured in the imputation model, it is recommended that a categorical variable be created in the 'data step' which combines the stratum and cluster codes provided by the data analyst. As outlined in (Greenland and Finkle, [117]), inverse probability weighting adjusts for non-response by weighting the outcomes of participants with non-missing information; by the inverse of the probability of having complete data (obtained by specifying a regression model for the missingness mechanism given fully observed covariates). In our approach, this procedure is adopted. In order to produce valid inference, two assumptions are considered. First, we assume that the missingness process is within the levels of the fully-observed covariates; which means, the data is missing at random. In addition, the regression model for the missingness process for inverse probability weighting should be specified correctly.

3.4 Simulation study

In this study, we conduct simulation studies and later apply the real data. The importance of the simulation study is to examine the techniques to handle incomplete observations, and explore the performance of such methods under different missing data conditions. In this study, the focus is on the intermittent missing data pattern.

3.4.1 Data generation, simulation designs and analysis of the simulated data

We simulated cross sectional binary datasets to mimic the original dataset and introduced different missing rates. For each of these different cases, we simulated 1000 datasets based on a logistic regression model scheme of the form (3.7) for sample sizes N= 100, 200, 500. The cross sectional binary outcomes were generated following a model with a linear

combination of the predictor as shown in the model below:

$$logit[P(Y_{ij}=1)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2). \tag{3.7}$$

where we assume an underlying binary response variable as is the case with the real application study. The null model effect is β_0, while β_1 and β_2 are the main effects for the variables x_1 and x_2 respectively for individual i with their interaction effect captured by β_3. Thus, x_1 is binary and x_2 is continuous respectively. We simulated two different covariates: x_2 from Bernoulli distribution with probability of success equals to 0.5, x_1 from Uniform distribution. For the purpose of simulation study, we used the parameters, $\beta_0 = -1, \beta_1 = 1, \beta_2 = 0.07$ and $\beta_3 = -0.25$. Thus the simulation model is explicitly written as

$$logit[P(Y_{ij}=1)] = \beta_0 + (1)x_1 + 0.07x_2 + (-0.25)(x_1 * x_2). \tag{3.8}$$

When the logit link function is inverted it leads to conditional binary logistic regression which is the probability of the event occurring as a function of covariates, thus equation (7) can be written equivalently as

$$P[P(Y_{ij}=1)] = \frac{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2))}. \tag{3.9}$$

We first generated a data set without any missing values followed by creation of incomplete data at different missingness rates. We assumed a MAR mechanism for the missingness model. The missingness problem was handled using the two approaches introduced in Sections 3.2.1 and 3.2.2 respectively namely IPW and multiple imputation.

For purposes of comparison, a larger number of imputations were necessary (Wood et al., [118]). Nevertheless, sufficient accuracy is possible even when M can be set to $3 \leqslant M \leqslant 5$. However, there is a caution that pegging on this range is risky (Schafer, [45]). On the other hand, the efficiency increments diminish as rapidly after the first M=2 imputations for a small fraction of missing information and after the first M=5 imputations for a larger fraction of missing information (Molenberghs and Verbeke, [33]). Furthermore, a rule of

thumb for choosing M is suggested (see White et al., [119]). Their suggestion is that M should be at least equal to the percentage of incomplete cases. The reason for the larger number of imputations is to render the final analysis reproducible, which is not always the case for a small number of imputations as also corroborated by (Van Buuren, [120] and White et al., [119]). A number of imputations say M=20 is readily possible given the current available computational power. This is necessary because if one wants to repeat the analysis for the same M then essentially the same results will be generated (Ivanova et al.. [121]). In our study, we performed M=20 imputations. This moderate value was chosen to account for the relatively large fraction of missing data and to limit the loss of power for testing any associations of interest (Kombo et al., [122]).

In order to compare the performance of the methods, we used bias and mean squared error (MSE). Bias is defined as the absolute difference between the average parameter estimate from a given number of replications.

3.4.2 Simulation results

Table 3.2 shows the outcome of the simulation study (based on 1000 simulated datasets and 20 imputations) to compare the MI and IPW missing data methods in terms of bias and MSEs, under N=100, 200 and 500 sample sizes. The missing rates of 10%, 30% and 50% represent low, moderate and high missing entries. In the table, large bias and MSE are shown in bold. Starting from N = 100 sample size we observe that for low and moderate missing rates MI produced less biased estimates than the IPW. In the case of MSEs, both methods were comparable, except for β_3 in MI that produced high values for all levels of missing rates. Furthermore, under the sample size of 200; with low missing rate the MI yielded more unbiased estimates than the IPW, except for the moderate missing rate. The performance of the MSEs were closer to each other. However, under the sample size of 500 for all levels of missing rates the MI produced more unbiased estimates, except for β_3 for each case. As we increased the sample size and the missing rate, it is observed that the

values for the MSEs also reduced. The findings reveal that the MI performs better than the IPW. The results also mean that studies should be carefully planned and designed putting remedial measures to reduce the rate of missing values.

Table 3.2: Bias and mean squared error (MSE) estimates for multiple imputation and inverse probability weighting methods, under MAR mechanism over 1000 samples: N=100,200 and 500 individuals.

Sample size	Missing rate	Parameter	MI		IPW	
			bias	MSE	bias	MSE
100	10%	β_0	-0.1557	0.1184	**0.1634**	0.1215
		β_1	-2.0161	6.4840	**-2.1418**	6.4843
		β_2	-0.0030	0.0020	**-0.0031**	0.0020
		β_3	**0.0856**	0.0634	0.0443	0.0466
	30%	β_0	-0.0171	0.0787	**0.0913**	0.1152
		β_1	**-2.1133**	6.4499	-1.6610	5.4304
		β_2	0.0006	0.0016	**-0.0015**	0.0020
		β_3	**0.0962**	0.0570	0.0301	0.0647
	50%	β_0	0.1358	0.1448	**0.1436**	0.1465
		β_1	-2.7397	10.3119	**-3.3642**	13.7106
		β_2	-0.0030	0.0028	-0.0030	0.0028
		β_3	**0.1387**	0.0818	0.0708	0.0594
200	10%	β_0	-0.0254	0.0685	**-0.0887**	0.0759
		β_1	-1.2168	3.6496	**-1.6535**	4.4260
		β_2	0.0009	0.0014	**0.0021**	0.0014
		β_3	**0.0244**	0.0538	0.0014	0.0413
	30%	β_0	**0.0214**	0.0746	-0.0736	0.0821
		β_1	**-1.7036**	4.4089	-1.6581	4.2745
		β_2	-0.0001	0.0015	**0.0018**	0.0016
		β_3	**0.1065**	0.0152	0.0021	0.0370
	50%	β_0	-0.0123	0.0957	**0.0160**	0.0905
		β_1	-1.0318	2.6379	**-1.0534**	3.2890
		β_2	**0.0008**	0.0020	-0.0003	0.0019
		β_3	**0.1568**	0.0582	0.0300	0.0529
500	10%	β_0	-0.0482	0.0446	**-0.0489**	0.0454
		β_1	-1.0479	2.3600	**0.0541**	1.1093
		β_2	0.0013	0.0009	**0.0012**	0.0009
		β_3	**0.0457**	0.0299	-0.0041	0.0272
	30%	β_0	-0.0405	0.0520	**-0.0934**	0.0575
		β_1	-1.5439	1.1148	**0.7541**	1.9190
		β_2	0.0012	0.0011	**0.0022**	0.0010
		β_3	**0.1035**	0.0345	-0.0249	0.0042
	50%	β_0	-0.1410	0.0714	**-0.1590**	0.0831
		β_1	-1.5502	1.1141	**2.4613**	1.6264
		β_2	0.0033	0.0011	0.0032	0.0012
		β_3	**0.1430**	0.0447	-0.6437	0.0439

3.5 Application results

3.5.1 Results from the application analysis

We present weighted and design-consistent estimates results for the risk factor of pulmonary disease prevalence obtained for multiple imputation and inverse probability weighting methods. In the case of multiple imputation, the analysis accounted for both the complex sampling design and the imputation process. The variability that is introduced by the imputation process and the variability that is accounted for in the complex sampling design are reflected by the variance estimates. For inverse probability weighting individuals are weighted by the inverse of the conditional probability of complete data given the fully observed covariates.

In Table 3.3, we present the results from each method. The overall estimates from the multiple imputation and inverse probability weighting methods are not very different, but the former produces smaller standard errors than the latter for all covariates. This displays the superior efficiency in the multiple imputation method over the inverse probability weighting method in real application. At the 5% level, covariates associated with single marital status and living with a partner/divorced/widowed are non-significant under the two methods, whereas the covariates associate with asthma, chronic bronchitis pneumonia and severe wheezing are significant under the multiple imputation method, but chronic bronchitis and severe wheezing are non-significant under the inverse probability weighting method.

3.5.2 Discussion of results

The results of the survey logistic regression (as the analysis model) with the parameter estimates and their standard errors pooled from the multiply imputed datasets and inverse probability weighting using the methods outlined in Sections 3.2.1 and 3.2.2 are presented respectively. The survey logistic regression model the variation in the pulmonary disease

Table 3.3: Overall and subgroup estimates, standard errors and $Pr > |t|$ of chronic obstructive pulmonary disease prevalence for (a) multiple imputation and (b) inverse probability weighting

Variable	(a)Multiple Imputations Analysis			(b)Inverse Probability Weighting		
	Estimate	$S.E$	$Pr > \|t\|$	Estimate	$S.E$	$Pr > \|t\|$
Overall	-0.1802	0.4512	0.6897	-0.6980	0.6519	0.2846
Gender						
Male	Ref					
Female	0.3131	0.2000	0.1174	0.3772	0.3550	0.2883
Marital Status						
Single	-0.4595	0.3297	0.1635	-0.2322	0.4691	0.6207
Married	-0.6671	0.2322	0.0035	-0.7781	0.3822	0.0421
Separated	Ref					
Living with a partner/						
divorced/widowed	-0.3850	0.3270	0.2399	-0.3003	0.5333	0.5735
Agegroup						
34-44	Ref					
45-54	0.2832	0.3000	0.3452	0.4192	0.4953	0.3076
55-64	0.6101	0.3001	0.0421	0.5523	0.4622	0.2324
65-74	0.6734	0.3184	0.0346	0.7258	0.5103	0.1553
Belief						
Religion	Ref					
No Religion	-0.1021	0.3292	0.7565	0.3643	0.5394	0.4996
Blood Cholesterol						
Yes	Ref					
No	-0.3015	0.2031	0.1375	-0.3331	0.3048	0.2747
Asthma						
Yes	Ref					
No	-0.6993	0.2271	0.0021	-0.5549	0.3198	0.0831
Chronic Bronchitis						
Yes	Ref					
No	-1.5337	0.2188	<.0001	-1.5904	0.3338	<.0001
Pneumonia						
Yes	Ref					
No	-1.4732	0.1988	<.0001	-1.0454	0.3220	0.0012
Severe Wheezing						
Yes	Ref					
No	-0.7438	0.2929	0.0137	-0.7487	0.4080	0.0668

prevalence, as a function of socio-demographic and illnesses variables which accounted for the complex sampling design. In Table 4, we display the adjusted odds ratio estimates of the logistic regression models for the two methods. The purpose of reference level was to ensure estimation and interpretation. Furthermore, the odds ratios helps to observe the multiplicative effect of each level and the possibility of having pulmonary disease as a predictor in relation to reference level that controls for the effect of the predictors in the model. The results show that the risk factors pulmonary disease is slightly lower among

singles (OR=0.595, 95% *CI*=0.312-1.134) under the MI than (OR=0.793, 95% *CI*=0.316-1.994) under the IPW. However, the effect is not significant, and the two approaches agree. It is also less statistically significantly among the married (*OR* = 0.462, 95% *CI* = 0.293-0.729) under the MI and (*OR*= 0.459, 95% *CI*=0.217-0.927) under the IPW. The odds of pulmonary disease is low among those living with a partner (OR=0.635 95% *CI*=0.338-1,195) under the MI and under the IPW (OR=0.741 95% *CI*=0.260-2.109). However this partner effect is not significant for the single marital status but significant for the married. The odds of pulmonary disease is significantly lower among those married compared to those who are separated. The results from the MI model show that those with and without asthma are not significantly different in the odds of pulmonary disease (OR=0.574, 95% *CI* = 0.306 - 1.075) but significantly different under the IPW analysis (OR=0.517, 95% *CI* = 0.335 - 0.797). Both methods show that religion is not significantly associated with pulmonary disease.

For both methods not having bronchitis is associated with lower odds of pulmonary disease (OR= 0.194, 95% *CI* = 0.127 - 0.295) under MI and (OR=0.204, 95% *CI* 0.106 - 0.393). Under both methods cholesterol status is not significantly associated with pulmonary disease. Likewise both methods show that age is not significantly associated with pulmonary disease. Age was also found not be significantly associated with pulmonary disease under both methods.

3.6 Discussion and conclusion

Missing data pose potential problems in the cross-sectional survey data, which quite often is composed of high rate of missing values due to a number of reasons. Complete case analysis is one of the methods used to deal with data with missing observations. This method excludes cases with missing values in the analysis, but the major concerns with this method is that such an approach can lead to substantially reduced power due to loss of information and can also lead to seriously biased estimates if the deleted sub-sample is significantly

Table 3.4: Adjusted odds ratio estimates for the survey logistic regression model under (a) multiple imputation analysis (b) inverse probability weighting

	(a)Multiple Imputations Analysis		(b)Inverse Probability Weighting	
Variable	OR	95%CL	OR	95%CL
Gender				
Male	Ref			
Female	1.315	(0.898,1.926)	1.458	(0.727,2.927)
Marital Status				
Single	0.595	(0.312,1.134)	0.793	(0.316,1.994)
Married	0.462	(0.293,0.729)	0.459	(0.217,0.972)
Separated	Ref			
Living with a partner				
/divorced/widowed	0.635	(0.338,1.195)	0.741	(0.260,2.109)
Agegroup				
34-44	Ref			
45-54	1.274	(0.717,2.263)	1.521	(0.575,4.020)
55-64	1.726	(0.977,3.050)	1.737	(0.701,4.302)
65-74	1.856	(1.020,3.378)	2.066	(0.759,5.626)
Belief				
Religion	Ref			
No Religion	0.968	(0.518,1.809)	1.439	(0.499,4.149)
Blood Cholesterol				
Yes	Ref			
No	0.752	(0.507,1.115)	0.717	(0.394,1.304)
Asthma				
Yes	Ref			
No	0.517	(0.335,0.797)	0.574	(0.306,1.075)
Chronic Bronchitis				
Yes	Ref			
No	0.194	(0.127,0.295)	0.204	(0.106,0.393)
Pneumonia				
Yes	Ref			
No	0.253	(0.173,0.370)	0.352	(0.187,1.066)
Severe Wheezing				
Yes	Ref			
No	0.485	(0.311,0.757)	0.473	(0.212,1.053)

different to that remaining. Alternatively, ad hoc methods can be used which are based on substituting the missing observations with plausible ones such as last observation carried forward, the mean and regression predictions (single imputation). However, the results obtained from such methods suffer potential loss of distributional relationships amongst the predictors and fail to produce measures of uncertainty introduced by the imputation procedure. For this reason, multiple imputation has emerged as one of the most powerful and reliable method of dealing with incomplete data such as the 2010 CESCAS data. Unbiased estimates of pulmonary disease prevalence are obtained that also accounted for the uncertainty about the missing observations themselves. We estimated the design-consistent

estimates of intercept and subgroup of pulmonary disease prevalence using multiple imputation and inverse probability weighting. The survey logistic regression model was fitted and results obtained show and results obtain show dissimilarity in the parameter estimates in the methods used. Nevertheless, multiple imputation may be preferred to inverse probability weighting. The results reveal that the IPW performs so closely to the MI using FCS. The strength of this research lies on the use of the MI method for imputing missing values in chronic obstructive pulmonary disease study. Missing data are unavoidable, pervasive and if not handled properly could give biased estimates. The use of an appropriate statistical techniques to estimate model parameters of interest and their variability are necessary for reliable inference. The key to an informative analysis for evidence based research lies in the design of the study generating the data for analysis. As much as possible the study design should such that it minimizes the occurrence of missing data. In particular for cross-sectional survey data there are many sources that may cause data to be missing at the individual and variable level. All effort needs to be put in places to control against such causes of missingness. In fact, none of these setbacks downplay the uniqueness a study designed by qualified statisticians with professional expertise in the survey methodology, to collect population-based information. A possible area of extension is the use of sensitivity analysis to analyze binary cross sectional survey data. It assesses the robustness of the results across different model assumptions, with the aim of identifying results that are most dependent on questionable or unsupported assumptions. In addition, further research should consider when one or both model of MI (FCS) and IPW are misspecified.

Chapter 4

Statistical methodologies for handling ordinal longitudinal responses with monotone dropout patterns using multiple imputation

Abstract

Missing data are common challenge in any longitudinal study such as clinical trials. Multiple imputation is one of the modern powerful methods of handling incomplete data. This approach is applicable to different missing data patterns but sometimes faced with complexity of the type of variables to be imputed and the mechanism underlying the missing values. In this study, we compare the performance of three methods under multiple imputation, namely expectation maximization, fully conditional specification and multivariate normal imputation in the presence of ordinal responses with monotone dropout. We demonstrated the usefulness of the ordinal negative binomial distribution for ordinal data generation through simulation studies and implementation. However, the real dataset application and simulation studies reveal that the three methods perform equally well, thus we conclude that any of the methods can handle ordinal outcomes with missing values.

4.1 Introduction

In longitudinal studies ordered categorical (ordinal) data are a common feature in health research particularly clinical trials and follow-up observational studies with multiple health outcomes. In most cases researchers are faced with data where the disease or health condition may fall into more than 2 levels of increasing disease severity. This gives rise to a typical ordinal outcome defining the different disease levels. From a clinical and diagnosis point of view it is important that an individual at any time of follow is classified into the correct disease category. For the sake of clarity, the ordinal data property requires that there is a clear order of the response outcome categories, but no existence of underlying interval scale between them. The methods developed for categorical data can be applied to the analysis of ordered categorical data, but such methods may result to loss of information. One of the advantages of using models and methods explicitly developed for ordinal data is that they take into account the natural ordering of the categories. In particular, models for ordered categorical data tend to be more parsimonious than their unordered counterparts, thus resulting in more efficient inferences with a clear interpretation of parameters as stated by (Ursino and Gasparini, [123]). There are two general areas of statistical inference that are of importance in modeling ordinal data. These are *association and regression* as stated by (Agresti, [124]). The development of logistic regression and loglinear models for categorical data occurred as far back as in the 1960s and 1970s. In those years ordinal data received some attention as in (Bock and Jones, [125] and Snell, [126]), but a stronger focus was inspired later by articles written by (McCullagh, [127]) on logit modeling of cumulative probabilities and (Goodman, [128]) on loglinear modeling of the odds ratios. The approaches for defining logits for an ordinal response are numerous, but there are three types that are prominent in biostatistics literature. These are the adjacent-categories logits, the continuation-ratio logits and the cumulative logits by (McCullagh, [127]) which is the most popular.

The methods for modeling ordinal data are based on the distribution of the ordered categories of the ordinal outcome normally represented by the first few integers $0, 1, ..., n$, interpreted simply as codes for the ordered categories without reference to odds ratios or their logarithms. It is essential that within these methods the metric properties of $1, ..., n$ play no dominant role in the interpretation of the results, since $1, ..., n$ are simply convenient labels for the ordered levels. The use of these methods range from a simple binomial distribution to more flexible distributions which allow for overdispersion, such as the beta-binomial (see for example Muniz-Terrera et al., [129] in which generalized additive models for location scale models are used to analyze cognitive test data), including the approach of (D'Elia and Piccolo, [130]) who proposed the use of a mixture of a binomial and a discrete uniform distribution. According to the study by (D'Elia and Piccolo, [130]), the psychometric point of view was taken into account in which the response is the result of the combination of feeling and uncertainty components which can be modeled using two parameters, obtaining nonetheless various possible shapes and behaviours of the response distribution. However, another area in which there has been a lot of research activity in recent times is the analysis of repeated measures data in the form of ordered categorical responses. Such data arise, from longitudinal studies, crossover experiments, studies of familial characteristics or kinship which all exhibit some sort of clustering. However, there are other well-developed statistical methods of analyzing count data which includes the Poisson, negative binomial (NB), hurdle and zero-inflated models. However, according to a study by (Dawson, [131]) the outcomes are measured using ordinal scales for reasons such as the need to reduce participant burden and limit error in cognitive recall. If counts are assessed as binned ordinal responses; this has advantage from the perspective of measurement (e.g., ease of burden and improved call) but introduces significant complexities in statistical modeling. The linear models and proportional odds mixed models (POMM) are commonly fitted to ordinal data.

When dealing with incomplete observations from a discrete variable (e.g. ordinal), the first

appealing method may be to treat the variable as continuous for the purpose of imputation, and then round off the imputed values to the nearest valid discrete value before going to fit the substantive model (Carpenter and Kenward, [132]). The intuition behind multiple imputation (MI) is to draw valid and efficient inferences by fitting an analysis model to the multiply imputed data. The imputed values should bear the structure of the data, and uncertainly about the structure and be sensitive to the process that led to missing observations as pointed out by (van Buuren, [133]). The approach of creating the imputed datasets depends on the missing data pattern. For the monotone dropout patterns, the parametric regression method that assumes multivariate normality or a nonparametric approach that employs propensity scores may be used (Molenberghs and Vebeke, [33]). Also the methods that assume normality have been successfully used by the authors (Choi et al., [134], Demirtas and Hedeker [135], Seitzman et al., [136]). On the other hand, imputations may be generated by performing a series of univariate regressions, instead of just a single large model (it becomes easier to estimate), and without assuming normality of the variables. However, researchers advised against handling ordinal response as a continuous or dichotomized variable for a number of purposes. These purposes include efficiency loss due to information loss, reduced statistical power and decreased generality of the analytic conclusions (Gameroff, [137]). Ideally, continuous models may produce predicted values outside the range of the ordinal variable and finally, and further a continuous model may yield correlated residuals and regressors when used for ordinal response and does not account for the ceiling and floor effects of the ordinal response. This may lead to biased estimates of the regression coefficients (Bauer and Sterba, [138]). There is still ongoing debate on the issue among researchers.

In our study, we simulated the ordinal variable from a discrete distribution namely the negative binomial distribution. In other words, we explicitly link the ordinal responses to an appropriate count distribution through known cut-points. The points are known simply because they are values that define the range of counts within each ordinal outcome. Thereafter, we introduced the concept of missing values through the monotone dropout pattern.

We also present the analysis of a real application. Finally, we discuss the results and the contributions of the proposed ordinal count model.

The paper is organized as follows. In Section 4.2, we give details of the ordinal count generation models. Section 4.3 gives a description of the imputation model followed by a simulation study and a real application in Section 4.4. The paper ends with a discussion in Section 4.5.

4.2 Ordinal negative binomial model (ONB)

In the ordinal outcomes study, the goal is generally to predict some underlying discrete outcome (e.g., classification of disease) as a function of a set of covariates. However, these discrete outcomes are measured with ordinal scores which collapse the counts from the parent discrete distribution into a series of response categories with known-cut points. From this procedure, we can assume that underlying the ordinal responses is a discrete count generated from the NB distribution. After generating the values we transform the outcome to ordinal scale with the use of cut-off chosen appropriately. (McGinley et al., [139]). Next, we discuss the procedure of linking the ordinal responses to the underlying count distribution.

In order to model the ordinal response as a function of an underlying NB distribution, we assume that underlying the ordinal response, Y_i for individual $i(i = 1,...,n)$, is an unobserved count latent variable, at four study visits, $j = 1,...,4$, Y^*_{ij}. In our study, the simulation of the ordinal outcome is from a latent NB random variable. Thus, we assume four repeated measurements per subject. Also at the each time point t_{ij} the observed outcome is y_{ij} based on the underlying NB variable Y^*_{ij}. The probability mass function (PMF) for the NB distribution in terms of Y^*_{ij} conditional on covariates, x_i, is

$$f(y^*_{ij}) = P(Y^*_{ij} = y^*_{ij}|x_i) = \frac{\Gamma(y^*_{ij} + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y^*_{ij} + 1)}(\alpha\mu_1)^{y^*_{ij}}(1 + \alpha\mu_i)^{-(y^*_{ij}+\alpha^{-1})}, y^*_{ij} = 0, 1, 2, ... \tag{4.1}$$

where $E(Y_{ij}^*) = \mu_i$, $Var(Y_{ij}^*) = \mu_i + \alpha\mu_i^2$, and α is the dispersion parameter. We adopt the use of the log link function to link the linear predictor to the mean of Y_{ij}^* as

$$log(\mu_{ij}) = x_{ij}^{'}\beta \tag{4.2}$$

where x_i is a $p \times 1$ vector of covariates (this includes '1' as the first element for the intercept) and β is a $p \times 1$ vector of regression coefficients.

The cumulative distribution function (CDF) for the NB distribution is simply the sum of the PMFs such that

$$F(y_{ij}^*) = \sum_{v=0}^{y_{ij}^*} f(v), \tag{4.3}$$

where the cumulative probability is evaluated at y_{ij}^*.

The next step is linking the ordinal response, Y_{ij}, to the unobserved latent variable, Y_{ij}^*. This is achieved by the use of a fixed set of cut-points to generate the categories of the ordinal response where

$$Y_{ij} = c \;\; if \; k_{c-1} < Y_{ij}^* \leq k_c \tag{4.4}$$

and the threshold k_c defines the upper bound of ordinal response category c ($c = 1, 2, ...M$). The probability of observing an outcome in category c can be expressed as a function of the cumulative probabilities of the underlying Y_{ij}^* distribution, that is

$$P(Y_{ij} = c|x_i) = P((k_{c-1} < Y_{ij}^* \leq k_c)|x_i) = F(k_c) - F(k_{c-1}). \tag{4.5}$$

In this case, $F(k_c)$ and $F(k_{c-1})$ designate the CDFs evaluated at the two consecutive cut-points c and $c-1$ for a NB distribution with a mean of μ_i and dispersion of α. The likelihood function over all individuals and categories for the ordinal data following underlying count distribution can be expressed as

$$L_{ONB} = \prod_{i=1}^{n} \left[\prod_{c=1}^{M} \left[F(k_c) - F(k_{c-1}) \right]^{y_{ic}} \right]. \tag{4.6}$$

Our proposed approach is different from the current practice of the proportional odds mixed models (POMM) because we use the ordinal responses of the underlying NB CDF as

against the logistic CDF. In the proposed ONB approach, we study the effect of covariates as if it were modeling the latent count responses directly, but the POMM handles the effect of covariates over the cumulative odds across response categories.

4.3 Imputation methods

When the pattern of missingness of a dataset is monotone, the variables with missing values are imputed sequentially with covariates obtained from their corresponding sets of preceding variables. Regression, predictive mean matching or propensity score methods are some of the various methods used to impute continuous variables. A logistic regression method may be used for binary or ordinal variables. On the other hand, a discriminant function for nominal or binary variables can be used. Using simulated and real incomplete ordinal datasets, we compared three multiple imputation procedures: the fully conditional specification (FCS) via chained equations (Van Buuren [133], Van Buuren [140], the multivariate normal imputation (MVNI) Schafer [45] and expectation maximum (EM)). The procedures for these methods are based on different theoretical assumptions and involve different computational techniques as pointed out by (Lee and Carlin, [141]).

4.3.1 Multivariate normal imputation (MVNI)

The methods of imputing multivariate data have been developed, for example, (Rubin and Schafer, [142]) gave approaches to effectively generate multivariate multiple imputations. This approach is based on Bayesian simulation algorithm draws from the posterior predictive distribution of the unobserved data given the observed data. The procedure assumes that the data are multivariate normally distributed and missing at random. This approach has been used by (Schafer, [45]) and derived imputation algorithms for multivariate numerical, categorical and mixed variable type data. The methodology describes the data by using a multivariate model to derive a posterior distribution and then draws imputations

based on the Gibbs sampling algorithm (hereafter referred to as data augmentation rather than the Gibbs sampler). The Markov chain Monte Carlo (MCMC) approach is used to draw imputed values from the estimated multivariate normal distribution.

Thus, data augmentation approach relies on Bayesian inference where missing data imputation is based on iterating between an imputation step (I-step) and Posterior step (P-step),as pointed out by (Tanner and Wong, [25]). The I- and P- steps are briefly outlined below assuming the ordinal response is multivariate normal that is $Y \sim \mathrm{N}(\mu, \Sigma)$

- *The imputation step* - With some estimated initial values for the mean vector μ and covariance matrix Σ, the I-step simulates values for missing data Y_m by randomly drawing it from the conditional predictive distribution of Y_m, that is, from a current estimate (rth iteration) $\theta^{(r)}$, of the parameter, a value Y_m^{r+1} of the missing data is drawn from the conditional distribution of Y_m given Y_o:

$$Y_m^{r+1} \sim P(Y_m|Y_o, \theta_r), \;\; \theta = (\mu, \Sigma) \tag{4.7}$$

- *The posterior step* - This step draws a value of the parameter θ from a complete-data posterior distribution:

$$\theta^{(r+1)} \sim P(\theta|Y_o, Y_m^{(r+1)}) \tag{4.8}$$

The new parameter θ is then used to update the I-step and the processes runs between the I and P step sequentially.

When Equations (4.7) and (4.8) are iterated from initial value $\theta^{(0)}$ the stochastic sequence $\{(\theta^{(r)}, Y_m^{(r)}); r = 1, 2, ...\}$is produced. The two steps are iterated for a sufficiently long time until the distribution of the estimates becomes stationary (Schafer, [45]). There is usually dependency across the steps because each current step depends on the previous one. Theoretically, the approach is good but may not be always realistic because of distributional assumptions (e.g. assuming normality for binary, ordinal and other non-normal variables). In the case of categorical variables, the MVNI methods draws imputations under the MVN

model and to accommodate the categorical nature of the data, one needs to round off the imputations to the nearest integer. However, there is need for caution as detailed in (Allison, [143]) about rounding off (binary case is cited) as such imputed values may lead to biased parameter estimates. Nonetheless, an argument was further raised by (Schafer, [45]) that inference from MVNI may be reasonable, even if multivariate normality does not hold especially for the cases of binary and categorical variables. For a detailed account of this procedure, see reference (Schafer, [45]).

4.3.2 Fully conditional specification (FCS)

Another option that is useful in the multivariate data is the fully conditional specification (FCS). The FCS approach is flexible and specifies the multivariate model by a series of univariate conditional models for each of the incomplete variables. The FCS does not depend on the multivariate normality assumption, and the univariate regression models can be utilized for the ordered logistic regression for ordinal variables if it is appropriately adopted. when Bayesian approach used, imputations are conducted in stepwise order starting with the variable with the smallest amount of missing observations and progresses sequentially until the variable with the largest missing value is finally captured. The imputations involve three stages: the imputation, analyses and pooling stages. In the first stage; the missing values are filled-in M times to generate M complete dataset. The second stage analyses each of the M complete dataset, and third stage combines estimates from the analyses from stage to provide single estimate. FCS deals with all different types of variable with missing values in a data set. Some continuous and some discrete including ordinal outcomes.

When the ordinal response variable Y has the features of a vector of unknown parameters $\theta = (\mu, \Sigma)$; with μ mean vector and covariance matrix Σ. Assume the complete data can be partitioned as, $Y = (Y_o, Y_m)$ where Y_o and Y_m are the observed complete incomplete components of the data respectively. As outlined in (Van Buuren et al., [144]) and also in (Van Buuren and Groothuis-Oudshoorn, [145]), multiple imputation via FCS proceeds as

follows:

- Calculate the posterior distribution of θ given the observed data, that is, $P(\theta|Y_o)$;
- Then θ^* is drawn from $P(\theta|Y_o)$;
- Then a value y^* from the conditional posterior distribution of y_m given $\theta = \theta^*$:

$$y^* \sim P(y_m|y_o, \theta = \theta^*). \tag{4.9}$$

The second and third steps are repeated depending on the number of imputations. These make the results to reliably simulate an approximately independent draws of the missing observations for an imputed dataset.

4.3.3 Expectation maximization (EM)

The expectation maximization (EM) algorithm by (Dempster et al., [24]), is a sound, general and iterative procedure for the maximum likelihood estimate (MLE) of a parametric distribution underlying some given data, where the data could be incomplete or has missing values. The EM algorithm handles incomplete data and the complications of estimates related to the MLE by attempting to solve smaller complete data problems which lead to parameter estimates for the entire dataset (incomplete and complete data). The EM algorithm deals with missing values using the following two steps: first, the missing data are imputed using the estimated values generated by MLE and secondly, the parameter estimates from the first step are re-estimated; this procedure is iteratively repeated until the final convergence step or when the current solution differs negligibly from the previous one. Each iteration of the EM algorithm consists of two steps - the expectation step and maximization step (Little and Rubin, [94]). Each step is completed once within each algorithm cycle, which means cycles are repeated until a suitable convergence criterion is satisfied. For more theoretical justification see (Dempster et al., [24], Little and Rubin

[94]). The fitted parameters (on convergence) are equal to a local maximum of a likelihood function which is the MLE in the case of a unique maximum as detailed in (Dempster et al. [24]). The EM algorithm has two disadvantages: first, it is slow to converge and second, it has no direct provision of a measure of precision for the MLEs. In order to overcome these challenges several techniques have been developed as detailed in (Louis [55], Baker [58], McLachlan and Krishnan [146]).

4.3.4 Software considerations

Valid inferences can be produced through a likelihood-based analysis without modeling the dropout process, especially when MAR is assumed. In this approach, the generalized linear mixed model is used as the analysis model. The implementation of this procedure may be done using the SAS procedures NLMIXED and GLIMMIX. If we must impute missing observations, describing of missing data pattern and multiple imputation is performed using the procedure PROC MI as pointed out by (Kombo et al., [122]). All types of variables may be considered. The procedure provides many techniques for imputation but rely on weather the variable is categorical or continuous. Here, our interest is to compare three different methods, EM, FCS and MVNI as implemented in PROC MI. To implement FCS, the fcs statement is specified in PROC MI. In order to run MVNI and EM, both methods require the use of the Markov Chain Monte Carlo (MCMC) simulation approach to draw the imputed values from the estimated multivariate normal distribution. In PROC MI, we specify the number of datasets to be imputed and the imputation model to use. After imputation, statistical algorithms are specified for the running of the analysis model of interest separately for each imputed dataset using the by `_Imputation_` statement, and the results are stored in an output file. In conclusion, a procedure called, PROC MIANALYZE, combines the estimates obtained from the analyses of the multiply imputed datasets to produce valid statistical inferences as also in the work by (Kombo et al., [122]). However, non-normal data analyses especially for categorical data, additional manipulations are

needed before PROC MIANALYZE is used (Ratitch et al., [147]). This is because Rubin rules (Rubin, [19]) for combining results assume that the statistics estimated are normally distributed. The estimates are regression coefficients and means which are approximately normally distributed, while others like the odds ratios, correlation coefficients and relative risks are non-normal.

4.4 Simulation study

Before the real data is analyzed, we conduct simulation studies to evaluate the properties and effectiveness of the methods of handling incomplete data. The importance of this is to investigate and explore the performance of such methods under different missing data conditions. In this study, the focus is on the monotone dropout.

4.4.1 Data generation, simulation designs and analysis of the simulated data

We conducted a simulation study to evaluate the performance of FCS, MNVI and EM. The ordinal simulated datasets are generated from an NB distribution using the approach described in Section 4.2 which are collapsed into ordinal responses with C categories generated at four study occasions, $j = 1, ...4$. We repeated the simulation for three different settings where C=3,4,5. We examined sample sizes of N=100,300 and 500. For NB distribution, we simulated 1000 datasets based on the generalised linear mixed model of the form (10). Furthermore, the model for longitudinal ordinal outcomes were based on the logit link function and the linear predictor below:

$$logit[P(Y_{ij}^{*} \leq c)] = \beta_0 + x'\beta + b_i, \;\; b_i \sim N(0,d), i = 1, ..., N. \tag{4.10}$$

The underlying latent variable (y^{*}) is assumed to be related to the observed response through the 'threshold concept' as indicated by the ordinal regression model. The response

is given based on the underlying unobserved categorical endpoint that follows a linear regression model which incorporates random effects and a predefined set of cut-off values (threshold values) α_C. Four different covariates were simulated: x_1 normally distributed with mean equals to 2 and σ set to 0.7, x_2 from Bernoulli distribution with probability of success equals to 0.5, x_3 from a uniform distribution and x_4 is a four-point assessment time. We used the following parameters for the purpose of simulation: $\beta_1 = 0.75$, $\beta_2 = 0.20$ $\beta_3 = 0.90$ and $\beta_4 = 0.40$. In our study, we did not use any interaction terms. The proportional odds logistic regression with random intercept for the response can be written as follows:

$$logit[P(Y_{ij}^* \leq c)] = \beta_0 + 0.75x_1 + 0.20x_2 + 0.90x_3 + 0.40x_4 + b_i.\ i = 1,...,N;\ j = 1,2,3,4. \tag{4.11}$$

When the logit link function is inverted it leads to the conditional ordinal logistic regression model whose equation can be written as

$$P(Y_{ij}^* \leq c) = \frac{exp(\beta_0 + x'\beta + b_i)}{1 + exp(\beta_0 + x'\beta + b_i)}. \tag{4.12}$$

Let $\phi_{ijc} = P(Y_{ij}^* \leq c)$, then ordinal outcome Y_{ij} (e.g., for $C = 4$) was obtained by setting the observation rule defined as

$$Y = \begin{cases} 0, & \text{if } \phi_{ij} \leq \tau_1, \\ 1, & \text{if } \tau_1 < \phi_{ij} \leq \tau_2, \\ 2, & \text{if } \tau_2 < \phi_{ij} \leq \tau_3, \\ 3, & \text{if } \phi_{ij} > \tau_3. \end{cases} \tag{4.13}$$

We imposed missing values on the full datasets, after which parameter estimates and standard errors were estimated by a likelihood based approach. Each estimate is a mean of 1000 estimates from the different simulated datasets. Here, we assumed the MAR mechanism in the missingness model, we adopted the approach of (Kombo et al., [122]) where individuals whose outcome was greater than some cut-off would miss at post baseline time points 2,3 and 4, that is, let dropout= $y_{ij} - y_{ij-1},\ j = 2,3,4$, producing values between 0 and 96; 0

and 107; 0 and 135 for the different choices of the categories of the ordinal outcome, that is C= 3,4 and 5 categories respectively. We ensured that approximately 30% of the outcome were missing and the correlation is 0.2. The probability of a value dropping depended on the immediate past history. Using the PROC MI approach, we imputed the missing entries using the EM, FCS and MVNI methods. The imputation of ordinal values were on continuous scale and rounded off to the specified categories under the MVNI method. We specified the maximum and minimum values necessary so as not create imputation values out of values. In the FCS procedure the fcs statement in PROC MI is used. The ordinal response was imputed using the ordinal logistic regression model as incorporated in the FCS approach. For each of this study, default for MCMC and FCS specification were used in the simulations. The algorithms converged to the correct posterior distributions and we had confidence that the imputed values in the various datasets were statistically independent. All the covariates were used to ensure that our imputation model was rich enough to try and satisfy the congeniality requirement under the MAR assumption. A larger number of imputations are needed for comparison of methods (Wood et al., [118]). Here, we performed $m = 20$ imputations. We chose the value to account for the missing values we generated and limit the loss of power for testing any associations of interest. Nonetheless, experts argue that m can be set to $3 \leq m \leq 5$ and still obtain accurate estimates. However, caution was raised that pegging on this range may be unnecessary (Schafer, [45]. In addition, (Molenberghs and Verbeke, [33]) showed that efficiency increments diminish rapidly after the first $m = 2$ imputations for a small fraction of missing information and after the first $m = 5$ imputations for larger fractions of missing information. The rule of choosing m was suggested by (White et al., [119]) where it is recommended that m should be at least equal to the percentage of the subject with missing values.

To asses the performance of the methods, we used standard errors (Std err), bias and mean squared error (MSE) of the estimates. Bias is defined as the absolute difference between the average parameter estimate from the analysis procedures (based on the 1000 simulated datasets) and the true value (i.e Bias= $| \bar{\hat{\beta}} - \beta |$).

4.4.2 Simulation results

We present the results of the simulation study in three Tables. Thus, Tables 4.1, 4.2 and 4.3 display results when the ordinal response variable has three, four and five levels respectively. In each Table, we compare and contrast the performance of the three methods; EM, FCS and MVNI under standard errors, bias and mean squared errors using different sample sizes. The sample sizes are N=100 300 and 500 respectively. The estimates with worst performance are shown in bold.

From Table 4.1, when the sample size is 100, MVNI performed better than EM and FCS in terms of standard errors. Specifically, errors were large in the estimates of EM and FCS than MVNI. In EM and FCS, lager errors were seen in the estimates (β_0 to β_4) and (β_1 and β_3) respectively. Bias was smallest in the estimates of FCS than EM and MVNI. In particular, the worst performance of EM and MVNI on bias pervaded through the estimates (β_1, β_2 and β_4) and (β_0 and β_3) which indicated a difference between average and true parameter. In terms of MSE, FCS outperformed EM and MVNI because it displayed smallest estimates. Under EM and MVNI, the worst MSE were on estimates (β_1, β_2 and β_4) and (β_0 and β_3) respectively.

Considering the sample size 300, under standard errors, FCS and MVNI performed better than EM. In details, the worst performance of EM on standard errors were obvious through the estimates (β_0 to β_4). In terms of bias, EM performed better than FCS and MVNI. The worst performance of FCS and MVNI on bias displayed through the estimates (β_1 and β_2) and (β_0 and β_3), and except β_4 for EM. Thus, FCS and MVNI performed better than EM in terms of MSE criterion because they were associated with smaller estimates. The worst performance of EM pervaded through the estimates (β_2 to β_4), except $\beta 1$ and β_0 for FCS and MVNI.

For sample size 500, based on standard errors, MVNI performed better than EM and FCS because MVNI produced the smallest estimates. The worst performance of EM and FCS on errors obvious through the estimates (β_0, β_1 and β_3) and (β_1, β_2 and β_4) respectively.

The estimates associated with EM and FCS were less biased than MVNI. Thus, MVNI estimates showed most bias in (β_0 to β_4), except β_4 and β_1 for EM and FCS. With respect to MSE, FCS estimates performed better than EM and MVNI. In particular, the worst performance were seen the estimates (β_2 and β_4) and (β_0 and β_3) for EM and MVNI, except β_1 for FCS.

In Table 4.2, under the sample size 100, estimates for standard errors were large in FCS and MVNI than EM. In details, the worst performance of FCS and MVNI on standard errors permeated through the estimates (β_3 and β_4) and (β_1 and β_2), except β_0 for EM. In terms of bias, EM and MVNI performed better than FCS. In particular, the worst performance of FCS on bias observed through the estimates (β_0 to β_4). The performance of EM was better than FCS and MVNI in terms of MSE. The worse performance were evident on the estimates (β_0, β_3 and β_4) and (β_1 and β_2) for FCS and MVNI respectively.

For sample size 300, EM and FCS performed better than MVNI in terms of standard errors. Specifically, the worst errors associated with MVNI through the estimates (β_0 to β_3), except β_3 for EM and β_4 for FCS. Bias were smaller in the estimates of EM and MVNI than FCS. In particular, the worst performance of FCS on bias were shown in the estimates (β_0, β_3 and β_4), except β_1 for EM and β_2 for MVNI respectively. In terms of MSE, EM and MVNI performed better than FCS. Particularly, the worst performance of EM recorded through the estimates (β_0, β_3 and β_4), and except (β_1) and (β_2) for EM and MVNI respectively

For sample size 500, the standard errors recorded for EM and FCS were smaller than MVNI. This indicated that EM and FCS performed better than MVNI. In details, the worst performance of MVNI on standard errors observed through the estimates (β_2, β_3 and β_4), and except β_0 and β_1 for EM and FCS respectively. In terms of bias, MVNI performed better than EM and FCS. In particular, the worse performance of EM and FCS permeated through the estimates (β_0, β_2 and β_3) and (β_1 and β_4) respectively. Under MSE, FCS and MVNI performed better than EM. The worst performance of EM on MSE pervaded through the estimates (β_0, β_2 and β_3), except β_1 and β_4 for FCS and MVNI respectively.

Table 4.1: Standard errors (Std Err), bias and mean squared error (MSE) estimates from Expectation maximization (EM), fully conditional specification (FCS) and multivariate normal (MVN), under MAR mechanism over 1000 samples: *N*=100, 300 and 500 individuals, for a 3-category ordinal outcome.

Sample	Par	Std Err			Bias			MSE		
		EM	FCS	MVNI	EM	FCS	MVNI	EM	FCS	MVNI
100	β_0	**0.0787**	0.0761	0.0432	0.1316	0.0856	**0.2594**	0.0960	0.0834	**0.1105**
	β_1	**0.0111**	**0.0111**	0.0074	**-0.7522**	-0.7520	-0.7496	**0.5769**	0.5766	0.5693
	β_2	**0.0161**	0.0154	0.0102	**-0.2039**	-0.2017	-0.2003	**0.0577**	0.0561	0.0503
	β_3	**0.0029**	**0.0029**	0.0016	-0.9007	-0.8999	**-0.9016**	0.8142	0.8127	**0.8145**
	β_4	**0.0076**	0.0074	0.0045	**-0.4832**	-0.4703	-0.4694	**0.2411**	0.2286	0.2248
300	β_0	**0.0790**	0.0743	0.0439	0.1411	0.0065	**0.2609**	0.0989	0.0743	**0.1120**
	β_1	**0.0120**	0.0112	0.0071	-0.7520	**-0.7537**	-0.7416	0.5775	**0.5793**	0.5571
	β_2	**0.0171**	0.0152	0.0097	-0.1969	**-0.2025**	-0.1937	**0.0559**	0.0562	0.0472
	β_3	**0.0030**	0.0027	0.0016	-0.9011	-0.9008	**-0.9018**	**0.8150**	0.8141	0.8148
	β_4	**0.0075**	0.0070	0.0043	**-0.4847**	-0.4720	-0.4739	**0.2424**	0.2298	0.2289
500	β_0	**0.0852**	0.0765	0.0435	0.0833	0.0653	**0.2482**	0.0921	0.0808	**0.1051**
	β_1	**0.0124**	**0.0124**	0.0069	-0.7373	**-0.7398**	**-0.7398**	0.5560	**0.5597**	0.5542
	β_2	0.0124	**0.0166**	0.0094	-0.2063	-0.2032	**-0.2108**	**0.0593**	0.0579	0.0538
	β_3	**0.0030**	0.0028	0.0015	-0.9000	-0.9001	**-0.9013**	0.8130	0.8130	**0.8138**
	β_4	0.0072	**0.0074**	0.0041	**-0.4815**	-0.4687	-0.4728	**0.2390**	0.2271	0.2276

Notes: The missing values are approximately (30%) on the outcome variable; MAR mechanism.

In Table 4.3, considering the sample size 100, EM and MVNI performed better than FCS in terms of standard errors. The worst performance of FCS on standard errors filtered through the estimates (β_0, β_2 and β_4), except β_3 and β_1 for EM and MVNI respectively. Bias were smaller in the estimates of EM and MVNI than FCS. Specifically,the worst performance of FCS on bias run through the estimates (β_0, β_2 and β_4), and except β_3 and β_1 for EM and

Table 4.2: Standard errors (Std Err), Bias and mean squared error (MSE) estimates from Expectation maximization (EM), fully conditional specification (FCS) and multivariate normal (MVN), under MAR mechanism over 1000 samples: *N*=100, 300 and 500 individuals, for a 4-category ordinal outcome.

Sample	Par	Std Err			Bias			MSE		
		EM	FCS	MVNI	EM	FCS	MVNI	EM	FCS	MVNI
100	β_0	**0.1584**	0.1540	0.1449	0.4584	**0.4932**	0.4276	0.3685	**0.3972**	0.3277
	β_1	0.0221	0.0216	**0.0250**	-0.7160	**-0.7208**	-0.7186	0.5348	0.5412	**0.5414**
	β_2	0.0328	0.0301	**0.0344**	-0.2425	**-0.2463**	-0.2426	0.0916	0.0908	**0.0933**
	β_3	0.0061	**0.0545**	0.0052	-0.8984	**-0.8985**	-0.8967	0.8132	**0.8618**	0.8093
	β_4	0.0139	**0.0163**	0.0146	-0.3156	**-0.3270**	-0.3179	0.1135	**0.1232**	0.1157
300	β_0	0.1538	0.1510	**0.1607**	0.4730	**0.5134**	0.4887	0.3775	**0.4146**	0.3995
	β_1	0.0213	0.0231	**0.0241**	**-0.7604**	-0.7568	-0.7568	**0.5995**	0.5958	0.5968
	β_2	0.0309	0.0320	**0.0352**	-0.2084	-0.1946	**-0.2156**	0.0743	0.0699	**0.0817**
	β_3	**0.0059**	0.0057	**0.0059**	-0.8954	**-0.8965**	-0.8959	0.8076	**0.8094**	0.8085
	β_4	0.0134	**0.0167**	0.0136	-0.3221	**-0.3375**	-0.3267	0.1171	**0.1306**	0.1203
500	β_0	**0.1556**	0.1513	0.1520	**0.4961**	0.4860	0.4662	**0.4017**	0.3875	0.3693
	β_1	0.0220	**0.0231**	0.0210	-0.7353	**-0.7364**	-0.7362	0.5627	**0.5654**	0.5630
	β_2	0.0307	0.0299	**0.0316**	**-0.2293**	-0.2241	-0.2240	**0.0833**	0.0801	0.0818
	β_3	0.0056	0.0054	**0.0057**	**-0.8992**	-0.8980	-0.8978	**0.8142**	0.8118	0.8117
	β_4	0.0144	0.0140	**0.0157**	-0.3079	**-0.3195**	-0.3123	0.1092	0.1161	**0.1132**

Notes: The missing values are approximately (30%) on the outcome variable; MAR mechanism.

MVNI. In the case of MSE, EM and MVNI performed better than FCS. The worst performance of FCS on MSE imbued through the estimates (β_0, β_2 and β_4), and except β_3 and β_1 for EM and MVNI.

For sample size 300, in terms of standard errors; EM performance was better than FCS and MVNI. In particular, the worse performance of FCS and MVNI on the standard errors

Table 4.3: Standard errors (Std Err), Bias and mean squared error (MSE) estimates from Expectation maximization (EM), fully conditional specification (FCS) and multivariate normal (MVN), under MAR mechanism over 1000 samples: *N*=100, 300 and 500 individuals, for a 5-category ordinal outcome.

Sample	Par	Std Err			Bias			MSE		
		EM	FCS	MVNI	EM	FCS	MVNI	EM	FCS	MVNI
100	β_0	0.2006	**0.2032**	0.1835	0.5981	**0.6306**	0.5592	0.5583	**0.6009**	0.4962
	β_1	0.0280	0.0316	**0.0317**	-0.7013	-0.7040	**-0.7046**	0.5198	0.5272	**0.5282**
	β_2	0.0415	**0.0456**	0.0435	-0.2589	**-0.2615**	-0.2590	0.1085	**0.1140**	0.1106
	β_3	**0.0077**	0.0073	0.0066	**-0.9005**	-0.9004	-0.8983	**0.8186**	0.8180	0.8135
	β_4	0.0175	**0.0188**	0.0185	-0.2964	**-0.3114**	-0.2992	0.1054	**0.1158**	0.1080
300	β_0	0.1956	0.1938	**0.2040**	0.5402	0.5551	**0.5602**	0.4870	0.5019	**0.5178**
	β_1	0.0270	**0.0321**	0.0306	-0.7515	**-0.7523**	-0.7469	0.5918	**0.5981**	0.5885
	β_2	0.0392	0.0415	**0.0447**	-0.1960	-0.1927	**-0.2051**	0.0776	0.0786	**0.0868**
	β_3	0.0075	**0.0076**	0.0075	-0.8940	-0.8935	**-0.8947**	0.8067	0.8059	**0.8080**
	β_4	0.0169	**0.0210**	0.0173	-0.3086	**-0.3261**	-0.3144	0.1121	**0.1273**	0.1161
500	β_0	0.1967	**0.2055**	0.1922	0.6657	**0.6967**	0.6278	0.6399	**0.6909**	0.5863
	β_1	0.0278	**0.0284**	0.0265	-0.7280	**-0.7304**	-0.7291	0.5578	**0.5619**	0.5581
	β_2	0.0389	**0.0448**	0.0400	**-0.2471**	-0.2393	-0.2404	0.1000	**0.1021**	0.0978
	β_3	0.0071	**0.0075**	0.0072	**-0.9018**	-0.9016	-0.9001	0.8203	**0.8204**	0.8174
	β_4	0.0182	0.0182	**0.0199**	-0.2895	**-0.3087**	-0.2951	0.1020	**0.1135**	0.1070

Notes: The missing values are approximately (30%) on the outcome variable; MAR mechanism.

were indicative through estimates (β_1, β_3 and β_4) and (β_0 and β_2) respectively. EM performed better than FCS and MVNI in terms of bias. Specifically, the worse performance of FCS and MVNI were obvious through the estimates (β_1 and β_4) and (β_0, β_2 and β_3). Under MSE, the performance of FCS and MVNI were worse compared to EM. In details, the worse performance on FCS and MVNI on MSE indicated through the estimates (β_0 and β_4) and (β_0, β_2 and β_3).

For sample size 500, EM and MVNI performed better than FCS in terms of standard errors. In details, the worst performance of FCS on standard errors permeated through the estimates (β_0 to β_3), and except β_4 for MVNI. In the case of bias, EM and FCS performed lower than MVNI. Specifically, the worse performance of EM and FCS on bias run through the estimates (β_2 and β_3) and (β_0, β_1 and β_4). In terms of MSE, EM and MVNI performed better than FCS. In particular, the worst performance of FCS on MSE permeated through estimates (β_0 to β_4).

4.4.3 Example: Lung HIV data

Data

The dataset used is from the national heart, lung and blood institute (NHBLI) longitudinal studies of HIV associated lung infections and complications (Lung HIV), Bruce et al., [148]. The cohorts study were conducted across eight different clinical centers. Before the study, a written consent of the participants were sought. The data are on 4760 patients (2889 males and 1871 females) and 12 patients were missing. The ages are between 17 and 77. The patients were followed up for 24 months (in 5 visits) and the axial emphysema distribution assessed was graded from 0 to 4, with high indicating worse. About 47% of the description of the axial distribution emphysema variable were missing. Some of the descriptive statistics of the dataset are summarized in Table 4.4.

The proportional odds assumption

Before the interpretation of the model results, it is essential to test the proportional odds assumption. This simply tests whether the parameters are the same across logits, simultaneously for all predictors. The assumption can be examined in STATA using Brant test or SAS using the score test. We conducted the test in SAS. A significant result indicates that proportional odds does not hold and suggests that separate parameters are needed across

Table 4.4: Descriptive statistics of the incomplete Lung HIV data

Lung HIV data	Description	Range	% miss	Freq
Baseline variables				
Sex	1=Male, 2=Female	1/2	0.25	
Age	Age patient at enrollment	19-77	0	
Time	Number of patients visits	1-5	0	823
Pulmonary disease	Pulmonary artery enlargement (0=No, Yes=1)	0	0	
Response variable				
Axi_emph	Best description of the axial distribution of emphysema	0-4	47.02	387

Note: Data missing on both the dependent and independent variables.

the logits for at least one predictor. The part of assumption results is presented in Table 4.5. The model of interest of the study was the main effects model. First, we analyzed the data without any alterations or attempts to impute the missing observations. This method is termed the direct likelihood (DL) approach. The parameter estimates from DL were chosen as reference for the real application dataset to check the relative performance of EM, FCS and MVNI when considering MI. This is because DL is valid under the same properties as multiple imputation. Thereafter, we conducted multiple imputations under the EM, FCS and MVNI and upon completion a similar marginal model was fitted to analyze the task. Finally, the SAS procedure MIANALYZE was employed to pool the results from multiple datasets.

Table 4.5: Score test for the proportional odds assumption

Chi-Square	df	Pr>Chisq
52.5569	12	<.0001

Table 4.6: Parameter estimates (Est), standard errors (SE), confidence limits (CL) obtained from the lung HIV data under the methods of direct likelihood (DL), expectation maximization (EM), fully conditional specification (FCS) and multivariate normal (MVNI) under MAR mechanism.

	DL			EM			FCS			MVNI		
Param	Est	SE	$Pr > \|t\|$	Est	SE	$Pr > \|t\|$	Est	SE	$Pr > \|t\|$	Est	SE	$Pr > \|t\|$
Intercept	1.2650	0.4655	0.0069	0.8912	0.2276	0.0008	0.8951	0.2080	0.0003	0.9678	0.2182	0.0002
Time	0.1333	0.1454	0.4131	0.0267	0.0349	0.4542	0.0360	0.0336	0.2963	0.0329	0.0472	0.4943
Pulm-disease	-0.7698	0.1741	$< .0001$	0.6493	0.2299	0.0106	0.6717	0.1804	0.0013	0.6970	0.1973	0.0021
Sex	0.0223	0.1066	0.8345	-0.0236	0.1397	0.8677	-0.0226	0.0798	0.7799	-0.0540	0.0887	0.5493
Age	0.2660	0.4412	0.5484	-0.0255	0.0531	0.6359	-0.0254	0.0447	0.5758	-0.0482	0.0489	0.3365

Notes: The missing values on both covariate and response variable are approximately 0.3% and 47% respectively.

Results

Table 4.6 shows the parameter estimates, standard errors and confidence limit of fixed effect using the imputation and likelihood methods. In application, EM, FCS and MVNI performed better than DL in terms of estimates, except for pulmonary disease. In the case of standard errors, DL displayed lager errors compared to EM, FCS and MVNI. In particular, the worst performance of DL on standard errors pervaded through (intercept, time and age), except pulmonary disease and sex for EM. Pulmonary disease is the only parameter that was significant under all the methods. This affirmed that pulmonary disease was common among the individuals that participated in the HIV-lung infection study.

4.5 Discussion

In our research, we have introduced a novel idea in the area of methodology where we simulated a discrete variable from the NB distribution and linked with the underlying ordinal responses, through the cumulative probabilities. This distribution is useful in analyzing

categorical data and has many response categories which represent the ranges of underlying count. In real application, it is common for analysts to be faced with both dropout and nonmissingness missing datasets, like the Lung-HIV data where the amount of dropout was high, while that of nonmissingness is much small. It is important to include all in the analyses as noted by (Molenberghs and Verbeke, [33]). One can opt for direct likelihood analysis or standard generalized estimating equations (GEE; Molenberghs and Verbeke [33], Diggle et al., [149], Liang and Zeger [61]). Weighted generalized estimating equations (WGEE; Robins et al., [79]) are possible but one has to find appropriate weights. On the other hand, one may make the missing patterns monotone by multiple imputation and go ahead to do the WGEE (Kombo et al., [122]). This study was to investigate and explore the performance of the EM, FCS and MVNI as MI methods. Each of the approach follows different theoretical assumptions, which involves different computational methods. In MVNI and EM, the specification of the imputation model is easy to use. But FCS requires an additional effort in model specification, and separate regression models must be fitted for each variable in the imputation model (Van Burren [144]). But in our situation, the conditional regressions were automatically specified because of the small number of variables involved and just two variables had missing values. The advantage of FCS is that it can handle ordinal variables naturally.

In this paper, missingness was on covariate and response variable, which were analyzed under three approaches of the MI methods. This study shows that FCS performs slightly fair than the other imputation methods used. This study stands as an extension to a study conducted by (Kombo et al., [122]) where missingness is on the outcome variable. Furthermore, in the study conducted in (Mcginley et al., [139]) missing data and different cut-points selections were not included. However, in our study we introduced missing data in the simulation studies and methods of handling it. Thus, we also introduced different cut-points selections. The approaches are extension to the study conducted in Mcginley et al. [139]. Actually, this does not limit the findings of this research; but further research should

be conducted to investigate the performance of ordinal-count models under misspecification of the latent response variable distribution, and alternative generating distributions.

Chapter 5

Conclusion and possible areas of further research

5.1 Conclusion

The focus of this research is to analyze both longitudinal and cross- sectional data, with interest on incomplete data. The study comprised both simulation and real data applications. In the real applications, different datasets are used; chronic obstructive pulmonary disease (COPD) dataset - a primary binary response dataset available in Excellence in Cardiovascular Health for the Southern America (CESCAS), United States of America 2010. Furthermore, HIV associated lung infection and complications (Lung-HIV) dataset - a clinical study trial provided by the national heart, lung and blood institute (NHBLI) 2013. Both data sets exhibit both monotone and non-monotone missing data patterns. One modeling framework: the marginal model was used due to the nature of the datasets.

In the analysis of incomplete longitudinal data, non-Gaussian data can be adopted to some extent under the less strict missing at random (MAR) assumption using standard statistical software packages. Thus, likelihood based approaches such as linear mixed models and generalized linear mixed models can be adopted for a Gaussian and non-Gaussian data respectively. Weighted generalized estimating equations can be used as an alternative. This is valid under MAR because of weighting. Furthermore, other methods can be used that do not need an explicit joint model for the missing data and the substantive analysis model, like

the expectation-maximization algorithm and multiple imputation (MI) and its extensions, MI-GEE and DR-GEE. All of these techniques can be implemented using SAS and other statistical packages. In SAS, the GLIMMIX macro and GLIMMIX procedure together with the GENMOD and GEE procedures are suitable for the generalized estimating equations. The NLIMIXED procedure also with the GLIMMIX can be used for the generalized linear mixed model. The application of these techniques and its strategies leave no reason for the use of ad hoc methods which are highly restrictive, such as complete case analysis, last observation carried forward, baseline observation carried forward, simple imputation techniques - it is actually simple and easy to implement. In this thesis, we are interested on investigating and comparing the performance of the analysis methods for incomplete data. This research was basically on how to deal with missing observations in either the covariates/outcome or both. This thesis started in Chapter 2, with an incomplete binary longitudinal outcome. Using a simulation study and life data application, the analysis methods compared the weighted generalized estimating equation (WGEE), generalized estimating equations based multiple imputation (MI-GEE) and generalized estimating equations based Doubly robust (DR-GEE), under different independent working correlation structures, such as exchangeable compound symmetry (CS), first order autoregressive (AR1) and toeplitz (TOEP). Furthermore, DR-GEE performed better than the other methods especially under TOEP correlation structure. In the simulation study, DR-GEE estimators are consistent and present small sample bias and smallest standard errors compared to the other techniques used..

Chapter 3 evaluated the performance of inverse probability weighting (IPW) and multiple imputation using fully conditional specification MI (FCS). These methods were assessed through a simulation study and a real data application. In the simulation study, datasets of different sample sizes ($N = 100, 200, 500$) were generated. This gave insights to the strength of the methods and what to expect when using real data. In the real life data application, the two methods performed creditably well, but MI (FCS) gave more reliable inferences.

Chapter 4 compared the performance of multiple imputation strategies, namely, multivariate normal imputation, fully specification condition and expectation-maximization algorithm applied to a discrete ordinal data. These methods were evaluated via simulation studies and a real data application example. In the simulation study, we explicitly used ordinal responses to a suitable underlying count distribution through NB. Furthermore, datasets of different sample sizes $(N = 100, 300, 500)$ were generated from negative binomial distribution which were collapsed into ordinal responses with different cut-points selections $(C = 3, 4, 5)$. The real application involved the HIV-Lung dataset. Incomplete data were present on both the covariate and outcome variable. The results showed that FCS displayed estimates that were slightly smaller than other methods.

5.2 Possible areas of further research

In the study design for data collection, a statistician should be among the designers to develop strategies that reduce missing data since the collection of data plays a prominent role in the problem of incomplete data for a specific study. This indicates that strategic planning can minimize the frequency of missing data; although there is no specified condition concerning the amount of incomplete data that can be acceptable. In the context of missing data, knowing the reasons leading to the incompleteness plays a pertinent role in choosing the appropriate statistical approach to handle the data. In fact, there is no general procedure for handling the incomplete data scenarios. However, proper study design and understanding the possible causes of missing data mechanism can assist to a considerable extent.
In this research, we investigated various modeling techniques with both monotone and non-monotone missingness pattern respectively. However, there are still some areas that need to be researched in the missing data scenario. These areas could serve as areas of further research. A research could still be conducted to investigate sensitivity analysis since it is of

importance to model MNAR incomplete longitudinal data. In this context, various sensitivity analysis frameworks can be compared. But our current study did not focus of shared parameter models and local and global influence which are other alternative techniques. In using identifying restrictions, our focus is limited on the continuous data setting. Research on the extension of this needed especially on the categorical data.

Finally, we acknowledge that not all the areas in missing data were covered. Other areas that need further research include the identifiability issues for local and global influence techniques, multiple imputation for recurrent event data and sensitivity of inference to data transformations. Furthermore, sensitivity analysis should be conducted on the latent-class mixture models which are an extension of the shared parameter mixture. This can also be used as another technique for sensitivity analysis.

References

[1] P. D. Allison. Multiple imputation for missing data: A cautionary tale. *Sociological methods & research*, 28(3):301–309, 2000.

[2] R. L. Carter. Solutions for missing data in structural equation modeling. *Research & Practice in Assessment*, 1, 2006.

[3] W. C. Regoeczi and M. Riedel. The application of missing data estimation models to the problem of unknown victim/offender relationships in homicide cases. *Journal of Quantitative Criminology*, 19(2):155–183, 2003.

[4] T. Rudas. Mixture models of missing data. *Quality & quantity*, 39(1):19–36, 2005.

[5] S. A. Stumpf. A note on handling missing data. *Journal of Management*, 4(1):65–73, 1978.

[6] P. E. McKnight, K. M. McKnight, S. Sidani, and A. J. Figueredo. *Missing data: A gentle introduction*. Guilford Press, 2007.

[7] J. Barnard and X. Meng. Applications of multiple imputation in medical studies: from aids to nhanes. *Statistical methods in medical research*, 8(1):17–36, 1999.

[8] E. D. De Leeuw and M. Huisman. Prevention and treatment of item nonresponse. *Journal of Official Statistics*, 19(2):153, 2003.

[9] C. M. Musil, C. B. Warner, P. K. Yobas, and S. L. Jones. A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research*, 24(7): 815–829, 2002.

[10] D. L. Streiner. The case of the missing data: methods of dealing with dropouts and other research vagaries. *The Canadian Journal of Psychiatry*, 47(1):70–77, 2002.

[11] L. M. Friedman, C. Furberg, D. L. DeMets, D. M. Reboussin, and C. B. Granger. *Fundamentals of clinical trials*, volume 3. Springer, 1998.

[12] S. Green, J. Benedetti, A. Smith, and J. Crowley. *Clinical trials in oncology*, volume 28. CRC press, 2012.

[13] S. Piantadosi. *Clinical trials: a methodologic perspective*. John Wiley & Sons, 2017.

[14] H. A. Kahn and C. T. Sempos. *Statistical methods in epidemiology*, volume 12. Monographs in Epidemiology & B, 1989.

[15] D. Clayton and M. Hills. *Statistical methods in epidemiology*. Oxford University Press, 1993.

[16] D. E. Lilienfeld and P. D. Stolley. *Foundations of epidemiology*. Oxford University Press, USA, 1994.

[17] S. Selvin. *Statistical analysis of epidemiologic data*. Oxford University Press, 2004.

[18] J. L. Schafer, M. Khare, and T. M. Ezzati-Rice. Multiple imputation of missing data in nhanes iii. In *Proceedings of the Annual Research Conference*, pages 459–487, 1993.

[19] D. B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[20] D. B. Rubin, H. S. Stern, and V. Vehovar. Handling “don’t know" survey responses: the case of the slovenian plebiscite. *Journal of the American Statistical Association*, 90(431):822–828, 1995.

[21] G. Molenberghs and G. Verbeke. Linear mixed models for gaussian longitudinal data. *Models for Discrete Longitudinal Data*, pages 35–43, 2005.

[22] A. A. Afifi and R. M. Elashoff. Missing observations in multivariate statistics i. review of the literature. *Journal of the American Statistical Association*, 61(315): 595–604, 1966.

[23] H. O. Hartley and R. R. Hocking. The analysis of incomplete data. *Biometrics*, pages 783–823, 1971.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[25] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987.

[26] J. L. Schafer and J. W. Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[27] P. D. Allison. Missing data: Sage university papers series on quantitative applications in the social sciences (07–136). *Thousand Oaks, CA*, 2001.

[28] E. D. De Leeuw. Reducing missing data in surveys: An overview of methods. *Quality & Quantity*, 35(2):147–160, 2001.

[29] R. J. A. Little and D. B. Rubin. Statistical analysis with missing data. 2002, hoboken.

[30] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[31] R. J. A. Little and D. B. Rubin. Statistical analysis with missing data, 2002.

[32] G. Molenberghs, C. Beunckens, C. Sotto, and M. G. Kenward. Every missing not at random model has got a missing at random bodyguard. *Journal of the Royal Statistical Society, Series B, Submitted*, 2007.

[33] G. Molenberghs and G. Verbeke. Models for discrete longitudinal data. 2005.

[34] G. Molenberghs and M. Kenward. *Missing data in clinical studies*, volume 61. John Wiley & Sons, 2007.

[35] R. J. A Little. Statistical analysis with missing data. *John A. Wiley & Sons, Inc., New York*, 85.

[36] H. Thijs, G. Molenberghs, B. Michiels, G. Verbeke, and D. Curran. Strategies to fit pattern-mixture models. *Biostatistics*, 3(2):245–265, 2002.

[37] T. R. Belin, G. J. Diffendal, S. Mack, D. B. Rubin, J. L. Schafer, and A. M. Zaslavsky. Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation. *Journal of the American Statistical Association*, 88(423):1149–1159, 1993.

[38] K. W. Wachter. Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation: Comment: Ignoring nonignorable effects. *Journal of the American Statistical Association*, 88(423):1161–1163, 1993.

[39] H. Demirtas and J. L. Schafer. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in medicine*, 22(16):2553–2575, 2003.

[40] G. Verbeke and G. Molenberghs. A model for longitudinal data. *Linear mixed models for longitudinal data*, pages 19–29, 2000.

[41] J. L. Schafer and M. K. Olsen. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4): 545–571, 1998.

[42] M. J. Van der Laan and J. M. Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

[43] A. Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[44] D. B. Rubin. Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association, 1978.

[45] J. L. Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

[46] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[47] G. Verbeke and G. Molenberghs. Linear mixed models for longitudinal data. In *Linear mixed models for longitudinal data.*, pages 240–268. Springer-Verlag New York, 2000.

[48] P. Diggle and M. G. Kenward. Informative drop-out in longitudinal data analysis. *Applied statistics*, pages 49–93, 1994.

[49] R. J. A. Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121, 1995.

[50] N. H. Nie, D. H. Bent, and C. H. Hull. Spss: Statistical package for the social sciences. Technical report, McGraw-Hill New York, 1970.

[51] C. H. Brown. Asymptotic comparison of missing data procedures for estimating factor loadings. *Psychometrika*, 48(2):269–291, 1983.

[52] M. G. Kenward, E. Lesaffre, and G. Molenberghs. An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics*, pages 945–953, 1994.

[53] F. Wang-Clow, M. Lange, N. M. Laird, and J. H. Ware. A simulation study of estimators for rates of change in longitudinal studies with attrition. *Statistics in Medicine*, 14(3):283–297, 1995.

[54] J. L. Arbuckle. Full information estimation in the presence of incomplete data. *Advanced structural equation modeling: Issues and techniques*, 243:277, 1996.

[55] T. A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233, 1982.

[56] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[57] D. B. Rubin. Em and beyond. *Psychometrika*, 56(2):241–254, 1991.

[58] S. G. Baker. A simple method for computing the observed information matrix when using the em algorithm with categorical data. *Journal of Computational and Graphical Statistics*, 1(1):63–76, 1992.

[59] W. D. Flanders and S. Greenland. Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, 10(5):739–747, 1991.

[60] L. P. Zhao and S. Lipsitz. Designs and analysis of two-stage studies. *Statistics in medicine*, 11(6):769–782, 1992.

[61] K. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[62] A. Rotnitzky, J. M. Robins, and D. O. Scharfstein. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association*, 93(444):1321–1339, 1998.

[63] M. D. Schluchter and K. L. Jackson. Log-linear analysis of censored survival data with partially observed covariates. *Journal of the American Statistical Association*, 84(405):42–52, 1989.

[64] J. G. Ibrahim. Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85(411):765–769, 1990.

[65] S. R. Lipsitz and J. G. Ibrahim. Using the em-algorithm for survival data with incomplete categorical covariates. *Lifetime Data Analysis*, 2(1):5–14, 1996.

[66] S. R. Lipsitz and J. G. Ibrahim. Estimating equations with incomplete categorical covariates in the cox model. *Biometrics*, pages 1002–1013, 1998.

[67] N. J. Horton and N. M. Laird. Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research*, 8(1):37–50, 1999.

[68] J. R. Carpenter and M. G. Kenward. A comparision of multiple imputation and inverse probability weighting for analysis with missing data. *Journal of the Royal Statistical Society, Series A*, 2005.

[69] S. R. Seaman and I. R. White. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, page 0962280210395740, 2011.

[70] G. Molenberghs, M. G. Kenward, and E. Lesaffre. The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, 84(1):33–44, 1997.

[71] B. Michiels, G. Molenberghs, and S. R. Lipsitz. Selection models and pattern-mixture models for incomplete data with covariates. *Biometrics*, 55(3):978–983, 1999.

[72] R. J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

[73] R. J. A. Little. A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3):471–483, 1994.

[74] G. Molenberghs, H. Thijs, I. Jansen, C. Beunckens, M. G. Kenward, C. Mallinckrodt, and R. J. Carroll. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, 5(3):445–464, 2004.

[75] W. Vach and M. Blettner. Logistic regression with incompletely observed categorical covariates-investigating the sensitivity against violation of the missing at random assumption. *Statistics in Medicine*, 14(12):1315–1329, 1995.

[76] G. Molenberghs, M. G. Kenward, and E. Goetghebeur. Sensitivity analysis for incomplete contingency tables: the slovenian plebiscite case. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(1):15–29, 2001.

[77] G. Verbeke, E. Lesaffre, and B. Spiessens. The practical use of different strategies to handle dropout in longitudinal studies. *Drug Information Journal*, 35(2):419–434, 2001.

[78] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. *Applied longitudinal analysis*, volume 998. John Wiley & Sons, 2012.

[79] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121, 1995.

[80] H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[81] J. R. Carpenter, M. G. Kenward, and S. Vansteelandt. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):571–584, 2006.

[82] S. Jolani, L. E. Frank, and S. Buuren. Dual imputation model for incomplete longitudinal data. *British Journal of Mathematical and Statistical Psychology*, 67(2): 197–212, 2014.

[83] J. L. P. da Silva, E. A. Colosimo, and F. N. Demarqui. Doubly robust-based generalized estimating equations for the analysis of longitudinal ordinal missing data. *arXiv preprint arXiv:1506.04451*, 2015.

[84] A. A. Tsiatis and M. Davidian. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 22(4):569, 2007.

[85] D. V. Mehrotra, X. Li, J. Liu, and K. Lu. Analysis of longitudinal clinical trials with missing data using multiple imputation in conjunction with robust regression. *Biometrics*, 68(4):1250–1259, 2012.

[86] C. Beunckens, C. Sotto, and G. Molenberghs. A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational statistics & data analysis*, 52(3):1533–1548, 2008.

[87] A. Satty, H. Mwambi, and G. Molenberghs. Different methods for handling incomplete longitudinal binary outcome due to missing at random dropout. *Statistical Methodology*, 24:12–27, 2015.

[88] A. Y. Toledano and C. Gatsonis. Generalized estimating equations for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate. *Biometrics*, 55(2):488–496, 1999.

[89] A. Donneau, M. Mauer, G. Molenberghs, and A. Albert. A simulation study comparing multiple imputation methods for incomplete longitudinal ordinal data. *Communications in Statistics-Simulation and Computation*, 44(5):1311–1338, 2015.

[90] G. Lin and R. N. Rodriguez. Weighted methods for analyzing missing data with the gee and causaltrt procedures. In *Proceedings of the SAS Global Forum 2014 Conference. URL http://support. sas. com/resources/papers/proceedings14/SAS166-2014. pdf*, 2014.

[91] J. S. Preisser, K. K. Lohman, and P. J. Rathouz. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in medicine*, 21(20):3035–3054, 2002.

[92] J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429): 122–129, 1995.

[93] G. M. Fitzmaurice, G. Molenberghs, and S. R. Lipsitz. Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 691–704, 1995.

[94] R. J. A Little and D. B. Rubin. Statistical analysis with missing data, 1987.

[95] X. Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, pages 538–558, 1994.

[96] S. Vansteelandt, J. Carpenter, and M. G. Kenward. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6 (1):37, 2010.

[97] K. Vermeulen and S. Vansteelandt. Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036, 2015.

[98] R. Scharfstein. Adjusting for nonignorable drop-out using semiparametric nonresponse models, with comments and rejoinder. *Journal of the American Statistical Association*, 94(448):1096–1146, 1999.

[99] R. Huang and K. C. Carriere. Comparison of methods for incomplete repeated measures data analysis in small samples. *Journal of statistical planning and inference*, 136(1):235–247, 2006.

[100] D. Clayton, D. Spiegelhalter, G. Dunn, and A. Pickles. Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):71–87, 1998.

[101] J. A.C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338:b2393, 2009.

[102] A. N. Baraldi and C. K. Enders. An introduction to modern missing data analyses. *Journal of school psychology*, 48(1):5–37, 2010.

[103] J. M. Brick and G. Kalton. Handling missing data in survey research. *Statistical methods in medical research*, 5(3):215–238, 1996.

[104] S. Lohr. *Sampling: design and analysis*. Nelson Education, 2009.

[105] A. Chinomona and H. Mwambi. Multiple imputation for non-response when estimating hiv prevalence using survey data. *BMC public health*, 15(1):1059, 2015.

[106] S. G. Heeringa, B. T. West, and P. A. Berglund. *Applied survey data analysis*. CRC Press, 2010.

[107] T. D. Pigott. A review of methods for missing data. *Educational research and evaluation*, 7(4):353–383, 2001.

[108] P. A. Berglund. Multiple imputation using the fully conditional specification method: a comparison of sas®, stata, iveware, and r. In *Proceedings of the SAS Global Forum 2015 Conference. Cary, NC: SAS Institute Inc*, pages 2081–2015, 2015.

[109] S. R. Seaman and I. R. White. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295, 2013.

[110] A. Bardach, L. Colantonio, J. Galanate, L. Gutierrez, and T. Poggio. Detection and follow-up of cardiovascular disease and risk factors in the southern cone of latin america. Centre of excellence in cardiovascular health for the southern cone, 2010.

[111] K. E. Wirth, E. J. T. Tchetgen, and M. Murray. Adjustment for missing data in complex surveys using doubly robust estimation: application to commercial sexual contact among indian men. *Epidemiology (Cambridge, Mass.)*, 21(6):863, 2010.

[112] M. A. Hernán, S. Hernández-Díaz, and J. M. Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, 2004.

[113] C. G. Moore, S. R. Lipsitz, C. L. Addy, J. R. Hussey, G. Fitzmaurice, and S. Natarajan. Logistic regression with incomplete covariate data in complex survey sampling: application of reweighted estimating equations. *Epidemiology*, 20(3):382–390, 2009.

[114] P. McCullagh and J. A. Nelder. Generalized linear models. *Journal of the Royal Statistical Society Series A*, 1972.

[115] P. McCullagh and J. A. Nelder. Generalized linear models, no. 37 in monograph on statistics and applied probability, 1989.

[116] P. A. Berglund and S. G. Heeringa. *Multiple imputation of missing data using SAS*. SAS Institute, 2014.

[117] S. Greenland and W. D. Finkle. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American journal of epidemiology*, 142(12):1255–1264, 1995.

[118] A. M. Wood, I. R. White, M. Hillsdon, and J. R. Carpenter. Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. *International journal of epidemiology*, 34(1):89–99, 2004.

[119] I. R. White, P. Royston, and A. M. Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2011.

[120] S. Van Buuren. *Flexible imputation of missing data*. CRC press, 2012.

[121] A. Ivanova, G. Molenberghs, and G. Verbeke. Mechanism for missing data incorporated in joint modelling of ordinal responses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2016.

[122] A. Y. Kombo, H. Mwambi, and G. Molenberghs. Multiple imputation for ordinal longitudinal data with monotone missing data patterns. *Journal of Applied Statistics*, 44(2):270–287, 2017.

[123] M. Ursino and M. Gasparini. A new parsimonious model for ordinal longitudinal data with application to subjective evaluations of a gastrointestinal disease. *Statistical methods in medical research*, page 0962280216661370, 2016.

[124] A. Agresti. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons, 2010.

[125] R. D. Bock and J. V. Jones. The measurement and prediction of judgment and choice. 1968.

[126] E. J. Snell. A scaling procedure for ordered categorical data. *Biometrics*, pages 592–607, 1964.

[127] P. McCullagh. Regression models for ordinal data (with discussion). *J. Roy. Statist. Soc., B*, 42:109–142, 1980.

[128] L. A. Goodman. Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74(367):537–552, 1979.

[129] G. Muniz-Terrera, A. van den. Hout, R. A. Rigby, and D. M. Stasinopoulos. Analysing cognitive test data: Distributions and non-parametric random effects. *Statistical methods in medical research*, 25(2):741–753, 2016.

[130] A. D'Elia and D. Piccolo. A mixture model for preferences data analysis. *Computational Statistics & Data Analysis*, 49(3):917–934, 2005.

[131] D. A. Dawson. Methodological issues in measuring alcohol use. *Alcohol Research and Health*, 27(1):18–29, 2003.

[132] J. R. Carpenter and M. G. Kenward. *Multiple imputation and its application*. John Wiley & Sons, 2012.

[133] S. Van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219–242, 2007.

[134] K. Choi, S. E. Gregorich, C. Hoff, O. Grinstead, C. Gomez, and W. Hussey. The efficacy of female condom skills training in hiv risk reduction among women: a randomized controlled trial. *American Journal of Public Health*, 98(10):1841–1848, 2008.

[135] H. Demirtas and D. Hedeker. An imputation strategy for incomplete longitudinal ordinal data. *Statistics in medicine*, 27(20):4086–4093, 2008.

[136] R. L. Seitzman, V. B. Mahajan, C. Mangione, J. A. Cauley, K. E. Ensrud, K. L. Stone, S. R. Cummings, M. C. Hochberg, T. A. Hillier, and J. S. Sinsheimer. Estrogen receptor alpha and matrix metalloproteinase 2 polymorphisms and age-related maculopathy in older women. *American journal of epidemiology*, 167(10):1217–1225, 2008.

[137] M. J. Gameroff. Using the proportional odds model for health-related outcomes: Why, when, and how with various sas® procedures. In *SUGI*, volume 30, pages 10–13, 2005.

[138] D. J. Bauer and S. K. Sterba. Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological methods*, 16(4):373, 2011.

[139] J. S. McGinley, P. J. Curran, and D. Hedeker. A novel modeling framework for ordinal data defined by collapsed counts. *Statistics in medicine*, 34(15):2312–2324, 2015.

[140] S. Van Buuren, H. C. Boshuizen, and D. L. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 18(6):681–694, 1999.

[141] K. J. Lee and J. B. Carlin. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American journal of epidemiology*, 171(5):624–632, 2010.

[142] D. B. Rubin and J. L. Schafer. Efficiently creating multiple imputations for incomplete multivariate normal data. In *Proceedings of the Statistical Computing Section of the American Statistical Association*, volume 83, page 88. American Statistical Association, 1990.

[143] P. D. Allison. Imputation of categorical variables with proc mi. *SUGI 30 proceedings*, 113(30):1–14, 2005.

[144] S. Van Buuren, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin. Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 76(12):1049–1064, 2006.

[145] S. Van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3), 2011.

[146] T. Krishnan and G. McLachlan. The em algorithm and extensions. *Wiley*, 1(1997): 58–60, 1997.

[147] I. Ratitch, B. Lipkovich, and M. Ó'Kelly. Combining analysis results from multiply imputed categorical data. *PharmaSUG 2013-Paper SP03*, pages 1–10, 2013.

[148] T. Bruce, L. Zhaoyu, and S. Arionna. Longitudinal studies of hiv-associated lung infections and complications (lung hiv). 2013.

[149] P. Diggle. *Analysis of longitudinal data*. Oxford University Press, 2002.

Appendices

Codes for Chapter 2

SAS macro code for WGEE

```
/****Macro code for dropout****/
%macro dropout(data=,id=,time=,response=,out=);
%if %bquote(&data)=%then %let data=&syslast;
proc freq data=&data noprint;
tables &id /out=freqid;
tables &time /out=freqtime;
run;
proc iml;
reset noprint;
use freqid;
read all var &id;
nsub = nrow(&id);
use freqtime;
read all var &time;
ntime = nrow(&time);
time=&time;
use &data;
read all var &id &time &response;
n = nrow(&response);
dropout =j(n,1,0);
ind = 1;
```

```
do while (ind <= nsub);
j=1;
if (&response[(ind-1)*ntime+j]=.)then print"First measurement is missing";
if (&response[(ind-1)*ntime+j]^=.)then
do;
j=ntime;
do until (j=1);
if (&response[(ind-1)*ntime+j]=.)then do;
dropout [(ind-1)*ntime+j]=1;
j=j-1;
end;
else j=1;
end;
end;
ind=ind+1;
end;
prev=j(n,1,1);
prev[2:n] = &response[1:n-1];
i=1;
do while(i<=n);
if &time[i]=time[1] then prev [i]=.; i=i+1; end;
create help var &id &time &response dropout prev;
append;
quit;
data &out;
merge &data help;
run;
%mend;
```

```
%dropout(data=forgenmod,id=id,time=time,response=y,out=wgee1);

/*** %dropout creates two variable***/
/*(1) "dropout":indicates missing or observed */
/*(2) "prev":previous observed measurement */

/*Step One: fit a logistic regression model for missingness to predict probabilities */
ods select ParameterEstimates;
proc genmod data=wgee1;
class ctime;
model dropout=ctime prev dose dose*prev/dist=bin;
output out=datapred pred=pred;
run;
ods select all;


/***** macro code for dropweight *****/
%macro dropwgt(data=,id=,time=,pred=,dropout=,out=);
%if %bquote(&data)= %then %let data=&syslast;
proc freq data=&data noprint;
tables &id /out=freqid;
tables &time /out=freqtime;
run;
proc iml;
reset noprint;
use freqid;
read all var &id;
nsub = nrow(&id);
use freqtime;
```

```
read all var &time;
ntime = nrow(&time);
time=&time;
use &data;
read all var &id &time &pred &dropout;
n = nrow(&pred);
wi=j(n,1,1);
ind = 1;
do while (ind <= nsub);
wihlp=1;
stay = 1;
/*first measurement*/
if (&dropout[(ind-1)*ntime+2]=1) then do;
wihlp=pred[(ind-1)*ntime+2];
stay=0;
end;
else if (&dropout[(ind-1)*ntime+2]=0) then wihlp = 1-pred[(ind-1)*ntime+2];
/*second to penultimate measurement*/
j=2;
do while ((j <= ntime-1) & stay);
if (&dropout[(ind-1)*ntime+j+1]=1) then do;
wihlp = wihlp*pred[(ind-1)*ntime+j+1];
stay=0;
end;
else if(&dropout[(ind-1)*ntime+j+1]=0) then wihlp = wihlp*(1-pred[(ind-1)*ntime+j+1]);
j=j+1;
end;
j=1;
```

```
do while (j <= ntime);
wi[(ind-1)*ntime+j]=wihlp;
j=j+1;
end;
ind=ind+1;
end;
create help var &id &time &pred &dropout wi;
append;
quit;
data &out;
merge &data help;
data &out;
set &out;
wi=1/wi;
run;
%mend;
%dropwgt(data=datapred,id=id,time=time,pred=pred,dropout=dropout,out=wgee2);

/* Step two: use predicted probabilities of the fitted missingness model to calculate the
weights(WI) */


/* Step three:WGEE analysis by specifying WI in weight statement*/
proc genmod data=wgee2 descending;
weight wi;
class id ctime;
model y=time dose time*time dose*time dose*time*time/dist=bin;
repeated subject=id/within=ctime type=AR(1) corrw;
run;
```

```
/*** Multiple Imputation (MI-GEE) ***/
proc mi data=amenorrhea out=amenorrhea_mi seed=123 nimpute=10;
class y;
var dose time y;
monotone logistic;
run;
proc print data= amenorrhea_mi;run;


/* fit the MI-GEE */ proc genmod data=amenorrhea_mi descending;
class id;
model y=time dose time*time dose*time dose*time*time/dist=bin;
repeated subject=id / type=AR(1) modelse; *modelse;*covb corrw;
run;


/** DR-GEE **/
/** Obtaining propensity score **/
proc logistic descending data = amenorrhea;
title 'Propensity Score Estimation';
model dose = time/lackfit   outroc = ps_r;
output out= ps_data  XBETA=ps_xb   STDXBETA= ps_sdxb   PREDICTED = ps_pred;
run;


/*** IPTW ***/
data ps_weight;
set ps_data;
```

```
if dose=1 then ps_weight=1/ps_pred;
else ps_weight=1/(1-ps_pred);
run;
proc print data=ps_weight; run;


/** Multiple Imputation **/
proc mi data=ps_weight out=ps_weight2 seed=127 nimpute=10;
class y;
var id time y;
monotone logistic;
run;
proc print data=ps_weight2(obs=5); run;


/** fit the GEE **/
proc genmod data=ps_weight2 descending;
class id;
model y=time dose time*time dose*time dose*time*time/dist=bin;
repeated subject=id/type=AR(1) modelse;
run;
```

Codes for Chapter 3

```
/* Complete data model*/
proc surveylogistic data= quad1;
*weight weight;
class trt (ref='1') /param=ref;
```

```
model y (event='1') = trt Age trt*Age;
output out=pred predicted=mhat;
run;
proc sort data=pred;
by id;
run;


/* missingness model*/
proc surveylogistic data= quad1;
*weight weight;
class trt (ref='1') /param=ref;
model y (event='1') = trt Age trt*Age;
output out=y_pred predicted=mhat;
run;
proc sort data=y_pred;
by id;
run;


/*create inverse probability weights and pseudo-outcomes*/
data quad1_final;
merge pred y_pred;
by id;
/*total population weights*/;
w=1/phat;
*w2=w*weight;
run;
proc print data=quad1_final(firstobs=1 obs=50);
```

```
run;
/*inverse probability weighting*/
proc surveylogistic data=quad1_final;
*weight w;
model y (event='1') = trt Age trt*Age;
run;

/* Mutiple imputation- MI(FCS)*/
proc mi nimpute=0 data=quad; run;
proc mi data=quad seed=234 nimpute=20 out=outfcs;
class trt y;
fcs nbiter=40 logistic (y/details);
var trt Age y;
run;

proc freq data=outfcs;
tables _imputation_ *y_imp / missing;
run;

proc surveylogistic data=outfcs;
class trt (ref='0') / param=ref;
model y (event='1')= trt Age trt*Age;
by _imputation_;
ods output parameterestimates=outparm;
run;
proc print data=outparm;
```

```
run;


proc mianalyze parms (classvar=classval)=outparm;
class trt;
modeleffects intercept trt Age trt*Age;
run;
```

Codes for Chapter 4

```
*****MI using multivariate normal distribution (MVN)****;
proc mi data=long nimpute=20 out=mi_mvn seed=368;
var x1 trt age time y_res;
run;
proc glm data=mi_mvn;
model y_res= x1 trt age time;
by _imputation_;
ods output ParameterEstimates=a_mvn;
run;
quit;
proc mianalyze parms=a_mvn;
modeleffects intercept x1 trt age time;
run;


*****MI using fully conditional specification (FCS)*****;
proc mi data= long nimpute=5 out=long_fcs;
class y_res;
```

```
*fcs plots=trace(mean std);
var x1 trt age time y_res;
fcs discrim(y_res /classeffects=include) nbiter =10;
run;


proc genmod data=long_fcs;
class y_res;
model y_res(event='1')=x1 trt age time /dist=normal;
by _imputation_;
ods output ParameterEstimates=gm_fcs;
run;


PROC MIANALYZE parms(classvar=level)=gm_fcs;
class y_res;
MODELEFFECTS INTERCEPT x1 trt age time;
RUN;


* Proportional odds test*;
proc logistic data=trial;
class AXI_EMPH / param=ref order=data;
model AXI_EMPH(descending)=Time PULM_ART SEX age;
output out=log predprobs=(c i);
run;


****** Direct Likelihood (DL) Analysis ********
proc mixed data=new;
```

```
class USUBJID AXI_EMPH Time PULM_ART;
model AXI_EMPH= Time age SEX PULM_ART /solution ddfm=kr;
parms/ols;
repeated Time/subject=USUBJID type=CS;
run;


********* EM (expectation maximum - MLE) ****** proc mi data=new nimpute=0; var
Time PULM_ART SEX age AXI_EMPH;
run;

proc mi data=new nimpute=4 seed=12 out=new1;
em;
var Time PULM_ART SEX age AXI_EMPH;
run;


proc reg data=new1 outest=regem covout noprint;
model AXI_EMPH=Time PULM_ART SEX age;
by _imputation_;
run;

proc print data=regem(obs=8);
var _Imputation_ _Type_ _Name_
Intercept time PULM_ART SEX age;
run;


proc mianalyze data=regem;
modeleffects intercept time PULM_ART SEX age;
```

```
run;
```

Simulation study of ordinal response with monotone dropout using Negative Binomial distribution

```
*Trial version*;
%let size=1000;
%let rep=500;
%let seed1=1;
%let seed2=2;
data countsim;
call streaminit(357);
do id=1 to &size;
do d=1 to &rep;
do time=1 to 4;
b0=1;
b1=0.75; b2=0.20; b3=0.90; b4=0.40; g0=0.1;
g1=-1.5; g2=0.80; g3=1; g4=-0.50; alpha=2;
*Simulating count data*;
bi=rand("Normal", 8, .5);
x1=rand("Normal", 2, .7);
trt=rand("Bernoulli", .5); *trt*;
age=ceil(20+10*ranuni(&seed2));
y=(b0+b1*x1+b2*trt+b3*age+b4*time+bi);
parm=1/(1+y*alpha);
res=rand('NEGB',parm,1/alpha);
pzero=cdf('LOGISTIC',g0+g1*x1+g2*trt+g3*age+g4*time+bi);
if ranuni(3299)<pzero then do;
y_zinb=res;
```

```
end;
else do;
y_zinb=0;
end;
output;end;
end;
end;
drop b0 b1 b2 b3 b4 g0 g1 g2 g3 g4 alpha bi;
run;
proc print data=countsim(obs=50);run;
data ordinal;set countsim;
if res=0 then y_res=0;
if 1 le res le 7 then y_res=1;
if 8 le res le 100 then y_res=2;
*if 81 le res le 140 then y_res=3;
if res ge 101 then y_res=3;
run;
proc print data=ordinal(obs=12);run;
data create;set ordinal;
by id d;
retain x11-x14 trt1-trt4 d1-d4 y_res1-y_res4 parm1-parm4;
array x1s(4) x11-x14;
array trts(4) trt1-trt4;
array sample(4) d1-d4;
array lin(4) parm1-parm4;
array resp(4) y_res1-y_res4;
if first.id then do;
do i=1 to 4;
```

```
x1s[i]=.;
trts[i]=.;
sample[i]=.;
lin[i]=.;
resp[i]=.;
end;
end;
x1s(time)=x1;
trts(time)=trt;
sample(time)=d;
lin(time)=parm;
resp(time)=y_res;
if last.id then output;
drop time i x1 trt d parm y_res;
run;
proc print data=create(obs=6);run;
data dropouts;set create; drop j miss;
array var4 y_res1-y_res4;
array lin4 y_res1-y_res4;
miss=0;
do j=2 to 4;
if linj eq 1 then miss=1; *generates approx 30%*;
if miss=1 then varj=.;
end;
run;
proc print data=dropouts(obs=15);run;
data long;set dropouts;
array myvariables1(4) x11-x14;
```

```
array myvariables2(4) trt1-trt4;
array myvariables3(4) d1-d4;
array myvariables4(4) parm1-parm4;
array myvariables5(4) y_res1-y_res4;
do time=1 to 4;
x1=myvariables1[time];
trt=myvariables2[time];
d=myvariables3[time];
parm=myvariables4[time];
y_res=myvariables5[time];
output;
end;
drop x11-x14;
drop trt1-trt4;
drop d1-d4;
drop parm1-parm4;
drop y_res1-y_res4;
run;
proc print data=long(obs=7);run;
proc sql;
create table percents as select nmiss(y_res) / count(*) as miss_y_pct from long;
quit; proc print; run;
```