UNIVERSITY OF
KWAZULU-NATAL

Statistical methods for handling incomplete longitudinal data with emphasis on discrete outcomes with application

A dissertation submitted in fulfilment of the academic requirements for the degree of Doctor of Philosophy in Statistics

by

Abdallah Yussuf Kombo

in the
School of Mathematics, Statistics and Computer Science
University of KwaZulu-Natal
South Africa

April 2017

Preface

This thesis and the work presented in it was done wholly while in candidature for a research degree in the School of Mathematics, Statistics and Computer Science at the University of KwaZulu-Natal, South Africa under the advice of Prof. Henry G. Mwambi and Prof. Geert Molenberghs of I-BioStat, Hasselt University and KU Leuven, Belgium. The work represents the author's original work and has not otherwise been submitted in any form for any degree or diploma to any university. Where use of the work of others has been made it has been duly acknowledged.

. . . .

Signed:(Abdallah Kombo) Date

As the candidate's supervisors, we agree to the submission of this thesis.

. . . .

Signed:(Prof. Henry G. Mwambi) Date

. . . .

Signed:(Prof. Geert Molenberghs) Date

Declaration

I, Abdallah Yussuf Kombo, declare that:

- The research reported in this thesis, except where otherwise indicated, is my original research.
- This thesis has not been submitted for any degree or examination at any other university.
- This thesis does not contain other persons' data, pictures, graphs, or other information, unless specifically acknowledged as being sourced from other persons.
- This thesis does not contain other person's writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
    (a) their words have been re-written, but the general information attributed to them has been referenced.
    (b) where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.
- Where I have reproduced a publication of which I am author, co-author or editor, I have indicated in detail which part of the publication was actually written by myself alone and have fully referenced such publications.
- This thesis does not contain text, graphics, or tables copied and pasted from the internet, unless specifically acknowledged and the source being detailed in the thesis and in the references section.

. . . .

Signed: Date

List of publications

1. Kombo, A. Y., Mwambi, H., and Molenberghs, G. (2016). Multiple imputation for ordinal longitudinal data with monotone missing data patterns. *Journal of Applied Statistics*, 1–18 (Chapter 4).

2. Kombo, A. Y., Mwambi, H., Molenberghs, G., Ngari, M., and Berkley, J. A. (under review for submission). Comparison of methods for the analysis of incomplete longitudinal ordinal outcomes: Application to a clinical study on childhood malnutrition. (Chapter 5).

3. Kombo, A. Y., Mwambi, H., and Molenberghs, G. (Ready for submission). Handling longitudinal continuous outcomes with dropout missing at random: A comparative analysis. (Chapter 1).

4. Kombo, A. Y., Mwambi, H., Molenberghs, G., and Berkley, J. A. (Ready for submission). A Simulation Study Comparing Weighted Estimating Equations with Multiple Imputation Based Estimating Equations in the Analysis of Correlated Count Data. (Chapter 3).

5. Kombo, A. Y., Mwambi, H., Molenberghs, G., Ngari, M., and Berkley, J. A. (Under review for submission). Fitting a transition model to incomplete longitudinal ordinal response data: Application to childhood malnutrition data. (Chapter 6).

The Prophet Muhammad (peace be upon him) said: "If anyone travels on a road in search of knowledge, God will cause him to travel on one of the roads of Paradise. The angels will lower their wings in their great pleasure with one who seeks knowledge. The inhabitants of the heavens and the earth and (even) the fish in the deep waters will ask forgiveness for the learned man. The superiority of the learned over the devout is like that of the moon, on the night when it is full, over the rest of the stars. The learned are the heirs of the Prophets, and the Prophets leave (no monetary inheritance), they leave only knowledge, and he who takes it takes an abundant portion."

Sunan of Abu-Dawood, Hadith 1631

Acknowledgements

In the name of the Almighty God, most Gracious, most Merciful. Praise be to God, the Cherisher and Sustainer of the world.

I am highly grateful to the Almighty God for His unmeasurable Mercies that saw me through my studies. It is such a great honour for me and humbling at the same time because it has been a long journey characterized by long hours of hard work and sacrifice. Balancing between family and school was not easy but I did my best and left the rest to God. I thank God for everything.

It is with pleasure and gratitude to acknowledge those who rendered assistance through the lengthy process of my study. I sincerely appreciate all the big and small contributions from everybody who either directly or indirectly helped make my studies less stressful and a success.

I am specifically grateful to my supervisor, Prof. Henry Mwambi for his presence, inspiration, dedication and his much guidance and help. I again thank him for his availability and readiness to read and correct all the many errors in my work. His help in and outside the research has been immense. Secondly, my thanks go to my Co-Supervisor, Prof. Geert Molenberghs for his powerful guidance. Much thanks for his quick responses whenever I consulted. I am deeply grateful for his willingness to work with me. Thank you both for sharing your enormous statistical knowledge with me.

I would also like to thank Kemri/Wellcome Trust Research Programme, Kilifi, Kenya for allowing me to use their data, and especially Prof. J.A. Berkley for facilitating the release of the data and his assistance in explaining the subject matter concerning the data.

A specially thanks goes to my wife, Ummu Sumayyah, who laboured faithfully for many long months, struggling with our children without my presence. Thank you for your understanding and enormous support in every way you could - sacrificing all the moments when I was supposed to be there. I sincerely apologise to my wife and children for being away when they needed me the most.

To the rest of my extended family, friends, fellow students, my group members, my professors, Taita Taveta University team, as well as the professionals whom I consulted in the course of my various researches. Their names may not appear here, but I won't forget their help. May the Almighty bless them all.

Abstract

In longitudinal studies, measurements are taken repeatedly over time on the same experimental unit. These measurements are thus correlated. The variances in repeated measures change with respect to time. Therefore, the variations together with the potential correlation patterns produce a complicated variance structure for the measures. Standard regression and analysis of variance techniques may result into invalid inference because they entail some mathematical assumptions that do not hold for repeated measures data.

Coupled with the repeated nature of the measurements, these datasets are often imbalanced due to missing data. Methods used should be capable of handling the incomplete nature of the data, with the ability to capture the reasons for missingness in the analysis. This thesis seeks to investigate and compare analysis methods for incomplete correlated data, with primary emphasis on discrete longitudinal data. The thesis adopts the general taxonomy of longitudinal models, including marginal, random effects, and transitional models.

Although the objective is to deal with discrete data, the thesis starts with one continuous data case. Chapter 2 presents a comparative analysis on how to handle longitudinal continuous outcomes with dropouts missing at random. Inverse probability weighted generalized estimating equations (GEEs) and multiple imputation (MI) are compared. In Chapter 3, the weighted GEE is compared to GEE after MI (MI-GEE) in the analysis of correlated count outcome data in a simulation study. Chapter 4 deals with MI in the handling of ordinal longitudinal data with dropouts on the outcome. MI strategies, namely multivariate normal imputation (MNI) and fully conditional specification (FCS) are compared both in a simulation study and a real data application. In Chapter 5, still focussing on ordinal outcomes, the thesis presents a simulation and real data application to compare complete case analysis with advanced methods; direct likelihood analysis, MNI, FCS and ordinal imputation method. Finally, in Chapter 6, cumulative logit ordinal transition models are utilized to investigate the influence of dependency of current incomplete responses on past responses. Transitions from one response state to another over time are of interest.

Contents

List of Figures

List of Tables

Chapter 1

Overall introduction

1.1 Longitudinal studies

In the medical, epidemiological and social sciences, studies are often designed to investigate changes in the response of interest observed or measured over time on each subject. These are called repeated measures or longitudinal studies. Here, the observations are ordered by time or even position in space. Repeated measures in space are common in say agriculture and ecological studies. Longitudinal studies are in contrast to cross-sectional studies in which the response of interest is measured only once at a given time for each sampled subject. The primary objective of longitudinal studies is often to examine changes in the responses over time as well as the factors that influence these changes. Thus, the methods used describe the dependence of the response variables on time, treatment effects of interest and other possible covariates.

Since the repeated measures are taken from the same subject over time then the data are typically correlated. This violates the usual independence assumption when dealing with cross-sectional samples. Therefore, statistical techniques which assume independence of observations, like the linear regression analysis and logistic regression cannot be directly applied. Advanced techniques are developed to account for the correlated nature of observations from each subject (Diggle, 1988; Diggle, Liang and Zeger, 2002; Zeger and Liang, 1992). The variability in the data comes in two ways such that: there is variability between the subjects (between-subject variability) and also variability within each subject (within-subject variability). Failure to account for this two-way variability may lead to: (i) incorrect inferences on the regression parameters due to underestimated standard errors of between-subject effects (like age, sex) and (ii) inefficient estimators where unnecessary larger standard errors of the within-subject effects (e.g., time) are obtained (Stokes, Davis and Koch, 2012).

A number of statistical methods exist for the analysis of longitudinal data. The choice of which will always depend on the type and nature of the data. (i) Longitudinal data may either be continuous or categorical. When the response is continuous and assumed to be Gaussian, there exists a general class of linear models that is suitable for the analyses. The linear mixed model is widely accepted as the unifying framework for a variety of correlated settings including longitudinal data (Verbeke and Molenberghs, 2009). However, when the response variable is categorical, fewer techniques are available. This is partly due to lack of a discrete analogue to the multivariate normal distribution (Aerts, et al., 2002); (ii) longitudinal data trajectories may be highly complicated, and there may be large variations between individuals; (iii) there are often missing data; (iv) some variables may be measured with error; (v) longitudinal data may be associated with time-to-event data, and joint modelling may be necessary; and (vi) in some studies the number of variables may be large while the sample sizes may be small. In longitudinal data analysis, new statistical methods are required to address one or more of the above problems since standard methods are not directly applicable.

To this effect, there has been extensive research for the analysis of longitudinal data in the last few decades. For a comprehensive review of various models and methods for the analysis of longitudinal data see, for example, Diggle et al. (2002) and Fitzmaurice et al. (2008), among others. Some of the commonly used models for longitudinal data include:

- Mixed effects models - these models include random effects to incorporate the between subject variation and the within-subject correlation in longitudinal data.
- Transitional models - in these models the within-individual correlation is modelled via Markov structures.
- Nonparametric and semiparametric models - In these models the mean structures are modelled semiparametrically or nonparametrically leading to partial or fully distributional free models. These models are more flexible than parametric longitudinal models. An example of the semi-parametric approaches is given by generalized estimating equations (GEE; Liang and Zeger, 1986).
- Bayesian models - Prior information or information from similar studies are incorporated for Bayesian inference. The advantage of Markov Chain Monte Carlo (MCMC) methods has led to rapid developments of these models.

Each of these modelling strategies has its own advantages and short comings and the choice of one will always depend on the nature of the data and the kind of analysis required. It is not the aim of this thesis to discuss all those modelling strategies as applied in longitudinal studies. However, some of them may feature in later chapters. In this thesis we seek to investigate the impact of missing data in longitudinal studies and the remedies to the missing data problem as it applies in longitudinal discrete outcome data. Methods for incomplete continuous longitudinal data will be briefly addressed as a precursor to the main focus of the thesis.

1.2 Missing data in longitudinal studies

Often longitudinal study designs are unbalanced due to attrition or failure to obtain all the required measurements for each subject at all occasions. A missing data value occurs when it is not observed but could have been observed. If for instance an examiner fails to record the test score of a student, the score is a missing data value. Missing data can occur on one or more of the variables of interest. They can occur on the predictors also known as covariates or on the outcome variable.

Thus preceding the original statistical analysis to be carried out to answer a research question of interest, there is the missing data problem to be solved. The reasons that lead to the missing data are varied and it is always necessary to reflect on the nature of missingness and its impact on inferences. This is especially important because some methods to handle missing data are specific to the structure of the missing values in the dataset and the reasons why the data values are missing. Below we briefly discuss these missing data patterns, the mechanisms and their impact on the missing data methods of choice.

1.2.1 Missing data patterns and mechanisms

Missing data patterns describe and explain the geography of the dataset, as in where in the dataset the values are observed and where the values are missing. They provide important information about the amount and structure of missing data. Understanding the missing data pattern is key because as will be seen in later chapters, some procedures to deal with missing data can be applied to any missing data pattern whereas other procedures are restricted to specific missing data patterns, and therefore having identified the variables that define the pattern, a suitable analysis procedure can be identified. First consider arranging a dataset in a rectangular or matrix form, where the rows correspond to observational units (subjects) and the columns correspond to variables. These variables, say $Y_{ij}, i = 1, \ldots, N; j = 1, \ldots, n$ (for n measurement occasions) may be ordered in such a way that if outcome Y_{ij} is missing for a unit i, then all subsequent variables $Y_{ik}, k > j$, are missing for that unit. This is termed a monotone missing data pattern. In longitudinal studies, monotone patterns (or dropout) may arise, where Y_{ij}

represents variables collected at the jth occasion for unit i. Figure 1.1(a) shows a monotone pattern. In practice, the missingness pattern is rarely monotone, but is often close to monotone. Otherwise, if a subject misses at a certain scheduled occasion but later returns into the study, then this is referred to as intermittent (non-monotone) missing data pattern. This is presented in figure 1.1(b). Figure 1.1(c) represents a special case called file matching. File matching occurs when variables are never observed together. Analyses of data with such type of patterns require making of strong assumptions about these partial associations. When estimating the association between two variables that are never jointly observed the implication is that some of these parameters will not be estimable from the data.

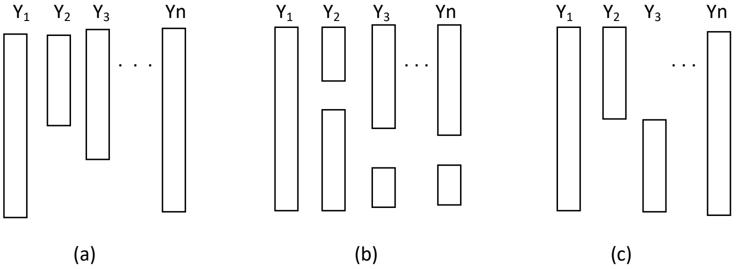


Figure 1.1: *Schematic presentation of missing data patterns: (a) monotone pattern, (b) arbitrary pattern and (c) file matching. Rows correspond to observational units and columns correspond to variables.*

The data may be missing due to varied reasons, known and unknown. Some of the reasons may be completely unrelated to the data at hand, while others may be closely related. These underlying reasons are generally known as missing data mechanisms. Rubin (1976) classified them into three namely: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Without loss of generality we focus on missing outcomes or responses in the definition of the mechanism. Data are said to be MCAR when the probability that a response is missing is unrelated to neither the specific values that in principle, should have been obtained nor the set of observed responses. This is equivalent to the assumption that the missing portion of the data happens to be a completely random sub-sample of the ”original complete data”. For instance when an examiner misplaces all the students’ scripts for a test, then we say that this data is missing completely at random. The above stated equivalence implies that, MCAR is a necessary condition for an analysis where incomplete cases of a data are entirely discarded and only the complete ones are analysed. Secondly, data are said to be MAR when the probability that responses are missing is related to the set of observed responses. More precise, MAR means that conditional on observed data, the probability of the value being missing is unrelated to its actual value. For example, consider a case where data are missing for students’ test results and missing mostly for the individuals performing below average in the dataset. In this case, the probability of missingness on test results is related to under performance. Under MAR, all the information about the missing data are contained in the observed data, but it occurs in a way that complicates statistical analysis. For valid analyses, all the observed data must be taken into account. MCAR is a special case of MAR, and occurs when the distribution for missingness does not depend on observed data either. In a practical sense, we can say that MAR is the most appealing mechanism to deal with the missing data problem. This is in spite of the fact that an individual's probabilities of response may be related only to their own measured information. This information can change from one individual to another. Hence it becomes worthwhile to make this assumption for analytic simplifications (Schafer and Graham, 2002). Finally, data are said to be MNAR when the probability that responses are missing depends on both observed responses and the specific missing values that, in principle, should have been obtained. MNAR is a complicated mechanism since the cause of dropout is related to subject's post dropout, unmeasured responses, even after allowing for the measured information. In such a situation, it becomes necessary to model the dropout concurrently with the response. Under this mechanism, the unobserved value can either be the unknown value of the missing data itself or other unobserved values (Brand, 1999). An example of an MNAR situation is when for instance a study is conducted on the efficacy of a teaching technique and individuals are enrolled to be followed up for some period. MNAR is evident when those obtaining very low points from a number of tests were likely to be missing at the end. In this case, an individual's anticipation of a low mark makes him/her drop out from follow up. These mechanisms will be formally defined in probability terms later in the thesis.

It is important to have a proper understanding of the missing data mechanisms because the performance of missing data handling procedures depend greatly on assumptions about the mechanisms. Although, its not possible to tell with certainty whether missing data are MAR or MNAR because there is no information about the missing data itself (Eekhout et al., 2012). In fact, this phenomenon is further discussed in Molenberghs et al. (2008) where they show that a formal distinction between MAR and MNAR is not possible. However, in spite of this, the impact on key parameter estimators and corresponding hypothesis tests can be considerable. Arguably, such a sensitivity analysis should virtually always be conducted. With the sensitivity analysis, the robustness violation of MAR is investigated to see if the questioned mechanisms lead to conclusions differing to those expected under MAR. Methods exist that can only distinguish whether data are MCAR or not MCAR by using a statistical model for the missingness probability where an MCAR model is nested within a MAR model.

1.2.2 Ignorability

Consider the assumption that the full data model parameters, say θ, and the missing data mechanism parameters, say ξ, are disjoint. From a Bayesian view-point, any joint prior distribution applied to (θ, ξ) can be factored out into independent marginal priors for θ and ξ. Taking a frequentist approach, it implies that the joint parameter space (θ, ξ) must be a Cartesian cross product of the individual parameter spaces for θ and ξ, i.e., $\Omega(\theta, \xi) = \Omega(\theta) \times \Omega(\xi)$. Essentially, under likelihood and Bayesian inferences, and provided the above stated regularity conditions hold, MCAR and MAR imply that the missing data mechanism can be ignored. With frequentist methods, the stronger MCAR is needed to automatically have ignorability. The consequence of ignorability is that then the missing data mechanism does not need to be modelled explicitly.

In essence, the MAR approach assumes that the missing observations are no longer random samples that are generated from the same sampling distribution as the observed values, hence the missing values must be modelled. Specifically, for data that have only missing response values, an MAR analysis assumes that the probability of a missing value can depend on some observed quantities but does not depend on any unobserved quantities hence you can model the probability of observing a missing outcome y_i by using the covariates x_i, but the probability is independent of the unobserved data value (which would be the actual y_i value). When data are missing on the covariate, MAR assumes that missingness is independent of unobserved data, conditional on both observed and modelled covariate data and on observed response data. This implies that responses that have similar observed characteristics (covariates x_i, for example) are comparable and that the missing values are independent of any unobserved quantities. This also implies that the missing data mechanism can be ignored and does not need to be taken into account as part of the modelling process.

In contrast, since the probability of missing data is related to at least some elements of the unobserved partition of the data, MNAR is often referred as non-ignorable (informative) missingness. The term non-ignorable refers to the fact that missing data mechanism cannot be ignored in the analysis, i.e., future unobserved responses cannot be predicted conditional on past observed responses; instead, we need to incorporate a model for the missingness mechanism (Nakai and Ke, 2011).

The effect of non-ignorable mechanism is unknown because normally there is not enough information from the data to allow modelling and investigation of the way data are missing. Hence, not feasible to conduct a satisfactory analysis Thijs et al. (2002). To assess the deviations from an ignorability mechanism, sensitivity analyses are investigated where models for the non-ignorable mechanism are investigated.

1.2.3 Missing data methods

When confronted with missing data, a number of approaches exist. They can be classified as simple/traditional (or commonly referred to as the ad hoc) approaches and advanced approaches. The simple approaches include the deletion methods, and single imputation methods. These approaches are simple and easily applicable in standard statistical software. They are also quite acceptable if dealing with small fractions of missing data but are seriously prejudiced when this fraction is large. The advanced ones are the likelihood based approaches and multiple imputation. Generally, they have an advantage over the simple traditional methods.

1.2.3.1 Simple methods for missing data

(i) **Deletion methods**

- Listwise deletion
  This method is also known as complete case analysis (or complete subject analysis). It is by far the most common treatment to missing data. Here, all incomplete cases are discarded and analysis carried out on what remains. It is very simple and easy to implement and standard statistical software can be employed for analysis. In fact, it is the default method in many statistical software packages. Under the assumption that data are missing completely at random and a small fraction of incomplete cases, it leads to valid unbiased parameter estimates. However, even when complete case analysis is valid, it can be very inefficient, such that it produces estimates with higher variance than would be obtained with other equally valid methods (Little and Rubin, 2014; Rubin, 1987), especially when we have to rule out a large number of cases. This consequently leads to the reduction of its statistical power. When data are not missing completely at random, results are biased. It is noted that the statistical analysis will be biased when the complete cases are systematically different from the incomplete ones. In essence, the disadvantages of listwise deletion outweigh its advantages. Nonetheless, the method is still used in some fields of research e.g., in medical and epidemiological research (Eekhout et al., 2012). This is in fact logical because different researches have different expectations. According to Schafer and Graham (2002), the impact of the missing data problem is minimal if only a small portion of the data set is missing, because then listwise deletion can be quite effective. But for all these arguments, leading researchers in the field are still hesitant in providing a definitive percentage of missing values below which it is still fine to use the method. It has proven to

be very difficult to map out a rule of thumb since the viability of using listwise deletion do not depend only on the missing data rate (Little and Rubin, 2014).

- Available case analysis

  With available case analysis or pairwise deletion, all available data are used to estimate the parameters in the model. Incomplete cases are deleted on an analysis by analysis basis in a sense that any given case may contribute to some analyses but not to others. Therefore, the sample size is not maintained from one analysis to another and so one cannot compare analyses because the sample is different each time. The method uses all information possible with each analysis and often an improvement over listwise deletion because it minimizes the number of cases discarded in any given analysis (Baraldi and Enders, 2010). However, its major drawback, like listwise deletion, is its reliance on the very strong, and in many cases unrealistic missing completely at random mechanism to produce unbiased and consistent parameter estimates. Another difficulty is that available case analysis can produce estimated covariance matrices that are implausible, such as estimating correlations outside of the range of -1.0 to 1.0. This estimation problem arises since differing numbers of observations are used to estimate components of the covariance matrix (Pigott, 2001).

  Generally, for the deletion methods, there should be no shock that such substantial loss of information may terribly impact the analysis, consequently reducing the precision of estimation in terms of larger standard errors, wider confidence intervals, smaller test statistics, or even larger p values. The reduced sample sizes lead to inefficient use of available data and the resulting analysis may lead to terribly biased estimates of effects of interest.

(ii) **Single imputation methods**

These are a collection of common traditional missing data techniques where one imputes (fills in) the missing data value with a seemingly suitable replacement value once. The value is usually estimated from the observed data. In effect, the dataset becomes complete hence complete data methods can be used for analysis. However, the disadvantage of these methods is that standard errors are underestimated, confidence intervals are unrealistically narrower and p-values in favour of type I error are obtained (Rubin and Schenker, 1991) indicating a higher precision and confidence than what can truly be inferred from the data. This is because of the fact that extra uncertainty due to missing data is not reflected as the filled in values were assumed and treated as if they were real values that would have been measured or observed.

A number of single imputation techniques exist. They include: mean imputation, regression imputation, indicator method, matching methods (Hot-deck, Last observation carried forward, Baseline observation carried forward), and stochastic regression imputation. We will briefly discuss only a few of these since the main idea is the same for all of them.

- Mean imputation
  This method can be classified into two: unconditional and conditional mean imputation (Greenland and Finkle, 1995; Schafer and Graham, 2002). Under unconditional mean imputation, the missing value is replaced by the overall variable mean or median from the observed data, or by a value randomly drawn from the subjects with observed data on that variable. Conditional mean imputation fills the missing value by the mean that is estimated from the specific subgroup to which the subject with missing data belongs. For a categorical variable the mode is used. The sample size is regained, and simple to use. However, the variability in the data is compromised, thus standard errors and variance estimates are underestimated. The method produces biased estimates regardless of the underlying missing data mechanism (Enders, 2010).

- Last Observation carried Forward (LOCF)
  Whenever a value is missing, the last value measured is substituted. Ordinarily, it is applied to settings where missingness is due to attrition (dropout). The gist of the method is that very strong and unfeasible assumptions need to be made for its validity: (1) for a longitudinal analysis or when the scientific problem is in terms of the last planned occasion, the analyst has to be convinced that a subject's measurement remains the same from the occasion of dropout onwards or during the period it misses in case of intermittent missingness. This consistency assumption is hardly possible or attainable. In clinical trials, for instance, one may believe that the subject's response profile changes as soon as they go off treatment. (2) Like other single imputation methods, LOCF has the propensity of treating the imputed and actually observed values as equal. Also, the general effect of the method is that both the mean and variance structures are gravely distorted and prejudiced such that no apparent simplification is possible.

1.2.3.2 Advanced methods for missing data

Because repeated measurements on an individual tend to be correlated, we recommend procedures that use all the available data for each participant, because missing information can then be partially recovered. More advanced methods have been developed, that are statistically justifiable and offer a better potential for precision and validity than the so-called traditional methods. They include the multiple imputation, maximum likelihood methods, Bayesian methods, and weighting methods among others. Longitudinal modelling by maximum likelihood can be a highly efficient way to use the available data. Multiple imputation of missing responses is also effective if we impute under a longitudinal model that borrows information from the observed data. In fact, the borrowing of information from observed data is a strategy that multiple imputation shares with maximum likelihood. On the other hand, weighting methods are equally valuable in some cases. Below we briefly discuss these methods.

- Multiple imputation

  Multiple imputation (MI) was initially proposed Rubin (1978a), and further elaborated in Rubin (1987); Little and Rubin (2014). Although initially proposed for public-use survey data, it has developed to general missing data problems. Formally, MI is described as a tri-step process: First, we estimate the conditional predictive distribution of the missing data given the observed data, and then (taking account of the uncertainty in the parameter estimates) we impute from this to create multiple complete datasets. Each of these complete datasets are then independently analysed using appropriate complete-data methods. Finally, the results are combined into a single inference, in a way that captures uncertainty regarding the imputation process.

  An important part of the imputation process is perhaps the evaluation of the imputation strategy, since it relies on untestable assumptions concerning the missingness process that created the partially observed measurements. MI usually assumes that the data are missing at random. The most important issue then is identifying variables that make this assumption viable. Ideally, the analysis model is pre-determined and the imputation method then can be evaluated simply by its ability to reproduce any complete data analysis. Chambers (2001) terms this phenomenon "preservation of analysis". He elaborates five performance requirements for an imputation method: predictive accuracy, ranking accuracy, distributional accuracy, estimation accuracy and imputation plausibility. These are in fact the generalization of those described by Allison (2000) and Rubin (1987, 1996). However, they note that these criteria are easily violated in practice, since the assumptions of missing at random are hardly met in practice. Although, it is possible to formulate and estimate data models that are not missing at random, these models are complex, untestable, and require specialized software with technical expertise. Hence, any general-purpose approach will necessarily invoke the missing at random assumption (Allison, 2000).

  There are various imputation models that can be used depending on the data and the missing data pattern. When missing data are monotone, predictive mean matching and propensity score methods may be used for continuous variables. For discrete variables, logistic regression and discriminant analysis can be used. In case of non-monotone missing data patterns Markov Chain Monte Carlo (MCMC) approaches have been proposed. It is noted that, methods for non-monotone patterns can be used for monotone patterns but the reverse is not true.

- Maximum Likelihood

  Maximum likelihood (ML) methods can be used to obtain the variance-covariance matrix for the variables in a model based on only available data. Using the obtained variance-covariance matrix, the regression model can then be estimated (Schafer, 1997). The ML methods are simpler and easy to implement in standard statistical software, e.g., SAS. The user needs to specify the model of interest and then proceed to indicate that they want to use ML (Yuan, 2010). There are two main ML methods:

  (a) Direct maximum likelihood

  Instead of deleting or imputing observations with missing values, the direct maximum likelihood (DL) and full-information maximum likelihood (FIML) methods use all the available information in all observations. Missing values are handled directly within an analysis model. The model is estimated by making use of all the available information. The procedure involves direct maximization of the multivariate normal likelihood function for the assumed linear model. FIML is oftenly used in structural equation models (SEMs) and multi-level models or growth models. When properly used, DL produces efficient estimates and correct standard errors. However, it involves specialized software, implying that it may be challenging and time consuming (Soley-Bori, 2013). Normally, MI and DL will produce similar results when data are missing on the outcome and the same information is used for both models (Collins, Schafer and Kam, 2011). DL will be revisited and used in later chapters.

  (b) Expectation maximization algorithm

  The expectation maximization (EM) algorithm (Dempster, Laird and Rubin 1977), is a a general iterative procedure that can be used to find the maximum likelihood estimates in the presence of missing data. The algorithm is useful when maximization from the complete data likelihood is straightforward but maximization based on the observed data is complicated and/or difficult to justify. Under an ignorable MAR assumption, the algorithm can be summarized as follows. Each iteration of the algorithm involves two steps: The expectation (E - step) and the maximization (M - step). The E-step determines the expected value of the log-likelihood conditional on the observed data and the current estimate of the missing data. This step is often reduced to simple sufficient statistics. Given the complete data log-likelihood, the M-step estimates the parameters that maximize the expected likelihood based on the E-step.

Known drawbacks of the EM algorithm are its initial inability to produce estimates of the covariance matrix of the maximum likelihood estimators. But advancements have lead to development of methods for such estimation to be incorporated into EM computational procedures. Another concern is its slowness to converge (depending on the amount of missing data), and in cases where E-step fails to settle to a closed form solution or the M-step failing to determine a unique maximum. These challenges have resulted in the development of modifications and extensions to the algorithm as well as many simulated based alternatives. One is the Stochastic EM (Celuex and Diebolt, 1985) among other variations. See also Baker (1992), Louis (1982), McLachlan and Krishna (2007), and Rubin (1991).

- Weighting methods

  Besides imputation, incomplete data may be handled by weighting methods. Weighting methods are based on observed measurements (Robins, Rotnitzky and Zhao, 1994). Under these methods, after ignoring all the missing values in the analysis, the observed values are weighted depending on how their distributions approximate the full sample or population. In this way, the predicted probability (weight) of each response is estimated from the measurements for the particular observed variable to correct for either the standard errors associated with the estimated parameters or the population variability. See Kalton and Flores-Cervantes (2003) for a discussion of the weighting methods. They provide a detailed review of the methods, and the stages involved in the weighting process.

  In the context of survey data, Rubin (1987) discusses a number of approaches for estimating and applying weights. When the number of nonrespondents is small, their responses are assumed relatively similar to those present who would be weighted to represent the excluded respondents.

  Robins, et al. (1995) proposed a weighted regression model that requires an explicit model for the missingness but relaxes some of the parametric assumptions in the measurements model. Based on the traditional GEE, they developed so-called weighted generalized estimating equations (WGEE) to deal with the bias caused by dropouts. In its original form, GEE relied on the stringiest MCAR mechanism. WGEE was developed to work on MAR as well as MNAR mechanisms, but requires the specification of a dropout model in terms of the observed responses and/or covariates. The idea behind the WGEE is to weight each individual by the probability of being observed. The weighting method enjoys elegant features of being less resource intensive. However, proper care must be taken when used because it can lead to large data loss when the rate of non-response is high (Lago and Clark, 2015). Generally speaking, weighting methods are a good alternative under certain circumstances, for instance, for monotone missingness patterns or in case of univariate analysis. The concept of WGEEs is tackled in later chapters in the context of continuous and count outcome data.

Since Robins, et al. (1995), a number of revisions and additions have been made on the original GEE. Birhanu et al. (2011) improved the WGEE to what are now known as the doubly-robust estimating equations (DREEs). Here, a predictive model for the unobserved responses conditional on observed ones is incorporated with the weighting. This made it more efficient and robust towards a broader set of deviations. However, the DREE method is hard to implement than the original (Van deer Laan and Robins, 2003). GEE can also be used after MI and hence the so-called MI-GEE approach Schafer (2003).

1.2.3.3 Selection and pattern mixture models

In terms of non-ignorable missingness, a completely sufficient data analysis that can be used in the process is not readily viable. Standard statistical models can result into very biased results. This is because the available observed measurements cannot provide sufficient information to confirm or refute ignorability. Researchers have proposed the inclusion of the missingness in the modelling process. They suggested modelling the missingness process jointly with the measurement process, and then proceed to apply likelihood-based approaches like the maximum likelihood or consider a Bayesian inference. Therefore, joint modelling of the measurement and missingness processes is necessary to account for informative nonresponse. In principle, one would consider the density of the full data as a joint distribution of the measurement and missingness processes. Two principal frameworks can be specified from the joint distribution. (1) *Selection models* (Little and Rubin, 2014), are based on the self-selection of individuals into observed and missingness ranks, where the missingness model is the density of the missingness process conditional on the measurements. In selection models, it is sensible to assume the ignorable, MAR assumption, but a number of modelling approaches have been proposed assuming non-ignorability. So far most missing data literature focuses on this class of modelling. This is because, the selection model seems more innate when concern is on parameters of the marginal distribution of the outcomes y averaged over missing data pattern (Little 1993).

(2) *Pattern mixture models* (Glynn, Laird and Rubin, 1993; Little, 1993) can be viewed

as a mixture of populations distinguished by the observed and missing data patterns. This means that inferences about the marginal parameters averaged over the missing data patters are obtained by conveying them as functions of the full data model parameters, say θ, and missing data mechanism parameters, say ξ. By this, it means that individuals are stratified according to their missing data patterns. Then, a separate model fit for each pattern and finally combining the results from the different "sub-models" to derive an average estimate of the model parameters. Both likelihood and Bayesian methods can be applied to these functions. However, an issue with this class of models is that they are under-identified (meaning they present inestimable parameters). Little (1993; 1994) proposed some approaches to deal with the under-identification. Verbeke and Molenberghs (2009) and Molenberghs and Kenward (2007) detail the approaches to circumnavigate the under-identification problem.

Beyond the above two modelling approaches under MNAR, a third framework exists. The *shared parameter model* (Wu and Carroll, 1988), relates or links the response model with the probability of missingness. It is an attractive framework for the joint modelling of the measurement and missingness process, which makes use of random effects to instigate the interdependence between the two processes. The underlying feature of these models is that the two processes are assumed independent, conditional on the random effects, meaning all association is brought about by random effects. Normally, the random effects are assumed to follow a normal distribution, and considered an important element in the design of the missing data process, implying that a misspecification of their distribution greatly jeopardises inference, thus producing wrong parameter estimates and standard errors (Tsonaka et al., 2009). However, Song et al. (2002), Tsiatis and Davidian (2004) and Wang and Taylor (2001) argue on the contrary. Their empirical results indicate that misspecification of the distribution of the random effects does not pose a serious impact on the parameters, save for exceptional cases, e.g., in discrete distributions. For further details on the shared-parameter framework, we recommend among others Albert and Follmann (2009), Rizopoulos et al.(2008) and Tsonaka et al. (2009).

1.3 Research Objectives

The main objective and focus of this thesis is to investigate or research on methods to handle incomplete longitudinal data, with principal interest falling on the non-Gaussian setting. Categorical (binary, ordinal) and counts outcomes are very common in real applied problems but missing data techniques for this type of data are less standard, because of the lack of a simple analogue to the normal distribution. However, we start with the common Gaussian type before embarking on the non-Gaussian case. Specific objectives include:

- To examine the comparative performance of multiple imputation and inverse probability weighting techniques when used for incomplete continuous outcome data subject to MAR dropouts.
- To compare two extensions of the generalized estimating equations, namely the weighted generalized estimating equations and multiple imputation based generalized estimating equations in the presence of incomplete count outcomes due to MAR dropouts.
- To investigate different multiple imputation strategies with applications on ordinal outcome data subject to both monotone and non-monotone missing data patterns.
- To examine the comparative performance of likelihood based methods and multiple imputation when presented with incomplete discrete data.
- To investigate the influence of dependence of current responses on past responses (history) in medical research. This approach is necessitated by the fact that the ordinal outcome categories are driven by an underlying disease or response process and the data is longitudinal. Thus transitions from one disease state to another over time are of interest.

1.4 Thesis outline

This thesis is a collection of 5 research papers which have been submitted for publication in international, accredited journals. Out of the papers, one is published and the rest are under review. These papers appear in Chapters 2 through 6, with each chapter presented as a stand alone and not a continuation of the previous one. However, the general ideas in these chapters are in a way interconnected in order to achieve the overall goal of the thesis. Chapter 1 served as an introduction and general overview to the thesis. The rest of the thesis is outlined as follows.

Although previously stated that the objective was to deal with discrete data, the thesis started with one continuous data case. In Chapter 2, the thesis presents a comparative analysis on how to handle longitudinal continuous outcomes with random dropout. Here, incomplete data methods, inverse probability weighted GEE and multiple imputation, which are valid under the MAR mechanism, are compared.

In Chapter 3, the weighted GEE is used again for discrete data. It is compared to GEE after multiple imputation (MI-GEE) in the analysis of correlated count outcome data.

This comparison is carried out using a simulation study.

Chapter 4, now deals with multiple imputation in the handling of ordinal longitudinal data with dropouts on the outcome. MI strategies, namely multivariate normal imputation and fully conditional specification are compared both in a simulation study and a real data application. The real application involves a dataset on patients who were under treatment for arthritis (the Arthritis data). In Chapter 5, still focussing on ordinal outcomes, the thesis presents a simulation and real data application study to compare complete case analysis with advanced methods, direct likelihood analysis and multiple imputation. For multiple imputation, three approaches, namely multivariate normal imputation, fully conditional specification and ordinal imputation method are contrasted. The real application is about nutritional status during recovery from severe malnutrition in children (RSCM). Then in Chapter 6 the thesis investigates the influence of history on current incomplete responses. Therefore, a transitional likelihood missing at random model is built, where we investigate the effects of conditioning on previous responses in addition to estimating the effects of measured covariates. The model is applied to the same data used in Chapter 5. This data were also used in Chapter 1.

In Chapter 7, a general conclusion to the thesis is presented. Finally, we present recommendations and point out areas for further research in this chapter. We give a consolidated references list at the end of the thesis.

Chapter 2

Handling longitudinal continuous outcomes with dropout missing at random: A comparative analysis

Abstract

Missing data is a prevalent problem in the analysis of data from longitudinal studies. Subjects may drop out before the end of the study, or be lost to follow-up in the sense that no further measurements can be obtained after the dropout time. The statistical methods to be used for handling incomplete data depends on the dropout mechanism assumed and probably the type of the data. This paper focuses on dropout missing at random. Two methods valid under a missing at random mechanism, namely multiple imputation and inverse probability weighting are compared through a simulation study and then applied to a real data set based on a continuous outcome. Specifically, we investigate the methods and evaluate their performance under various dropout rates and sample sizes in the simulation study. The simulation studies reveal that the multiple imputation approaches have higher efficiency and less bias. The real longitudinal data is from a study on childhood malnutrition.

2.1 Introduction

Longitudinal studies are designed to collect data on every individual within a sample at each measurement occasion. However, it is quite common that missing data arises.

Incompleteness for longitudinal data occurs as dropout (monotone missing data pattern), which is when individuals leave the study prematurely, before the end of follow up for some reason(s), known or unknown. Alternatively, an individual may miss a measurement occasion but appear at subsequent occasions, resulting in intermittent missing measurements. With missing data, a number of issues arise in the analysis: (1) the analysis now becomes more complicated, (2) there is risk of efficiency loss, and (3) there is an issue of bias, because the observed measurements may not necessarily be the same as the unobserved ones (Barnard and Meng, 1999).

In clinical trials, it is quite possible that the actual reasons for missingness are not known. Discarding the incomplete participants of the study and only analysing the measured cases, namely complete case (CC) analysis, may lead to biased estimates, hence erroneous and imprecise inferences. Determining the appropriate analysis method for incomplete datasets is key to valid parameter estimation and reliable study conclusions.

Over time, researchers have been working on developing relevant methods to handle incomplete data, ranging from simple, easy to use ad hoc ones to more advanced, methodologically challenging approaches (Rubin 1976; Ibrahim 1990; Robins, Rotnitzky and Zhao 1994, 1995; Carpenter, Kenward and White, 2007; Little and Rubin 2014). Two common, attractive methods amongst the different methods that have been proposed, which are based on multiple imputations and inverse probability weighting. These two methods assume that the data are missing at random (Rubin 1976, 1987). The missing at random assumption implies that the probability of missingness is only related to the fully observed variables and not on the unobserved or partially observed variables.

Multiple Imputation (MI), initially proposed by Rubin (1978) and later detailed in Rubin (1987) has so far been recognized as an influential and very practical approach in dealing with incomplete data problems for both discrete and continuous outcomes. MI replaces missing values with estimated values multiple times and then analysis is carried out independently on the now "completed" datasets. The technique has captured the interest of many researchers and concise expositions have been presented. See Rubin (1996); Schafer (1997, 1999); Horton and Lipsitz (2001); Carpenter and Kenward, 2013; Little and Rubin (2014). Inverse Probability Weighted (IPW) estimating equations is another powerful approach. First described by Robins, Rotnitzky and Zhao (1995), the approach traces its roots from survey analysis, presented by Horvitz and Thompson (1952). It was later improved by a number of researchers (Robins and Rotnitzky, 1995; Scharfstein, Rotnitzky and Robins, 1999). Wider literature exists that describes the IPW approach (Fitzmaurice, Molenberghs and Lipsitz, 1995; Yi and Cook, 2002a, 2002b; Carpenter, Kenward and Vansteelandt, 2006; Molenberghs and Kenward, 2007; Seaman and White, 2011).

The main difference between IPW and MI is that IPW needs a model for the missingness mechanism, whereas MI needs the analyst to specify which variables are to be used as regressors in the imputation model. In addition, unless a monotone missing data pattern is used, the missingness model for IPW can only use complete variables. However, IPW's good side is that the approach does not require a complete specification of the joint distribution of the longitudinal responses but rather is based on the specification of the first two moments. Both methods can be used for all types of outcomes, but a great deal of work has been devoted to binary response data particularly under the IPW approach. But, it is not surprising that essentially little has been done in terms of comparing them for continuous response data because the two methods come from two different schools of thought. A recent comparison of these methods in a cross-sectional setting found the performances of these methods to be similar, with MI only slightly more efficient than IPW (Carpenter, Kenward and Vansteelandt, 2006). In the context of survey data, Seaman and White (2011) compared the performance of MI with IPW. In their paper based on a binary outcome data, they illustrated why, despite MI generally being more efficient, IPW may sometimes be preferred. Using marginal structural models, a comparison of these approaches found that MI was slightly less biased and considerably less variable than IPW (Moodie et al., 2008).

In this paper, we compare the performance of MI and IPW in the analysis of incomplete continuous outcome (longitudinal) data under different dropout rates and sample sizes while assuming that the data are missing at random.

Because a so-called direct likelihood (DL) or ignorable likelihood analysis is valid under the missing at random mechanism (Mallinckrodt et al., 2003a, 2003b; Verbeke and Molenberghs, 2009), its results will be presented and used as reference against which IPW and MI will be contrasted. In the DL method, the observed data are used without weighting nor imputation. The strength of this method lies in the accurate formulation of the likelihood of the data as it is and it works for both intermittent and monotone missingness patterns. For incomplete longitudinal data, a linear mixed model (LMM) only needs the missing at random assumption to hold. See Verbeke and Molenberghs (2009) for a detailed discussion of the LMM approach.

The rest of the paper is organised as follows. In Section 2.2, we present the notation and concepts of possible mechanisms that can lead to missing data. In Section 2.3, statistical approaches to be compared are considered in detail. Section 2.4 contains the simulation study. In this section multiple datasets of various sizes are simulated then droputs caused and missing data methods applied. We present the results of the simulation study and discussion thereof. In Section 2.5, we present a real data application. We use a clinical study dataset (which is also incomplete) to elucidate the comparative performance of the competing methods. Section 2.6 provides a discussion and conclusion to the paper.

2.2 Dropout mechanisms in longitudinal studies

Suppose that N individuals are to be observed at n occasions. For the ith individual $(i = 1, 2, \ldots, N)$ we can have a series of measurements $Y_i = (Y_{i1}, \ldots, Y_{in})'$, where Y_{ij} is the jth outcome for individual i. Y_{ij} can either be continuous or discrete depending on the study problem. Each individual has a covariate matrix X_i. The covariates may be time stationary and/or time varying. In longitudinal studies, individuals may not be observed at all n occasions on account of some stochastic missing data mechanism. For this reason we can assume that an individual i contributes $n_i \leq n$ repeated observations, that are not necessarily equal over all individuals. We define an indicator variable R_{ij} to be 1 if the outcome Y_{ij} is observed and equal to 0 if unobserved. The full data information for the ith subject is given jointly by Y_i and R_i, with a joint distribution that can be expressed as:

$$f(Y_i, R_i|X_i, \theta, \gamma) = f_r(R_i|Y_i, X_i, \gamma) f_y(Y_i|X_i, \theta), \tag{2.1}$$

where θ and γ are vectors that govern the joint distribution, with γ parameterizing the misssing data mechanism and θ comprising the parameters that relate the outcome of interest and covariates. In general, the missing data mechanism can depend on the full vector of responses, Y_i and the covariate matrix X_i. Let Y_i^o denote the vector of observed responses and Y_i^m denote the vector of unobserved responses for subject i. Following Rubin's taxonomy Rubin (1976, 1987), first, data are missing completely at random (MCAR) if the missingness process does not depend on Y_i; $f(R_i|Y_i^o, Y_i^m, X_i, \gamma) = f(R_i|X_i, \gamma)$. Second, the missing data are said to be missing at random (MAR) if the missingness process depends on the observed responses and probably on measured covariates but not on the unobserved responses; $f(R_i|Y_i^o, Y_i^m, X_i, \gamma) = f(R_i|Y_i^o, X_i, \gamma)$. Finally, data are missing not at random (MNAR) when the probability of missingness is related to the values that should have been observed, in addition to the ones actually observed; $f(R_i|Y_i^o, Y_i^m, X_i, \gamma) = f(R_i|Y_i^o, Y_i^m, X_i, \gamma)$. Under likelihood and Bayesian inferences, and provided regularity conditions hold, MCAR and MAR imply that the missing data mechanism can be ignored. With frequentist methods, this is generally true only under MCAR. Notice also if missingness depends on (possibly time varying) X_i, it is not MCAR. If X_i are only baseline covariates it is sometimes called covariate dependent MAR.

The focus in this paper is on missing data due to subject dropouts. For all components of Y_{ij} that are missing, the corresponding components of R_{ij} will be 0. The dropout time for the ith subject can be defined by introducing a discrete integer valued variable:

$$D_i = 1 + \sum_{j=1}^{n} R_{ij}. \tag{2.2}$$

The model for the dropout process can therefore be written:

$$f(R_i|Y_i, X_i, \gamma) = Pr(D_i = d_i|R_i, X_i, \gamma) \tag{2.3}$$

where d_i is a realization of the variable D_i. In (2), it is assumed that all subjects are observed on the first occasion so that D_i takes values between 2 and $(n + 1)$. The maximum value $(n + 1)$ corresponds to a complete measurement sequence.

2.3 Statistical methods to be compared

2.3.1 Multiple imputation

Multiple imputation (MI) is a simulation-based approach that fills missing values multiple times to create complete data sets. Standard MI procedures assume that the data are MAR. Extensions towards MNAR are possible. Because of the fundamental untestable nature of the assumptions that need to be made, such extensions have their place in so-called sensitivity analysis. However, this is not the focus of the current study. The MI process involves three distinct stages. First, the missing values are filled in $M \geq 2$ times to generate M complete data sets. In the filling-in process, a joint distribution for the complete data set (including observed and unobserved data) and a prior distribution of parameters are assumed for the data augmentation algorithm to simulate random draws from a missing data distribution. That is, M independent random values can, given the observed values, be generated from a stationary conditional distribution of the missing values as in the Bayesian estimation technique. After the imputation step, M complete data sets are obtained. Each of the M complete data sets are then analysed using appropriate standard procedures, depending on the types of response and assumptions used for the analysis model. Finally, the estimates from the M analyses are pooled to produce a single set of estimates that incorporates the usual sampling variability as well as the variability due to the missing data.

The quality of the imputation model will influence the quality of the analysis model results, so it is important to carefully consider the design of the imputation model. In some but not all cases, the MI inference assumes that the analysis model is the same as the imputation model (Meng, 1994), meaning that all variables appearing in the imputation model should be included in the analysis model. However, practically, the two models need not necessarily be the same. Therefore, to obtain high quality imputations for a particular variable, Van Buuren, Boshuizen and Knook (1999) recommended the inclusion of the following covariates in the imputation model: variables that are in the analysis model, variables associated with missingness of the imputed variable, and variables correlated with the imputed variable. One can include auxiliary variables which may or may not have missing values. While it is almost always impossible to test the MAR assumption, including auxiliary variables in the imputation model can minimise bias as well as making the MAR assumption more viable.

Now, to formally describe MI, we consider the process as presented in Verbeke and Molenberghs (2009). Thus under the MAR assumption, MI imputes Y_i^m by drawing from the conditional distribution $f(Y_i^m|Y_i^o, \gamma)$. Since γ is unknown, we estimate it from the data to yield $\hat{\gamma}$, and use estimated version of the distribution $f(Y_i^m|Y_i^o, \hat{\gamma})$. Since $\hat{\gamma}$ is a random variable, its variability is taken into account when drawing the imputations. In a Bayesian sense, γ is a random variable whose distribution depends on the data. First, obtain the posterior distribution of γ from the data, a distribution which is a function $\hat{\gamma}$. After formulating the posterior distribution of γ, the following imputation algorithm can be used: (1) Draw γ^* from the posterior distribution of $\gamma, f(\gamma|X_i, Y_i^o)$. If needed, approximate this posterior distribution by the normal distribution. (2) Draw Y_i^m from $f(Y_i^m|X_i, Y_i^o, \gamma^*)$. (3) Use the completed data Y_i and the model to estimate the parameter of interest (β^*) and its variance ($V(\beta^*)$), called the within imputation variance. The steps described above are repeated independently M times, resulting in $\beta_k^*, V(\beta^*), k = 1, \ldots, M$. Steps 1 and 2 are referred to as the imputation task, and step 3 is the estimation task. Finally, combine the estimates obtained after M imputations. The overall estimated parameter vector is the average of all individual estimates:

$$\beta^* = \frac{1}{M}\sum_{k=1}^{M}\beta_k^*. \tag{2.4}$$

We obtain the variance as a sum of the within-imputation variances and the between-imputations variability:

$$V^* = W + \left(\frac{M+1}{M}\right)B, \tag{2.5}$$

where

$$W = \frac{1}{M}\sum_{k=1}^{M}V(\beta_k^*), \qquad \text{and} \qquad B = \frac{1}{M-1}\sum_{k=1}^{M}(\beta_k^* - \beta^*)(\beta_k^* - \beta^*)'. \tag{2.6}$$

Here, W measures the within-imputation variability while B measures the between-imputation variability.

2.3.2 Inverse probability weighting

Inverse probability weighting (IPW) is a standard method used for handling dropouts. This method is valid under the MAR assumption (Robins, Rotnitzky and Zhao, 1995), but requires specification of a dropout model in terms of observed outcomes and/or covariates. IPW is more frequently used in marginal models for discrete outcomes rather than continuous outcomes. However, in this paper, it is adopted for dealing with continuous outcomes. The primary idea behind IPW is that if individual i has a probability λ_{ij} of being observed at occasion j then this individual should be given a weight, ω_{ij} say, so as to minimize the bias caused by dropout in the analysis. The weight ω_{ij} for the ith individual at time j is assigned the inverse of the cumulative product of fitted probabilities: $\hat{\omega}_{ij}(\hat{\alpha}) = [\hat{\lambda_{i1}}(\hat{\alpha}) \times \hat{\lambda_{i2}}(\hat{\alpha}) \times \cdots \times \hat{\lambda_{ij}}(\hat{\alpha})]^{-1}$ where α is a vector of unknown parameters. Note here that you need a monotone dropout model for this where the vector α is common for each occasion j.

In longitudinal data settings, IPW can be incorporated into Liang and Zeger's (1986) conventional generalized estimating equation (GEE) method. The GEE methodology generalizes the usual univariate likelihood equations by introducing the covariance matrix of the response vector, Y_i. The GEE methodology is used to model the marginal expectation of responses as a function of a set of covariates. We introduce the classical form of GEE.

Let $X_i = (x_{i1}, \ldots, x_{in_i})'$ denote an $(n_i \times p)$ matrix of covariates where, $x_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ is the $(p \times 1)$ covariate vector associated with y_{ij}. Let $y_i = (y_{i1}, \ldots, y_{in_i})'$ be an $(n_i \times 1)$ observed response vector, and $\mu_{ij} = E(y_{ij}), i = 1, \ldots, N;\ j = 1, \ldots, n_i$. Now, assume the marginal regression model is given as:

$$g(\mu_{ij}) = x'_{ij}\beta, \tag{2.7}$$

where β is a $(p \times 1)$ vector of the regression parameters of interest and $g(.)$ is a link function. Let the $(n_i \times n_i)$ covariance matrix for Y_i be $V_i(\varphi) = \phi A_i^{\frac{1}{2}} R_i(\rho) A_i^{\frac{1}{2}}$, where A_i is a diagonal matrix of variance functions, $R_i(\rho)$ is a working correlation matrix of Y_i, as a function of ρ, the correlation parameter, and ϕ is a dispersion parameter. Then, the GEE estimators for the regression parameters are the solutions to the equation

$$\sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta} V_i(\varphi)^{-1}\ (Y_i - \mu_i) = 0, \tag{2.8}$$

where $\frac{\partial \mu_i}{\partial \beta}$ is the derivative matrix of the mean vector μ_i with respect to β.
The GEE methodology has traditionally been used for the analysis of marginal models for discrete responses. In this paper, it is adopted for a continuous response. Consequently, the following assumptions can be made for the marginal models with the continuous outcome, Y_{ij}.

- The response mean is related to the covariates by an identity link function: $\mu_{ij} = \eta_{ij} = x'_{ij}\beta$. The link function $g(.)$ generally relates the expected values, μ_i of the response vector, Y_i to the covariate matrix X_i. It takes the general form $g(\mu_i) = \eta_i = X_i\beta$, where η_i denotes the linear predictor vector whose jth row is $g(\mu_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp}$. The function $g(.)$, should be monotone and differentiable. When monotonocity holds, the inverse function $g(.)^{-1}$ can be defined by the relation $g^{-1}(g(\mu_i)) = \mu_i$. For a continuous response with normality assumption, the link function is an identity link: $g(\mu_i) = \mu_i$ and the inverse will simply be $\mu_i = g(\mu_i)$. Under this identity link, the expected value of the response is simply a linear function of the covariates multiplied by their regression coefficients.

- The variance of each Y_i, conditional on the effects of the covariates, is ϕ and does not depend on the mean response. Here, $v(\mu_{ij}) = 1$ is a known "variance function", thus implying $\mathrm{Var}(Y_i) = \phi v(\mu_i) = \phi$, with ϕ denoting the variance of the conditional normal distribution of the response, given the covariates. The assumption that the variance is constant over time may be unrealistic. To relax it, a separate scale parameter, ϕ_j could be estimated at the jth occasion if the longitudinal design is balanced on time.

- The within-individual correlation among repeated responses is modelled by assuming, for example, a first-order autoregressive AR(1) covariance structure:
$\mathrm{Corr}(Y_{ij}, Y_{ik}) = \rho^{|k-j|}$, which indicates the pairwise correlation between observations, for all j and k and $0 \leq \rho \leq 1$. The AR(1) structure implies homogeneous variances. In addition, it specifies that the correlations between observations on the same subject are not equal, but decrease towards zero with increasing length of the time interval between observations.

For marginal models with an identity link function, the generalized least square estimator of β can be considered as a special case of the GEE. Therefore, the estimates of parameters in the marginal model for continuous response with an identity link are

$$\hat{\beta} = \left\{ \sum_{i=1}^{N} X'_i \hat{V}_i^{-1} X_i \right\}^{-1} \sum_{i=1}^{N} \left(X'_i \hat{V}_i^{-1} Y_i \right), \tag{2.9}$$

where $\hat{V}_i$ is the maximum likelihood estimator that can be used to find the best unbiased estimates of V_i (Verbeke and Molenberghs, 2009) and

$$\text{Cov}(\hat{\beta}) = \left\{ \sum_{i=1}^{N} X_i' \hat{V}_i^{-1} X_i \right\}^{-1} \sum_{i=1}^{N} \left(X_i' \hat{V}_i^{-1} \hat{\text{Var}}(Y_i) \hat{V}_i^{-1} X_i \right) \left\{ \sum_{i=1}^{N} X_i' \hat{V}_i^{-1} X_i \right\}^{-1}. \quad (2.10)$$

Here, $\hat{\text{Var}}(Y_i)$ is an estimate of $\text{Var}(Y_i)$ which yields a robust estimator of $\text{Cov}(\hat{\beta})$ when substituted in equation (2.10).

With incomplete data that are MAR, the GEE method provides inconsistent estimates of model parameters (Liang and Zeger, 1986). This is because GEE is based on MCAR therefore cannot handle incomplete data that are MAR without further modification. In weighted generalized estimating equations, a subject's contribution to standard GEE is weighted by the inverse of the probability of dropout at particular time point, given the subject did not miss in any of the previous occasions. Therefore, incorporating all the assumptions herein made, valid parameter estimates in longitudinal studies with MAR dropout are obtained by solving the weighted estimating equations:

$$\sum_{i=1}^{N} (Y_i - X_i\beta)' V_i^{-1} W_i(\hat{\alpha}) (Y_i - X_i\beta) = 0, \quad (2.11)$$

where $W_i(\hat{\alpha}) = diag[R_{i1}\hat{\omega}_{i1}(\hat{\alpha}), \ldots, R_{in_i}\hat{\omega}_{in_i}(\hat{\alpha})]$, is a diagonal matrix containing inverse probability weights for the ith subject, for $j = 2, \ldots, n_i$, and $\hat{\alpha}$ is a vector of nuisance parameters handled by the introduction of a working correlation matrix. The difference between equations (2.11) and (2.8) is that (2.11) have weights while (2.8) do not have. In (2.11), $\hat{\omega}_{i1} = 1$, and $V_i = A_i^{\frac{1}{2}} R_i(\rho) A_i^{\frac{1}{2}}$ is a $n_i \times n_i$ working covariance matrix for Y_i in which $R_i(\rho)$ is an $n_i \times n_i$ working correlation matrix. Now as stated earlier, ω_{ij} is often unknown and needs to be estimated. It requires modelling the missing data process in order to obtain the weights ω_{ij}.

Let $\lambda_{ij}(\alpha) = P(R_{ij} = 1 | R_{i(j-1)} = 1, X_i, Y_i, \alpha)$ be the probability of a response being observed at time point j for the ith subject given that the subject was observed at time point $j - 1$. If MAR holds, the model for $\lambda_{ij}(\alpha)$ can include the observed history such that:

$$\lambda_{ij}(\alpha) = P(R_{ij} = 1 | R_{i,j-1} = 1, X_i, h_{ij}, \alpha), \qquad h_{ij} = Y_{i1}, \ldots, Y_{i,j-1}. \quad (2.12)$$

The missingness mechanism only depends on observed data and may be specified up to a $(q \times 1)$ vector of unknown parameters, α. Here, λ_{ij} can be modelled as a logistic regression model with Z'_{ij}, a vector of predictors, which may include missingness indicator variables, covariates and previous responses such that

$$\text{logit}[\lambda_{ij}(\alpha)] = Z'_{ij}\alpha. \tag{2.13}$$

The log partial likelihood for ith subject can then be expressed as:

$$\ell(\alpha) = \sum_{i=1}^{N}\sum_{j=2}^{n_i} R_{i(j-1)}\ln\{\lambda_{ij}(\alpha)^{R_{ij}}[1-\lambda_{ij}(\alpha)]^{1-R_{ij}}\}. \tag{2.14}$$

Differentiating (2.14) with respect to α gives the estimating equations

$$S_i(\alpha) = \left\{\sum_{i=1}^{N}\sum_{j=2}^{n_i} R_{i(j-1)}[R_{ij}-\lambda_{ij}(\alpha)]\right\}. \tag{2.15}$$

Setting (2.15) equal to zero, yields $\hat{\alpha}$, and consequently $\hat{\lambda}_{ij}(\hat{\alpha})$ can be obtained as an estimate of $\lambda_{ij}(\alpha)$. Consistent parameter estimates can be obtained conditional on two assumptions (Hogan, Roy and Korkontzelou, 2004):
(1) *Non-zero probability of remaining in the study*: Given past history, the probability that individual i is still in the study at time j is bounded away from zero; $P[R_{ij} = 1|R_{i,j-1} = 1, X_i, h_{ij}] > 0$.
(2) *Correct specification of dropout model*: The probability of dropout at time j must be correctly specified: $\nu_{ij}(\alpha) = P[R_{ij} = 0|R_{i,j-1} = 1, X_i, Y_{i,j-1}]$. Under monotone missingness, the probabilities of remaining in the study is therefore:

$$P[R_{ij} = 1|R_{i(j-1)} = 1, X_i, h_{ij}, \alpha] = \prod_{k=1}^{j}\{1-\nu_{ik}(\alpha)\} = \prod_{k=1}^{j}\hat{\lambda}_{ik}(\hat{\alpha}). \tag{2.16}$$

Thus, the weight $\hat{\omega}_{ij}(\hat{\alpha})$, the inverse of the unconditional probability of being observed at time j, can be calculated:

$$\hat{\omega}_{ij}(\hat{\alpha}) = \frac{1}{1\times(\hat{\lambda}_{i1}(\hat{\alpha}))\times\cdots\times(\hat{\lambda}_{ij}(\hat{\alpha}))}, \qquad j = 2,\ldots,n_i, \tag{2.17}$$

where $\hat{\omega}_{ij}(\hat{\alpha}) = 1$ for $j = 1$. Therefore, if the above two assumptions hold, and if dropout follows an MAR mechanism, the estimators of the parameters $\hat{\beta}$ in the weighted marginal model for a continuous response (with an identity link) will be of the form

$$\hat{\beta} = \{\sum_{i=1}^{N} X'_i\hat{V}_i^{-1}W_i(\hat{\alpha})X_i\}^{-1}\sum_{i=1}^{N}(X'_i\hat{V}_i^{-1}W_i(\hat{\alpha})Y_i), \tag{2.18}$$

and

$$Cov(\hat{\beta}) = \left\{\sum_{i=1}^{N} X_i' V_i^{-1} W_i(\hat{\alpha}) X_i\right\}^{-1} \left(\sum_{i=1}^{N} X_i' V_i^{-1} W_i(\hat{\alpha}) W_i(\hat{\alpha})' X_i\right) \left\{\sum_{i=1}^{N} X_i' V_i^{-1} W_i(\hat{\alpha}) X_i\right\}^{-1}, \tag{2.19}$$

where $\hat{\beta}$ is consistent for β, and $\hat{\alpha}$ is a consistent estimator of α under a correctly specified model, $\lambda_{ij}(\alpha)$.

2.4 Simulation study

2.4.1 Data generation

In this section, we present a simulation study to illustrate the comparative performance of IPW-GEE, MI and DL. We generated data to mimic a typical longitudinal study. In particular, we are interested in modelling a continuous outcome Y as a function of predictors, X. The outcome of interest was generated at 6 study occasions, $j = 1, 2, \ldots, 6$. The vector of responses Y_i for the ith subject is assumed to be normally distributed. In essence, we performed simulations based on a linear mixed model for Y, with a linear predictor of the form

$$E[Y_{ij}|x_{ij}] = \beta_0 + x_{ij}'\beta + b_i, \qquad b_i \sim N(0, \sigma^2). \tag{2.20}$$

where x_{ij} contains x_{ij1} and x_{ij2} denoting binary group effects gender and site of study, respectively; x_{ij3} is a continuous variable denoting age of the subject and x_{ij4} is a continuous time variable. The x's were generated using random number generators following their respective distributions. The regression coefficients were fixed at $\beta' = (\beta_0,\ \beta_1,\ \beta_2,\ \beta_3,\ \beta_4) = (10,\ -0.2,\ -0.2,\ 0.05,\ 0.23)$. The random effects b_i account for the individual to individual variability and assumed to be independent and identically, normally distributed estimated such that, $b_i \sim \mathrm{N}(0, 0.76^2)$. The choice of these values for simulation was informed by a preliminary exploratory study carried out on a child malnutrition dataset. The goal of this simulation was to simulate correlated longitudinal data via the use of random effects. We demonstrate how to handle incomplete data after simulating dropout from the complete simulated datasets. For simplicity we did not include interaction terms. Using model (2.20), we performed $S = 300$ simulation replications, for each of sample sizes: $N = 125,\ 250,\ 500$. Here, $N = 125$ corresponds to a moderately small sample size, $N = 250$ to a moderate sample size and $N = 500$ represents a large sample size.

First, from the full datasets generated, we carried out a repeated measures likelihood analysis that employs a linear marginal model combined with a variance-covariance model that incorporates correlations. In particular, we assumed a compound symmetry structure. We ensured that the number and timing of observations are equal for all subjects. The data were assumed Gaussian and thus a direct likelihood method was used to estimate the model parameters. We present these results in Table 2.1.

Note here that the R-side random effect with a compound symmetry (CS) structure only is analogous to the random effects model when the distribution is Gaussian and the G-matrix is positive definite. But, note that the CS symmetry only follows when the random-effects structure consists of a single random intercept as the case used in the simulation model equation (2.20).

Mixed modelling analysis using SAS procedures like MIXED is computationally intensive, requiring substantial amount of memory and execution time. A number of recommendations have been presented to circumvent the memory issues and to reduce execution times. See, for example, Kiernan et al. (2012); Tao et al. (2015). Together with other efficient coding techniques, the choice of a simpler covariance structure can be beneficial. However, we caution that as much as one would want to use the simpler structure, it should be guided by expert knowledge or investigated to see if the structure is supported by the data.

Table 2.1: Maximum likelihood parameter estimates (Est), standard errors (S.E.) and p-values for the full datasets simulated at different sample sizes: $N = 125,\ 250,\ 500$ and $S = 300$ simulation replications. True values: $\beta_0 = 10.0,\ \beta_1 = -0.2,\ \beta_2 = -0.2,\ \beta_3 = 0.05,\ \beta_4 = 0.23$.

	N = 125		N = 250		N = 500	
Param	Est (S.E.)	p-value	Est (S.E.)	p-value	Est (S.E.)	p-value
β_0	9.7770 (0.0110)	<.0001	9.7759 (0.0073)	<.0001	9.7936 (0.0053)	<.0001
β_1	-0.2129 (0.0075)	<.0001	-0.2067 (0.0054)	<.0001	-0.2028 (0.0037)	<.0001
β_2	-0.1952 (0.0086)	<.0001	-0.1893 (0.0060)	<.0001	-0.1961 (0.0042)	<.0001
β_3	0.0509 (0.0003)	<.0001	0.0505 (0.0002)	<.0001	0.0501 (0.0001)	<.0001
β_4	0.2300 (0.0000)	<.0001	0.2300 (0.0000)	<.0001	0.2300 (0.0000)	<.0001

Examining Table 2.1, we notice that the parameter estimates obtained get closer to the true values as the sample size is increased. This is true for all parameters except β_2 and β_4. For β_2, $N = 125$ produces an estimate closer to the true value than $N = 250$. For β_4, the true value is reproduced for all three sample sizes. With regard to standard errors, they get smaller as the sample size increases, but for β_4, the same standard error is produced for all sample sizes. These results indicate the gain in having a larger sample size for simulation studies. But, note here that this gain may depend on the question being answered, and may not generally be true in all situations.

Next we created dropouts on the outcome variable y according to a simple mechanism: Missing at random, dependent on the continuous variable x_{ij3}.

In other words, let $drp = x_{ij3}$, where x_{ij3} is as previously described in equation (2.20). Then, let $den = \max x_{ij3}$, i.e., the largest value amongst x_{ij3} values in the data. Next, calculate the probability of dropout at the j^{th} occasion: $pdrpt_{ij} = drp/den$. Finally, if the probability of dropout at j is greater than some uniformly distributed random value (u), then a value of the outcome y drops at occasion $j+1$, i.e., if $pdrpt_{ij} > u, \quad u \sim unif[0,1]$, then $y_{i,j+1}$ misses. With this approach MAR monotone missingness patterns were achieved at the following approximate rates: $8\%, 19\%$ and 33%. These rates denote low, moderate and high dropout rates, respectively. We ensured ignorability by not allowing dropout at occasion j to depend on y_{ij} itself. Different dropout rates were achieved by varying the occasion where dropouts started. These dropout rates indicated the percentages of data missing by the end of the study follow up.

2.4.2 Parameter estimation

The generated incomplete datasets were subjected to the three analysis methods, namely IPW-GEE, MI and DL.

MI was carried out using SAS PROC MI, by assuming multivariate normality on the variables. For valid results, the imputation model and the analysis model should be congenial. For congeniality, it means that the imputation model must contain at least all the variables that are intended to be included in the analysis model. It is recommended that variables that are predictive of the missingness are included in the imputation procedure. In this way, MAR can be satisfied. In this simulation study, the imputation and analysis model were the same and the default 25 imputations were used in SAS version 9.4. However, with advanced computer systems nowadays, higher numbers of imputations do not pose a big problem in terms of space and time requirements. Suggestions have been made regarding the choice of number of imputations. See, for example, Schafer (1997); White, Royston, and Wood (2011). Nonetheless, we concur with Kombo, Mwambi and Molenberghs (2016) that analyst's discretion on this matter is highly important, based on the problem at hand.

To draw the imputations, we used the Markov Chain Monte Carlo (MCMC) approach with default SAS specifications for iterations, Jeffreys prior and the expectation-maximization posterior mode.

On the other side, IPW-GEE was implemented using the SAS macros provided by Molenberghs and Verbeke (2005). In particular, the macros "DROPOUT" and "DROPWGT" were used to create the dataset for IPW-GEE analysis. The macro DROPOUT was used to estimate the probabilities of dropout and the macro DROPWGT passed the weights (predicted probabilities) to be used in the weighted estimating equations. Unlike MI, the IPW-GEE method requires specification of a model for the dropout. For dropout in longitudinal settings, the dropout model takes the form of a discrete hazard model such that; at each measurement occasion, the occurrence of a dropout is regressed on previous and current values of the outcome as well as the covariates. In principle, for continuous outcomes, the dropout model can be easily generalized by including the full history (say, $H_{ij} = y_{1j}, \ldots, y_{ij-1}$), and/or covariates and also allowing interactions with time (Molenberghs et al., 2014). Based on our experience in this study, we assumed a logistic regression model (2.21), in which $y_{i,j-1}$ is the subject's previous outcome. The variables x_3 and x_4 were also used as covariates for the dropout model:

$$\text{logit}[P(D_i = j|D_i \geq j)] = \psi_0 + \psi_1 y_{i,j-1} + \psi_2 x_{ij3} + \psi_3 x_{ij4} + \psi_4 y_{i,j-1} * x_{ij4}. \tag{2.21}$$

Here, $\psi_0, \psi_1, \psi_2, \psi_3, \psi_4$ are regression parameters to be estimated. The variable x_4 (assessment time points) is used as a continuous variable in both the dropout model and the main analysis model.

Essentially, obtaining the weighted GEE estimates for regression parameters, β, is a two-step algorithm: (1) Fit a logistic regression to estimate the weights, and (2) estimate β by specifying the estimated weights in the WEIGHT statement in a SAS procedure, say GENMOD. See Molenberghs and Verbeke (2005) for details of implementing the IPW method macros.

In the direct likelihood analysis, data are analysed the way they are without imputation nor deletion of the incomplete cases. Because we assumed MAR, a direct-likelihood ignorable analysis was conducted and parameter estimates obtained by specifying method = ML, and the GAUSS estimation algorithm used.

In MI and DL analyses, the SAS procedure MIXED was used. We specified a REPEATED statement for the DL approach. Note here that, the repeated statement indicates how PROC MIXED should order observations for a given subject. Without the repeated statement, the procedure assumes that the observations for a given subject are listed in an appropriate order within the data and have no missing values. Then, different results may be obtained (with and without the repeated effect listed) for certain covariance structures.

Notice here that parameters from a marginal model (e.g., the GEE) and a hierarchical model (e.g., the generalized linear mixed model (GLMM)), in case of non-Gaussian outcomes have to be interpreted differently. This is because the fixed effects in GLMMs are interpreted conditional on the random effects. In our case, this issue does not arise since there is no difference in the interpretation of parameter estimates from the two model formulations in the Gaussian outcome case. But to make the models be in the same class we fitted marginal models in the DL analysis and the GEE analysis after MI. In other words, a mixed-effects model was not fitted in either of the cases.

To asses the comparative performance of the DL, MI and IPW-GEE methods, we used relative bias and efficiency. We defined relative bias as the difference between the true value, β_T, and the average parameter estimate from the DL, MI and IPW-GEE, $\hat{\beta}_M$, (based on the 300 data replications) divided by the true value, i.e., Relative Bias $= (\beta_T - \hat{\beta}_M)/\beta_T$. Efficiency is the variability of an estimate around the true population parameter. We compute it as the average width of the 95% confidence interval – which is usually approximately four times the magnitude of the standard error.

2.4.3 Simulation results

Here, we present the results of the simulation study. Relative bias and efficiency estimates are presented for incomplete data methods namely; DL, MI and IPW-GEE. We also present results for the full datasets (FD), i.e., the datasets before creating dropouts. The results are based on 300 simulated dataset replications. A better method is expected to produce parameter estimates closer or similar to the true values used to simulate the complete datasets, hence a small relative bias. Likewise, a small efficiency value denotes a better or precise method. Results are presented in Tables 2.2, 2.3 and 2.4, for $8\%, 19\%$ and 33% dropout rates respectively. In the three tables, the largest relative bias and efficiency values are presented in boldface.

Table 2.2: Relative bias and efficiency estimates for MI, DL and IPW-GEE methods: Dropout rate = 8%. Simulation replications, $S = 300$. We also present estimates for the full datasets (FD).

Sample		Relative bias				Efficiency			
size	Par	FD	DL	MI	IPW-GEE	FD	DL	MI	IPW-GEE
N = 125	β_0	**0.0223**	**0.0223**	0.0027	-0.0119	0.0440	0.0438	0.0653	**1.1192**
	β_1	-0.0645	-0.0710	-0.0743	**-0.7570**	0.0302	0.0303	0.0293	**0.4508**
	β_2	0.0240	0.0200	0.0221	**-0.2015**	0.0344	0.0341	0.0322	**0.5444**
	β_3	-0.0174	-0.0174	-0.0172	**0.0420**	0.0011	0.0011	0.0011	**0.0160**
	β_4	0.0000	**-0.0026**	-0.0019	0.0000	0.0000	0.0013	0.0026	**0.0112**
N = 250	β_0	**0.0224**	**0.0224**	0.0033	-0.0049	0.0291	0.0291	0.0463	**0.8124**
	β_1	-0.0335	-0.0365	-0.0369	**-0.2365**	0.0216	0.0215	0.0208	**0.3668**
	β_2	0.0535	0.0510	0.0496	**0.0750**	0.0238	0.0238	0.0226	**0.4036**
	β_2	-0.0100	-0.0096	-0.0094	**0.0700**	0.0008	0.0008	0.0008	**0.0132**
	β_4	0.0000	**-0.0013**	**-0.0013**	0.0000	0.0000	0.0001	0.0020	**0.0080**
N = 500	β_0	0.0206	**0.0207**	0.0011	0.0157	0.0214	0.0215	0.0327	**0.6004**
	β_1	-0.0140	-0.0160	0.0157	**0.0590**	0.0148	0.0150	0.0147	**0.2672**
	β_2	0.0195	0.0175	0.0199	**0.2285**	0.0168	0.0170	0.0160	**0.2896**
	β_3	-0.0028	-0.0026	-0.0028	**-0.0140**	0.0006	0.0006	0.0005	**0.0100**
	β_4	0.0000	**-0.0022**	-0.0019	0.0000	0.0000	0.0007	0.0014	**0.0036**

Examining Table 2.2, and considering relative bias, we notice that largest values are produced by IPW-GEE and DL, where DL produced the largest values for β_0 and β_4. IPW-GEE produced largest values for β_1, β_2 and β_3. This was consistent for all sample sizes. In most cases the relative biases produced by FD, DL and MI are very close, and in some cases equal for two methods. Regarding efficiency, all largest values were produced by IPW-GEE method.

Table 2.3: Relative bias and efficiency estimates for MI, DL and IPW methods: Dropout rate = 19%. Simulation replications, $S = 300$. Estimates for the full datasets (FD) are also provided.

Sample		Relative bias				Efficiency			
size	Par	FD	DL	MI	IPW	FD	DL	MI	IPW
N = 125	β_0	**0.0223**	0.0220	0.0021	-0.0049	0.0440	0.0449	0.0616	**1.0428**
	β_1	-0.0645	-0.0645	-0.0653	**-0.5605**	0.0302	0.0308	0.0268	**0.4436**
	β_2	**0.0240**	-0.0010	0.0015	-0.0010	0.0344	0.0350	0.0295	**0.5224**
	β_3	-0.0174	-0.0180	-0.0180	**0.0520**	0.0011	0.0011	0.0010	**0.0156**
	β_4	0.0000	0.0000	**-0.0002**	0.0000	0.0000	0.0029	**0.0049**	0.0048
N = 250	β_0	**0.0224**	0.0223	0.0031	-0.0095	0.0291	0.0294	0.0438	**0.7780**
	β_1	-0.0335	**-0.0345**	-0.0341	-0.0185	0.0216	0.0215	0.0188	**0.3640**
	β_2	**0.0535**	0.0420	0.0414	-0.0315	0.0238	0.0244	0.0209	**0.3964**
	β_3	-0.0100	-0.0096	-0.0094	**0.0900**	0.0008	0.0008	0.0007	**0.0128**
	β_4	0.0000	-0.0013	**-0.0020**	0.0000	0.0000	0.0021	0.0030	**0.0080**
N = 500	β_0	**0.0206**	0.0205	0.0009	0.0189	0.0214	0.0218	0.0313	**0.6056**
	β_1	-0.0140	-0.0145	0.0159	**0.0830**	0.0148	0.0151	0.0135	**0.2760**
	β_2	0.0195	0.0150	0.0165	**0.3420**	0.0168	0.0170	0.0149	**0.2948**
	β_3	-0.0028	-0.0028	-0.0030	**-0.0060**	0.0006	0.0006	0.0005	**0.0100**
	β_4	0.0000	0.0004	**0.0006**	0.0000	0.0000	0.0015	**0.0024**	0.0004

Shifting focus to Table 2.3, with a 19% dropout rate, the scenario observed in Table 2.2 is slightly changed. Here, MI produced the largest relative bias values for β_4 in all sample sizes while DL produced one largest value for $\beta_1(N = 250)$. Although slightly different from its performance in Table 2.2, IPW-GEE also here produced most of the largest bias values. Looking at efficiency, IPW-GEE produced the largest values for all cases except for β_4 $(N = 125,\ 500)$ which were produced by MI.

In Table 2.4, the trends are largely similar to what was observed in Table 2.2. IPW-GEE produced the largest relative bias for β_1, β_2 and β_3 for all sample sizes, while DL produced largest values for $\beta_0(N = 125,\ 500)$ and $\beta_4(N = 500)$. Regarding efficiency, again here largest values were produced by IPW-GEE for all cases except for β_4, where they were produced by MI.

Generally, IPW-GEE produced the most biased estimates relative to the other methods. Similarly, largest efficiency estimates were also produced by IPW-GEE. This was probably not strange since the IPW-GEE's standard errors were notably larger, hence wider 95% confidence intervals. Overall, we notice that FD, DL and MI are very close to each other in performance while IPW-GEE performs slightly different from the other methods.

Table 2.4: Relative bias and efficiency estimates for MI, DL and IPW methods: Dropout rate = 33%. Simulation replications, $S = 300$. We also provide estimates for the full dataset (FD).

Sample		Relative bias				Efficiency			
size	Par	FD	DL	MI	IPW	FD	DL	MI	IPW
N = 125	β_0	0.0223	**0.0226**	0.0029	-0.0117	0.0440	0.0462	0.0569	**1.0444**
	β_1	-0.0645	-0.0610	-0.0648	**-0.9815**	0.0302	0.0316	0.0251	**0.4580**
	β_2	0.0240	0.0185	0.0201	**-0.1170**	0.0344	0.0351	0.0262	**0.5360**
	β_3	-0.0174	-0.0174	-0.0172	**0.0480**	0.0011	0.0012	0.0009	**0.0160**
	β_4	0.0000	0.0000	**-0.0076**	0.0000	0.0000	0.0050	**0.0055**	0.0008
N = 250	β_0	**0.0224**	0.0223	0.0033	-0.0092	0.0291	0.0309	0.0408	**0.7828**
	β_1	-0.0335	-0.0300	-0.0031	**-0.7430**	0.0216	0.0225	0.0173	**0.3804**
	β_2	0.0535	0.0520	0.0517	**0.0840**	0.0238	0.0240	0.0189	**0.4144**
	β_3	-0.0100	-0.0082	-0.0080	**0.0840**	0.0008	0.0008	0.0006	**0.0136**
	β_4	0.0000	-0.0035	**-0.0039**	0.0000	0.0000	0.0036	**0.0039**	0.0008
N = 500	β_0	**0.0206**	**0.0206**	0.0010	0.0137	0.0214	0.0222	0.0300	**0.6180**
	β_1	-0.0140	-0.0135	-0.0145	**-0.5405**	0.0148	0.0153	0.0128	**0.2892**
	β_2	0.0195	0.0160	0.0196	**0.3390**	0.0168	0.0173	0.0134	**0.3076**
	β_3	-0.0028	-0.0022	-0.0080	**0.0100**	0.0006	0.0006	0.0005	**0.0108**
	β_4	0.0000	**-0.0030**	-0.0029	0.0000	0.0000	0.0026	**0.0028**	0.0004

2.5 Application

2.5.1 Data: Recovery from severe childhood malnutrition (RSCM)

This section aims at elucidating on the findings of the simulation study conducted in Section 2.4 using a real data application. The application involves a longitudinal study of Kenyan children recovering from severe childhood malnutrition (RSCM). The RSCM study is a clinical trial, which was conducted by KEMRI/Wellcome Trust Research Programme, Kilifi, Kenya. The data were collected for 1778 children in total, aged 2 to 59 months in 4 different hospitals in Kenya. All were recruited in hospital where they had been admitted with severe, acute malnutrition. The children were enrolled shortly prior to discharge and followed up for one year. Children who died or for other reasons (e.g., deformity), full or complete sequence measurements were not possible (meaning one or more variables will always be missing) were excluded, leaving 1138 children who satisfied the inclusion criteria for this analysis. Participants were allocated using a computer generated randomization method in permuted blocks of 20, stratified per hospital and age (younger or older than 6 months). Treatment was concealed and patients, family and all trial staff were masked to the treatment assignment. The participants were given the recommended medical care and feeding, and followed for 12 months. After the initial visit, subjects were followed for 9 more scheduled visits. The primary endpoint was mortality, assessed each for 6 months, then for every 2 months for the last 6 months. Secondary endpoints were nutritional recovery, readmission to hospital, and illness episodes treated as an outpatient. Analysis was via intention to treat (Berkley et al. 2016). At the initial visit, baseline data on health, anthropometry, and socioeconomic status of probable prognostic importance were obtained by the study teams at each site. Children received standard care for severe acute malnutrition (SAM) and other medical conditions according to WHO guidelines for complicated SAM.

In this study, about 40% of the children had one or more anthropometric data points missing meaning that not all measurements were taken for every child at each scheduled visit. The proportion of missing data amongst anthropometric variables was 9.8%. The variables' names and descriptions are as follows: *sex*: [1=Male, 0=Female]; *age*: age in months calculated from date of enrolment and date of birth; *site* [1=rural, 2=urban]: the four hospitals (Kilifi, Malindi, Mombasa and Mbagathi) where the trial was conducted. We combined Kilifi and Malindi as rural while Mombasa and Mbagathi (Nairobi) were combined as urban; *muac*: mid-upper arm circumference in centimetres; *zhc*: head circumference; *zwei*: weight for age; *zlen*: length for age; *zwfl*: weight for length. The anthropometric variables; *zhc, zwei, zlen*, and *zwfl* are continuous, in form of Z scores calculated using the World Health Organization (WHO) macro for STATA (2006) while *muac* are raw values. Further trial details may be accessed at Berkley et al. (2016).

2.5.2 Analysis results

In this application, we consider *muac* to be the outcome of interest in the longitudinal data analysis. We model *muac* as a linear function of predictors, say X, accounting for the correlations among multiple observations within a subject. The outcome *muac* is not fully observed, but unlike the simulation study in Section 2.4 where we imposed monotone missingness patterns, here the patterns are non-monotone. The outcome variable *muac* has about 8% intermittent missing values, but the predictors are fully observed. However, this is not a problem since both DL analysis and MI using MCMC method can handle non-monotone/arbitrary missing values with no issues. In fact, methods designed for arbitrary missing data patterns can also handle monotone missing data patterns but the reverse is not true. Some methods are restricted to monotone patterns only.

For prediction regarding measures of malnutrition, information about *zhc, zwei, zlen*, and *zwfl* may be considered extraneous to the regression model, but may be associated with the *muac* as a measure of malnutrition, and consequently make the MAR assumption more viable. We incorporated these auxiliary variables in the imputation model but not in the analytical model.

On the other side, to avoid complexity in the analysis, for IPW-GEE, we first monotonized the arbitrary patterns. We therefore filled in about half of the missing values (using mcmc impute=monotone statement in PROC MI, and the default 25 imputations), then performed the weighted GEE analysis. The results of the methods are presented in Table 2.5.

Notice here that one does not have to model the correlation structure of the response model correctly; one only needs to use a working correlation structure to produce consistent estimates, i.e., the GEE approach will produce consistent estimator of even when the working correlation structure is far from the true structure. However, one convenient way is to use the independence structure which has been shown to maintain high efficiency in many cases (Zeger, Liang and Albert, 1988). We used the independence working correlation structure. We note the poor inference for the sex and site parameters with IPW-GEE which could possibly be attributed to a mis-specification of the weight model, or difficulties of handling interim missigness.

Table 2.5: Parameter estimates (Est), standard errors (Std Err) and p-values from MI, DL and IPW-GEE methods from the RSCM dataset. Missing data; non-monotone 8% on the outcome variable. Standard errors are in brackets.

	DL		MI		IPW-GEE	
Effect	Est (Std Err)	P-value	Est (Std Err)	P-value	Est (Std Err)	P-value
Intercept	10.4749 (0.0751)	<.0001	10.4843 (0.0756)	<.0001	10.2772 (0.0615)	<.0001
sex (F)	-0.0500 (0.1031)	0.6280	-0.0510 (0.1040)	0.6242	-0.2361 (0.0865)	0.0063
site (R)	-0.2312 (0.0646)	0.0004	-0.2258 (0.0651)	0.0005	-0.1547 (0.0627)	0.0136
age	0.0400 (0.0053)	<.0001	0.0400 (0.0053)	<.0001	0.0204 (0.0049)	<.0001
month	0.2159 (0.0037)	<.0001	0.2116 (0.0036)	<.0001	0.2943 (0.0048)	<.0001
age*sex (F)	-0.0027 (0.0076)	0.7225	-0.0025 (0.0077)	0.7427	0.0482 (0.0068)	<.0001

Note: F = female R = rural

From the application results, it is clear that the values of DL and MI are closer to each other than the IPW-GEE. The performance is consistent for all parameters. Although the values produced by IPW-GEE appear to be slightly different from the other two methods, they are not very far. In fact, the difference between IPW-GEE and the other methods is less than 0.09 except for two cases (Intercept, sex). For sex and age-by-sex interaction, DL and MI estimates are insignificant, while those of IPW-GEE are significant. Also, the interaction effect in IPW-GEE is opposite in direction to those produced by MI and DL. The MI and DL methods give consistent interpretation of effects across all model terms while the IPW-GEE disagreed on sex and the interaction term. Therefore, we checked the performance of the methods when a month-by-sex interaction is included, see Table 2.6. We noticed that concerning parameter estimates and standard errors, the performance is similar to what was observed in Table 2.5. However, in this table, the month-by-sex interaction is significant for all the three methods, unlike what was observed for age-by-sex interaction in Table 2.5, where an insignificant interaction effect was produced by MI and DL when it was highly significant in IPW-GEE.

Table 2.6: Parameter estimates (Est), standard errors (Std Err) and p-values from MI, DL and IPW-GEE methods. The RSCM dataset. Missing data; non-monotone 8% on the outcome variable. Sex-by-time(month) interaction included.

	DL		MI		IPW-GEE	
Effect	Est (Std Err)	P-value	Est (Std Err)	P-value	Est (Std Err)	P-value
Intercept	10.4472 (0.0644)	<.0001	10.4555 (0.0649)	<.0001	9.9785 (0.0514)	<.0001
sex (F)	0.0079 (0.0658)	0.9043	0.0084 (0.0662)	0.8995	0.3755 (0.0508)	<.0001
site (R)	-0.2334 (0.0644)	0.0003	-0.2268 (0.0650)	0.0005	-0.2045 (0.0631)	0.0012
age	0.0387 (0.0038)	<.0001	0.0388 (0.0038)	<.0001	0.0450 (0.0034)	<.0001
month	0.2240 (0.0051)	<.0001	0.2194 (0.0051)	<.0001	0.3064 (0.0070)	<.0001
month*sex (F)	-0.0168 (0.0073)	<.0211	-0.0159 (0.0072)	0.0281	-0.0253 (0.0097)	0.0091

Note: F = female R = rural

Further, the inclusion of both age-by-sex and month-by-sex interaction terms in the model did not yield interpretation contrary to what could generally be inferred from Tables 2.5 and 2.6, for all cases except for sex by IPW-GEE. Here, the sex effect is insignificant while it has been significant in Tables 2.5 and 2.6. Results that include the two interaction terms are presented in Table 2.7.

Table 2.7: Parameter estimates (Est), standard errors (Std Err) and p-values from MI, DL and IPW-GEE methods. The RSCM dataset. Missing data; non-monotone 8% on the outcome variable. Sex-by-time(month) and age-by-sex interactions included.

	DL		MI		IPW-GEE	
Effect	Est (Std Err)	P-value	Est (Std Err)	P-value	Est (Std Err)	P-value
Intercept	10.4326 (0.0772)	<.0001	10.4413 (0.0778)	<.0001	10.2429 (0.0624)	<.0001
sex (F)	0.0380 (0.1099)	0.7292	0.0376 (0.1109)	0.7345	-0.1620 (0.0898)	0.0712
site (R)	-0.2322 (0.0645)	0.0003	-0.2257 (0.0651)	0.0005	-0.1643 (0.0625)	0.0086
age	0.0400 (0.0053)	<.0001	0.0400 (0.0053)	<.0001	0.0197 (0.0049)	<.0001
month	0.2240 (0.0051)	<.0001	0.2194 (0.0051)	<.0001	0.3090 (0.0070)	<.0001
age*sex (F)	-0.0026 (0.0076)	0.7321	-0.0025 (0.0077)	0.7423	0.0489 (0.0068)	<.0001
month*sex (F)	-0.0168 (0.0073)	0.0212	-0.0159 (0.0072)	0.0281	-0.0284 (0.0096)	0.0032

Note: F = female R = rural

Even after varying the assumptions in the analysis model, we note that the performance in Tables 2.5 – 2.7 was consistent. That is, MI and DL are closer or equal to each other compared to IPW-GEE with any of the aforementioned methods. Overall, we note that there is a gain in using MI and DL (and any of the two can be used) compared to IPW-GEE. For correct inference from IPW-GEE one needs to pay attention in correct specification of the weight model and the handling of interim missingness if present.

2.6 Discussion and conclusion

We assessed the performance of IPW-GEE and MI in handling incomplete continuous outcomes. The results of the two methods were also checked against the DL analysis method, which is valid under the MAR assumption. First, dropouts were created at different rates on simulated datasets of various sample sizes and the three methods applied to these incomplete datasets. Then, the same methods were used on the *RSCM* dataset as an application to real data. In this paper, the dropout rates were diverse, ranging from 8% to 33% in order to investigate the performance of the methods when different amount of data are missing.

Generally, the results showed that all the methods can be satisfactorily used for incomplete continuous outcomes with the assumption of a MAR mechanism.

Specifically, when we consider both relative bias and efficiency, a better performance was observed for MI and DL over IPW-GEE in the simulation study. This performance was, however, somehow expected since as reported in Schafer and Graham (2002), IPW-GEE method can be less powerful compared to a Bayesian approach like the MI. Also, IPW-GEE is more commonly used in marginal models for discrete outcomes than continuous outcomes data (Fitzmaurice, Molenberghs and Lipsitz, 1995; Robins, Rotnitzky and Zhao, 1995). The comparative simulation study together with the real application results tend to agree with this view.

Considering the performance of MI and DL in the simulation study, we noticed that they performed very close to each other. This scenario can happen because as reported by Collins, Schafer and Kam (2001), MI and DL analysis can produce similar results when data are missing on the outcome and the same information is used for both models. We also observed from the results that the DL and MI estimates were close to the FD estimates. It has been found by other researchers that DL can produce unbiased estimates that are comparable to those of the full data analysis (Kadengye et al., 2012; Molenberghs and Verbeke, 2005). However, there are situations where MI can be more justified (Molenberghs and Verbeke, 2005). In this simulation study, we did not find proof to confidently claim that MI was better than the DL method. An important note is that the use of direct likelihood methods is attractive to analyze incomplete data and might be presented as the analyst's ideal option, but, such methods have computational complexity particularly considering the longitudinal nature of the data. Care has to be taken in the way DL is implemented.

On the RSCM data application, it was also observed that the MI and DL analysis produced parameter estimates and corresponding standard errors equal or very close to each other.

Generally, our results suggested that, although IPW-GEE was traditionally found to be attractive and specific to longitudinal discrete binary outcomes, it may also be used for continuous outcomes, subject to MAR dropouts. However, it may be slightly less efficient compared to DL and MI. An interesting observation is on the standard errors. IPW-GEE (in the application) produced slightly lower standard errors compared to the two other methods. This performance opens up IPW-GEE for further investigation under continuous outcome scenario through carefully planned simulation studies combined with theoretical examination covering wider possible alternatives and assumptions.

Chapter 3

A Simulation Study Comparing Weighted Estimating Equations with Multiple Imputation Based Estimating Equations in the Analysis of Correlated Count Data

Abstract

A frequent statistical problem with the analysis of longitudinal data is that subjects may drop out of the study before the end of the follow-up period. This is in addition to the patent feature that multiple observations taken from the same individual are correlated. If the mechanism leading to dropout or missing data in general is not ignorable, one has to be careful for biased estimates of the parameters of interest. However, if the dropout process is ignorable and maximum likelihood or Bayesian estimation is chosen, then unbiased estimators follow (Little and Rubin 2014). This paper focuses on dropouts missing at random for longitudinal count data. Using a simulation study, semi-parametric methods, namely weighted estimating equations and multiple imputation followed by generalized estimating equations are compared. Over time, several papers have been written regarding this comparison Clayton et al., 1998; Beunckens, Sotto and Molenberghs, 2008; Satty, Mwambi and Molenberghs, 2015) but focus was

mostly on binary data. Results show that multiple imputation based generalized estimating equations outperforms the weighted generalized estimating equations in estimating regression coefficients.

3.1 Introduction

Longitudinal data are often encountered in epidemiological, social sciences and medical problems to address various research questions. However, a challenge may arise in the analysis if subjects drop out of the study before the end of the follow-up period. A subject is called a dropout when the response variable is observed through to a certain visit and is missing for all subsequent visits (Diggle et al., 2002; Fielding, Fayers and Ramsay, 2009; Carpenter and Kenward 2013). This dropout, and the fact that the observations themselves are bound to be correlated have to be taken into consideration for valid inferences. In the presence of dropout, appropriate statistical methods have to be chosen since some methods are suitable only for certain missing data mechanisms. It is therefore imperative to consider the mechanism that govern the missingness. Rubin (1976; 1987) and Little and Rubin (2014) classified these mechanisms into three such that: data is said to be missing at random (MAR) if conditional on observed outcomes and probably on the design factors, the distribution of missingness does not depend on unobserved data. Missing completely at random (MCAR), if the missingness distribution is independent of both observed and unobserved data and they are missing not at random (MNAR) for any violation of MAR, so that the it may depend on the unobserved data and possibly on covariates and/or observed outcomes.

Correlation may arise when an outcome is measured repeatedly over a period of time on the same subject (e.g., longitudinal studies) or when multiple outcomes taken one or more times but on the same subject, such as in clinical trials for multiple investigative endpoints. The key idea in longitudinal data analyses is that a correction structure is necessary to account for the within-subject correlations (i.e., the correlated errors). A popular way to deal with correlation is the use of linear mixed models to analyze continuous longitudinal data (Diggle et al., 2002; Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2009) and generalized linear mixed models (GLMMs; Zeger and Karim, 1991; Breslow and Clayton 1993; Wolfinger and O'onnell, 1993; Molenberghs and Verbeke, 2005) for analysis of discrete longitudinal data. In GLMMs, individual-specific random effects are incorporated to explicitly acknowledge the correlation induced by the between-subject variation. GLMMs are useful in the accommodation of nonnormally distributed responses where a nonlinear link is specified between the response mean and the predictor variables which includes subject specific random effects. Generally, they rest upon two building blocks: random effects to account for individual to individual variability and generalized linear models (GLM; McCullagh and Nelder, 1989) for nonnormal data using a link function appropriate for the exponential family member (Bolker et al., 2009).

Initially, advanced methods to deal with dropout were focused mostly on continuous longitudinal data but over time work on discrete longitudinal data, in particular counts and multinomial data types is gaining momentum. Discrete binary or Bernoulli data have also been studied extensively (for example, Preisser, Lohmann and Rathouz, 2002; Schafer, 2003; Ali and Talukder, 2005; Molenberghs and Verbeke, 2005; Smith and Smith, 2006; Beunckens, Sotto and Molenberghs, 2008; Yi, He and Liang, 2011a; Yi, Zeng and Cook, 2011b; Goncalves, Cabral and Azzalin, 2012).

In the analysis of discrete data, the normality assumption in the model is no longer valid, and one has to look for an alternative route. This may call for specification of a full joint distribution for the set of measurements $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})$ per subject in the study. Consequently, the need to specify all moments ensues (Verbeke, Molenberghs and Rizopoulos, 2010). In some cases, when observations are not taken at constant time points for all subjects or where we have longer sequences, specification of a full likelihood and inferences on the parameters may be burdensome. In response, Liang and Zeger (1986) proposed so-called generalized estimating equations (GEEs). They are a common approach to fit marginal models to longitudinal data, particularly for discrete outcomes (Preisser, 2013). These models allow for the correlation structure in the data due to repeated observations on the same subject over time. They require only the correct specification of a univariate marginal distribution and adoption of an assumed working correlation structure. However, GEE based inferences are only valid when data are missing completely at random (Liang and Zeger, 1986). Typically, frequentist methods require the stronger MCAR assumption to yield valid inferences. Robins, Rotnitzky and Zhao (1995) proposed a class of weighted estimating equations (hence WGEE) that make the method valid under the MAR assumption provided that a regression model for dropout is correctly specified even if the repeated measures model is misspecified. From the estimated dropout probabilities, weights are formed and applied to GEE to address the potential bias due to dropout. This approach has so far been studied most for binary outcomes while count data has received less attention. Alternatively, Schafer (2003) proposed the use of multiple imputation (MI) for the missing response values from a fully parametric model then analyzed by a method of choice, whether parametric or not. He argues that MI does interact well with a variety of semi and nonparametric estimation procedures like the marginal GEE which then leads to MI-GEE. See Carpenter and Kenward (2013) for a discussion of nonparametric MI. In this paper, we focus on the comparison of WGEE with MI-GEE in the analysis of correlated longitudinal count data.

In Section 3.2, we introduce the model for correlated longitudinal data and describe the notation used in the paper. The dropout concept is also described here. Section 3.3 deals with the statistical methods used for analysis in the paper. We describe a simulation study, analysis procedures involved, and results in Section 3.4. Section 3.5 draws conclusions about the paper and point out possible areas for further research.

3.2 Model formulation and dropout concepts

3.2.1 The model

In this study, we are interested in longitudinal data from a discrete distribution. In particular, we are dealing with repeated counts from a multivariate Poisson distribution. Although univariate discrete distributions have been studied extensively, multivariate counterparts have not received attention to the same scale. This is due to the computational complexity involved, specifically regarding the calculation of the probabilities. The Poisson distribution falls in this category. Researchers have thus proposed various ways to analyse correlated count data. One approach is the generalized linear mixed model (GLMM). The GLM generalizes the linear regression model where the linear component which is expressed in terms of covariates is related to the response variable via a link function (e.g., the logit link function for binary data and the log link for count data). For GLMMs random effects have to be included in the linear predictor.

Normally, repeated measures within a subject are by design expected to show correlation compared to observations between subjects. This correlation can be captured by means of random effects. Hence, the GLMM is a candidate model. Suppose we have repeated counts, $Y_{ij}, j = 1, 2, \dots n_i$ from subject $i = 1, 2 \dots N$. We can express the GLMM as

$$\ln(\lambda_{ij}|b_i) = X'_{ij}\beta + Z'_{ij}b_i, \tag{3.1}$$

whose conditional mean model would therefore be

$$\mathrm{E}[Y_{ij}|b_i] = \lambda_{ij}|b_i = \exp(X'_{ij}\beta + Z'_{ij}b_i), \qquad b_i \sim \mathrm{N}(0, G) \tag{3.2}$$

where $Y_{ij} \sim \mathrm{Pois}(\lambda_{ij}|b_i)$ is the conditional distribution of the j^{th} observation given the random effects vector b_i for a design vector Z_{ij}, with a rate parameter λ_{ij}. The parameter β is a vector of regression coefficients of interest, with fixed covariates X_{ij}.

Note that throughout this paper models will be considered conditional on the random effects vector b_i or marginal with respect to random effects vector b_i. Also, both conditional and marginal models are conditional on the covariate vector X_{ij} but for simplicity X_{ij} may be suppressed from notation. Furthermore, we assume that the covariates are independent of the random effects. Inference on this model for count data is based on the marginal likelihood (3.3) obtained by integrating the random effects out of the conditional likelihood of every subject such that

$$L(\beta, \Theta; y) = \int f(y_i|b_i) f(b_i) db_i, \tag{3.3}$$

where $f(y_i|b_i)$ is the conditional distribution of the response measure given the random effect, $f(b_i)$ is a distribution of the random effects and Θ denotes unknown parameters of variances/covariances. Integrating out random effects induces a marginal correlation between responses through the same subject (Laird and Ware, 1982). The estimates are thus obtained by integrating out the random effects and maximizing the marginal likelihood. These parameter estimates are those that maximise the marginal likelihood function. Unfortunately, closed forms do not exist for all, but at least for some models (Schabenberger, 2005). Molenberghs et al. (2010) derived the marginal mean and variance specific to Poisson data such that

$$\begin{aligned} \mu_{ij} &= \ln(\lambda_{ij}) = X_{ij}\beta + 0.5 Z'_{ij} D Z_{ij} \\ \text{Var}(Y_i) &= M_i + M'_i (e^{Z_i D Z'_i} - K_i) M_i \end{aligned} \tag{3.4}$$

where K_i is a matrix of 1s and M_i is a diagonal matrix with the elements μ_{ij} along the diagonal.

3.2.2 The dropout

In the complete data vector $Y_i = (Y_{i1} \dots Y_{in_i})'$, $i = 1, \dots, N$, Y_{ij} is the j^{th} response for a subject i and a complete covariate vector X_{ij} at the observation level. Let R_i be a $(n_i \times 1)$ binary random vector where $R_{ij} = 1$ if the i^{th} subject's response is observed at time j and 0 otherwise. With the occurrence of missing values, we will view the complete data set as $Y_i = (Y_i^o, Y_i^m)$, where Y_i^o denotes the set of the actually observed partition and Y_i^m is for the missing data partition. An individual's full data information is jointly distributed as:

$$f(y_i, r_i|X_i, \theta, \psi) = f_y(y_i|X_i, \theta) f_r(r_i|y_i, X_i, \psi), \tag{3.5}$$

where $f_r(r_i|y_i, X_i, \psi)$ is referred to as the missing data model whose parameters are contained in ψ. Note that r_i is a vector of the observed value of the missingness indicator vector R_i for the n_i repeated measurement occasions for individual i. The ψ parameters are generally unknown to the analyst and commonly have no intrinsic scientific value. The full data model of interest is parameterized by θ.

The distribution of R may depend on Y_i. In probability terms we may define these distributions such that the data is said to be missing at random (MAR) if $f(R_i \mid Y_i^0, Y_i^m, X_i, \psi) = f(R_i \mid Y_i^0, X_i, \psi)$. Missing completely at random (MCAR), if $f(R_i \mid Y_i^0, Y_i^m, X_i, \psi) = f(R_i \mid X_i, \psi)$, and they are missing not at random (MNAR) for any violation of MAR, so that $f(R_i \mid Y_i^0, Y_i^m, X_i, \psi) = f(R_i \mid Y_i^0, Y_i^m, X_i, \psi)$. Parameter separability means that the parameters θ and ψ are distinct in the sense that the joint parameter space, $\Omega(\theta, \psi) = \Omega(\theta) \times \Omega(\psi)$. If this holds, we make use of likelihood inference, and the missing data mechanism is MAR, then so-called ignorability assumption (Rubin, 1976; Little and Rubin, 2014) holds. The consequence of ignorability is that then the missing data mechanism does not need to be modelled explicitly. Evidently, because MCAR is a special case of MAR, it also falls under ignorability. For MNAR mechanisms, or when the mechanism is MAR but frequentist inference is used, then ignorability cannot be invoked automatically.

3.3 Statistical methods for handling incomplete correlated data

3.3.1 Generalized estimating equations

GEEs provide a means to conveniently analyze repeated count data with reasonable statistical efficiency (Liang and Zeger, 1986; Smith and Smith, 2006). The method estimates model parameters by iteratively solving a system of equations based on extended quasi-likelihood where the extension to the generalized linear model is towards incorporating correlations. It focuses on the correct specification of the mean, thus avoiding full modelling of the association structure while still obtaining valid inferences. The marginal expectation $E[Y_{ij}] = \mu_{ij}$ can be modelled in terms of covariates through some link function, $g(\mu_{ij}) = \mathbf{x}'_{ij}\beta$. Here, μ_{ij} is the mean response of subject i at time j and $\boldsymbol{\beta}$ is a vector of regression parameters. On the other hand, the marginal variance depends on the marginal mean such that $\text{Var}(Y_{ij}) = \phi\nu(\mu_{ij})$, where ϕ is a scaling parameter. Following Liang and Zeger (1986); Molenberghs and Verbeke (2005) and Birhanu et al. (2011), the generalized estimating equations for the vector $\boldsymbol{\beta}$ have the form:

$$S(\beta) = \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta'} V_i^{-1}(y_i - \mu_i) = 0, \tag{3.6}$$

where $V_i = A_i^{\frac{1}{2}} C_i(\alpha) A_i^{\frac{1}{2}}$ is a covariance matrix of Y_i in which A_i is a diagonal matrix of the marginal variances and $C_i(\alpha)$ expresses the marginal correlation between the repeated measures. Here, $\boldsymbol{\alpha}$ is a vector of nuisance parameters which may be handled by the introduction of a working correlation structure. such as independence, autoregressive of the first order (AR(1)), exchangeable, or unstructured. In the exchangeable structure, the correlations between any two measurements are assumed to be the same regardless of the time from one period to another. In the unstructured case, every pair of measurements is given its own association parameters. On the other hand, for AR(1) the correlations decline exponentially with distance between the measures, i.e., $\text{Corr}(Y_{i,j}, Y_{i,h}) = \rho^{|j-h|}$. Under independence, the identity matrix serves as the working correlation matrix.

When the marginal mean, μ_i is correctly modelled, then under mild regularity conditions the estimator $\hat{\beta}$, solution to (3.6), satisfies:

$$\hat{\beta} \sim \text{AN}(\beta \ , \ \text{Var}(\hat{\beta})), \qquad \text{with} \qquad \text{Var}(\hat{\beta}) = \Psi_0^{-1}\Psi_1\Psi_0^{-1}, \tag{3.7}$$

where,

$$\Psi_0 = \left(\sum_{i=i}^{n} \frac{\partial \mu_i'}{\partial \beta} \ V_i^{-1} \ \frac{\partial \mu_i}{\partial \beta'} \right) \qquad \text{and} \qquad \Psi_1 = \left(\sum_{i=i}^{n} \frac{\partial \mu_i'}{\partial \boldsymbol{\beta}} \ V_i^{-1} \ \text{Var}(y_i) \ V_i^{-1} \ \frac{\partial \mu_i}{\partial \boldsymbol{\beta}'} \right). \tag{3.8}$$

3.3.1.1 Weighted generalized estimating equations

When data are incomplete, GEE suffers bias from its frequentist nature and it is generally valid only under the strong assumption of MCAR (Birhanu et al., 2011). As a remedial measure, the weighted generalized estimating equations (WGEE; Robins, Rotnitzky and Zhao, 1994; 1995), effectively remove bias and provides valid statistical inferences to regression parameter estimates for marginal models in the incomplete longitudinal data scenario by allowing it to be MAR.

The idea of the weighting method is to weight the contributions from subjects with different missingness patterns to the usual GEE formulation by the inverse of the probability that a subject drops out at the time they dropped. Consistent with the definition of the binary indicator variable R_{ij} in Section (3.2.2), such a weight can be written as:

$$\begin{aligned} \omega_{ij} \equiv P(D_i = j) = \prod_{t=2}^{j-1} & [1 - P(R_{it} = 0 | R_{i2} = \ldots = R_{it-1} = 1)] \\ & \times P(R_{ij} = 0 | R_{i2} = \ldots = R_{ij-1} = 1)^{I\{j \le n_i\}}, \end{aligned} \tag{3.9}$$

where $j = 2, 3, \ldots, n_i+1$. Here, D_i is a dropout indicator for the time at which a subject drops out, i.e., $D_i = \sum_{j=1}^{n_i} R_{ij}+1$, and whose realization is d_i. Clearly, $\omega_{i2} = P(R_{i2} = 0)$ because of the assumption that $R_{i1} = 1$ is a sure event since we assume that the all subjects are observed on the first time point so that $2 \leq D_i \leq n_i + 1$. This implies $D_i = n_i + 1$ represents a complete sequence of observations. Thus, in the WGEE approach, a consistent estimate of $\boldsymbol{\beta}$ may be obtained from

$$S(\beta) = \sum_{i=1}^{N} W_i^{-1} \frac{\partial \mu_i}{\partial \beta'} (A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}})^{-1} (y_i - \mu_i) = 0, \tag{3.10}$$

where $W_i = \text{diag}\{R_{i1}\omega_{i1}, \ \ldots, \ R_{in_i}\omega_{in_i}\}$, a diagonal matrix of event specific weights. The weight is given by ω_{ij} when $R_{ij} = 1$ and 0 otherwise. Equivalently, following Molenberghs and Verbeke (2005),

$$S(\beta) = \sum_{i=1}^{N} \sum_{d=2}^{n_i+1} \frac{I(D_i = d)}{\omega_{id}} \frac{\partial \mu_i}{\partial \beta'}(d)(A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}})^{-1}(d) \ \{y_i(d) - \mu_i(d)\} = 0, \tag{3.11}$$

where $y_i(d)$ and $\mu_i(d)$ are respectively the first $d - 1$ elements of y_i and μ_i.
The square root of the diagonal matrix of the variance matrix, say $S(\hat{\beta})$ yields the standard errors. GEE provides two versions of the standard errors i.e., model based and empirical or robust standard errors. Now, although parameter estimates under GEE are valid even if the structure of the covariance matrix is mis-specified, but in such a case the standard errors will not be good and some data based (empirical) adjustments need to be done for more reliable standard errors. In this study, we used the empirical standard errors.

3.3.1.2 Multiple imputation-generalized estimating equations (MI-GEEs)

Under multiple imputation (Rubin, 1978; 1987), each missing value in the data is replaced independently by a vector of $m \geq 2$ plausible values drawn from the conditional distribution of the unobserved values given the observed ones. The variability among the m imputations reflects the uncertainty with which the missing values can be predicted from the observed ones and this is captured by the conditional distribution. After MI, m complete data sets are thus created (imputation stage) each of which can be analyzed by standard complete data methods (analysis stage). Each analysis will produce regression coefficients and corresponding standard errors. These results are then combined into a single inference (pooling stage) using Rubin's (1987) simple rules, thus combining the variation within and across the m imputed data sets.
Notice that MI can be highly efficient even for a smaller value of m provided the proportion of missing values is not exceedingly high. According to Schafer and Olsen (1998),

$m = 3$ to 5 can suffice for adequate results. However, nowadays larger numbers of imputations do not present a problem given the highly efficient computational resources available. Suggestions have been made regarding the choice of number of imputations. See, for example, Schafer (1997); White, Royston, and Wood (2011).

Let $\hat{\beta}_l$ denote the estimate of a parameter of interest β from the l^{th} imputed data set, then the overall estimate from MI is given by

$$\bar{\hat{\beta}} = \frac{1}{m}\sum_{l=1}^{m}\hat{\beta}_l \tag{3.12}$$

and the variance associated with $\bar{\hat{\beta}}$ is thus

$$V = W + \left(\frac{1+m}{m}\right)B, \tag{3.13}$$

where

$$W = \frac{1}{m}\sum_{l=1}^{m} W_l \qquad \text{and} \qquad B = \frac{1}{m-1}\sum_{l=1}^{m}\left(\hat{\beta}_l - \bar{\hat{\beta}}\right)\left(\hat{\beta}_l - \bar{\hat{\beta}}\right)'. \tag{3.14}$$

Here, W measures the within-imputation variability while B measures the between-imputation variability.

As Schafer (2003) stated, MI can be used to create the imputations from a fully parametric model. Then, one analyzes the imputed datasets by a semi-parametric or non-parametric estimation procedure to achieve greater robustness. Paik (1997); Beunckens, Sotto and Molenberghs (2008); Satty, Mwambi and Molenberghs (2015) used MI to fill in missing values for GEE analysis in data that are MAR but for binary outcomes. So GEE can be used after multiple imputation, leading to a hybrid method named MI-GEE (Schafer, 2003). Typically, the missing data mechanism can be further ignored given that the MAR condition holds.

In a simulation study, Beunckens, Sotto and Molenberghs (2008) showed that MI-GEE has good robustness properties against model misspecification compared to WGEE for longitudinal binary data. Satty, Mwambi and Molenberghs (2015) showed that MI-GEE perfomed better when compared to WGEE and GLMM for longitudinal binary data in the presence of dropout. In the present study, we consider count data.

3.3.2 Working correlation structure in GEE

Keeping with previous notation of GEE, if $C_i(\alpha)$ is the true correlation matrix of Y_i, then V_i is the true covariance matrix of Y_i. Usually, the unknown parameters of the working correlation matrix are estimated in an iterating procedure using the current value of parameter vector $\boldsymbol{\beta}$ to compute approximate functions of the Pearson residuals:

$$r_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{\nu_i(\mu_{ij})}}.$$

Selecting an appropriate working structure is pertinent. If the structure is correctly specified, the GEE method yields a best asymptotically normal (BAN) estimator of mean parameters (Gosho, 2014). We note that consistent estimates of regression coefficients β and their standard errors are yielded even with misspecification of the structure. This is because loss of efficiency diminishes as the number of subjects in a study becomes large. However, Rotnitzky and Jewell (1990); Fitzmaurice (1995); Sutradhar and Das (2000); Wang and Cary (2003) indicate that relative efficiency of parameter estimates in GEE is lowered when the correlation structure is misspecified. Because of these reasons for potential loss of efficiency, researchers have developed interest in statistical techniques for selecting a suitable working correlation structure. Some of these strategies include: the quasi-likelihood under the independence model criterion (QIC; Pan, 2001); correlation information criterion (CIC; Hin and Wang, 2009); the Rotnitzky and Jewell criterion (RJC; Rotnitzky and Jewell, 1990) and DEW (Gosho, Hamada and Yoshimura, 2011). Below we briefly describe some of these techniques.

Rotnitzky and Jewell (1990) proposed a selection approach where they define some test statistics as follows (noting that Ψ_0 and Ψ_1 are as previously defined in (3.8)):

$$\Phi_0 = \frac{1}{N}\Psi_0, \qquad \Phi_1 = \frac{1}{N}\Psi_1 \qquad \text{and} \qquad \Phi = \Phi_0^{-1}\Phi_1.$$

They further checked the adequacy of the working correlation structure and noted that when a working correlation structure is correctly specified then Φ should be close to the identity matrix. The idea is that if the working correlation structure is close to the true structure, the model based estimate $\hat{\Phi}_0$ of the covariance matrix of $\hat{\beta}$ and the robust ('sandwich') estimate $\hat{\Phi}_1$ should be similar, so that, $\Phi = \Phi_0^{-1}\Phi_1$ should be close to an identity matrix. Based on the proposal, Hin, Carey and Wang (2007) redefined the technique:

$$\text{RJC(C)} = \left[\left(\frac{p - tr(\Phi)}{p}\right)^2 + \left(\frac{p - tr(\Phi^2)}{p}\right)^2\right]^{\frac{1}{2}}, \tag{3.15}$$

where tr defines the trace of the matrix and p is the rank of the model.

Pan (2001) proposed a selection criterion based on quasi-likelihood as an improvement to the likelihood reliant Akaike's information criterion (AIC). Following the model specification as in previous sections, $E(Y) = \mu$ and $\text{Var}(Y) = \phi\nu(\mu)$. But for simplicity of notation, let us suppose that ϕ is known and thus ignored in the quasi-likelihood function. Now, suppose we have a true model M_1 and a candidate model M_2 and each model can be indexed by the parameter vector β. The two models can be separated by using the Kullbeck-Leibler distance (Kullbeck and Leibler, 1951), also known as cross entropy. Pan managed the separation and subsequently obtained its approximation by a Taylor series expansion to the second order partial derivative. By circumventing the first order partial derivative which is difficult, Pan expressed the quasi-likelihood under independence such that

$$\mathrm{QIC(C)} = -2\sum_{i=1}^{N}\sum_{j=1}^{n_i} Q\left[\beta;(Y_{ij},X_{ij})\right] + 2tr\left[\Omega_{ind}\mathrm{Var}(\hat{\beta})\right]. \tag{3.16}$$

Hin and Wang (2009) proposed a selection criterion that improves the performance of Pan's (2001) QIC. They ignore the first part of QIC, (3.16) to compare different correlation structures. The first part is the sum of quasi-likelihood functions for the independent observations in the longitudinal data. It does not depend on the specified correlation matrix. The authors proposed using the second part of (3.16) which denotes the penalty term in QIC. It can better reflect the efficiency impairment of parameter estimation. Their selection technique is expressed as

$$\mathrm{CIC(C)} = 2tr\{\Omega_{ind}\mathrm{Var}(\hat{\beta})\}. \tag{3.17}$$

Gosho, Hamada and Yoshimura (2011) proposed a criterion that measures the discrepancy between the covariance matrix estimator and the specified working correlation matrix. It evaluates the appropriateness of the working correlation structure. Gosho (2014) calls this criterion DEW and we will stick to the nomenclature. Gosho's technique selects the criterion that minimizes (3.18).

$$\mathrm{DEW(C)} = tr\left[\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{Var}(y_i)\right)\left(\frac{1}{N}\sum_{i=1}^{N}V_i\right)^{-1} - I\right]^2, \tag{3.18}$$

where I is the identity matrix.

It is the very concept of GEE that the working correlation structure can be misspecified, otherwise we are back to conventional modelling of this structure. Nonetheless, there is indeed some value in having a reasonably well specified structure, for efficiency reasons, but the way it is introduced, and used by many people, is almost as if it needs to be correct, defying the very concept of GEE.

3.4 Simulation study

3.4.1 Data generation

To investigate the comparative performance of WGEE and MI-GEE, we generated repeated correlated counts from a Poisson distribution. Repeated measures within a subject by design are expected to show strong correlation compared to observations between subjects. The correlation structure of the longitudinal data can be modelled by means of random subject effects. The subject-specific effects account for the degree of subject to subject variation that exists in the population. In this case, the generalized linear mixed model is a candidate. Note here that we generate the data from a model that conforms to a random effects formulation, although the final target is a marginal process (GEE). For the specific case of count data (as will be seen), all parameters except the intercept will retain their meaning. This intercept can easily be adjusted to have the intended marginal interpretation.

In the initial simulations, complete longitudinal count data were generated. Then 500 samples of different sizes were randomly drawn. We assumed that the subjects are equally randomized into two treatment (T_i) arms (coded treatment = 1 and placebo = 0). We also assumed that the subjects were followed up for four time points, $\mathrm{time}_j, j = 1, 2, 3, 4$. In essence, we generated longitudinal data following a generalized linear model with a linear predictor of the form (3.1) and whose conditional mean model would consequently be:

$$E[Y_{ij}|b_i] = \lambda_{ij}|b_i = \exp(\beta_0 + \beta_1 \mathrm{T}_i + \beta_2 \mathrm{time}_j + \beta_3 \mathrm{T}_i\, {}^*\mathrm{time}_j + b_i), \tag{3.19}$$

where the outcome $Y_{ij} \sim \mathrm{Pois}(\lambda_{ij}|b_i)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ and $b_i \sim N(0, \sigma^2)$ are i.i.d random effects to account for variability between subjects. The parameter values used in the simulations are $\beta_0 = 2.3, \beta_1 = 0.1, \beta_2 = -0.3, \beta_3 = 0.2$ and b_i are drawn as i.i.d $N(0, 0.5^2)$. The sample sizes studied were $N = 100$ and 250.

Then for each of the simulated complete datasets, dropout was introduced assuming an MAR mechanism. We simulated dropout after the first time point. Threshold values were set such that a subject dropped if they fulfilled the criteria resulting into four missingness patterns: dropout at second time point; dropout at third time point; dropout at fourth time point or no dropout implying a complete observation. This induced approximately $12\%, 28\%$ and 45% dropout rates depicting low, medium and high missingness rates. A monotone missingness pattern was adopted such that if a subject dropped at time j then $Y_{ij'}$ will be missing for all $j' > j \geq 2$. In the simulation study, we assumed a dropout model where subjects whose outcome met some dropout criterion would miss at post baseline time point 2, 3 or 4, that is, for $j > 1$, $dr = y_j - y_{(j-1)},\ j = 1,\ 2,\ 3,\ 4.$

This yielded some negative and positive values. We randomly selected 3 values as our cut-off points such that: if $dr < 1, dr = 0$ or $dr > 1$ then y_j misses this consequently causing the $12\%, 28\%$ and 45% dropout rates respectively. The dropout at measurement occasion j, depended on the observed value at $j - 1$. Ideally, this would mean the dropout followed a MAR mechanism. However, we did not in any way attempt to substantiate the mechanism from the data. Hence we cannot be completely certain that our model was a good fit to the true missingness mechanism.

3.4.2 Analysis procedures

All calculations were carried out in the SAS system for windows, version 9.3. The data were first analyzed as a generalized linear mixed model. However, to recover marginalized or population averaged parameters, one needs to integrate out the random effects. SAS procedures NLMIXED and GLIMMIX can be used to directly fit nonlinear GLMMs. The analysis is valid as long as the missing values are MAR and the mild regularity conditions for ignorability are satisfied. The aforementioned SAS procedures incorporate random effects in the model and so allow for subject-specific and population-averaged (marginal) inference. They are largely equivalent but for the purpose of the current paper, procedure GLIMMIX with adaptive quadrature was adequate. PROC GLIMMIX fits statistical models to correlated data where the response variable is not necessarily Gaussian, i.e., GLMMs. Like linear mixed models, the GLMMs assume Gaussian random effects. In the GLIMMIX procedure one selects the distribution of the response variable conditional on normally distributed random effects (in the absence of which it fits generalized linear models).

For all cases, 20 quadrature points are chosen and the Newton-Raphson technique used for optimization. Here we agree with Molenberghs et al. (2010) who noted that adaptive quadrature and Newton-Raphson iteration produce the most reliable results when contrasted with non-adaptive quadrature and quasi-Newton technique.

Another SAS procedure, GENMOD fits generalized linear models. Additionally, the procedure fits models to correlated responses by the GEE method. It specifies the response distribution in its “DIST” option and a link function in the “LINK” option. It can be used to fit models with most of the correlation structures as discussed in Liang and Zeger (1986) for GEEs. In this study, the GEE results are based on the compound symmetry for the working correlation structure and empirical standard errors.

MI requires coherence between the imputation and analysis models. This means that the imputation model must contain at least all the variables that are intended to be included in the analysis model. This may include all transformations and possible interactions to variables that are needed in the intended tests. Alternatively, a bigger model can be chosen for the imputation than the analysis model. This may be achieved by including auxiliary variables, that we feel may predict the missingness or are related to the missing variable(s). The auxiliary variables are not of interest in the analysis model but are included in the imputation model to increase the estimation power as well as try to make the MAR assumption more plausible. In our study, all variables used in the imputation model were also used in the analysis model. Here, we created SAS code to generate predictions for observations with missing values using coefficients from a Poisson model. Then using these estimates as starting values we generated the imputations. This procedure reduced computing times considerably since the starting values were closer to the final estimates. After this, we used PROC GENMOD for the GEE part. While MI-GEE uses imputations to fill in the missing data, in WGEE we used weights which were generated as the inverse of the probability of dropout (taken from a dropout model). We adopted the WGEE macros as described in Molenberghs and Verbeke (2005) In particular, we employed the macro "DROPOUT" to estimate the probabilities for dropout and the macro "DROPWGT" to pass weights to be used in WGEE. We refer the reader to Molenberghs and Verbeke (2005) for a detailed review of WGEE and its application. But recently a new GEE procedure in SAS/STAT®13.2, PROC GEE (Lin and Rodriguez, 2015) implements the weighted GEE method directly. However, at the time of preparing this manuscript the software was not available to the authors hence the reason why GEE was implemented using Proc GENMOD in SAS.

Notice that here we used a likelihood based and a quasi-likelihood method i.e., the GLMM and the GEE. These approaches are based on two different formulations. In GLMMs the correlation between repeated observations is modelled through the inclusion of random effects, conditionally on which the observations are assumed to be independent. On the other hand, this association is modelled through a working correlation matrix for the GEE method. On this note, the integration of the GLMM was necessary in order to have a marginal model. Again we note that the relationship between the parameter estimates β_R from a random effects model and β_M from a marginal model (GEE) is not straightforward. These two parameter vectors have completely different interpretations. The random effects parameter estimates need to be adjusted so as to have marginal interpretation that can be comparable to their GEE correlatives.

From (3.19) we see that

$$\begin{aligned}\ln(E[Y_{ij}|b_i]) &= \beta_0 + \beta_1 \mathrm{T}_i + \beta_2 \mathrm{time}_j + \beta_3 \mathrm{T}_i\ {}^*\mathrm{time}_j + b_i \\ &= \lambda(X, \beta) + b_i, \qquad (3.20)\end{aligned}$$

where the random vector b_i is independent of X. We find the marginal mean for $Y \sim \text{Poi}(\lambda)$ such that

$$E[Y_{ij}] = \int e^{\lambda(X,\beta)+b_i} dF(b_i). \tag{3.21}$$

If $\int e^{b_i} dF(b_i)$ exists and is finite, the marginal mean equals the conditional mean plus a constant. The constant depends upon the parameters indexing the distribution of the random effects and it is captivated into the intercept of the marginal mean model, provided $\lambda(X, \beta)$ is a linear transformation (Ritz and Spiegelman, 2004). Therefore, for $Y \sim \text{Poi}(\lambda|b_i)$, where $b_i \sim N(0, \sigma_b^2)$ with CDF $F_b(b_i)$, the marginal mean can be defined by (3.22), i.e., integrating out individual heterogeneities:

$$E[Y_{ij}] = \int_0^\infty e^{\lambda+b_i} dF(b_i) = e^{\lambda+\sigma_b^2/2}. \tag{3.22}$$

Suppose $\lambda(X, \beta) = \beta_{0C} + X\beta$, then the marginal mean function is $\beta_{0M} + X\beta$, where $\beta_{0M} = \beta_{0C} + \sigma_b^2/2$.

We asses the performance of the GLMM, WGEE and MI-GEE in terms of bias, efficiency and mean squared error (MSE). We define bias as the absolute difference between true value and the estimate from incomplete data method; i.e., Bias $= \beta - \bar{\hat{\beta}}$, where $\bar{\hat{\beta}}$ is the average of the estimates from $S = 500$ simulation replications of the dataset. The "true" value refers to the coefficient (β) inference from the complete datasets, before the introduction of dropouts. Efficiency is the variability of an estimate around the true population parameter. Here, we compute it as the average width of the 95% confidence interval- which is usually approximately four times the magnitude of the standard error. Finally, the mean squared error is defined as: MSE $= \text{Bias}^2(\bar{\hat{\beta}}) + \text{Var}(\bar{\hat{\beta}})$. Smaller values of these assessment criteria are preferred.

3.4.3 Results

Simulation results of GLMM, WGEE and MI-GEE are presented in Table 3.1. Results are presented for $N = 100$ and 250 for $S = 500$ simulation runs. Under MI-GEE, $m = 20$ imputations are used. Note that the primary focus was to compare WGEE with MI-GEE, but we generated our data from a standard GLMM model setting. To our advantage we therefore extend the results to include those from a conditional and marginal route of inference.

Considering bias, smallest values are observed for MI-GEE while WGEE produced the largest values. This is the same for $N = 100$ and $N = 250$. This behaviour is the same for all parameters except β_0 where largest values are recorded from GLMM. The worst case for the WGEE was for β_1 when $N = 100$ for 12% dropout rate. The value estimated was of opposite sign to that of the perceived true value. In fact, under WGEE only β_0 (for all cases), β_2 $(12\%, N = 100)$ and β_3 $(12\%,\ N = 250)$ fall below the acceptable 10% bias.

Turning to efficiency, GLMM gives the smallest values followed by MI-GEE. As we would expect again the largest cases are for WGEE. This is not strange since the standard errors were notably larger (hence wider 95% confidence intervals). This is explained by the fact that additional sources of uncertainty due to missingness are taken into account coupled with the reality that marginal standard errors would almost always be slightly larger than their conditional counterpart (note that this does not mean larger bias). In this case we somehow expected GLMM to perform better in terms of efficiency because of the way efficiency has been defined. Nonetheless, the differences between the GLMM and MI-GEE values are not very large. At two decimal places the two values would be equal in exactly 12 cases. On one point β_3, 28%, $N = 250$ MI-GEE estimate is less than GLMM.

On the MSE, small values are obtained for MI-GEE. In some cases, WGEE produced smaller values than GLMM such as, for example, β_0 in all cases. This is not surprising because of the bias-variance trade-off. In a number of places, equal values are obtained between GLMM and MI-GEE at 4 decimal places. Specifically, for β_1, β_2, β_3 (12%, $N = 100$) and β_2, β_3 $(12\%, N = 250)$ for GLMM and MI-GEE.

In general, we see from Table 3.1 that MI-GEE is favourable as compared to both GLMM and WGEE. Its performance is consistent regardless of the dropout rate. On the other hand, in all other cases except for when the missingness rate was 12% ($N = 100$), under WGEE the treatment effect (T) and its interaction with time (T*time) are not significant at 5% level (results not shown in Table 3.1). This could mainly be due to notable increases in the standard errors (reflecting the lost information in the partially observed data that WGEE could not fully capture) in comparison with GLMM and MI-GEE.

3.5 Conclusion

The method of generalized estimating equations presents a unique way of modelling correlated data. The approach is attractively applicable in estimation of models where the response variable is continuous, dichotomous, or counts, the latter being the focus of this paper. It has the ability to account for correlations in repeated measures data where conditional independence across observations is unlikely and the possibility of a better understanding of the empirical properties of such dependencies. An important point is

Table 3.1: Bias, Efficiency and Mean Squared Error (MSE) of GLMM, WGEE and MI-GEE incomplete data methods. Missingness rates of 12%, 28% and 45% on the response variable under a MAR mechanism. True parameter values: $\beta_0 = 2.3,\ \beta_2 = 0.1,\ \beta_3 = -0.3,$ and $\beta_4 = 0.2$. Sample sizes: N = 100 and 250 for 500 dataset replications.

		Bias			Efficiency			MSE		
Dropout	Param	GLMM	WGEE	MI-GEE	GLMM	WGEE	MI-GEE	GLMM	WGEE	MI-GEE
		N=100								
12%	β_0	**0.4984**	0.1778	0.1220	0.0268	**1.0880**	0.0340	**0.2484**	0.1056	0.0150
	β_1	0.0033	**1.0111**	0.0018	0.0373	**1.5664**	0.0436	0.0001	**1.1758**	0.0001
	β_2	0.0047	**0.2325**	0.0021	0.0034	**0.4060**	0.0143	0.0000	**0.0644**	0.0000
	β_3	0.0035	**0.3575**	0.0021	0.0042	**0.4444**	0.0177	0.0000	**0.1401**	0.0000
28%	β_0	**0.5216**	0.0845	0.1135	0.0272	**2.0316**	0.0343	**0.2721**	0.2651	0.0130
	β_1	0.0260	**0.1608**	0.0151	0.0378	**2.0408**	0.0441	0.0008	**0.2862**	0.0003
	β_2	0.0401	**0.1140**	0.0158	0.0039	**0.2240**	0.0144	0.0016	**0.0161**	0.0003
	β_3	0.0238	**0.1412**	0.0078	0.0060	**0.3236**	0.0180	0.0006	**0.0265**	0.0001
45%	β_0	**0.6753**	0.1913	0.1132	0.0276	**0.4516**	0.0336	**0.4561**	0.0493	0.0129
	β_1	0.0462	**0.1186**	0.0015	0.0383	**0.6764**	0.0432	0.0022	**0.0427**	0.0001
	β_2	0.1577	**0.1695**	0.0036	0.0083	**0.2572**	0.0143	0.0249	**0.0329**	0.0000
	β_3	0.0348	**0.1417**	0.0008	0.0095	**0.2944**	0.0177	0.0012	**0.0255**	0.0000
		N=250								
12%	β_0	**0.4956**	0.0954	0.1223	0.0168	**1.0252**	0.0216	**0.2456**	0.0748	0.0150
	β_1	0.0012	**0.1858**	0.0011	0.0236	**1.7000**	0.0276	0.0000	**0.2151**	0.0001
	β_2	0.0043	**0.0432**	0.0019	0.0021	**0.3404**	0.0090	0.0000	0.0091	0.0000
	β_3	0.0031	**0.0114**	0.0017	0.0027	**0.5000**	0.0112	0.0000	0.0158	0.0000
28%	β_0	**0.5180**	0.1388	0.1134	0.0172	**1.4580**	0.0216	**0.2683**	0.1521	0.0129
	β_1	0.0224	**0.2633**	0.0149	0.0239	**1.4812**	0.0279	0.0005	**0.2064**	0.0003
	β_2	0.0395	**0.1193**	0.0156	0.0025	**0.2068**	0.0092	0.0016	**0.0169**	0.0002
	β_3	0.0264	**0.1453**	0.0073	0.0387	**0.2196**	0.0114	0.0007	**0.0241**	0.0001
45%	β_0	**0.6694**	0.2142	0.1130	0.0175	**0.3424**	0.0212	**0.4481**	0.0532	0.0128
	β_1	0.0379	**0.1294**	0.0022	0.0242	**0.4464**	0.0274	0.0015	**0.0292**	0.0001
	β_2	0.1562	**0.2310**	0.0035	0.0052	**0.1576**	0.0090	0.0244	**0.0549**	0.0000
	β_3	0.0324	**0.1242**	0.0007	0.0060	**0.2000**	0.0112	0.0011	**0.0179**	0.0000

Note: Param = parameter, GLMM = generalized linear mixed model, WGEE = weighted generalized estimating equations, MI-GEE = multiple imputation based generalized estimating equations. Largest values for bias, efficiency and mean squared error are presented in **bold**.

that, as long as the mean structure is correct, the parameter estimates are consistent as the number of subjects becomes large (Liang and Zeger, 1986). Nonetheless, more efficient estimates are obtained for specification closer to the true structure. Notice that the regression coefficients from a GEE analysis correspond to the average of individual regression lines, and thus these estimates are 'population averaged' (Zeger and Liang, 1986). If population averaged estimates are of interest then GEE analysis may probably provide the most valid results. The main challenge in the analysis is accounting for the missingness mechanism. The purpose of this paper was to compare the performance of WGEE and MI-GEE analyses in the presence of incomplete correlated count data.

Great precaution must be taken as the analysis is not straightforward. First, the analyst must be aware of potential bias in multiple imputation arising from rounding data imputed under normal assumptions when the data are truly not normal (e.g., count variables). Rounding to make imputed values "plausible" can actually cause more bias than using the original seemingly "implausible" imputation value. Although Schafer (2007) argues that slight departures from normality will yield robust inferences, this may not be definitive. Horton, Lipsitz and Parzen (2003) do not recommend the use of the method for discrete data. We argue that rounding should not be used indiscriminately. Horton, Lipsitz and Parzen (2003) suggest that the analyst should impute a discrete variable directly from a discrete distribution. To this effect, for comparison we first imputed the incomplete count data using a Poisson regression model and then performed the GEE analysis on the now complete data sets. On the other hand, an important note is that WGEE works on the concept of weighing contributions by the inverse of the probability of being observed (weights)-thus inverse probability weighting (IPW). These probabilities must be hemmed away from zero so as to avoid hitches of division by zero (Hogan, Roy and Korkontzelou, 2004; Satty, Mwambi and Molenberghs, 2015). The method is ingenious and boasts of good properties, but requires specification of a model for the weights (a dropout model). If the weights are correctly specified, WGEE provides consistent model parameter estimates given that the MAR mechanism is satisfied (Robins, Rotnitzky and Zhao, 1995).

Previous studies have shown the good performance of MI relative to IPW in different data types (e.g., Clayton et al., 1998; Beunckens, Sotto and Molenberghs, 2008). In their works, they used binary data. This paper has contributed to the evidence. Using simulations, we have demonstrated that MI-GEE is actually stronger than WGEE when used for correlated count data. We can argue that when subjects drop out, bias in the marginal model estimates cannot be removed by assigning weights for completers to compensate for dropouts. In this case, the dropouts should be handled more cautiously, such as multiple imputing those dropouts. Nonetheless, we do not claim to have performed a perfectly definitive analysis in this paper because the very large standard errors from the WGEE analysis are a cause for worry. Of course, one should be careful not to extrapolate our findings too much beyond the simulation settings considered.

Chapter 4

Multiple imputation for ordinal longitudinal data with monotone missing data patterns

Abstract

Missing data often complicate the analysis of scientific data. Multiple imputation is a general purpose technique for analysis of datasets with missing values. The approach is applicable to a variety of missing data patterns but often complicated by some restrictions like the type of variables to be imputed and the mechanism underlying the missing data. In this paper, the authors compare the performance of two multiple imputation methods, namely fully conditional specification and multivariate normal imputation in the presence of ordinal outcomes with monotone missing data patterns. Through a simulation study and an empirical example, the authors show that the two methods are indeed comparable meaning any of the two may be used when faced with scenarios, at least, as the ones presented here.

[1]

[1] Kombo, A. Y., Mwambi, H., and Molenberghs, G. (2016). Multiple imputation for ordinal longitudinal data with monotone missing data patterns. *Journal of Applied Statistics*, 1-18.
DOI: 10.1080/02664763.2016.1168370
To link to this article: http://dx.doi.org/10.1080/02664763.2016.1168370

4.1 Introduction

Longitudinal studies are an important source of information in health sciences and other areas but often have the problem of missing data. Ordinal outcomes are increasingly becoming common in these studies. However, analysts are challenged if they need to impute missing values for such outcomes due to their hierarchical nature (Carpenter and Kenward, 2013; Chen et al., 2005). Missing values in longitudinal studies occur when not all of the planned measurements of a subject outcome vector are actually observed. This turns the statistical analysis into a missing data problem. For example, a subject may terminate early from a scheduled sequence of clinical visits for a number of reasons, both known and unknown. This type of missing pattern is termed dropout (monotone missing data pattern). Alternatively, a subject may miss a scheduled visit but appear at the next occasion. This is referred to as an arbitrarily (intermittent) missing data pattern. In this study, we focus on the former pattern of missingness. The reasons that lead to missingness are varied and it is always necessary to reflect on the nature of missingness and its impact on inferences. In Rubin (1976), these reasons are classified into three categories. Data are said to be missing completely at random (MCAR) if the probability of missingness is independent of both the observed and unobserved measurements, missing at random (MAR) if, conditional on the observed data the probability of missingness is independent of the unobserved measurements and missing not at random (MNAR) for a violation of the above scenarios. Under the unrealistic MCAR, simple incomplete data methods such as last observation carried forward (LOCF), complete case analysis and available case analysis may be employed. However, even under the strong MCAR assumption it is not guaranteed that LOCF analysis is valid. In fact, LOCF is not recommended, not even when missingness is MCAR and there is a (potential) treatment effect. Indeed, analysts see it unscientific to use the ad hoc methods when broadly valid likelihood analyses can be easily implemented with standard software (Beunckens, Molenberghs and Kenward, 2005). Generally speaking, the MAR assumption represents the most general condition under which valid inferences can be obtained without reference to the missing data mechanism, given inferences are likelihood based or Bayesian (Beunckens, Molenberghs and Kenward, 2005; Kenward and Carpenter, 2007).

Recent advances in computational statistics have produced a new billow of flexible and formally justifiable procedures with sound statistical basis like multiple imputation (MI). MI, initially proposed by Rubin (1977) and further detailed in Rubin (1987) and Schafer (1997), has become one of the most popular approaches in handling missing data. MI can be used not only with continuous variables but also with binary and categorical variables. It provides a way of accounting for uncertainty associated with imputations. This is a major strength against a number of existing single imputation methods. MI replaces each of the missing values with $m \geq 2$ plausible values generated under an appropriate imputation model to obtain m complete datasets. This replication captures the uncertainty about the missing data. The resulting m multiply imputed datasets are then analysed separately using an appropriate well-known standard method for complete data. The third stage is to combine the m analysis results into one for inferences, where the standard errors of estimates take account of the variation within and between the m imputations (Rubin, 1987).

MI is a viable candidate for handling missing data in multivariate analysis. This is because it introduces appropriate random error into the imputation process and makes it possible to produce unbiased estimates of all parameters (Allison, 2000; Rubin, 1987). It can be used with any kind of data and any kind of analysis without specialized software (Allison, 2000). However, one key feature of MI is that, for correct and valid inferences, the imputation model should be correctly specified. It is agreed that the analysis and the imputation model should be congenial in the sense that the imputation model should be able to reproduce the major features of the analysis model (Rubin, 1987; Meng, 1994; Allison, 2001). In this paper, the imputation model includes the same variables that are in the analysis model. Regarding MI, it is also important to note that standard MI procedures assume that the data are MAR. While it is almost always impossible to test this assumption, including auxiliary variables in the imputation model that predict the missingness, together with variables that are correlated that will be included in the analysis model, can minimise bias. It also makes the MAR assumption more viable (Collins, Schafer and Kam, 2001; Schafer, 2003). On the other hand, it is also possible to use MI procedures to impute data that are MNAR, but this requires making additional assumptions about the missingness mechanism.

This paper is concerned primarily with the comparison of two MI methods namely fully conditional specification (FCS) and multivariate normal imputation (MVNI) as applied to ordinal outcome variables with a monotone missing data pattern. Moreover, for the purpose of this paper, we focus on one ordinal outcome variable over time but the ideas presented here are applicable to other ordinal forms and data settings.

The paper is organised as follows. In Section 4.2, we give the key definitions and necessary notation. A description of the imputation methods is given in Section 4.3 followed by a simulation study and application in Section 4.4.

4.2 Definitions and notation

4.2.1 Missing data model

Suppose that for the ith subject in the study, a sequence of measurements Y_{ij} is expected to be measured at occasions $j = 1, \ldots, n_i$. Due to some reasons, some values of $Y_i =$

$(Y_{i1}, \ldots, Y_{in_i})'$ are not observed. Then Y_i, can be partitioned into two subvectors such that $Y_{i,o}$ contains the observed measurements and $Y_{i,m}$ the unobserved measurements. Now, if we let Y to be the complete set of observations, then Y can be partitioned such that $Y = (Y_o, Y_m)$. We define a random vector $R_i = (R_{i1}, R_{i2}, \ldots, R_{in_i})$ compatible with the vector of observations Y_i such that $R_{ij} = 1$ if the outcome Y_{ij} is observed and 0 otherwise. Using Heckman (1977), the joint distribution of the full data Y and the indicator vector variable R can be factorized as

$$f(Y, R|\theta, \psi) = f(Y|\theta)P(R|Y, \psi), \tag{4.1}$$

where ψ denotes a vector of parameters governing the missingness mechanism and θ denotes the measurement process model parameters. The conditional distribution of the missing data mechanism can be equivalently expressed as $f(R|Y_o, Y_m, \psi)$. Diggle and Kenward (1994) propose modelling the probability of missingness at a particular measurement occasion as a linear function of the response values at previous occasions. For simplicity, we assume that this dropout depends only on the observed response just before the time it fails to be recorded and the unobserved response at the missing point. However, this model can be extended to include measured or observed covariates. If we denote by Y_{ij}, the response at measurement occasion j, the missing data model can be written as

$$\text{logit}[P_j(R_{ij} = 0|y_{i1}, y_{i2}, \ldots, y_{i(j-1)}, y_{ij})] = \psi_0 + \psi_1 y_{i(j-1)} + \psi_2 y_{ij}, \tag{4.2}$$

where $P_j(R_{ij} = 0|y_{i1}, y_{i2}, \ldots, y_{i(j-1)}, y_{ij})$ is the conditional probability of missingness at occasion j, given the history of responses, $y_{i1}, y_{i2}, \ldots, y_{i(j-1)}, y_{ij}$, the response subject to missingness, y_{ij} and ψ_0, ψ_1 and ψ_2 are the model parameters to be estimated. The model reduces to a MAR model if $\psi_2 = 0$. MCAR if $\psi_1 = \psi_2 = 0$. If $\psi_2 \neq 0$, then we cannot rule out MNAR but note that the test for $\psi_2 = 0$ versus $\psi_2 \neq 0$ (MAR versus MNAR) relies on untestable assumptions such as the distributional form (Kenward, 1998; Molenberghs and Kenward, 2007; Newson, Jones and Hofer, 2012; Rhoads, 2012). In fact, Molenberghs et al. (2008) show that a formal distinction between MAR and MNAR is not possible because for any MNAR model there exists a MAR counterpart that fits the data equally well.

4.2.2 Ordinal responses

There are cases where the outcome variable can be polytomous. While the typical logistic regression analysis models a binary response, logistic regression can also be applied to multilevel cases. If the response variable takes on values that have no inherent order (e.g. voting party A, B, C, or D), then the response is nominal. If it takes on intrinsic values like the levels of agreement (e.g. strongly agree to strongly disagree), then the response is ordinal. Then, for ordered categorical variables, the binary logistic regression extends to polytomous logistic regression. A number of logistic regression models have been studied for ordinal response variables (Agresti, 1989; Armstrong and Sloan, 1989; Cox, 1995; Liu and Agresti, 2005; McCullagh, 1980). When there is need to consider several factors, special multivariate analysis for ordinal data is the natural alternative (Das and Rahman, 2011), although other methods, like mixed models may be used. However, ordinal logistic regression models have been most useful (McCullagh, 1980; Ananth and Kleinbaum, 1997). Several ordinal logistic regression models exist, namely the proportional odds model, partial proportional odds model (PPOM), continuous ratio model and the stereotype regression model. The most common among the ordinal logistic regression models is the proportional odds model (Bender and Grouven, 1998). The proportional odds model (a specific form of cumulative odds model), is a logit model that allows ordered data to be modelled by analysing it as a number of dichotomies. A binary logistic regression model compares one dichotomy (yes/no) whereas the proportional odds model compares a number of dichotomies by arranging the ordered categories into a series of binary comparisons. Here, the assumption is made that the effect of each explanatory variable is the same for each binary comparison (logit). This is the proportional odds assumption, also referred to as the parallel lines assumption (or equal slopes assumption). It leads to parsimony of the model, because it means that the effect of a predictor variable on the ordinal response is explained by one parameter. However, it may pose a restriction on the flexibility of the model, which may or may not be adequate for the data. Then before any model statistics are interpreted, it is important to test the assumption, a violation of which may lead to incorrect interpretation of results (Ananth and Kleinbaum, 1997). The assumption is commonly used with the cumulative logit link. On the other hand, mixed effects models have also been found very useful for longitudinal categorical (nominal or ordinal response) data. The main reason why random effects are used is to take account of correlated data due to clustering as a result of repeated measures from the same individual.

In medical and clinical research, it is not easy to get a continuous outcome for that kind of information you need. More often, the variable of interest has a natural ordering, say no disease, mild and severe. In this case using an ordinal outcome for the disease model may make sense other than ‘no disease’ and ‘diseased’, that is, collapsing the ordinal levels to binary ones. If this is done, one has to find an appropriate correlation structure of the dichotomized data, and then inflate the correlations intentionally in order to make them what they should have been. This means that one follows the ordinal - binary - Gaussian - ordinal - binary conversion scheme. This scheme is applicable when presented with correlated ordinal outcomes data. The ordinal levels are collapsed to binary ones, and then converting the correlated binary outcomes to multivariate normal outcomes in such a way that re-conversion to binary and then back to the original ordinal scale after performing multiple imputation, yields the desired original marginal distributions and correlations. The conversion strategy ensures that the correlations are transformed reasonably which enables the user to take advantage of well-established imputation strategies for Gaussian outcomes. A key methodological focus then turns out to be conducting multiple imputation under multivariate normality assumption with re-conversion to the original ordinal scale while preserving key distributional assumptions. This strategy, however, may not be applicable in every scenario (Demirtas and Hedeker, 2008). The polytomous logistic regression model may be employed for the ordered categorical variable, but fails to make proper use of the information about the ordering. One way of taking advantage of the ordering is the use of 'cumulative odds', 'cumulative probabilities' and 'cumulative logits'.

Now, suppose that our data comprise of a set of $i = 1, \ldots, N$ independent clusters (subjects in our longitudinal data context) where the ith subject consists of n_i observations. As before, let Y_{ij} denote the jth ($j = 1 \ldots, n_i$) response in subject i. This response may fall in any of $c = 1, \ldots, C$ distinct ordered categories for $C \geq 2$. Further, let x_{ij} denote a vector of predictor variables for the jth observation in the ith subject. Then Y_{ij} will have a multinomial distribution with parameter vector $\boldsymbol{\pi}$. In this case, π_{jc} is the probability of the jth measurement falling into category c so that we have our cumulative probabilities given as

$$P(Y_{ij} \leq c) = P(Y_{ij} \leq c|x_{ij}) = \pi_{i1} + \cdots + \pi_{ic}. \tag{4.3}$$

Now using a logit link, we will have a cumulative logit model defined as

$$\text{logit}(P(Y_{ij} \leq c)) = \log\left[\frac{P(Y_{ij} \leq c)}{1 - P(Y_{ij} \leq c)}\right] = \alpha_c - x'_{ij}\boldsymbol{\beta}, \tag{4.4}$$

where $P(Y_{ij} \leq c)$ is the probability of being at or below category c, given a set of predictors. Here, $c = 1, \ldots, C - 1$ for the C categories of the ordinal outcome, α_c gives the threshold parameters (intercept terms that depend on the categories). These parameters, however, are seldom of practical importance except for computing response probabilities. The regression parameters, $\boldsymbol{\beta}$, reflect the association between the predictor variables and the outcome variable. Notice that, while the regression coefficients do not vary (i.e. β has the same effect for each of the $C-1$ cumulative logits, implying that $x'_{ij}\boldsymbol{\beta}$ is independent of c), a different intercept exists for each level of the cumulative model. Given that the regression parameters ($\boldsymbol{\beta}$) are subtracted (model (4.4)), this means that a unit increase in the predictor variable will increase the log-odds of being in category greater than c. In other words, it means that the higher the value of $X'_{ij}\boldsymbol{\beta}$, the higher the probability of response falling in a category at the upper end of the response scale. But note that β itself can be estimated as negative which will give an increasing effect of the odds in categories less than or equal to c. The model describes the cumulative logits across $c-1$ response categories. One can transform the cumulative logits to obtain estimated cumulative odds and also the cumulative probabilities of being at or below category c.

4.3 Imputation methods

When the dataset has a monotone missingness pattern, variables with missing values are imputed sequentially with covariates obtained from their corresponding sets of preceding variables. To impute continuous variables, a regression method, a predictive mean matching method or a propensity score method may be used. A logistic regression method may be used for a binary or ordinal variable. Alternatively, a discriminant function for nominal or binary variables can be used. For real and simulated incomplete ordinal datasets, we contrast two multiple imputation procedures the fully conditional specification (FCS) via chained equations (Van Buuren, 2007; Van Buuren, Boshuizen and Knook, 1999). and the multivariate normal imputation (MVNI; Schafer, 1997). These approaches are based on different theoretical assumptions and involve very different computational methods (Lee and Carlin, 2010).

4.3.1 Multivariate normal imputation

Approaches to imputing multivariate data have been developed. For example, Rubin and Schafer (1990) provided procedures for generating multivariate multiple imputation. This Bayesian simulation algorithm draws imputations from the posterior predictive distribution of the unobserved data given the observed data. The method assumes that the data are multivariate normally distributed and missing at random. Schafer (1997) used this underlying approach and derived imputation algorithms for multivariate numerical, categorical and mixed data. The methodology describes the data by encompassing a multivariate model and derive a posterior distribution and then draw imputations from these by Gibbs sampling (here after referred to as data augmentation rather than Gibbs sampling). It uses the Markov chain Monte Carlo (MCMC) approach to draw imputed values from the estimated multivariate normal distribution.
Given our ordinal response variable $Y \sim \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, data augmentation (Tanner and Wong, 1987) in Bayesian inference with missing data is based on iterating between an imputation step (I-step) and a Posterior step (P-step).

- *The imputation step*– With some estimated initial values for the mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$, the I-step simulates a value 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:

$$Y_m^{(r+1)} \sim P(Y_m|Y_o, \theta^r), \qquad \boldsymbol{\theta} = (\boldsymbol{\mu}, \mathbf{\Sigma}). \tag{4.5}$$

• *The posterior step–* This step draws a value of the parameter $\boldsymbol{\theta}$ from a complete-data posterior distribution:

$$\theta^{(r+1)} \sim P(\theta|Y_o, Y_m^{(r+1)}). \tag{4.6}$$

The updated estimates are then used in the imputation step.
Iterating equations (4.5) and (4.6) from initial value $\theta^{(0)}$ will yield a stochastic sequence $\{(\theta^{(r)}, Y_m^{(r)}); \quad r = 1, 2, \ldots\}$. The two steps are iterated sufficiently long until the distribution of the estimates becomes stationary (Schafer, 1997). Each step depends on the previous one, meaning that there is dependency across the steps. This approach is theoretically sound but based on distributional assumptions that may not always be realistic (e.g., assuming normality for binary, ordinal variables). For categorical variables, the MVNI method draws imputations under the MVN model and so we need to round off the imputations to the nearest integer to accommodate the categorical nature of the data. Allison (2005), however, cautions about rounding (he cites the binary case) because the rounded imputed values may lead to biased parameter estimates. Nonetheless, Schafer (1997) had already argued that inference from MVNI may be reasonable even if multivariate normality does not hold, for example, in the cases of binary and categorical variables. We refer the reader to Schafer (1997) for a detailed account of this procedure.

4.3.2 Fully conditional specification

An alternative option, applicable to multivariate data, is the fully conditional specification (FCS) approach. FCS is a flexible method that specifies the multivariate model by a series of conditional models for each of the incomplete variables. Unlike MVNI, it does not necessarily rely on the multivariate normality assumption and thus univariate regression models can be appropriately tailored to be used for ordered logistic regression for ordinal variables. Using a Bayesian approach, imputations are done stepwise starting with the variable with the least amount of missing values and progressing like that until the variable with the most missing data is finally handled. It involves two phases in each imputation: the filled-in stage and the imputation stage. During every stage, draws are randomly done from both the posterior distribution of the parameters and posterior

distribution of the missing values. At the filled-in stage, the missing values are filled in sequentially over the variables, one after the other with preceding variables serving as covariates. The filled-in values are then used as starting values for the imputation stage. At the imputation stage, the filled-in values are replaced with imputed values for each variable sequentially at each iteration.
Let the ordinal response variable Y be characterized by a vector of unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$; $\boldsymbol{\mu}$ is a mean vector while $\boldsymbol{\Sigma}$ is a covariance matrix. As before, $Y = (Y_o, Y_m)$. Following Van Buuren et al. (2006) and also in Van Buuren and Groothuis-Oudshoorn (2011), multiple imputation via FCS proceeds as follows:

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

$$y^* \sim P(y_m|y_o, \theta = \theta^*). \tag{4.7}$$

Repeat the second and third steps depending on the number of imputations. The steps are repeated long enough for the results to reliably simulate an approximately independent draw of the missing values for an imputed dataset.

4.3.3 Software considerations

When we assume MAR, valid inferences can be obtained through likelihood-based analysis without modelling the dropout process. Consequently, the generalized linear mixed model - as the analysis model - is used. This approach may be implemented by using SAS procedures NLMIXED and GLIMMIX. If we need to impute missing values, both the description of missing data patterns and multiple imputation is performed using the procedure PROC MI. It may be used for all types of variables. The procedure offers several methods for imputation depending on whether the variable is continuous or categorical. Here we are interested in comparing MVNI and FCS as implemented in PROC MI. For MVNI, it uses the Markov Chain Monte Carlo (MCMC) approach to draw imputed values from the estimated multivariate normal distribution. To use it, the user calls it by specifying the mcmc statement in the MI procedure. To run FCS, the fcs statement is specified in PROC MI. In PROC MI, the imputation model to be used and the number of imputed datasets to be created are specified. After imputation, statistical procedures run the analytic model of interest separately for each imputation using _Imputation_ as a BY variable, and the results are stored in an output file. Finally, a procedure call, PROC MIANALYZE combines the estimates obtained from the analyses for multiply imputed data to produce valid statistical inferences. However, for some complete data analyses, like those for categorical data, additional manipulations are needed before PROC MIANALYZE is used (Ratitch, Lipkovich and O'kelly, 2013). This is because Rubin (1987) rules for combining results assume that the statistics estimated are normally distributed. Such estimates, like regression coefficients and means, are approximately normally distributed, while others like the odds ratios, correlation coefficients and relative risks are nonnormal. If interest is on the latter group of estimates, they can first be normalized before applying Rubin's combination rules to the transformed estimates. In Van Buuren (2012) some transformations to various types of estimated statistics are suggested.

By default, the SAS procedure LOGISTIC fits the proportional odds model combined with the cumulative logit link. When the assumption of the common slopes is valid for some variables but not for others, PROC GENMOD may be used to fit the PPOM. Alternatively, PROC LOGISTIC may also be used but with a specification of the UNEQUALSLOPES option in the model. PROC CATMOD can be used in case of a nonproportional odds model.

4.4 Simulation study

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

We conducted a simulation study to examine the performance of FCS and MVNI. The datasets were generated using a scenario that mimics common longitudinal studies. The simulated datasets are based on an ordinal outcome with C categories which are generated at four study occasions, $j = 1, \ldots, 4$. The setting was repeated for three different settings where $C = 3, 4, 5$. For each of the different scenarios, we simulated 1000 datasets based on a generalised linear mixed model scheme of the form (4.8) for sample sizes $N = 100, 250, 500$. Consequently, longitudinal ordinal variables were generated following a model with a linear predictor:

$$\text{logit}[P(Y_{ij}^* \leq c)] = \alpha_c + \boldsymbol{x'}\boldsymbol{\beta} + b_i, \qquad b_i \sim N(0, d). \tag{4.8}$$

An ordinal regression model was motivated by assuming an underlying latent variable (y^*) which is related to the actual response through the 'threshold concept'. The response is defined based on some underlying unobserved continuous endpoint that follows a linear regression model incorporating random effects and a prespecified set of cut-off values (threshold values) α_c. The data were generated by assuming a vector of predictor variables $\boldsymbol{x'} = (x_1, x_2, x_3, x_4)$, which is a combination of both continuous and binary variables. Here, x_1 and x_3 are binary group effects (i.e. $x = 0, 1$) representing a treatment group indicator and gender respectively, x_2 is a continuous variable representing exposure period and x_4 is a four-point assessment time. For the simulations we used the parameters, $\beta_1 = 0.9, \beta_2 = 0.2, \beta_3 = 0.5$ and $\beta_4 = 0.8$. For simplicity of the simulations in this paper, we did not assume any interaction of terms. In this case, our simulation model is explicitly written as

$$\text{logit}[P(Y_{ij}^* \leq c)] = \alpha_c + 0.9x_1 + 0.2x_2 + 0.5x_3 + 0.8x_4 + b_i. \qquad b_i \sim N(0, 1.8^2) \tag{4.9}$$

By inverting the logit link function, it leads to the conditional ordinal logistic regression model, noting that equation (4.8) can be equivalently written as

$$P(Y_{ij}^* \leq c) = \frac{\exp(\alpha_c + \boldsymbol{x'\beta} + b_i)}{1 + \exp(\alpha_c + \boldsymbol{x'\beta} + b_i)}. \tag{4.10}$$

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

$$Y = \begin{cases} 1 & \text{if} \quad \phi_{ij} \leq \tau_1, \\ 2 & \text{if} \quad \tau_1 < \phi_{ij} \leq \tau_2, \\ 3 & \text{if} \quad \tau_2 < \phi_{ij} \leq \tau_3, \\ 4 & \text{if} \quad \phi_{ij} > \tau_3. \end{cases} \tag{4.11}$$

First from the full datasets without imposing any missing values, parameters and standard errors were estimated by a likelihood based approach. Each estimate is an average of 1000 estimates from the different simulated datasets. Then, we assumed a rather simple MAR model of missingness, where subjects whose outcome was greater than some cut-off probability would miss at post baseline time points 3 and 4, that is, let $drp = y_{ij} - y_{ij-1}, j = 2, 3, 4$, yielding values between -2 and 2; -3 and 3; and -4 and 4 for the different choices of the categories of the ordinal outcome, that is, for $C = 3, 4$ and 5 categories respectively. Then we normalized these values by defining $ndrp = (drp + (C - 1))/2C$ in order to confine them to the range $[0, 1]$. Finally, if $ndrp > \gamma + 0.6u$ (where $u \sim [0, 1]$ is a uniformly distributed random number) then $y_{i(j+1)}$ misses. We held (for the C = 3, 4, 5 categories, respectively) $\gamma = 0.4$ so as to ensure that about 30% of the response data were missing by the end of the study. The probability of a value dropping depended merely on the immediate history.

Then, the missing entries were imputed using FCS and MVNI as carried out in PROC MI. We used the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) to obtain the starting values for our imputations. MVNI was performed using the SAS PROC MI with a specification of the MCMC statement. The ordinal values were imputed on the continuous scale and rounded off to the required categories. Maximum and minimum values were specified based on the scale of the response options of the dataset. These specifications were necessary so as to ensure that imputations were not created outside the range of the response values. FCS was carried out using fcs statement in PROC MI. The ordinal response was imputed using the ordinal logistic regression model as incorporated in the FCS procedure. For all cases in the study, default values for MCMC and FCS specifications were used in the simulations. We realized that the algorithms still converged to the correct posterior distributions and were confident that the imputed values in the different datasets were statistically independent. All the other predictor variables were used to ensure that our imputation model was rich enough to try and satisfy the congeniality requirement under the MAR assumption. For simplicity, throughout the analyses in this paper the categorical time was treated as continuous.

For comparison of methods, a larger number of imputations are necessary (Wood et al., 2005). We performed $m = 20$ imputations. This relatively high 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. Nonetheless, researchers argue that m can be set to $3 \leq m \leq 5$ and still get sufficient accuracy. However, Schafer (1997) cautions that pegging on this range might be risky. On the other hand, Molenberghs and Verbeke (2005) 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. However, a rule of thumb for choosing m is suggested (see White, Royston and Wood, 2011). They suggest that m should be at least equal to the percentage of incomplete cases. Nevertheless, we caution the reader that still discretion is necessary, based on the problem at hand.

To compare the performance, we used bias and mean squared error (MSE) of the parameter estimates. We defined bias as the absolute difference between the average parameter estimate from the analysis procedures (based on the 1000 data replications) and the true value (i.e., Bias $= |\bar{\hat{\beta}} - \beta|$).

4.4.2 Simulation results

Results of the simulation study (based on 1000 simulated datasets and 20 imputations) are presented. We present three tables, where Table 4.1 represents results when the ordinal outcome variable has three categories/levels, Table 4.2, the variable has four levels and Table 4.3 when the variable has five levels. The results are presented for MVNI, FCS, direct likelihood (DL) and full data analysis (FDA). In this paper, full

data refer to the simulated dataset that has no missing values. Although the original idea of the paper was to contrast the performance MVNI and FCS, DL is presented as an additional approach because of its known ability to handle incomplete data. Rather than imputing missing measurements, Mallinckrodt et al. (2003), suggested the use of a direct likelihood approach to deal with incomplete correlated data under the ignorable assumption. Here, the observed cases are analysed without any analyst's adjustments, that is, without imputation nor deletion, by the use of models that provide a framework where clustered data can be analysed by including both fixed and random effects in the model (in case of GLMMs for non-Gaussian data) (Kadengye et al., 2012). The authors in Kadengye et al. (2012) further showed that DL analysis of incomplete datasets produced unbiased parameter estimates that were comparable to those from a full data analysis. These arguments were echoed by Molenberghs and Verbeke (2005), who also pointed out cases where MI is justified.
For clarity, results are presented here for regression coefficients only and not the intercepts. In all tables, larger values depicting worst cases are in bold.
In Table 4.1, considering bias, we notice that the largest values are obtained for MVNI. These are followed by FDA in all cases except β_4 where FCS produces larger values than FDA. The trend is the same for all sample sizes. The FCS and DL values are very close to each other with one case ($\beta_3, N = 500$) where they are the same. Looking at MSE, we observe a similar situation as for bias, that is, bigger values for MVNI followed by FDA except β_4 where FCS produces larger values than FDA. Comparing DL and FCS, we see equal values for all cases save for $\beta_1, \beta_3, \beta_4$ for $N = 100$, and β_4 in $N = 250$. However, these values are very close such that in $3-$decimal places, they are equal. Looking at standard errors, largest values are observed for DL consistently except β_4. MVNI produces the smallest values in all the other cases except β_4, for $N = 100, 250$.

Now shifting focus to Table 4.2, the scenario we observed in Table 4.1 changes. We notice that largest bias are recorded for FDA for all β's except β_4 where MVNI gives the largest bias. Exactly, the same trend is produced under MSE. Looking at standard errors, here the same scenario as in Table 4.1 is reproduced. Again, DL and FCS produce the same or very close values.

In Table 4.3, the previous trends observed for standard errors are replicated here. For bias and MSE the trends change slightly. Now, the largest biases are recorded for MVNI in all cases except β_2 for all sample sizes, and β_3 for $N = 250$. The same set-up is produced under MSE. Like before very close or equal values are observed for FCS and DL.

In terms of bias MVNI seems to be more biased than FCS. If one is interested in smaller standard errors then MVNI has mostly smaller values than FCS or at times they are

Table 4.1: Standard errors (Std Err), Bias and mean squared error (MSE) estimates from fully conditional specification (FCS) and multivariate normal imputation methods (MVNI)

		Std Err				**Bias**				**MSE**			
Sample	**Par**	FDA	DL	FCS	MVNI	FDA	DL	FCS	MVNI	FDA	DL	FCS	MVNI
N=100	β_1	0.0114	**0.0130**	0.0111	0.0106	0.0614	0.0154	0.0144	**0.0965**	0.0039	0.0004	0.0003	**0.0093**
	β_2	0.0007	**0.0008**	0.0007	0.0007	0.0108	0.0042	0.0043	**0.0212**	0.0001	0.0000	0.0000	**0.0004**
	β_3	0.0128	**0.0158**	0.0125	0.0120	0.0269	0.0100	0.0074	**0.0556**	0.0009	0.0003	0.0002	**0.0032**
	β_4	0.0030	0.0035	**0.0040**	**0.0040**	0.0559	0.1820	0.1826	**0.2354**	0.0031	0.0331	0.0334	**0.0554**
N=250	β_1	0.0076	**0.0091**	0.0077	0.0075	0.0540	0.0191	0.0193	**0.1004**	0.0030	0.0004	0.0004	**0.0101**
	β_2	0.0004	**0.0006**	0.0005	0.0004	0.0113	0.0048	0.0049	**0.0217**	0.0001	0.0000	0.0000	**0.0005**
	β_3	0.0082	**0.0094**	0.0077	0.0074	0.0356	0.0157	0.0167	**0.0620**	0.0013	0.0003	0.0003	**0.0039**
	β_4	0.0018	0.0022	0.0024	**0.0028**	0.0547	0.1821	0.1824	**0.2350**	0.0030	0.0332	0.0333	**0.0552**
N=500	β_1	0.0056	**0.0067**	0.0053	0.0059	0.0611	0.0274	0.0272	**0.1086**	0.0038	0.0008	0.0008	**0.0118**
	β_2	0.0003	**0.0004**	0.0003	0.0003	0.0111	0.0048	0.0049	**0.0218**	0.0001	0.0000	0.0000	**0.0005**
	β_3	0.0052	**0.0065**	0.0053	0.0052	0.0440	0.0246	0.0246	**0.0693**	0.0020	0.0006	0.0006	**0.0048**
	β_4	0.0013	0.0016	**0.0018**	0.0017	0.0554	0.1843	0.1844	**0.2369**	0.0031	0.0340	0.0340	**0.0561**

Notes: Also estimates from full data analysis (FDA) and direct likelihood (DL) method. Missing values, approximately (30%) on the response variable; MAR mechanism. A case where ordinal variable has $C = 3$ levels.

Table 4.2: Standard errors (Std Err), Bias and mean squared error (MSE) estimates from fully conditional specification (FCS) and multivariate normal imputation methods (MVNI).

		Std Err				**Bias**				**MSE**			
Sample	**Par**	FDA	DL	FCS	MVNI	FDA	DL	FCS	MVNI	FDA	DL	FCS	MVNI
N=100	β_1	0.0100	**0.0118**	0.0100	0.0099	**0.2382**	0.1598	0.1595	0.1933	**0.0568**	0.0257	0.0255	0.0375
	β_2	0.0006	**0.0007**	0.0006	0.0006	**0.0517**	0.0314	0.0315	0.0358	**0.0027**	0.0010	0.0010	0.0013
	β_3	0.0102	**0.0123**	0.0106	0.0097	**0.1214**	0.0879	0.0894	0.1082	**0.0148**	0.0079	0.0081	0.0118
	β_4	0.0023	0.0033	**0.0036**	**0.0036**	0.2141	0.3929	0.3930	**0.4007**	0.0458	0.1544	0.1545	**0.1606**
N=250	β_1	0.0063	**0.0079**	0.0065	0.0064	**0.2329**	0.1556	0.1563	0.1914	**0.0543**	0.0243	0.0245	0.0367
	β_2	0.0003	**0.0004**	**0.0004**	0.0003	**0.0520**	0.0316	0.0316	0.0359	**0.0027**	0.0010	0.0010	0.0013
	β_3	0.0066	**0.0082**	0.0066	0.0066	**0.1290**	0.0868	0.0871	0.1082	**0.0167**	0.0076	0.0076	0.0118
	β_4	0.0015	0.0020	0.0021	**0.0022**	0.2123	0.3884	0.3885	**0.3972**	0.0451	0.1509	0.1509	**0.1578**
N=500	β_1	0.0044	**0.0056**	0.0047	0.0045	**0.2384**	0.1611	0.1614	0.1958	**0.0569**	0.0260	0.0261	0.0384
	β_2	0.0002	**0.0003**	**0.0003**	0.0002	**0.0521**	0.0316	0.0317	0.0360	**0.0027**	0.0010	0.0010	0.0012
	β_3	0.0042	**0.0054**	0.0047	0.0045	**0.1385**	0.0952	0.0951	0.1164	**0.0192**	0.0091	0.0091	0.0136
	β_4	0.0010	0.0015	**0.0017**	0.0014	0.2131	0.3910	0.3913	**0.3989**	0.0454	0.1529	0.1531	**0.1591**

Notes: Also estimates from full data analysis (FDA) and direct likelihood (DL) method. Missing values, approximately (30%) on the response variable; MAR mechanism. A case where ordinal variable has $C = 4$ levels.

Table 4.3: Standard errors (Std Err), Bias and mean squared error (MSE) estimates from fully conditional specification (FCS) and multivariate normal imputation methods (MVNI).

		Std Err				**Bias**				**MSE**			
Sample	**Par**	FDA	DL	FCS	MVNI	FDA	DL	FCS	MVNI	FDA	DL	FCS	MVNI
N=100	β_1	0.0103	**0.0114**	0.0093	0.0096	0.2500	0.2295	0.2294	**0.2565**	0.0626	0.0528	0.0527	**0.0659**
	β_2	0.0005	**0.0006**	0.0005	0.0005	**0.0546**	0.0499	0.0499	0.0514	**0.0030**	0.0025	0.0025	0.0026
	β_3	0.0106	**0.0115**	0.0097	0.0094	0.1354	0.1253	0.1247	**0.1415**	0.0184	0.0158	0.0156	**0.0201**
	β_4	0.0024	0.0030	**0.0034**	0.0033	0.2228	0.2908	0.2904	**0.3221**	0.0496	0.0846	0.0843	**0.1038**
N=250	β_1	0.0061	**0.0069**	0.0059	0.0056	0.2433	0.2279	0.2304	**0.2562**	0.0592	0.0520	0.0531	**0.0657**
	β_2	0.0003	0.0003	0.0003	0.0003	**0.0549**	0.0502	0.0503	0.0518	**0.0030**	0.0025	0.0025	0.0027
	β_3	0.0068	**0.0076**	0.0060	0.0064	**0.1409**	0.1255	0.1255	0.1408	**0.0199**	0.0158	0.0158	**0.0199**
	β_4	0.0013	0.0018	0.0018	**0.0019**	0.2230	0.2903	0.2904	**0.3228**	0.0497	0.0843	0.0843	**0.1042**
N=500	β_1	0.0044	**0.0052**	0.0043	0.0041	0.2497	0.2320	0.2317	**0.2597**	0.0624	0.0539	0.0537	**0.0675**
	β_2	0.0002	**0.0003**	0.0002	0.0002	**0.0549**	0.0503	0.0503	0.0519	**0.0030**	0.0025	0.0025	0.0027
	β_3	0.0041	**0.0049**	0.0039	0.0041	0.1465	0.1328	0.1323	**0.1482**	0.0215	0.0177	0.0175	**0.0220**
	β_4	0.0010	0.0013	**0.0014**	**0.0014**	0.2244	0.2918	0.2922	**0.3241**	0.0504	0.0851	0.0854	**0.1050**

Notes: Also estimates from full data analysis (FDA) and direct likelihood (DL) method. Missing values, approximately (30%) on the response variable; MAR mechanism. A case where ordinal variable has $C = 5$ levels.

equal. Generally, FCS may seem slightly better than MVNI, but both methods seem to perform equally well. DL is another favourable alternative in case one is not well conversant with the imputation methods. Faster and easily implemented in standard statistical software.

4.4.3 Example: arthritis data

4.4.3.1 Data

The dataset used are from a homoeopathic clinic in Dublin, made available in Pawitan (2001). The data are on 60 patients (12 males and 48 females) between the ages of 18 and 88 who were under treatment for arthritis. The patients were followed up for a month (in 12 visits) and their pain scores assessed. Only two patients had all the scores for the 12 visits. The score was graded from 1 to 6, with high indicating worse. Only those with a baseline score greater than 3 and a minimum of six visits are reported. About 36% of the pain score data were missing. Of the 60 patients 27 had RA-type arthritis where 5 were males and 22 were females, while 33 had type OA. Seven of these were males. Some descriptive statistics of the dataset are summarized in Table 4.4 and Figure 4.1.

Table 4.4: Descriptive statistics of the incomplete arthritis data.

Arthritis data: Variable	Description	Range	% miss	Mean	Mode	Std Dev.
Baseline variables						
Sex	1= Male, 0=Female	1/0	0			
Age	Age of the patient	18 - 88	0	59.5	57	12.6
Time	Number of patient visits	1 - 12	0			
Type	Arthritis type (RA =1, OA = 0)	1/0	0			
Years	Number of years with symptom	0 - 57	0	10.7	1	12.2
Response variable						
pain_scores	Scores on the arthritis pain	1 - 6	35.56%		4	

Note: Data missing on the dependent variable. [a] Arthritis type (RA =rheumathoid arthritis, OA = ostheo-arthritis). [b] Std Dev = standard deviation.

Looking at Figure 4.1, it is apparent that many patients missed their visits towards the end of the follow up. After the sixth visit the missing data were more than 30% on every visit.

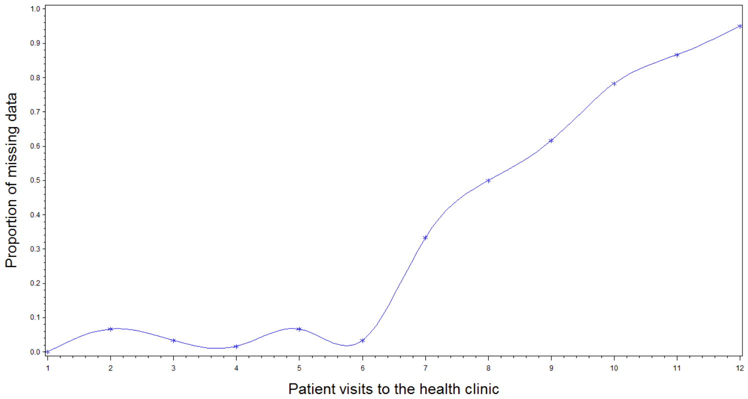


Figure 4.1: The proportion of missing data per scheduled visit to the clinic.

4.4.3.2 The proportional odds assumption

Before the model statistics can be interpreted, it is very important to test the proportional odds assumption. The assumption was examined using the Brant test in STATA. A non-significant omnibus test provides informal evidence that the assumption is not violated. Table 4.5 gives part of the assumption results. The assumption was upheld for age, type and years. The same cannot be said for sex and time.

In case the proportional odds assumption is not satisfied for some variables but satisfied for others, then a partial proportional odds model (PPOM) can be fit. However, the PPOM is just an extension of the proportional odds model (POM). Both PPOM and POM can be adequate for data analysis. The most important aspect with regards to interpretation of analysis results involving ordinal data is that the interpretation should take the odds proportionality into account, i.e., the odds of being in high or lower category depending on the case. Using the PPOM or for simplicity using the POM, would not change the overall final inference. For simplicity, our results did not consider the PPOM.

Table 4.5: Brant test of proportional odds assumption.

Variable	chi2	p>chi2	df
All	73.87	0.000	20
Sex	34.88	0.000	4
Age	7.59	0.108	4
Time	30.55	0.000	4
Type	8.35	0.079	4
Years	6.00	0.199	4

A model of interest for the study was the main effects model. Only the dependent variable had missing values. At first, the data were analysed without any alterations or attempts to impute the missing values. This was under the direct likelihood (DL) approach. We chose the DL parameter estimates as reference for the real application dataset against which we can check the relative performance of MVNI versus FCS when considering MI. Because direct likelihood is valid under the same properties as multiple imputation, we expect the two approaches to produce similar parameter estimates or somehow close to each other. After the direct likelihood analysis we conducted the multiple imputations under FCS and MVNI where upon imputation, a similar marginal model as the direct likelihood analysis was fitted in the analysis task. Finally, the SAS procedure MIANALYZE was employed to pool the results from multiple datasets.

4.4.3.3 Results

Table 4.6 shows the parameter estimates, standard errors and 95% confidence limits of fixed effect estimates by the imputation methods and direct likelihood analysis. These analysis results showed similar trends to those from the simulated data for most cases. The results indicate that the parameter estimates by MVNI were comparable to those of direct likelihood in more cases than FCS. In three cases, MVNI values were closer to those from the direct likelihood method compared to two FCS cases. Moreover, MVNI resulted in smaller standard errors than the FCS method for age, time and type. Equal values are observed for years. MVNI gives a larger standard error than FCS for sex. This may be attributed to the fact that both sex and years were highly insignificant predictors by both MVNI and FCS, as is evidenced in the confidence limits. Both methods seem to perform fairly well in general. Looking at the direct likelihood method, it gives smaller standard errors than the imputation methods for all parameters except time. It is equally a favourable alternative method when faced with incomplete ordinal data and may be used whenever one is not sure about what imputation method to use or not having necessary know how on imputation methods.

4.5 Discussion

The idea behind MI is to draw valid and efficient inferences by fitting analysis models to multiply imputed data. We ensured that the imputed values bear the structure of the data, and uncertainty about the structure and included any knowledge about the process that led to the missing data (van Buuren, 2007). The method of choice to create the imputed datasets depends on the missing data pattern. For monotone missing patterns a parametric regression method that assumes multivariate normality or a non-parametric method that employs propensity scores may be be used (Molenberghs and Verbeke, 2005). Alternatively, one may generate imputations by performing a series of univariate regressions, rather than just a single large model (making it somewhat easier to estimate), and without assuming normality of the variables.

Table 4.6: Parameter estimates, standard errors (StdErr) and confidence limits (C. L.) obtained from the arthritis data under the methods of direct likelihood (DL), fully conditional specification (FCS) and multivariate normal imputation (MVNI).

	DL			**MVNI**			**FCS**		
Param	Est	StdErr	95% C. L.	Est	StdErr	95%C. L.	Est	StdErr	95% C. L.
sex	0.2130	0.1543	(-0.0895, 0.5154)	0.2192	0.2273	(-0.2282, 0.6667)	0.1991	0.2212	(-0.2361, 0.6342)
age	0.0262	0.0062	(0.0140, 0.0383)	0.0260	0.0070	(0.0127, 0.0393)	0.0255	0.0072	(0.0114, 0.0396)
time	-0.2218	0.0610	(-0.3414, -0.1023)	-0.2004	0.0284	(-0.2565, -0.1444)	-0.2212	0.0304	(-0.2812, -0.1612)
type	0.9868	0.1275	(0.7369, 1.2366)	0.9620	0.1677	(0.6326, 1.2914)	0.9799	0.1770	(0.6319, 1.3280)
years	0.0107	0.0045	(0.0018, 0.0195)	0.0106	0.0070	(-0.0032, 0.0243)	0.0099	0.0070	(-0.0038, 0.0236)

Note: Missing values about (36%) on the response variable.

When faced with a discrete variable (e.g. ordinal), an appealing approach at first sight may be to treat ordinal variables as continuous for the purpose of imputation, and then round the imputed data values to the nearest valid discrete value before continuing to fit the substantive model (Carpenter and Kenward, 2013). However, researchers caution the analyst from analysing ordinal outcome as a continuous or dichotomized variable for a number of reasons. First, comparing an ordinal to a continuous outcome or dichotomizing it to run a binary logistic regression may lead to efficiency loss due to information loss, reduced statistical power and decreased generality of the analytic conclusions (Gameroff, 2005). Logically, continuous models can yield predicted values outside the range of the ordinal variable and finally, a continuous model may produce correlated residuals and regressors when used for ordinal outcomes and does not account for the ceiling and floor effects of the ordinal outcome. This may lead to biased estimates of the regression coefficients (Bauer and Sterba, 2011). This issue has created a lot of debate among researchers. Schafer (1997) argues that methods assuming multivariate normality may be used in cases where the normality assumption does not hold. Furthermore, these methods have also been successfully used by the authors in Choi et al. (2008); Demirtas, Frees and Yucel (2008); Seitzman et al. (2008). This is therefore still an active area of further research. However, apart from imputing the ordinal variable directly as a continuous variable, another option is to use a set of indicators. The values are imputed as continuous, and then assign imputed values into categories based on the mean indicators imputed in a separate round of imputation. In Lee et al. (2012), this strategy of comparing methods for imputing ordinal data using methods that assume multivariate normality is discussed.

More often analysts are faced with datasets with both dropouts and nonmonotone missingness, like the arthritis data where the amount of dropout was considerable, while that of nonmonotone missingness is much smaller. It is heedful to include all in the analyses as noted by Molenberghs and Verbeke (2005). One can undisputedly opt for direct likelihood analysis or standard generalized estimating equation (GEE; Diggle et al., 2002; Liang and Zeger, 1986; Molenberghs and Verbeke, 2005). Weighted generalized estimating equation (WGEE; Robins, Rotnitzky and Zhao, 1995) is possible but one has to find appropriate weights. Alternatively, one may make the missing patterns monotone by multiple imputation and go ahead to do the WGEE.

The primary goal for this study was to investigate the performance of MVNI and FCS as MI methods. These two approaches follow different theoretical assumptions and thus involve different computational methods. Each of the methods comes with its own specifications. MVNI is appealing because of its ease of specification of the imputation model. Conversely, FCS requires an added effort in model specification, and separate regression models must be fitted for each variable in the imputation model (van Buuren

2007). But in our problem these conditional regressions were automatically specified because of our small number of variables and only one variable had missing values. On the other hand, an added advantage of FCS again is the natural handling of ordinal variables. For MVNI we had to handle the ordinal variables under a continuous scale in order to take advantage of the well-established imputation procedures for Gaussian outcomes, and then rounded to the required categories post-estimation. Basically, this assumption has been the major stumbling block in the working of MVNI and a number of researchers have reported FCS being better than MVNI, for example, Van Buuren (2007); Yu, Burton and Riviero-Arias (2007). In this study, we did not find a strong reason to support this. Specifically speaking, the MVNI approach is equally appropriate as is FCS when faced with missingness in ordinal variables, at least of the type presented. Similarly, Lee and Carlin (2010) are in support of the findings. We notice that the conclusions for comparing the two methods differed among researchers. This was probably due to differences between their simulation studies, and the way they rounded off the continuous values e.g., Lee and Carlin (2010) who used adaptive rounding with MVN. However, without doubt, further comparisons on these two methods, where more settings will be considered is incumbent.

In this paper, we focussed on MAR mechanisms for monotone missing data patterns. The methods of FCS and MVNI can be extended to non-monotone missing data patterns (UCLA, 2015). Although the authors doubt the suitability of the MAR assumption for non-monotone missing data, Robins and Gill (1997) present a new strategy of ignorable non-monotone missing data models, called the randomised monotone missingness (RMM), which is a subset of MAR. They argue that the RMM is the only plausible non-monotone MAR mechanism that is not MCAR, but they caution the user not to analyse non-monotone missing data assuming that the missingness is ignorable if a statistical test has rejected the hypothesis that the missingness process can be represented as RMM. We recommend interested readers to Robins and Gill (1997) for further details on RMM and Daniel and Kenward (2012) who reiterate the RMM idea and extend it to a Markov randomised monotone missingness (MRMM). MRMM is a specific subset of RMM. The authors present a clear theoretical framework and applicability in non-monotone missingness patterns. We therefore state that the methods employed in our paper can further be extended to non-monotone cases. These methods are valid under MAR. When faced with non-monotone missingness, one may take the Daniel and Kenward (2012); Robins and Gill (1997) routes as one of the options that exist in the literature. If under any circumstances, it happens that the MAR is not a sensible assumption for non-monotone missing cases, as an outset, sensitivity analyses are advised. However, a shift from MAR to possibly MNAR is not a worry, because as pointed out by Molenberghs et al. (2008) the price to pay is minimal as no formal distinction exists between MAR and MNAR. This is because for any MNAR model there exists an MAR counterpart that fits the data very well.

In this paper, missingness was only on the outcome variable. This does not limit the applicability of FCS and MVNI to that case only. The methods can be extended to situations where data are missing for outcomes and covariates. A lot of work has been done on this. In the papers, (Royston, 2004; Van Buuren, Boshuizen and Knook, 1999), MICE alias FCS was used to fill missing values in incomplete covariates. The assumption of multivariate normality has been used to impute in covariates and responses. We cite Schafer (1997); Schafer and Yucel (2002); Seaman, Bartlett and White (2012) among many works in the literature.

Chapter 5

Comparison of methods for the analysis of incomplete longitudinal ordinal outcomes: Application to a clinical study on childhood malnutrition

Abstract

Ordinal responses are often encountered in longitudinal studies, especially in clinical trials. Apart from failing to meet the usual normality assumption for analysis and inference, these data are prone to missingness. Thus, using ordinary least squares regression for such type of data could produce biased and inefficient estimates. In addition, failure to deal with incomplete information jeopardizes the validity of inferences, while some of the available methods for dealing with incomplete data may not meet the distributional assumptions of the data. This paper presents likelihood estimation methods for longitudinal ordinal data, focussing on the cumulative logit model, and also compares three methods for incomplete ordinal data subject to both monotone (dropout) and non-monotone missingness. In particular, complete case analysis and direct maximum likelihood analysis of the incomplete data are contrasted with multiple imputation strategies, namely: the full conditional specification, multivariate normal imputation and the ordinal imputation method. Applications are based on the analysis of longitudinal nutritional data from a clinical study conducted in four study sites in Kenya. The findings

showed that for incomplete ordinal outcome variables, direct maximum likelihood analysis and multiple imputation strategies produced comparable results. Complete case analysis generally gave poor results.

[1]

5.1 Introduction

Longitudinal studies, including clinical trials, are designed to record observations repeatedly over time. Missing response data is very common in these studies due to study dropout, mistimed measurements, or generally loss to follow up. Subjects may drop out of the study prematurely resulting in a monotone missingness pattern, also termed dropout, or they may miss one follow up time and then be measured at the next follow up time. The latter results in an intermittent (non-monotone) missingness pattern. When data are incomplete, the validity of any analysis approach will require that certain assumptions about the reasons for missingness are tenable. Rubin (1987) and Little and Rubin (2014) classified the mechanisms in three categories. A) Data are missing completely at random (MCAR) if the probability of missingness is independent of responses observed or unobserved, or any other variables in the analysis; here, any analysis valid for the whole dataset is valid for the observed data. B) Data are missing at random (MAR) when the probability of missingness is dependent only on observed responses. C) Data are missing not at random (MNAR) when the probability of missingness is dependent on unobserved responses and potentially on observed information.

When data are missing only on the response variable, an MAR analysis assumes that the probability of a value missing may depend on observed measurements and covariates but, given these, is independent of any unobserved measurements. For missing covariates, MAR assumes that missingness is independent of missing outcomes and covariates, given observed outcomes and covariates. In other words, MAR assumes that responses that have similar observed characteristics are comparable and that the missing values are independent of any unobserved measurements. Conventionally, there are a number of methods to handle the missing data problem. Under the the stringent MCAR assumption, one can opt for the complete case analysis (CCA) by discarding cases with missing observations and proceed with the analysis using only the observed data. If MCAR does not hold, CCA will be severely biased. But, even if the MCAR holds, a CCA analysis is less efficient compared to analyses that use all available data because in the latter all the information is available to draw inferences. Little and Rubin (2014) state that, likelihood-based inference is valid whenever the mechanism is MAR and provided the

[1] A. Kombo et al. (under review for submission). Comparison of methods for the analysis of incomplete longitudinal ordinal outcomes: Application to a clinical study on childhood malnutrition.

technical condition holds that the parameters describing the missingness mechanism are distinct from the measurement model parameters. The expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977), is a general iterative procedure that can be used to find the maximum likelihood estimates in missing data problems. The multiple imputation (MI) procedure (Rubin, 1978b;1987) replaces each of the missing value with a set of $M \geq 2$ plausible values, i.e., values drawn from the distribution of the missing data given the observed data, that account for the uncertainty about the right value to impute. The imputed data sets are then analyzed by using standard procedures for complete data and combining the results from these analyses. Correct imputation leads to valid large sample inferences and produces estimators with good large sample properties (Little and Rubin, 2014). Work on semiparametric approaches for all types of missingness based on weighted estimating equations (WEE) include Lipsitz, Ibrahim and Zhao (1999); Robins, Rotnitzky and Zhao (1994, 1995a). Alternatively, Schafer (2003) proposed the use of multiple imputation for the missing response values from a fully parametric model, then followed by GEE, leading to a hybrid method indicated by MI-GEE. The approaches are computationally efficient and can produce robust estimates that are consistent in more relaxed settings.

So far most methodological work has been carried out on continuous and binary outcomes. For discrete data, comparisons are mostly on binary response data. Comparisons of analysis methods are needed in other categorical types of outcomes.

In clinical trials, it is common for analysts to encounter response measures that are categorical in nature with more than two categories. The responses represent categories of outcome information rather than the usual interval scale. If the response variable takes on values that cannot be ordered inherently then the response is nominal. If the response takes values that can be ordered naturally (e.g., nutritional status: severe, moderate, at risk, and well nourished), then the response is ordinal. Health related outcomes are often ordinal, but fail to satisfy the preconditions usually needed to perform an ordinary least squares (OLS) regression. When the outcomes are highly non-Gaussian, as is the case when most of the respondents' score is skewed at the very top or bottom of the scale, ordinal regression can be more justified, and perhaps more informative than the ordinary least squares regression. In fact, when the response is categorical, OLS cannot produce the best linear unbiased estimator (BLUE), i.e., it is biased and inefficient. Ordinal regression analysis provides sensible ways of estimating parameters for ordinal response data, regardless of whether accompanying predictor variables are also categorical or continuous. Ordinal regression takes advantage of the ordering by use of the "cumulative odds","cumulative probabilities", and "cumulative logits" concept. The situation is more complicated when some of the respondents are not measured on some follow up occasions and hence lead to missing data.

This paper presents a comparison of different analysis methods for longitudinal ordinal response data with missing values. In particular, we compare the traditional complete case analysis method with advanced methods like the direct maximum likelihood (DL) analysis (Beunckens, Molenberghs and Molenberghs, 2005) and multiple imputation in its different imputation paradigms, namely: a full conditional specification (FCS; Van Buuren et al. 2006; Van Buuren, 20017), multivariate normal imputation (MNI; Schafer, 1997) and the ordinal imputation method (OIM; Donneau, Mauer and Molenberghs, 2015).

The analysis here only deals with approaches under MAR, while MNAR is beyond the scope of the paper. However, we note that the MAR assumption cannot be fully substantiated from the data, and that MAR and MNAR cannot be distinguished on formal statistical grounds. One only suspects that the data are not MAR but nothing from the data will indicate whether or not that is true (Allison, 2014). Note that standard multiple imputation implementations almost all assume MAR to hold. An exception is SAS procedure MI, which allows for MNAR missingness as of version 9.4. The procedure offers a variety of options to conduct sensitivity analysis and examine how MNAR mechanisms could jeopardize the MAR results. This is important because Molenberghs et al. (2008) have shown that MAR and MNAR cannot be formally separated. However, in spite of this, the impact on key parameter estimators and corresponding hypothesis tests can be considerable. Arguably, such a sensitivity analysis should virtually always be conducted.

This paper is structured as follows. In Section 5.2, we introduce the data setting and a brief description of the extended formulation of binary logistic regression to ordinal response variables. We also describe incompleteness in this section and give a model for dropout. In Section 5.3, statistical methods for incomplete longitudinal data are discussed. A motivating example dataset (RSCM data) is presented in Section 5.4, where data analyses are also carried out. First we carry out a comparative analysis using a subset of the dataset to compare methods for the incomplete longitudinal ordinal outcome data. Next, we conduct a simulation study by generating datasets that mimic the RSCM data and repeat the comparative analysis of the methods of interest. We describe the findings of the analyses and discussions thereof. In Section 5.5, we apply the same methods to the whole original incomplete RSCM dataset. We conclude and point out areas for further research in Section 5.6.

5.2 Data setting and modelling framework

For each individual $i = 1, \ldots, N$ in a study, we consider a series of measurements $Y_i = (Y_{i1}, \ldots, Y_{in_i})'$, along with fixed covariate matrix $X_i = (x_{i1}, \ldots, x_{in_i})$ which may include measurement occasions $(t_{i1}, \ldots, t_{in_i})$, where $x_{ij}, j = 1, 2, \ldots, n_i$ is a p-dimensional vector of covariates at time t_{ij}.

If the response variable Y_{ij} takes on two values, say 'event = 1' or 'nonevent = 0', then the conditions for linear regression are not met. It implies that the errors are binary and not normally distributed. Then, binary logistic regression may be used. If we let the probability of an event be $\pi_{ij} = Pr(Y_{ij} = 1)$, then the logistic model can be written as:

$$\text{logit}(\pi_{ij}) \equiv \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha + x'_{ij}\beta = \alpha + x_{ij1}\beta_1 + x_{ij2}\beta_2 + \cdots + x_{ijp}\beta_p, \qquad (5.1)$$

where α is the intercept parameter and β is the vector of slope parameters. The linear predictor $\alpha + x'_{ij}\beta$ models the log odds of the event of interest as a function of covariates. It is key to note here that in equation (5.1), it is necessary to account for the fact that observations from the same subject are correlated.

In many studies, the response can have more than two levels. Logistic regression can be extended to more than two response levels i.e., to the so-called polytomous logistic regression. Specifically, in ordinal response variables (where the responses possess an intrinsic ordering) it extends to the ordinal response model; the proportional odds model (McCullagh, 1980) which is combined with a cumulative logit link. The proportional odds model, also known as the cumulative logit model is likely the most common ordinal logistic regression model (Bender and Grouven, 1998).

Suppose the ordinal response variable Y has $C, (c = 1, 2, \ldots, C)$ levels. We use a multinomial distribution and a cumulative logit link to address the nature of the response, and $C - 1$ separate linear predictors to model the probabilities. In the context of longitudinal ordinal observations, we write $\pi_{ijc} = Pr(Y_{ij} = c)$, the probability of observation Y_{ij} taking level c. If we let the cumulative probability be $\phi_{ijc} = Pr(Y_{ij} \leq c|x_{ij})$, the probability of being at or below category c, given a set of predictors, we define the general cumulative logit link model as:

$$\begin{aligned}\text{logit}(\phi_{ijc}) = \log\left(\frac{\phi_{ijc}}{1-\phi_{ijc}}\right) &= \log\left[\frac{\pi_{ij1} + \cdots + \pi_{ijc}}{\pi_{ij,c+1} + \cdots + \pi_{ijC}}\right] \\ &= \alpha_c + x'_{ij}\beta_c, \qquad c = 1, 2, \ldots C - 1, \qquad (5.2)\end{aligned}$$

where α_c gives the threshold parameters (intercept terms that depend on the ordinal levels), and c indexes the $C - 1$ logits. The regression parameters, β_c, reflect the association between the predictor variables and the outcome variable specific to a given response function.

A key simplifying assumption in equation (5.2) is to impose a restriction on the linear predictors by assuming the same slope parameters for each of the response logits (this is referred to as the proportional odds assumption) and by restricting α_c to monotocity (i.e., $\alpha_1 < \alpha_2 < \alpha_3 < \cdots < \alpha_{C-1}$), so that;

$$\text{logit}(\phi_{ijc}) = \alpha_c + x'_{ij}\beta, \qquad c = 1, 2, \ldots C - 1. \tag{5.3}$$

This model has been studied by many researchers. Aitchison and Silvey (1957) used a normit scale to obtain a maximum likelihood analysis. Cox and Snell (1989) on the other hand employed the log-odds scale. For the log-odds scale, the cumulative logit model is often referred to as the proportional odds model. The proportional odds model has $C - 1 + p$ parameters to be estimated. Otherwise, the most general model (5.2) will require $(C-1)(1+p)$ parameters, allowing different slope coefficients for the $C - 1$ response logits.
The cumulative logits in (5.3), can be exponentiated to obtain cumulative odds:

$$\frac{\phi_{ijc}}{1 - \phi_{ijc}} = \exp(\alpha_c + x'_{ij}\beta), \tag{5.4}$$

from which cumulative probabilities can be solved such that

$$\phi_{ijc} = \frac{\exp(x'_{ij}\beta)}{1 + \exp(x'_{ij}\beta)}. \tag{5.5}$$

However, it should be noted that the proportional odds model is the result of the somehow stringent assumption of proportionality of odds, which may not be automatically valid for all ordinal response variables. If proportionality is valid for one set of coefficients and does not hold for some, model (5.3) may be rewritten as:

$$\text{logit}(\phi_{ijc}) = \alpha_c + x'_{ij}\beta + z'_{ij}\gamma_c, \tag{5.6}$$

where x_{ij} represents the predictor variables with equal slopes, z_{ij} represents the predictor variables with unequal slopes, β represents the regression parameters for x_{ij} and γ_c represents the regression parameters for z_{ij} for a given c. For further and wider details on ordinal variables modelling, we recommend among others; Agresti (1989, 2007, 2010); Armstrong and Sloan (1989); Greenland (1994); Lee (1992); Molenberghs and Verbeke (2005); Stokes, Davis and Koch (2012).

In case Y_i is not completely observed we write $Y_i = (Y_{i,o}, Y_{i,m})$, where $Y_{i,o}$ and $Y_{i,m}$ denotes the observed and missing components of Y_i respectively. We define a vector of missingness indicators $R_{in_i} = (R_{i1}, \ldots R_{in_i})'$, where $R_{ij} = 1$, if Y_{ij} is observed and 0 otherwise $(j = 1, \ldots n_i)$. In our case, we consider dropout only on the outcome variable. Covariates x_i are thus assumed fully observed. Certainly, the approaches we take in this paper can also be employed to non-monotone and incomplete covariate data settings. Therefore, we can define the full data as a combination of the processes generating

Y_i (measurement process) and R_i (missingness process). The full data density can be represented by:

$$f(y_i, r_i|x_i, \theta, \psi), \tag{5.7}$$

where the parameters θ and ψ represent the measurement and missingness processes respectively. The full data density (5.7) can be factorized as

$$f(y_i, r_i|x_i, \theta, \psi) = f(y_i|x_i, \theta) f(r_i|y_i, \psi). \tag{5.8}$$

The conditional distribution of the missing data mechanism can be equivalently expressed as $f(r_i|y_{io}, y_{im}, \psi)$. Since in this part of the paper we confine the analysis to dropout, we prefer to use a scalar variable D_i rather than R_i. We define D_i to be the occasion at which a dropout occurs, and denote it: $D_i = 1 + \sum_{j=1}^{n_i} R_{ij}$. The model for the dropout process is based on a logistic regression for the probability of dropout at occasion j, given that the subject was in the study up to occasion $j-1$. We denote this probability by $P(h_{ij}, y_{ij})$, and express the outcome history as $h_{ij} = (y_{i1}, y_{i2}, \ldots, y_{i,j-1})$. For simplicity, we assumed that the dropout depends only on the current observed measurement (y_{ij}) and the immediately preceding measurement $(y_{i,j-1})$. We therefore assume the dropout model to be

$$\begin{aligned} \text{logit}[P(h_{ij}, y_{ij})] &= \text{logit}[Pr(D_i = j | D_i \geq j, h_{ij}, y_{ij})] \\ &= \psi_0 + \psi_1 y_{i,j-1} + \psi_2 y_{ij}, \end{aligned} \tag{5.9}$$

where ψ_0 denotes the intercept of the regression, and the coefficients ψ_1 and ψ_2 are the effects of $y_{i,j-1}$ and y_{ij} respectively. The model reduces to a MAR model if $\psi_2 = 0$, i.e., the missingness process is related to the observed outcome prior to dropout. MCAR applies if $\psi_1 = \psi_2 = 0$, implying the missingness is independent of the previous and current measurement. If $\psi_2 \neq 0$, then the missingness depends on the missing data at the dropout occasion. Hence, we cannot rule out MNAR and the missingness process cannot be ignored. Notice here that the test for $\psi_2 = 0$ versus $\psi_2 \neq 0$ (MAR versus MNAR) relies on untestable assumptions such as the distributional form (See, for example, Kenward, 1998; Molenberghs and Kenward, 2007; Newsom, Jones and Hofer, 2012; Rhoads, 2012). In fact, Molenberghs et al. (2008) show that a formal distinction between MAR and MNAR is not possible.

5.3 Statistical approaches to incomplete data

5.3.1 Traditionally used approach

A complete case analysis (CCA) discards all incomplete cases and analysis is carried out on what remains. Its apparent advantage is that it is very simple and easy to implement in standard statistical software without extra toil. In fact, it is the default method in many statistical software packages. Under the most stringent MCAR assumption, it leads to valid unbiased parameter estimates. However, even when it is valid, the method suffers from major drawbacks. In fact, it can be very inefficient, such that it produces estimates with higher variance than would be obtained with other equally valid methods, especially when we have to rule out a large number of cases (Rubin, 1987; Little and Rubin, 2014). This consequently impairs its statistical power and precision. When the missingness process is MAR but not MCAR and a CCA analysis used, results are severely biased. Also, the statistical analysis may be biased when the complete cases are systematically different from the incomplete ones. However, complete case analysis can have an auxiliary analysis role, particularly if it relates to a scientific question (Beunckens, Molenberghs and Kenward, 2005). Thus in this paper the CCA method is not of primary interest.

5.3.2 Direct maximum likelihood analysis

It is important to consider approaches for handling missing data based on methods that are valid under the less restrictive MAR assumption. The likelihood-based MAR analysis (also termed likelihood-based ignorable analysis), or direct maximum likelihood (DL) analysis, is one where the observed data are used without deletion nor imputation. Because of this, appropriate and automatic adjustments, i.e., validity under MAR, are made to parameters at times when data are incomplete, due to the within-subject correlation. DL uses information on all subjects, including information from early dropouts (Beunckens, Molenberghs and Kenward, 2005).

From the full data likelihood contribution for the i^{th} subject, $f(y_i, r_i|\theta, \psi)$, we view the observed data likelihood L contribution for the sequence y_i as:

$$L(\theta, \psi|y_i, r_i) \propto f(y_i^o, r_i|\theta, \psi),$$

where

$$\begin{aligned} f(y_i^o, r_i|\theta, \psi) &= \int f(y_i, r_i|\theta, \psi) dy_i^m \\ &= \int f(y_i^o, y_i^m|\theta) f(r_i|y_i^o, y_i^m, \psi) dy_i^m, \end{aligned} \tag{5.10}$$

with all parameters and variables as described in Section 5.2. Here, we make the key assumption that the response distribution and the missing data mechanism model are correctly specified. Then, under the MAR assumption, (5.10) simplifies to:

$$\begin{aligned} f(y_i^o, r_i|\theta, \psi) &= \int f(y_i^o, y_i^m|\theta) f(r_i|y_i^o, y_i^m, \psi) dy_i^m \\ &= f(y_i^o|\theta) f(r_i|y_i^o, \psi). \end{aligned} \tag{5.11}$$

Moreover, if parameter separability holds, meaning the parameters θ and ψ are distinct in the sense that the joint parameter space is: $\Omega(\theta, \psi) = \Omega(\theta) \times \Omega(\psi)$, we make use of likelihood inference for the parameter of interest θ, which is thus based on the marginal density of the observed data only. In this case, the missing data mechanism is termed ignorable (Little and Rubin, 2014; Rubin, 1976). The consequence of this is that the missing data mechanism does not need to be modelled explicitly.

5.3.3 Multiple Imputation

The idea is to fill the missing values by randomly drawing plausible values from the conditional distribution of the missing observations given the observed ones. Multiple imputation involves three steps.

Imputation step: Instead of filling in a single value, the conditional distribution of the missing data is used to generate multiple (i.e., $M \geq 2$) values that reflect the uncertainty around the actual value. The missing data are filled in with the estimated values and a complete dataset created. In this way, M complete datasets are obtained.

Analysis step: Each of the M complete datasets from the first step is then analyzed using an appropriate analysis model.

Pooling step: Finally, the parameter estimates obtained from the M complete data analyses are combined for inference.

When imputing one or many variables, the analyst has to consider a number of decisions. The imputation procedure chosen will depend on the missing data pattern as well as the type of variables with missing values or type of distribution under which the variables are imputed. However, most of the imputation software packages (like the norm module in R, and SAS PROC MI) assume a fully parametric multivariate normal distribution on the partially observed variables. When normality does not hold, like in categorical

variables (binary and ordinal) analysts still choose to use methods assuming multivariate normality as an approximation and to round the imputed values to the nearest observed integer. Thus, there is ongoing debate on the appropriateness of this rounding operation. In this paper, we focus on three imputation strategies. The first is to treat the ordinal missing data as normally distributed and use models that assume normality to impute the missing values. Second, we will impute based on models designed for categorical data (e.g., linear discriminant analysis, or logistic regression models). Specifically, the proportional odds logistic regression model will be used. The binary and proportional odds logistic regression models for imputation are available in SAS PROC MI, IVEware and MICE in R. Finally, we use the ordinal imputation method.
Notice that the standard applicability of MI approaches assumes the MAR assumption, but this does not limit its applicability strictly to this mechanism. The MI approaches can be extended to MNAR provided the user is willing to make some additional assumptions about the mechanism. Also, MI can be used for both monotone and non-monotone missing data patterns, and in situations where missingness is in both the outcome as well as covariates. Allison (2012) has handled cases where missingness is in both the dependent variable as well as the predictor variables using maximum likelihood and multiple imputation. But note that, for users of SAS, there is no procedure that does maximum likelihood analysis for logistic regression with missing data on the predictors! Below the three statistical formulation of the three methods are briefly presented.

5.3.3.1 Multivariate normal multiple imputation

Multivariate normal imputation assumes that the data are sampled from a multivariate normal distribution. The idea is to generate plausible imputations that account for between-imputation variability. Such imputations are based on the data augmentation algorithm (Tanner and Wong, 1987), and are obtained by iteratively alternating between two steps: an imputation step (I-step) and a posterior step (P-step). For our ordinal response variable, each $(X_{ij}, Y_{ij}), i = 1\ldots, N; j = 1, \ldots n_i$ is assumed to have been randomly sampled from a multivariate normal distribution with mean vector μ and covariance matrix Σ. Let $\theta = (\mu, \Sigma)$. The covariates are suppressed from notation. Missingness is also assumed only on the response variable Y. Ideally, the ordinal data is not normally distributed, the normality assumption is just one way of tackling the problem and this approach should be considered an approximation.
The I-step: Given starting values for θ, a value for missing data Y^m is randomly drawn from the conditional multivariate normal distribution of $Y^m|Y^o$; $f(Y^m|Y^o, \theta)$. Denote the mean vector of the variable in the observed and in the missing parts of the dataset as $\mu = (\mu_o, \mu_m)$. But, note here that this partitioning is not at the level of the dataset,

but rather at the level of the subject. However, we suppress the index i for simplicity of notation. Similarly, the covariance matrix is partitioned such that

$$\Sigma = \begin{pmatrix} \Sigma_o & \Sigma_{o,m} \\ \Sigma'_{o,m} & \Sigma_m \end{pmatrix},$$

where (due to symmetry) $\Sigma_{o,m} = \Sigma'_{o,m}$ denotes the covariance matrix between Y^o and Y^m, Σ_o and Σ_m represent the variance matrix for Y^o and Y^m, respectively. The conditional mean $\mu_{m|o}$, and the conditional covariance matrix $\Sigma_{m|o}$, must be derived. Following Donneau et al. (2015) and Schafer (1997), it is assumed that $f(Y^m|Y^o, \theta)$ follows a normal distribution with conditional mean $\mu_{m|o}$ and conditional covariance matrix $\Sigma_{m|o}$, i.e.,

$$Y^m|Y^o, \theta \sim \mathrm{N}(\mu_{m|o}, \Sigma_{m|o}),$$

where,

$$\Sigma_{m|o} = \Sigma_m - \Sigma'_{o,m}\Sigma_o^{-1}\Sigma_{o,m} \qquad \text{and} \qquad \mu_{m|o} = \mu_m + \Sigma'_{o,m}\Sigma_o^{-1}(Y^o - \mu_o). \tag{5.12}$$

The P-step: After the first iteration, new values for $\theta^* = (\mu^*, \Sigma^*)$ are drawn from its posterior distribution (typically from a normal-Wishart family, given a normal-inverse-Wishart prior distributions). Assuming an objective prior distribution for θ^*, its posterior at the rth iteration will therefore be expressed as

$$\mu_{|\Sigma}^{(r)} \sim \mathrm{N}\left(\bar{Y}, \frac{1}{N}\Sigma^{(r)}\right), \qquad \Sigma^{(r)} \sim U^{-1}(N-1, (N-1)S), \tag{5.13}$$

where

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_{ij}, j = 1, \ldots, n_i, \quad S \ = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})(y_i - \bar{Y})' \quad \text{and} \quad U = \frac{1}{M}\sum_{l=1}^{M} U_l,$$

with U measuring the within-imputation variability, U_l being the corresponding variance-covariance matrix for the l^{th} imputed dataset, $l = 1, \ldots, M$.

Note here that $\bar{Y}$ and S are governed by the observed data and the missing data imputed at the last imputation step. The two steps are repeated sequentially thus creating a Markov chain of pairs $(Y^m_{(1)}, \theta_{(1)}), (Y^m_{(2)}, \theta_{(2)}), \ldots$. Each step depends on the previous one, creating dependency across the steps. The two steps are iterated long enough until convergence.

The imputed values obtained in this fashion are not discrete and they need to be rounded off to the nearest integer value. We have to first impute at the continuous variable scale using the normal data assumption and then discretize based on estimated thresholds.

Importantly, minimum and maximum values must be provided to capture the scope of the ordinal variable imputed. Yucel and Zaslavsky (2004) provide practical suggestions on rounding in multiple imputation.

5.3.3.2 Full conditional specification

For the categorical models, imputation can also be done using the MCMC algorithm with chained equation imputation. Full conditional specification also known as multiple imputation via chained equations (Brand, 1999; Van Buuren and Groothuis-Oudshoorn, 2011), or partially incompatible MCMC (Rubin, 2003) is a practical method to generating multiple imputations, one for each partially recorded variable in the dataset. It can handle different variable types (continuous, binary, nominal and ordinal) since each variable is imputed using its own imputation model. The variables with missing values are imputed sequentially one after the other. The first variable, say y_1, with missing data is regressed on all other variables $y_2, y_3, \ldots, y_n$, limited to subjects with the observed y_1. Then the process is repeated for the next variable with missing values, but this variable also uses the imputed values of y_1. It continues until all variables with missing values are exhausted. This result is called a cycle. To stabilize the results, the procedure is iterated for a number of cycles to produce a single imputed dataset. The entire procedure is repeated M times to produce M complete datasets.
Being specific to the incomplete ordinal response variable, Y with $C > 2$ categories, and using the cumulative proportional odds model (5.3), β and α are estimated by maximum likelihood. Values β^* and α^* are drawn from a normal approximation to their posterior distribution. Let the estimated probability that an observation falls in category $c = 1, \ldots C$ be given by

$$p_{ijc} = Pr(y_{ij} \leq c|x_{ij}; \beta^*, \alpha^*) - Pr(y_{ij} \leq c-1|x_{ij}; \beta^*, \alpha^*).$$

For each missing observation Y_{ij}^m, let $p_{ijc}^* = Pr(y_{ij} = c|x_{ij}; \beta^*)$ be the drawn category membership probabilities, and $\phi_{ijc} = \sum_{c'=1}^{c} p_{ijc'}^*$. Then each imputed observation y_{ij}^* is obtained by

$$Y_{ij}^* = 1 + \sum_{c=1}^{C-1} I(u_{ij} > \phi_{ijc}),$$

where u_{ij} is a random draw from a uniform distribution, $u_{ij} \sim \mathrm{u}(0,1)$ for $I = 1$ if $u_{ij} > \phi_{ijc}$ and 0 otherwise.
FCS can be used with arbitrary missing data patterns, with the advantage that it does not require as many iterations as MCMC.

5.3.3.3 Ordinal imputation method

Sometimes it is important to impute incomplete ordinal data by a model that is consistent with the data type. The ordinal imputation model (OIM) serves as a congenial alternative to the ordinal type of data. Here, we briefly describe the OIM algorithm as presented by Donneau et al. (2015).

The OIM is applicable to monotone missingness patterns. For incomplete longitudinal data with these patterns, multiple imputation generally considers completely measured assessment occasions as covariates to sequentially apply approaches designed for univariate data. The OIM strategy employed in SAS PROC MI considers estimating the probability for each category using a cumulative logistic regression model, and then imputes each category for missing values based on the estimated probabilities. The model links the ordinal outcome to a set of q covariates. Considering a longitudinal set-up, these covariates may comprise of the covariates of the substantive model, say $X_{ij}, (i = 1, \ldots, N; j = 1, \ldots, n_i)$, possible auxiliary covariates (A_{ij}) and the vector of previous outcomes $h_{ij} = (Y_{i1}, Y_{i2}, \ldots, Y_{i,j-1})'$. Let $X_i^* = (X_{ij}, A_{ij}, h_{ij})$. We define a proportional odds model:

$$\text{logit}[Pr(Y_{ij} \leq c)|x_{ij}^*] = \lambda_{0c} + x_{ij}'^*\lambda, \tag{5.14}$$

where $\hat{\Lambda} = (\lambda_0', \lambda')'$ are regression coefficient estimates, with $\lambda_0 = (\lambda_{01}, \ldots, \lambda_{0,C-1})$, and the corresponding covariance matrix $V = V(\hat{\Lambda})$. These estimates are obtained by fitting (5.14) to the observed data. Starting from these estimates, the OIM algorithm proceeds (in summary) as follows:

1. Draw new values for Λ, say Λ^*, by assuming a large-sample normal approximation, $\text{N}(\hat{\Lambda}, V(\hat{\Lambda}))$ of its posterior distribution from a noninformative prior $Pr(\Lambda) \propto k, k$ is a constant. In essence,

$$\Lambda^* = \hat{\Lambda} + \nabla' Z,$$

   where ∇' is the upper triangle matrix of the Cholesky decomposition, where $V = \nabla'\nabla$ and Z is a $(C-1)+q$ vector of independent random normal variates.

2. For an observation with missing values Y_{ij}^m and corresponding covariates X_{ij}^*, by using (5.14) calculate the expected probabilities, $P_c = Pr[Y_{ij} = c|x_{ij}^*], c = 1, \ldots C$.

3. Then, for an observation with missing values Y_{ij}^m, draw a random variate from a multinomial distribution with vector of probabilities $(P_1, \ldots, P_C)$ derived in step 2.

4. Steps 1 to 3 are repeated M times to obtain a sequence of imputed values $(Y_{ij}^{(1)}, Y_{ij}^{(2)}, \ldots, Y_{ij}^{(M)})$, for $(i = 1, \ldots, N; j = 1, \ldots, n_i)$.

5.4 Brief description of the data and simulation studies

5.4.1 The data: Recovery from Severe Childhood Malnutrition (RSCM)

This paper uses data from a clinical trial on nutritional status during recovery from severe malnutrition in children. This longitudinal study was conducted by the KEMRI/Wellcome Trust Research Programme, Kilifi, Kenya. The data were collected for 1778 children in total aged 2 to 59 months in 4 different hospitals in Kenya. All were recruited in hospital and had been admitted with severe, acute malnutrition. The children were enrolled shortly prior to discharge and followed up for one year. Follow up was for 10 scheduled occasions; at $0, 1, 2, 3, 4, 5, 6, 8, 10$ and 12 months. Children who died, withdrew before the end of the study, or for other reasons (e.g., deformity), full or complete sequence measurements were not possible (meaning one or more variables will always be missing) were excluded from this analysis, leaving 1138 children who satisfied the inclusion criteria for this analysis.
Overall, about 60% of participants had 100% complete data and 40% had one or more anthropometric data points missing. The proportion of missing data amongst anthropometric variables was 9.8%. The missing values were due to follow up visits being missed completely, or anthropometric measurement data were incomplete if conducted at a home visit without all the anthropometry equipment. In this instance the missingness may depend on unobserved responses of interest and thus assumed to be nonrandom. However, extreme care has to be taken when interpreting evidence for or against MNAR using only the data under study. Trial details may be accessed at Berkley et al. (2016).

5.4.2 Preliminary simulation study I

For the applications presented in this section, we extracted the RSCM participants with no missing data and hence had 729 subjects with complete information from the whole dataset (of 1138 children). The variables' names and descriptions are as follows: sex: sex of the subject (Female or Male); age: this is the age in months calculated from date of enrolment and date of birth; site: the 4 hospitals where the trial was conducted in Kilifi, Malindi, Mbagathi (Nairobi) and Mombasa. In the analysis, the variable site is dichotomized such that Mombasa and Mbagathi are put together as urban, while Kilifi and Malindi are grouped as rural. muac: mid-upper arm circumference in centimetres; zhc: head circumference; zwei: weight for age; zlen: length for age; zwfl: weight for

length.
The anthropometric variables are continuous, as Z scores calculated from the World Health Organization (WHO) reference population (2006) (WHO and Unicef, 2009), except muac which are raw values. zhc, zwei, zlen, and zwfl are the Z scores.
In this study we use muac as the response variable. We use it because it is the best predictor of mortality as it combines the other anthropometric measures and age. Alternatively, one may opt to consider all the measures of malnutrition in a multivariate analysis model, hence, investigate their dependence noting that they share the same covariates. Here, one has to consider correlations among the response variables and between subjects as well.

First, we carry out exploratory analyses on the data based on the continuous outcome muac. The exploratory analysis shows that muac in its original continuous form is normally distributed. Figure 5.1 displays the QQ-plot for the continuous outcome variable muac.

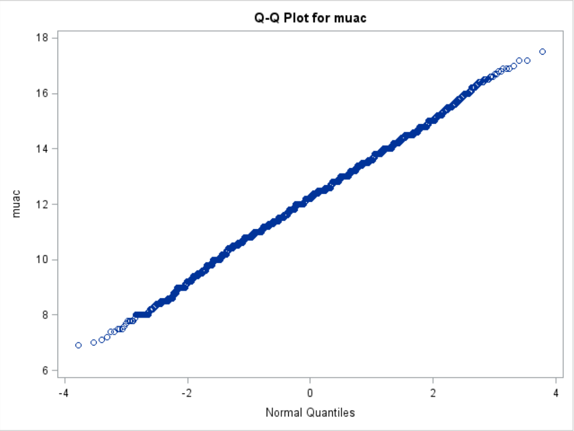


Figure 5.1: A QQ plot for the continuous outcome variable muac

On further exploration, we noted that there is evidence of variability within and between subjects. This is supported by the spaghetti plots or the subject specific profile plots over time. There was a shift, as expected from a lower status to a better status of malnutrition during recovery and the muac increases over follow up time implying treatment was generally effective. This is evident from Figure 5.2.

Next we carried out a repeated measures likelihood analysis that employs a linear model combined with a variance-covariance model that incorporates correlations for all the observations arising from the same subject. In essence we assume that the observations are ordered similarly for each subject, meaning all subjects were measured at the same intervals. Based on AIC values, a model with an unstructured correlation matrix provided the better fit for the data. The data is assumed to be Gaussian and therefore

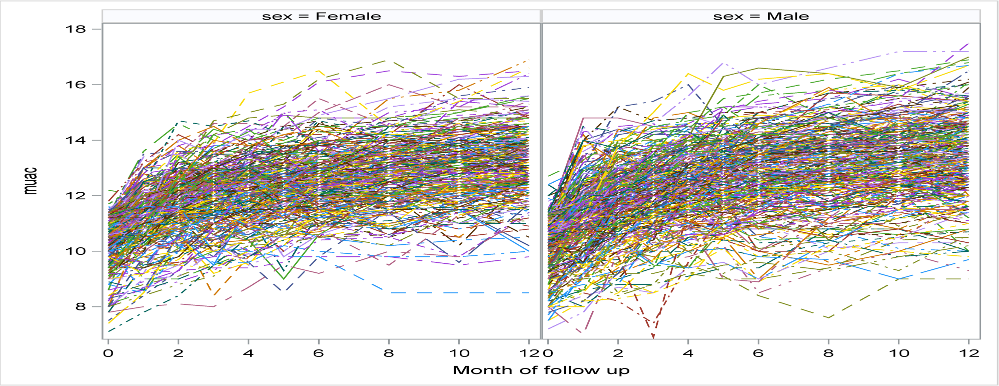


Figure 5.2: "Spaghetti plots" of response curves for subjects

the likelihood maximized to estimate the model parameters. Table 5.1 presents the parameter estimates.

Table 5.1: Parameter estimates, standard errors (SE) and P-values using the continuous outcome for malnutrition.

Effect	Estimate	SE	Pr> $\|t\|$
Intercept	10.2473	0.0807	<.001
sex(female)	0.1188	0.1237	0.337
site: rural	-0.1359	0.0605	0.025
age	0.0449	0.0053	<.001
month†	0.1808	0.0036	<.001
sex*age (female)	-0.0143	0.0092	0.120

†Month of follow up.

Larger values of muac imply better status of nutrition. From Table 5.1 it can be seen that a unit increase in the age of a child increases muac by 0.0448 units. It is also clear that the recovery rate is different for rural and urban children. The mean muac for rural children is 0.1359 lower compared to that of their urban counterparts. The results also show that there is no significant difference in mean muac for males and females. There is also a significant time effect showing that for a 1 month increase, muac increases by 0.1808 units.

5.4.2.1 The ordinal outcome and methods

The primary aim of this paper is to analyse an ordinal outcome from a longitudinal (follow up) study and ultimately compare and contrast the previously discussed methods for incomplete ordinal data. In particular, we fitted an ordinal logistic regression model. But note that, initially, the outcome variable was continuous. We categorized the malnutrition variable, according to WHO guidelines as follows: Children whose muac is less than 11.5 cm are categorized in the severe level; those with muac greater than or equal to 11.5 but less than 12.5 cm fall in moderate; those with muac greater than or equal to 12.5 but less than 13.5 cm are at risk; and finally muac 13.5 cm or more are categorized in the high level, well nourished.

After categorization, the distribution of the outcomes is now negatively skewed thus violating the normality assumption. Table 5.2 displays the distribution of the malnutrition status among the children per site.

Table 5.2: Distribution of the malnutrition status by sex and site.

		Level of outcome*			
Sex	Site	1	2	3	4
Female	Rural	235	255	182	88
	Urban	906	737	771	516
Male	Rural	392	325	286	277
	Urban	846	656	741	707

Note: * 1 = severe; 2 = moderate; 3 = at risk; and 4 = well nourished.

The first step to comparing the methods was deciding the model to use for the data. The 4−level malnutrition measure was modelled using four predictor variables: sex of subjects, site of follow up, age of the subjects and the follow up period. We also include an interaction between the age and sex. Essentially, based on the cumulative logit model (5.3) in Section 5.2, the working model is explicitly written as:

$$\begin{aligned}\text{logit}[Pr(\text{malnut} \leq c)] &= \alpha_c + \beta_1\text{sex} + \beta_2\text{site} + \beta_3\text{age} + \beta_4\text{month} \\ &\quad + \beta_5\text{sex*age}, \quad c = 1, 2, \ldots C - 1, \end{aligned} \tag{5.15}$$

We fitted a standard cumulative odds model in SAS version 9.3. It should be noted here that for the multinomial distribution, when one wants to fit a GEE e.g., in SAS PROC GENMOD, a restrictive fact is that only an independence working assumption is allowed. But, this is not a disturbing hindrance since valid parameter estimates and empirically corrected standard errors can be obtained regardless of the working structure used. However, this is true when analyzing full data (and possibly complete imputed datasets). When the data are incomplete and MAR, consistency and robustness to the choice of correlation are usually lost. But in this paper, we present maximum likelihood estimates.

After the model was fitted, the parameter estimates were recorded as the “true” parameter values. Afterwards, a percentage of responses were dropped randomly at the following approximate rates: 15%, 28% and 38%. First we performed a complete case analysis. Next DL and MI were performed. After imputing, the same model previously used was fitted. The new parameter estimates from DL and MI were recorded and compared to the “true” parameter values. Following, we discuss the dropout and models involved in MI.

5.4.2.2 The dropout

It is logical to assume that the probability of a subject i in a clinical trial to drop out at measurement occasion j depends on the history h_{ij}, $h_{ij} = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,j-1})'$, implying that the MAR assumption holds. Therefore, we constructed a dropout mechanism such that the probability of dropping out at any given time occasion j was a function of the malnutrition response recorded the previous time occasion.In particular, the probability that a subject dropped in the time occasion j given that he or she responded in time occasion $(j-1)$ was determined by

$$P_j = \frac{1}{1 + \exp(-13.5 - \text{muac}_{j-1})}, \tag{5.16}$$

where muac_{j-1} is the observed response at occasion $j-1$. This strategy means that those who recorded an improvement in the malnutrition status were likely to drop out. This dropout mechanism satisfies the missing at random assumption. The number of cases still present in each of the 10 time points of follow up simulated at three different dropout rates are presented in Table 5.3.

Table 5.3: Number of cases still present under different simulated dropout rates using a missing at random strategy.

Freq Miss* = 1174 (15%)		Freq Miss = 2182 (28%)		Freq Miss = 3945 (38%)	
Month	N	Month	N	Month	N
0	792	0	792	0	792
1	792	1	792	1	792
2	792	2	792	2	792
3	792	3	792	3	792
4	792	4	792	4	460
5	792	5	792	5	364
6	792	6	349	6	298
8	792	8	277	8	246
10	237	10	202	10	190
12	173	12	158	12	149

Note: Month refers to month of follow up.
*Freq Miss refers to the frequency of missing values by the end of the study.

5.4.2.3 The imputation and analysis models

To perform MI, one must choose an imputation model for the imputation stage implemented in PROC MI. Then the imputed data sets are subjected to a common analysis model at the analysis stage. However, for the results to be correct, the imputation and the analysis model should be congenial. For congeniality or coherence, it means that the imputation model must contain at least all the variables that are intended to be included in the analysis model. This may include all transformations and possible interactions to variables that are needed in the intended tests. Alternatively, a bigger model can be chosen for the imputation than the analysis model. This may be achieved by including auxiliary variables, that we feel may predict the missingness or are related to the missing variable(s). The auxiliary variables are not of interest in the analysis model but are included in the imputation model to increase the estimation power as well as try to make the MAR assumption more plausible. We estimated the bivariate correlations among the covariates and to the outcome variable to be imputed. In our setting, variables zwfl and zwei are included in the imputation model but are removed from the analysis model. A recommended auxiliary variable is when the coefficient of correlation, $r > .4$. However, this is still an area of active research currently. Allison (2012) believes that including these types of terms introduces unnecessary error into the imputation model. On the other hand, other researchers do not see any harm on the practice, e.g., Enders (2010). Therefore, researcher discretion is advised. A good auxiliary variable can have missing information or not and be just as effective in reducing bias (Enders, 2010). At times, values are missing on a covariate. In such a situation the dependent variable is also used in the imputation model. If it is ignored from the imputation model, there is a possibility of reducing the strength of the correlation between the predictors and the dependent variable, meaning the imputed values and the observed values will not have the same correlation towards the dependent variable.

On the other hand, DL does not create any possibility of a conflict between the imputation and analysis model. Everything is done under the same model. Every variable in the analysis model will be used in dealing with the missing data. In case of interactions or nonlinearities, they will automatically be integrated into the method for handling the missing data (Allison, 2012).

5.4.2.4 Results

In this section we present the analysis results of the ordinal outcome datasets. First we present the results for our reference dataset (herein referred to as full dataset). This is the dataset before introducing dropouts. Table 5.4 gives the standard cumulative logit regression results.

Probabilities are cumulated over the lower ordered values. From Table 5.4, we notice that there is evidence that age at enrolment affects malnutrition status differently for males and females. This is depicted by the significant (at $\alpha = 0.05$) p-value $= 0.0105$, for

Table 5.4: Maximum likelihood parameter estimates (Estimates, Std Error, 95% Confidence Limits and P-Values) for the reference dataset

Par		Est	Std Err	Wald 95% C. L.		Pr >ChiSq.
Intercept1		1.2299	0.0553	1.1214	1.3383	<.001
Intercept2		2.6290	0.0613	2.5090	2.7491	<.001
Intercept3		4.1761	0.0714	4.0362	4.3160	<.001
sex	Female	0.0651	0.0419	-0.0171	0.1472	0.1205
site	rural	0.1469	0.0246	0.0987	0.1952	<.001
age		-0.0548	0.0031	-0.0609	-0.0487	<.001
month[†]		-0.3233	0.0066	-0.3363	-0.3104	<.001
age*sex	Female	0.0079	0.0031	0.0018	0.0139	0.0105

[†] month of follow up.

the age-by-sex interaction. We further infer that the decrease with age in the estimated cumulative odds of malnutrition status below any level c of the ordinal outcome is stronger for males than for females i.e., for male children, the estimated cumulative odds decrease by a factor of $\exp(-0.0548) = 0.9447$ for every unit increase in age, compared to a decrease of $\exp(-0.0548 + 0.0079) = 0.9542$ for the female children. In particular, with a significant coefficient for the interaction term (estimate = 0.0076, SE= 0.0031) and likelihood ratio test statistic = 6.55 (not shown in the table), 1 df : p-value = 0.0105, the decrease in estimated odds can be regarded as different between female and male children. Generally, we notice from the results that younger and female children are identified as more malnourished than male children. Similar results were reported in Berkley et a. (2005) where the authors noted that muac as a measure of malnutrition tends to identify younger and female subjects malnourished more frequently than with Z score approaches. Further, in line with what was observed in Table 5.1, the nutritional status of rural children is lower compared to urban children. In particular, the cumulative odds of severe malnutrition for rural children is $\exp(0.1469) = 1.1582$ times that of their urban counterparts.

Next, we present results obtained after applying complete case analysis (CCA), direct maximum likelihood (DL), full conditional specification (FCS), multivariate normal imputation (MNI) and ordinal multiple imputation (OIM). We provide three different tables; for $15\%, 28\%$ and 38% dropout rates respectively. Maximum likelihood parameter estimates were obtained. In this paper, to implement the MNI, expectation-maximization algorithm for maximum likelihood estimates was used. Notice that, for the current PROC MI in SAS, maximum likelihood estimation in MNI and OIM is not possible for categorical data. Dummy variables must be created to include such variables in the estimation. Under the MI strategies, 20 imputations were conducted with default SAS specifications (for number of iterations) for the various methods.
For comparison, we define our own measure, $RAD(\hat{\beta}) = |\hat{\beta}_F - \hat{\beta}_M|/S.E.(\hat{\beta}_F)$. The

measure $RAD(\hat{\beta})$ is an absolute difference between $\hat{\beta}_F$ and $\hat{\beta}_M$ divided by the standard error of $\hat{\beta}_F$. Here $\hat{\beta}_F$ is an estimate from the created full dataset and $\hat{\beta}_M$ is the estimate from the other model in the presence of missing data, namely CCA, DL, FCS, MNI and OIM. The smaller the value of $AD(\hat{\beta})$, the better the method. Smallest values are in bold in all three tables. Therefore, CCA, DL and MI approaches were considered to perform effectively if obtained parameter estimates are close or similar to those of full dataset analysis. The tables are presented as: Table 5.5 for 15% dropout, Table 5.6 for 28% dropout and finally Table 5.7 for 38% dropout.

Table 5.5: $RAD(\hat{\beta})$ measures for CCA, DL, FCS, MNI, and OIM. Dropout rate: 15%.

Parameter		CCA	DL	FCS	MNI	OIM
Intercept1		27.7396	1.8535	**0.1790**	4.5805	2.5009
Intercept2		52.3230	1.8401	**0.3638**	3.8760	2.0767
Intercept3		76.1513	2.9104	**1.0070**	3.0896	3.1653
sex	Female	16.7327	1.0286	**0.1575**	0.2005	0.2554
site	rural	12.9228	0.9431	**0.2358**	4.8171	5.2520
age		5.6129	0.3871	0.3548	0.2903	**0.2581**
month†		4.6970	6.6212	**0.5455**	5.0152	0.8485
age*sex	Female	6.7097	1.1290	0.5161	**0.3226**	0.5484

† month of follow up

Examining Table 5.5, we notice that smallest values were produced by FCS in all variables except age and age by sex interaction. OIM produced the smallest value for age while MNI had the smallest value for the interaction effect. Throughout, we find that CCA produced largest values as expected.

Table 5.6: $RAD(\hat{\beta})$ measures for CCA, DL, FCS, MNI, and OIM. Dropout rate: 28%.

Parameter		CCA	DL	FCS	MNI	OIM
Intercept1		31.1863	4.6166	**0.5118**	5.6944	3.2315
Intercept2		60.1615	2.9625	**0.6558**	3.1746	1.8956
Intercept3		85.5588	2.6401	0.8473	**0.4678**	3.1106
sex	Female	17.1146	1.9570	0.1193	**0.1050**	0.1623
site	rural	14.3699	1.9146	**0.1098**	4.8130	5.8130
age		2.8387	0.1290	0.2258	0.6452	**0.0323**
month†		**0.6970**	19.9848	4.0606	0.8636	3.8788
age*sex	Female	7.5484	1.1613	0.2581	**0.1613**	0.2258

† month of follow up.

From 28% dropout, Table 5.6, FCS is now challenged by MNI. FCS produced the smallest values for Intercept1, Intercept2 and site. MNI had the smallest value for Intercept3, sex and age by sex interaction. Like in Table 5.5 OIM produced the smallest value for age. Unusually, CCA produced the smallest value for month. But even with this, in all the other remaining parameters, CCA recorded the largest values.

For 38% dropout rate, Table 5.7, the challenge is between FCS and OIM. Each of the two produced smallest values for 3 parameters. MNI had the smallest value for sex. Like in Table 5.6, CCA here also produced smallest value for month, but as in the other dropout rates, it had largest values for all other remaining variables.

Table 5.7: $RAD(\hat{\beta})$ measures for CCA, DL, FCS, MNI, and OIM. Dropout rate: 38%.

Parameter		CCA	DL	FCS	MNI	OIM
Intercept1		34.5226	2.9277	**0.1917**	5.0452	2.7450
Intercept2		67.7047	2.3328	3.2300	1.0375	**0.5938**
Intercept3		85.9090	6.4580	2.6162	8.0840	**0.1485**
sex	Female	17.5107	2.9475	0.2578	**0.0167**	0.0382
site	rural	15.9878	3.1585	**0.3902**	5.5691	6.6060
age		3.4194	1.1935	**0.0000**	0.8065	0.0645
month†		**1.3939**	24.3788	5.0152	9.7879	5.1212
age*sex	Female	9.4839	1.1613	0.4516	0.6129	**0.3226**

† month of follow up.

Generally, the largest values were recorded for CCA for all variables except for month where DL gave the largest value. This was consistent for all three dropout rates. Although DL did not record the smallest value for all dropout rates, but it is ranked second from smallest for site in all the droupout rates. It also gave second smallest values for Intercept1, Intercept2 and Intercept3 under 15% and Intercept3 for 28% dropout rates. Overall, FCS produced the highest number of smallest values under the different dropout rates. It is also clear that CCA is a poor method in all the dropout rates. But, to get a precise picture of how the methods are ranked in performance, we applied the Mahalanobis distance (MD) statistic, defined such that:

$$MD = (\beta_F - \beta_M)' * VarCov(\beta_F)^{-1} * (\beta_F - \beta_M).$$

VarCov is the variance-covariance matrix of the full dataset parameter estimates. Table 5.8 presents the Mahalanobis distance estimates for each method under the three dropout rates.

Table 5.8: Mahalanobis distance measures for CCA, DL, FCS, MNI, and OIM. Dropout rates: 15%, 28% and 38%.

Drop	Method				
Rate	CCA	DL	FCS	MNI	OIM
15%	31274	67.69523	9.981807	92.48802	131.246
28%	45524.25	1071.064	62.73536	254.8666	92.95468
38%	47409.17	3769.034	289.815	1195.495	154.9534

Examining Table 5.8, we notice that the smallest MD is produced by FCS for 15% and 28% dropout rates, and OIM for 38% while CCA has the largest value throughout. Although we have found our results to be likely favourable, but we used a subset of the data and performed a limited simulation study. This, therefore, does not necessarily reflect an accurate and comprehensive comparison because the conclusion drawn is only based on a single replication. The fact is that virtually all simulation studies borrow their strength and conclusive power from replication, where the same data generating and corresponding analysis mechanism is repeated a number of times. In the next section, replication-based simulations are conducted.

5.4.3 Simulation study II

After seeing the performance of the methods on the limited simulation study in Section 5.4.2 (only one dataset), we repeated the comparisons but now utilizing the power of replication (and law of large numbers). We generated S=500 samples, each of size N=1000 subjects. An ordinal regression model was motivated by assuming an underlying latent variable, say $y*$, which is related to the actual ordinal response through the threshold concept. The response was therefore based on some underlying continuous endpoint that follows a linear regression model incorporating random effects and a specified set of threshold values α_c. This data generation mimics the 60% full data subset of the RSCM dataset (as in Subsection 5.4.2) and so the choices made here are in line with the study protocol. We assumed a vector of predictor variables $x' = (x_1, x_2, x_3, x_4)$ which is a combination of both binary and continuous variables. Here, x_1 and x_2 represent binary group effects sex and site respectively while x_3 and x_4 are continuous variables representing age and a 10-point observation time, respectively. We used as starting values for this simulation study, the direct maximum likelihood analysis results for the 60% RSCM dataset (see Table 5.4), such that $\beta_1 = 0.0651, \beta_2 = 0.1469, \beta_3 = -0.0548, \beta_4 = -0.3233,$ and $\beta_5 = 0.0079$). We therefore defined the simulation model explicitly as

$$\begin{aligned}\text{logit}[P(y^*_{ij} \leq c|x)] = \alpha_c + 0.0651x_1 + 0.1469x_2 - 0.0548x_3 - 0.3233x_4 \\ + 0.0079x_1x_3 + b_i, \quad b_i \sim N(0, 0.7564^2). \qquad (5.17)\end{aligned}$$

By letting $\phi_{ij} = P(y^*_{ij} \leq c|x)$, we obtained the corresponding ordinal response Y_{ij} (with C = 4 levels) by using an observation rule defined as

$$Y_{ij} = \begin{cases} 1 & \text{if} \quad 0 \leq \phi_{ij} < \tau_1, \\ 2 & \text{if} \quad \tau_1 \leq \phi_{ij} < \tau_2, \\ 3 & \text{if} \quad \tau_2 \leq \phi_{ij} < \tau_3, \\ 4 & \text{if} \quad \tau_3 \leq \phi_{ij} \leq 1. \end{cases} \tag{5.18}$$

From the full datasets, before imposing any dropouts, parameters and standard errors were estimated using a likelihood-based analysis. Each estimate is an average of 500 estimates from the different simulated datasets. We use these parameter estimates as our "true values". Then, assuming a dropout mechanism similar to (5.16), we generated dropouts at approximate rates of 21% and 40%. Dropout is only on the outcome variable. To ensure ignorability, we did not allow dropout on y to depend on y itself. These incomplete datasets were subjected to the methods under investigation, with specification for each method as used in previous sections. In this study we used 20 imputations. To compare the performances of the methods, relative biases and mean squared errors (MSE) are presented in Table 5.9. For convenience we present only estimates for regression parameters, intercepts are not included. Smallest relative bias and MSE values are presented in boldface.

Considering 21% dropout rate, most smallest relative bias values are produced by OIM. We notice that in most cases the values are equal for two or more methods, or very close. Strangely, for 40% dropout rate, CCA produces 3 smallest relative bias values. This may be attributed to the amount of data that was available for analysis. Even after deleting the incomplete cases, there was enough data to reproduce the full data results. This should not be regarded as a proof for validity of CCA in this case. It has previously been found to perform poorly in the RSCM subset, and the weakest method for 21% dropout above.

Regarding MSE, DL, FCS and OIM seem to perform close to each other. In many cases very close or equal values are obtained between DL and one or more of MI strategies or amongst the MI strategies. Generally, when MSE values are considered, it can be observed that DL and MI strategies are equally viable for the scenarios, at least, as discussed in this study, with only a slight gain of FCS and OIM over MNI.

Overall, the findings in these simulations are in agreement with the RSCM subset where the DL and MI strategies performed closely and preferable while CCA performed poorly. If in some cases CCA was found to perform well, this was perhaps due to chance. The method has severely been discouraged by leading researchers like Rubin (1987); Little

Table 5.9: Relative bias and mean squared error estimates (MSE) from the simulated datasets; for S=500 samples and N=1000 subjects. Dropout rates: 21% and 40%.

Drop	Param	Relative bias					MSE				
rate		CCA	DL	FCS	MNI	OIM	CCA	DL	FCS	MNI	OIM
21%	β_1	-0.8855	-0.8041	-2.6118	-2.3029	**-0.8028**	0.0049	**0.0040**	0.0422	0.0328	**0.0040**
	β_2	0.6052	0.5703	0.1404	0.2033	**0.1399**	0.0131	0.0116	**0.0007**	0.0015	**0.0007**
	β_3	**0.9649**	0.9967	0.9951	0.9967	0.9967	**0.0033**	0.0036	0.0035	0.0036	0.0036
	β_4	-0.1519	**0.1358**	0.1360	0.4253	**0.1358**	0.0039	**0.0031**	**0.0031**	0.0305	**0.0031**
	β_5	1.0198	**1.0000**	1.0099	**1.0000**	**1.0000**	0.0001	0.0001	0.0001	0.0001	0.0001
40%	β_1	**-0.4758**	-0.8333	-0.8282	-2.2153	-0.8321	**0.0016**	0.0043	0.0042	0.0303	0.0043
	β_2	0.5761	0.5698	0.5703	0.2346	**0.1399**	0.0119	0.0116	0.0116	0.0020	**0.0007**
	β_3	**0.9515**	0.9983	0.9967	0.9983	0.9967	**0.0032**	0.0036	0.0036	0.0036	0.0036
	β_4	-0.2920	**0.1392**	0.1394	0.4740	**0.1392**	0.0144	**0.0033**	**0.0033**	0.0379	**0.0033**
	β_5	**0.9109**	1.0099	1.0099	1.0000	1.0099	0.0001	0.0001	0.0001	0.0001	0.0001

and Rubin (2014); Molenberghs and Verbeke (2005) among others, and we do not intend to uphold it because of perhaps a few fair cases in this study.

5.5 Application: The original dataset

After examining the performance of our methods in the simulation studies, we now use the same methods on the whole, original RSCM dataset. The dataset has 8% missing values on the ordinal dependent variable (malnut). Note that, malnut is the ordinal form of the original muac. The distribution of the missing data is arbitrary, i.e., a mixture of monotone and non-monotone patterns. When the pattern of missingness is arbitrary, one can create MI datasets by imputation via the multivariate normal model (MNI) or using the chained equations approach (FCS). DL can also be used to handle arbitrary missingness. To use OIM, the missing values have to be filled in sequentially, by first making the pattern of missingness monotone, then proceed to implement the OIM. Notice that methods capable of handling arbitrary pattern of missingness can also handle monotone missingness. However, the opposite is not true. There are methods whose strength lies in handling monotone patterns and not arbitrary patterns. We present in Table 5.10 the results of CCA, DL and MI strategies on the original childhood malnutrition dataset.

Table 5.10: Parameter estimates, Standard errors and P-values from the original dataset, 100% Recovery from severe childhood malnutrition (RSCM) dataset: Arbitrary missingness rate: 8%. Standard errors are presented in parenthesis.

	Parameter							
Method	Intercept1	Intercept2	Intercept3	sex(F)	site(R)	age	month†	age*sex(F)
CCA	1.5117 (0.0943) <.0001	3.0066 (0.1064) <.0001	4.5671 (0.1252) <.0001	-0.0420 (0.0707) 0.5523	0.2766 (0.0442) <.0001	-0.0731 (0.0051) <.0001	-0.3408 (0.0115) <.0001	0.0133 (0.0051) 0.0086
DL	1.2837 (0.0489) <.0001	2.7149 (0.0543) <.0001	4.2745 (0.0634) <.0001	0.0700 (0.0369) 0.0581	0.1730 (0.0225) <.0001	-0.0547 (0.0028) <.0001	-0.3316 (0.0058) <.0001	0.0045 (0.0028) 0.0996
FCS	1.2959 (0.0498) <.0001	2.7150 (0.0551) <.0001	4.2605 (0.0648) <.0001	0.0715 (0.0372) 0.0548	0.1740 (0.0230) <.0001	-0.0550 (0.0028) <.0001	-0.3303 (0.0061) <.0001	0.0044 (0.0028) 0.1122
MNI	1.0624 (0.0462) <.0001	2.5280 (0.0521) <.0001	4.1284 (0.0608) <.0001	0.0718 (0.0364) 0.0483	0.3352 (0.0449) <.0001	-0.0540 (0.0028) <.0001	-0.3270 (0.0057) <.0001	0.0043 (0.0027) 0.1118
OIM	1.1086 (0.0470) <.0001	2.5410 (0.0521) <.0001	4.0990 (0.0620) <.0001	0.0703 (0.0381) 0.0654	0.3421 (0.0453) <.0001	-0.0547 (0.0028) <.0001	-0.3305 (0.0061) <.0001	0.0045 (0.0028) 0.1086

† month of follow up, F = female R = rural

Investigating the results in Table 5.10, we notice that the DL and MI methods are performing very close to each other. Generally, considering the P-values, we observe that the results are consistent across the DL and MI methods in terms of the final conclusions and inference.

Examining standard errors, we notice that all methods produce the same or very similar values, with the exception of CCA (whose errors are generally larger than the other methods). In fact, in most cases the values (for DL and MI methods) were equal to three decimal places. From this table of results we conclude that the DL and MI methods perform more or less equally. Overall, we realize here that these results are in line with the general inference in the simulation studies. We also noticed that the results in Table 5.10 are consistent with those from the extracted full dataset (60% study completers of RSCM dataset) before we introduced dropouts in Section 5.4.2.2, see results in Table 5.4. But we somehow expected this consistency because as it has been reported by Kadengye et al. (2012); Molenberghs and Verbeke (2005), DL can produce unbiased estimates that are comparable to those of the full data analysis , and that MI and DL can produce similar results when data are missing on the outcome and the same information is used for both models (Collins, Schafer and Karim, 2001).

5.6 Conclusion

In longitudinal studies, outcomes are often measured on an ordinal scale. Analysing such data as equally spaced points on a continuum (as the case of ordinary least squares regression) may lead to erroneous inferences. One may be tempted to dichotomize the ordinal outcome and run a binary logistic regression, but much information is lost and the statistical power jeopardized. Furthermore, the resulting odds ratios may depend on arbitrarily chosen cut-off points used to dichotomize the ordinal outcome. The ordinal logistic regression is an appropriate and user-oriented model for ordered categorical outcomes. However, it was not the purpose of this paper to spell out the working and implementation of the proportional odds model. Neither, was it the purpose to compare it to the ordinary least squares regression or with alternative ordinal logistic regression models such as the continuation-ratio model, or the adjacent categories model. These other models may be more appropriate than the proportional odds model in certain situations. In this paper, we were concerned with demonstrating different types of methods applicable when the ordinal outcome is incomplete. We compared the performance of CCA, and DL analyses and different forms of MI; namely FCS, OIM and MNI.

Generally, we favour the methods that use full information by multiply imputing the missing observations or just use the available information by direct maximum likelihood. This is in contrast with the method that deletes incomplete cases to force a complete dataset. From our findings, we conclude that multiple imputation methods slightly outperformed direct maximum likelihood (DL) estimation for the $15\%, 28\%$ and 38% dropout rate that we imposed on a portion of the dataset. This should not be regarded as unusual. Although DL is a powerful and easy to implement method that takes advantage of the more relaxed MAR assumption, there are cases where MI would be preferred instead of DL. An obvious example where missingness is in both covariates and outcomes. In the medical field where this data comes from, every observation is important and removing individuals with incomplete information from the study is potentially misleading. The analyst might be removing important contributors of the study. Multiple imputation provides methods that fill the empty spaces in the data while capturing uncertainty between and within the number of imputations. It takes care of the probability of filling in with the correct information by observing the behaviour of the measured individuals. Specifically, as demonstrated in this paper (for a subset of the RSCM data), FCS and OIM should be given upper hand when dealing with incomplete ordinal outcomes. This should in a way be expected because the two methods are sensitive to the distribution of the outcomes. However, when we took the whole dataset, we had 8% non-monotone missingness patterns and found that DL, FCS, MNI and OIM perform very similar. This similarity was also observed in the replicated simulation study, where the missingness was monotone at 21% and 40% on the response variable. As a result, it becomes very difficult to decide which one is better than the other for the variety of missing data rates and patterns.

In this study, we assumed and worked under the assumption of a MAR mechanism. At the same time, we cannot totally rule out a reflection on MNAR approaches. In realistic settings, the reasons for dropout are diverse and this makes it difficult to entirely justify on priori grounds the assumption of MAR. Nonetheless, both maximum likelihood and multiple imputation can be done when data are MNAR, but to do this, a missingness model must be specified; that is, a model explaining how missingness depends on both observed and unobserved values. This brings about three issues (Allison, 2014), which include: (1) For any dataset, there are an infinite number of MNAR models, (2) Inferences will rely on the selected model, and (3) it is not possible to tell from the data which of the models is better than the other. Because of these issues, it becomes important to perform a sensitivity analysis. In this case, the sensitivity of inferences to departures from the MAR assumption are investigated where the missing values are imputed assuming a feasible MNAR scheme and the results examined (National Research Council, 2010; Rodriguez and Stokes, 2014). It is beneficial that the SAS procedure MI (Version 9.4, SAS/STAT 13.1) has a number of options for carrying out sensitivity analyses based

on multiple imputation; a subset of these make use of pattern-mixture models.
Also, here we assumed missingness only on the outcome. Missing data can also occur on the covariates or both outcome and covariates at the same time. Investigations of such settings are incumbent. However, we note that, although the methods here discussed can be applied to non-monotone missingness, OIM is a monotone (dropout) pattern method. In case of non-monotone missingness, the missing values need to be imputed sequentially, i.e., impute the non-monotone cases to make it monotone and then proceed to carry out the OIM method. In essence, we view this as a double effort and more research on OIM is recommended to avoid the monotization step in its implementation.

Chapter 6

Fitting a transition model to incomplete longitudinal ordinal response data: Application to childhood malnutrition data

Abstract

Ordinal responses are frequently encountered in longitudinal studies, especially in clinical trials. In an ordinal response, the levels may represent stages (or states) of a disease. The transitions from one state to another are often important. A number of methods exist that deal with the ordinal responses in longitudinal studies. In the analysis of such data, the dependence between responses coming from the same individual has to be taken into account. Furthermore, in addition to the ordinal nature of the responses, the problem of incomplete data may arise. In this paper, a transitional likelihood missing at random model is built, and we investigate the effects of conditioning on previous responses in addition to estimating the effects of measured covariates. The model is applied to the analysis of incomplete childhood malnutrition data recorded from a longitudinal study carried out by Kemri-Wellcome Trust Research Programme, Kenya. Our analysis found that dependence on past responses had some effect on current response, and that the influence diminished with distance from the current response. In this dataset from a clinical trial, the odds of current severe malnutrition was negatively related to better nutritional status at previous occasions. Urban children showed better improvement than rural children over time. Age was negatively related to severe malnutrition and female children had lower nutritional status than male children. To account for the incomplete

nature of the data, direct likelihood and multiple imputation analyses produced similar results.

6.1 Introduction

In medical research, data may be collected repeatedly over time. These data may be presented to have an intrinsic order, hence ordinal in nature (e.g., disease status: severe, mild, well). For such ordered categorical responses, researchers may be interested in how the phenomenon of interest (the ordinal response) is transiting from one status to another between two time points. Researchers may also consider investigating how the ordinal responses obtained from the same individual are correlated or if there is dependence between the current subject's response and previous responses (and covariates). In such longitudinal studies, various approaches may be followed to account for the outcome's dependence on previous outcomes. One way is to implement a marginal model to investigate population averaged state transition over the study period. Here, the marginal mean of the response variable at a given time is modelled directly as a function of the predictor variables, similar to cross-sectional analysis. Generalized estimating equations (GEEs; Liang and Zeger, 1986) are usually used in the context of marginal modelling, where a working correlation is used, together with the sandwich estimator, to capture the dependence. Moreover, when this route is followed, correction is needed for incomplete data, such as weighted GEE and multiple imputation based GEE (MI-GEE). Alternatively, a conditional (random-effects) model may be adopted to deduce the subject's behaviour. Conditional random effects models are used to infer on the presence of variability between subjects. But, if the key interest of the study is to investigate how transitions from one state of the response to another occur over time, then transition models become more appropriate (Ganjali, 2010). When this route is taken, fitting is possible using ordinary (ordinal) logistic regression, therefore no need for GEE. In transition models, the conditional distribution of the response for a subject at a follow-up occasion is a function of the subject's covariates and response history (previous responses) or perhaps a subset of the most recent responses (Diggle et al., 2002). In essence, lagged responses, too, are used as covariates. Since the probability of the response is conditioned on history, the transition model is alternatively referred to as an autoregressive model. The order of a transition model is the length of the history, i.e., the number of past responses upon which the current response is conditioned or perceived to depend on. Special members of this class of models include the Markov models (Feller, 1968).

In longitudinal studies subjects may withdraw before the end of the follow-up period, leading to missing data. This type of missingness is termed dropout (or monotone missingness). Alternatively, some subjects may miss intermittently, i.e., be missing in some occasions and resurface later leading to a non-monotone or intermittent type of missingness. The reasons for missingness are varied, and may be known or unknown to the researcher. Rubin (1987) and Little and Rubin (2014) make important distinctions between the reasons (also referred to as missingness mechanisms/processes): A missingness process is termed missing completely at random (MCAR) if the process is independent of responses observed or unobserved, or any other variables in the analysis. When faced with such a mechanism, any analysis valid for the whole dataset is valid for the observed complete cases. A missingness mechanism is called missing at random (MAR) when the probability of an outcome being missing is independent of any unobserved responses given observed data. Finally, a missingness process that is dependent on unobserved responses and probably on observed information is termed missing not at random (MNAR).

An important issue is that missingness causes the data to be unbalanced. If care is not taken in the way the data are handled, the missing data may lead to biased inferences. For example, in clinical trials, subjects who improve may tend to miss scheduled visits more than those who do not. In such a case, the group with more improved subjects will have more missing data. Analysis of complete data will therefore be biased against the arm with more improved cases. However, there are situations where valid inference can still be obtained, even in the presence of missing data (Tipa et al., 1996). Specifically, this is true in situations where parameter separability is possible. That is, the measurement process parameters, say θ, and the missing data process parameters, say ξ, are distinct in the sense that the joint parameter space, $\Omega(\theta, \xi)$ is such that: $\Omega(\theta, \xi) = \Omega(\theta) \times \Omega(\xi)$. Here, we can make use of likelihood-based inference of the parameters of interest θ, based on the marginal density of the observed data only. From a Bayesian view-point, any joint prior distribution applied to (θ, ξ) can be factored out into independent marginal priors for θ and ξ. In such cases, Rubin (1976) terms the missing data process as ignorable. Essentially, when using likelihood and Bayesian inference, a MAR process leads to ignorability; for frequentist inferences, the stronger MCAR is needed to automatically have ignorability. The consequence of ignorabilty is that the missing data process does not need to be modelled explicitly (Rubin, 1976; Little and Rubin, 2014). In contrast, for non-ignorability, the missing data process cannot be ignored in the analysis, i.e., future unobserved responses cannot be predicted conditional only on past responses; instead, we need to incorporate a model for the missing data process (Nakai and Ke, 2011). It should be emphasized here that, in the application to missing data classifications, ignorability does not imply that the analyst can ignore missing values. It refers to the fact that aspects that cause missingness are unrelated or weakly related with the estimated effects of interest. In a restricted sense, ignorability refers to whether the missing data process must be modelled as part of the analysis process or not (Allison, 2001).

In this paper, we assume ignorable dropout in a cumulative logit transition model. Nonetheless, we note that data in longitudinal studies may be intermittently missing. For such a situation, the estimation assumptions and SAS code adopted for the analysis in this paper can be generalized to cover the intermittent missingness case. Uranga and Molenberghs (2014) describe this generalization and provide a set of SAS macros that can be used to fit a conditional model for Gaussian data. In particular, they use a matrix-oriented programming language, namely the Interactive Matrix Language (IML) in SAS. The macros can be adapted and tailored for non-Gaussian data and other assumptions as the case may demand.

The rest of the paper is structured as follows. In Section 6.2, a motivating example dataset is briefly described and prepared for use. In Section 6.3, the transition model of interest is discussed in the context of cumulative logits for longitudinal data. This model will be applied in Section 6.4, where exploratory analyses on the dataset are conducted. Then the appropriate transition model will be fitted and results thereof discussed. Section 6.5 provides the discussion and conclusion to the paper.

6.2 The data: Recovery from severe childhood malnutrition (RSCM)

The motivating example in this paper involves a longitudinal study to asses the nutritional status of Kenyan children who were recovering from a severe childhood malnutrition. The study was conducted by Wellcome-Trust Research Programme, Kilifi, Kenya. We refer to the data from this study henceforth as the RSCM data. These data were collected for 1778 children aged 2 to 59 months in 4 different hospitals in Kenya. All participants were recruited in hospital where they had been admitted with severe, acute malnutrition. The children were enrolled shortly prior to discharge and followed up for one year. At the initial visit (time point 0), baseline covariate information were recorded - such as *age* at enrolment (calculated from date of birth and date of enrolment), *sex* (Female/Male), *site* (Mombasa, Malindi, Mbagathi and Kilifi). In this analysis, the variable site is dichotomized such that Mombasa and Mbagathi are combined as urban, while Kilifi and Malindi are grouped as rural. Initial malnutrition status was recorded for each individual. Other variables included *muac*: mid-upper arm circumference in centimetres; *zhc*: head circumference; *zwei*: weight for age; *zlen*: length for age; *zwfl*: weight for length. The anthropometric variables *zhc, zwei, zlen*, and *zwfl* are continuous, in terms of Z scores calculated using the World Health Organisation

(WHO) macro for STATA (2006) (WHO and Unicef, 2009) while *muac* are raw values. After enrolment, individuals were followed-up for another nine occasions scheduled for months $1, 2, 3, 4, 5, 6, 8, 10$ and 12. In this paper, the measure of interest is the child's malnutrition status. We use the continuous variable *muac* to create the four levels of the ordinal response or outcome.

The proposed transition model is basically an extension of the classical cumulative logit model (Ananth and Kleinbaum, 1997). These models are generally referred to as latent variable models. This is because, they are usually applied in cases where a set of ordinal variables are used as indicators or representatives of an underlying latent variable, where the latent variable is the main variable of interest. Because the latent variable cannot be measured directly or for a clear clinical interpretation of effects (e.g., a treatment effect), it is appropriate to transform it to an ordinal variable. Therefore, to implement the cumulative logit transition model, *muac* was categorized based on WHO recommended categories of malnutrition as an ordinal scale: severe [1] = 'less than 11.5cm'; moderate [2] = 'more than or equal to 11.5 cm but less than 12.5cm'; at risk [3] = 'more than or equal to 12.5 cm but less than 13.5cm'; and well nourished [4] = 'more than or equal to 13.5cm'.

Children who died, withdrew before the end of the study, or for other reasons (e.g., deformity), full or complete sequence measurements were not possible (meaning one or more variables will always be missing) were excluded from this analysis, leaving 1138 children who satisfied the inclusion criteria for this analysis. About 8% of the data were missing intermittently on the outcome variable. More details of the trial from which we extracted the RSCM dataset may be accessed at Berkley et al. (2016).

6.3 The cumulative logits transition model

For each individual $i = 1, \dots, N$ in a study, we consider a series of measurements $Y_i = (Y_{i1}, \dots, Y_{in_i})'$, along with a matrix of covariates $X_i = (x_{i1}, \dots, x_{in_i})'$ which may include measurement occasions $(t_{i1}, \dots, t_{in_i})$ and other possible predictor information.

In the context of longitudinal ordinal responses, the cumulative logit regression model (McCullagh, 1980) is perhaps the most popular ordinal logistic regression model. Now, suppose the response variable has K ordered categories $(c = 1, 2, \dots, K)$, then the cumulative logit model estimates the effects of explanatory variables on the log odds of selecting lower, rather than higher response categories. Let $\phi_{ijc} = Pr(Y_{ij} \leq c|x_{ij})$ denote the probability of being at or below category c, given a set of predictors. We define the general cumulative logit model as:

$$\log\left(\frac{\phi_{ijc}}{1-\phi_{ijc}}\right) = \alpha_c + x'_{ij}\beta, \qquad c = 1, 2, \ldots K-1, \tag{6.1}$$

where α_c gives the intercept terms that depend on the ordinal categories, c indexes the $K-1$ logits, and x_{ij} is a vector of covariates at occasion $j, (j = 1, 2, \ldots, n_i)$ for the i^{th} individual. The regression coefficients, β, reflect the association between the predictor variables and the response variable. The cumulative odds model assumes the same slope parameters for each of the response logits, i.e., the effects of the different predictor variables is identical across the $K-1$ logits. When this model fits well, it estimates a single vector of parameters rather than $K-1$ different vectors of parameters to describe the effect of the predictor variables. This property is called the proportional odds assumption of model (6.1) or equivalently the equal slopes assumption.

In the context of transition models, a response Y_{ij} in a longitudinal sequence is a function of covariates x_{ij} (if there are any available), and its history, $h_{ij} = (Y_{i1}, \ldots, Y_{i,j-1})'$. Assuming a general transition model for an ordered response variable with K categories over the time points and monotone missing data patterns, we present and discuss two model types namely a purely marginal model and a model that includes the history of the observed outcomes given by:

$$\log\left[\frac{Pr(Y_{i1} \leq c|x_{ij})}{Pr(Y_{i1} > c|x_{ij})}\right] = \alpha_{c1} + x'_{i1}\beta, \qquad c = 1, 2, \ldots K-1, \tag{6.2}$$

$$\log\left[\frac{Pr(Y_{ij} \leq c|h_{ij}, x_{ij})}{Pr(Y_{ij} > c|h_{ij}, x_{ij})}\right] = \alpha_{cj} + \beta x'_{ij} + \lambda' h_{ij}, \qquad c = 1, 2, \ldots K-1, j = 2, \ldots, n_i. \tag{6.3}$$

Here, Y_{i1} represents the response at the initial time point where there is no history, and Y_{ij} is the response for the next $n_i - 1$ follow-up times, $j = 2, \ldots, n_i$. The parameters α_c and β are as earlier on defined and λ is a vector of the autoregressive parameters. Under the transition model, we assume that the cumulative response logit function (6.3) is a linear function of both the history and covariates (h_{ij}, X_{ij}), where the parameters α, β and λ can take any values on the real line $[-\infty, \infty]$. For convenience of notation, we assume that the time points (read measurement occasions) are equally spaced. If they are not, then robust assumptions need to be made about the distributional form of the time dependence. Otherwise, the transition model (6.3) is a well paused one and the effects on Y_{ij}, given h_{ij} extend the class of generalized linear model formulation (Molenberghs and Verbeke, 2005). Relying on the law of total probability and assuming random dropout, the joint probability, $f(y_{i1}, \ldots, y_{in_i})$ of the responses $Y_{ij},\ \ j = 1, \ldots, n_i$, of the i^{th} subject can be expressed as:

$$f(y_{i1},\dots,y_{in_i}) = f(y_{i1})\cdot f(y_{i2}|y_{i1})\cdot f(y_{i3}|y_{i2},y_{i1})\cdot \quad \dots \quad \cdot f(y_{in_i}|y_{i1},\dots,y_{in_i-1}),$$

where $f(\cdot|\cdot)$ denotes the transition probability capturing the conditional dependence of the current outcome on previous ones. The joint distribution can then be equivalently simplified such that

$$f(y_{i1},\dots,y_{in_i}) = f(y_{i1})\cdot\prod_{j=2}^{n_i} f(y_{ij}|h_{ij}), \tag{6.4}$$

where generally, $f(y_{ij}|h_{ij}) = f(y_{ij}|y_{i,j-1}, y_{i,j-2},\dots,y_{i,j-q})$ implies a $q-$order transition model, which is of order one if $q = 1$, in which case the term $f(y_{i1})$ is replaced with $f(y_{i1}, y_{i2},\dots,y_{iq})$ and the product starts from $j = q+1$ to n_i. Such a model implies the observations made after the first q are independent conditionally on the past q observations. Equation (6.4) simply states that the joint likelihood contribution for an individual i is the marginal probability density function (pdf) of the initial outcome Y_{i1} times the product of all subsequent conditional pdfs of an outcome given its history. If the dependence on covariates is also included in the formulation, the joint likelihood for a single subject becomes

$$f(y_{i1},\dots,y_{in_i}|X_i;\theta) = f(y_{i1}|X_i;\theta)\cdot\prod_{j=2}^{n_i} f(y_{ij}|h_{ij},X_i;\theta),$$

implying that the full likelihood function for all subjects will subsequently be:

$$L(\theta;y,X) = \prod_{i=1}^{N}\{f(y_{i1}|X_i;\theta)\prod_{j=2}^{n_i} f(y_{ij}|h_{ij},X_i;\theta)\}.$$

Here, θ denotes the parameters of interest which include the regression coefficients for measured covariates β and λ which capture the outcome history dependence in the transition model. The maximization of the above likelihood can be carried out in any standard statistical software package to obtain the parameters of interest. Note here that, because $y_{i1}, i = 1,\dots,N$ is almost always observed for all subjects, then $f(y_{i1})$ may or may not be of major importance but rather the focus is more on later contributions $f(y_{ij}|h_{ij})$ for $j \geq 2$. Thus, the partial likelihood contribution from individual i given by $\prod_{j=2}^{n_i} f(y_{ij}|h_{ij})$ can suffice or for a general q-order transition model the partial likelihood is $\prod_{j=q+1}^{n_i} f(y_{ij}|h_{ij})$. However, the baseline response can have a significant effect to future outcomes in some situations. We refer interested readers to Agresti (2010); Diggle et al. (2002); Ghahroodi et al. (2009); Lee (1992); McCullagh (1980); Molenbeghs and Verbeke (2005); Noorian and Ganjali (2012) and references therein, as well as Stokes et al. (2012) for a detailed exposition of the idea of transition modelling and other forms of longitudinal ordinal response data models.

6.3.1 Estimation of the cumulative logits transition model

Both marginal and random-effects models can be used to analyze the RSCM dataset. But, these methods would require that assumptions on the covariance structure about the repeated measures on the outcome variable be made. Considering our outcome is ordinal (taken at specific fixed time points) and that the measurement times were not even, specifying a working correlation structure would not be easy. Also, because of the multi-state nature of the outcome variable, realizing that subjects' current state may be influenced by past experience, we opted to analyze the outcome via the transition model reflecting on the influence of explanatory variables (covariates and previous outcomes) on the current outcomes. The cumulative logits transition model is simply an extension of the polytomous (ordinal) logistic regression model. Accordingly, the parameter estimates can be obtained directly by maximizing the ordinal likelihood function, that is, estimating parameters using maximum likelihood by treating past outcomes as additional explanatory variables.

In the case of incomplete outcomes (both monotone and non-monotone), a direct likelihood approach may be taken. Here, the incomplete data are analyzed directly without deletion or need to impute the missed values. We find this rather straightforward, so we relied on this approach for our results. Alternatively, one may opt to multiply impute the incomplete cases then proceed with the fitting of a transition model. We will take this route as well and compare its results with the aforementioned approach. It should be noted here that it is not the aim of this paper to compare the strength of direct likelihood and multiple imputation, but rather to make sure we deal with the incomplete data problem adequately and to counter check our results in the context of transition modelling. These methods are valid under the less stringent missing at random mechanism. We will assume this mechanism and try as much to adhere to the conditions for its validity as discussed by Rubin (1987; 1996). Nonetheless, these conditions are often violated in practice, and most of the time the mechanism is merely assumed to hold but may not. Unfortunately, very little can be done to definitively establish the missing data mechanism's nature since it is not possible to differentiate between data that are missing at random from that which are missing not at random using the observed data only. Therefore, several authors have argued in favor of sensitivity analysis, where the impact of varying missing data assumptions on the target inferences is examined (for example, Carpenter and Kenward, 2013; Carpenter, Kenward and White, 2007; Molenberghs et al., 2003; Molenberghs and Verbeke, 2005; National Research Council, 2010; Rodriguez and Stokes, 2014; Satty and Mwambi, 2013). This is outside the scope of this paper. While models for data that are not missing at random can be formulated and estimated, theses models are normally complex and untestable, and require specialized software and expertise (Allison, 2000). Consequently, any general purpose method will resort to the missing at random mechanism and results would be satisfactory. In our dataset, the percentage of missing values is about 8% thus we expect a minimal impact on parameter estimates from both routes.

6.4 Application to the RSCM data

6.4.1 Exploratory data analysis

First, we use *muac* as the outcome of interest in its original continuous form. A plot of mean profiles per sex grouping over follow up time is presented in Figure 6.1. It is observed that male children started with low *muac* values. But, at the second observational timepoint, the profiles switch and the profile for male children remains above that for female children until the end of follow up. The switch shows that male children responded better to treatment over time than the female children because larger values of *muac* imply better status of nutrition. However, this kind of trend was expected because as reported in Berkley et al. (2005), *muac* as a measure of malnutrition tends to identify female subjects as malnourished more frequently unlike with other measures like the Z score approaches, thus implying that transformation to Z scores helps to have a better informative measure. We however, use *muac* because it is the best predictor of mortality as it correlates better with the the other anthropometric measures and age.

Note that repeated longitudinal measurements from the same subject are correlated. That is, within-subject measurements are not independent. Correlation coefficients and regression models can be used to investigate the relationship among variables that have ordinal, interval or ratio level scales. On examining the correlation coefficients, we noticed a decaying structure with time lag. That is, observations closer in time tended to be more correlated than observations far apart in time. Figures 6.2 (a) and (b) show scatter plot matrices of ordinary least squares (OLS) means. The plots confirm the suggestion of decaying correlations with time lags. This consequently supports the use of an autoregressive structure dependence.

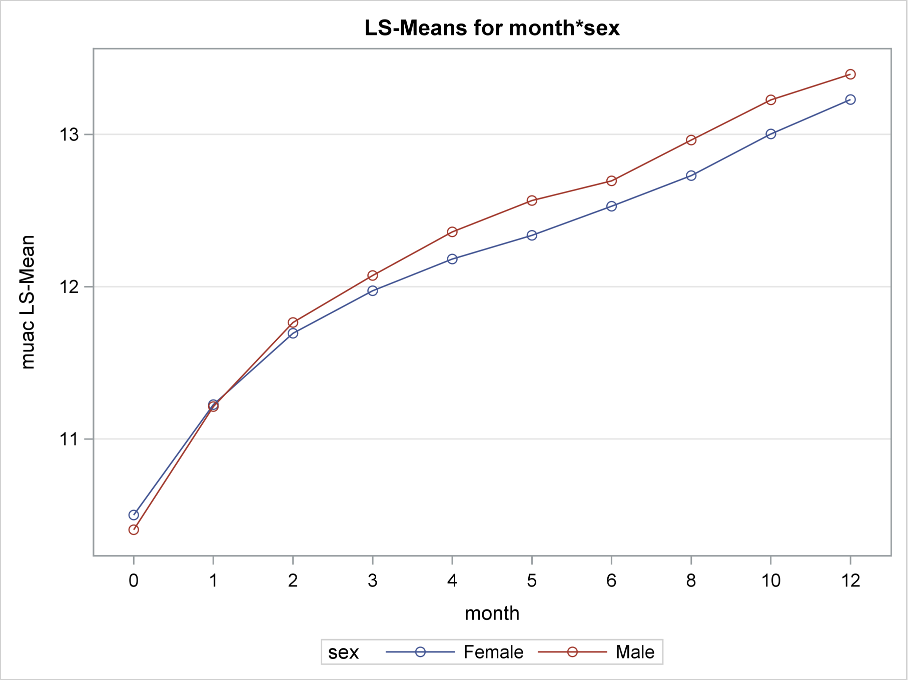
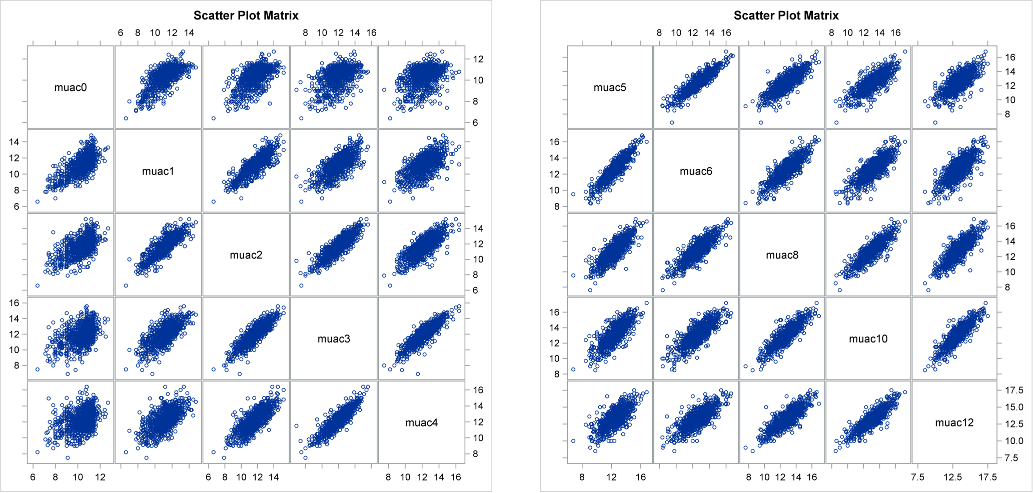


Figure 6.1: Mean muac by sex grouping at every measurement occasion



(a) A scatter plot matrix for the first 5 time points

(b) A scatter plot matrix for the last 5 time points

Figure 6.2: OLS means for the outcome variable *muac*

6.4.2 Model fitting

The exploratory analysis in Section 6.4.1 suggests the use of an autoregressive transition model. To fit such models, successive measurements conditioned on their past outcomes (history) are assumed independent of each other. For this reason standard generalized linear models statistical software can be used (Molenberghs and Verbeke, 2005). In particular, the autoregressive transition model can be easily fit using logistic regression and parameters estimated by maximum likelihood by treating history as additional explanatory variables. In this study, we used the SAS procedure GENMOD and employed the functionality of the %dropout macro to create lags to ensure that the history is captured as part of the explanatory variables in the analysis model. However, we note that the macro here stated is valid under the monotone missing data patterns. To implement it for non-monotone missingness patterns, it is appropriate to make the missingness pattern monotone, and then proceed with the macro (Uranga and Molenberghs, 2014). Our outcome variable had 8% values missing arbitrarily. Using SAS PROC MI, and a specification of the MCMC method, we monotonised the missingness patterns and thus remained with approximately 4% missing cases. But, before exploring the transition model, we first fitted a marginal cumulative logit model (non-transitional model) initially, without an interaction term (here called Model 1- eq (6.5)) then, including a sex-by-age interaction term, Model 2 - eq (6.6):

$$\text{logit}[Pr(Y_{ij} \leq c|x_{ij})] = \alpha_c + \beta_1 sex + \beta_2 site + \beta_3 age + \beta_4 month, \tag{6.5}$$

$$\text{logit}[Pr(Y_{ij} \leq c|x_{ij})] = \alpha_c + \beta_1 sex + \beta_2 site + \beta_3 age + \beta_4 month + \beta_5 sex * age, \tag{6.6}$$

where $\alpha_c, c = 1, \ldots, K - 1$ are as defined in Section 6.3. Table 6.1 presents the results of the two competing models.

Examining the model fit criteria, we realize that the AIC value for Model 1 is 25625.0699 and for Model 2 is 25624.0694. From the AIC values, Model 2 provided a slightly lower value than Model 1 although the interaction term itself is statistically insignificant at 5% level. Also, the introduction of the interaction makes the sex main effect insignificant while it is highly significant in Model 1. Ideally, marginally a model with significant main effects would probably be preferable. But, with our competing models the gain from Model 1 to Model 2 is very minimal. The AIC values show a difference of only 1 unit, indicating that Model 2 is not necessarily better than Model 1. For this reason, Model 1 is used for further analysis in this paper.

Table 6.1: Parameter estimates, standard errors (StdErr) and P-values obtained from fitting Model 1 and Model 2.

	Model 1			Model 2		
Parameter	Estimate	StdErr	Pr> $\|Z\|$	Estimate	StdErr	Pr> $\|Z\|$
Intercept1	1.0945	0.0457	<.0001	1.0912	0.0457	<.0001
Intercept2	2.5175	0.0507	<.0001	2.5144	0.0507	<.0001
Intercept3	4.0806	0.0595	<.0001	4.0779	0.0595	<.0001
sex(Female)	0.1146	0.0182	<.0001	0.0601	0.0363	0.0975
site: rural	0.3702	0.0443	<.0001	0.3667	0.0443	<.0001
age	-0.0552	0.0027	<.0001	-0.0549	0.0028	<.0001
month†	-0.3294	0.0057	<.0001	-0.3294	0.0057	<.0001
sex*age (Female)				0.0047	0.0027	0.0832
AIC*	25625.0699			25624.0694		

† Month of follow up: Indicator variable for the 10 follow-up time points.

* AIC: Akaike Information Criterion for model fit. Model with smaller statistic is preferred.

Probabilities are accumulated over the lower ordered values. From Table 6.1 (and considering Model 1), we observe that females have an approximately 12% (odds ratio = 1.1214) increase in the odds of higher malnutrition compared to males. On the other hand, with each unit increase in age, the odds of higher malnutrition decreases by about 5% (odds ratio = 0.9463), whereas with each month of follow up, the odds of higher malnutrition decreases by 28% (odds ratio = 0.7194). All these changes are significant. Further, we observe that the nutritional status of rural children is lower compared to their urban counterparts. In particular the cumulative odds of higher malnutrition for rural children is 1.4480 times that of the urban children. With these results, it becomes interesting to investigate whether the current malnutrition status of a child depends on the previous statuses or the history of malnutrition.

Now following the findings from the exploratory analysis, we fit a transition model based on the assumptions of Model 1. However, as it was not very clear to determine the order of the transition model, a saturated transition model was first fitted. This model included all the covariates together with nine lagged responses. Lags that appeared to have little or no relationship with the current malnutrition status were dropped at the 5% significance level. Preferably, variables that are related to the dependent variable are of interest, and the size and strength of the correlation are investigated. On investigation, a strong dependence was observed of the current outcome of interest on previous malnutrition values particularly at lags $1, 2$, and 5, i.e., $y_{i,j-1}$, $y_{i,j-2}$, and $y_{i,j-5}$. These will henceforth be referred to as lag 1, lag 2 and lag 5 effects, respectively. When higher lags were tried, like lag 9, it was realized that some parameters could not be estimated. This was probably because of too many parameters to estimate with limited data available. Therefore, the saturated model was reduced to model (6.7) below, with β signifying effects for the predictor variables, and λ for history parameters:

$$\begin{aligned}\text{logit}[Pr(Y_{ij} \leq c|x_{ij}, h_{ij})] = \alpha_c + \beta_1 sex + \beta_2 site + \beta_3 age + \beta_4 month \\ + \lambda_1 y_{i,j-1} + \lambda_2 y_{i,j-2} + \lambda_5 y_{i,j-5},\end{aligned} \tag{6.7}$$

Here, $\lambda = (\lambda_1, \lambda_2, \lambda_5)$, thus indicating a transition model of order 5. Table ?? shows results of fitting model, equation (6.7).

Examining Table ??, it is observed that when lag 1 is introduced the behaviour is as it was observed for Model 1 (model with no lagged responses), i.e., all variables are significant. When lag 2 is introduced, the month which was previously significant now becomes highly insignificant. When dependence is extended to lag 5, month becomes significant again but age and sex, which were initially significant now become insignificant.

Regarding the influence of conditioning on history on the outcome of interest, we can generally infer that individuals are improving with time. The negative estimate of the history parameters justify this. The negative coefficients tell us that the cumulative odds of malnutrition against better nutrition decrease for a child with a healthier history. That is, the better the nutrition history, the lower the odds of current malnutrition. This trend was also depicted in Figure 6.2 which shows positive correlations between time points. Looking at the AIC values we note that with the introduction of lags, the model fit became better. In fact, the model with longer history, namely lag 5, fits the data better than the first two shorter history models.

For the results in Table ??, the missing data in RSCM were properly accommodated by the direct likelihood method. However, we also corrected the missing values by multiple imputation. We used 20 imputations to impute the original continuous variable *muac* by assuming multivariate normal imputation model, before categorizing it to the ordinal outcome. However, we noticed that the two approaches produced similar results. But this was in a way expected, probably because of two reasons. First, generally, multiple imputation and direct likelihood analyses will produce similar results when data are missing on the outcome and the same information is used for both models (Collins, Schafer and Kam, 2011). Secondly, the percentage of missing values was relatively small (8%) hence a minimal impact was expected on the parameter estimates. Results after multiple imputation are displayed in Table ??. The overall inference from Tables ?? and ?? is the same. Both direct likelihood and multiple imputation can be used on equal footing as a way of correcting incompleteness.

Table 6.2: Parameter estimates, standard errors (StdErr) and P-values obtained from fitting the transition model, equation (6.7). Missing values corrected by direct likelihood approach.

Effect	Par	Est (StdErr)	Pr> $\|Z\|$	Est (StdErr)	Pr> $\|Z\|$	Est (StdErr)	Pr> $\|Z\|$	Est (StdErr)	Pr> $\|Z\|$
Intercept1	α_1	1.0945 (0.0457)	<.0001	3.4053 (0.0676)	<.0001	3.6376 (0.0800)	<.0001	4.2306 (0.1420)	<.0001
Intercept2	α_2	2.5175 (0.0507)	<.0001	6.1602 (0.0906)	<.0001	6.5585 (0.1039)	<.0001	7.3058 (0.1689)	<.0001
Intercept3	α_3	4.0806 (0.0595)	<.0001	9.2167 (0.1210)	<.0001	9.8298 (0.1381)	<.0001	10.6948 (0.2091)	<.0001
sex(F)	β_1	0.1146 (0.0182)	<.0001	0.0571 (0.0212)	0.0071	0.0637 (0.0230)	0.0057	0.0516 (0.0306)	0.0917
site: rural	β_2	0.3702 (0.0443)	<.0001	0.2519 (0.0511)	<.0001	0.2715 (0.0554)	<.0001	0.3126 (0.0728)	<.0001
age	β_3	-0.0552 (0.0027)	<.0001	-0.0265 (0.0032)	<.0001	-0.0146 (0.0034)	<.0001	-0.0046 (0.0046)	0.3211
month†	β_4	-0.3294 (0.0057)	<.0001	-0.0285 (0.0072)	<.0001	-0.0057 (0.0082)	0.4874	-0.0513 (0.0137)	0.0002
$y_{i,j-1}$	λ_1			-2.6017 (0.0370)	<.0001	-2.3665 (0.0460)	<.0001	-2.2041 (0.0606)	<.0001
$y_{i,j-2}$	λ_2					-0.5920 (0.0403)	<.0001	-0.8275 (0.0568)	<.0001
$y_{i,j-5}$	λ_5							-0.1893 (0.0489)	<.0001
AIC*		25625.0699		16197.7185		13696.6918		7782.5993	

† Month of follow up. F = female children

Table 6.3: Parameter estimates, standard errors (StdErr) and p-values obtained from fitting the transition model 6.7 after multiply imputing the missing values.

Effect	Par	Est (StdErr)	Pr> $\|Z\|$	Est (StdErr)	Pr> $\|Z\|$	Est (StdErr)	Pr> $\|Z\|$	Est (StdErr)	Pr> $\|Z\|$
Intercept1	α_1	1.0748 (0.0465)	<.0001	3.4246 (0.0690)	<.0001	3.6537 (0.0830)	<.0001	4.2010 (0.1457)	<.0001
Intercept2	α_2	2.4932 (0.0512)	<.0001	6.1921 (0.0922)	<.0001	6.5783 (0.1070)	<.0001	7.2687 (0.1734)	<.0001
Intercept3	α_3	4.0295 (0.0611)	<.0001	9.2306 (0.1234)	<.0001	9.8246 (0.1429)	<.0001	10.6094 (0.2172)	<.0001
sex(F)	β_1	0.0948 (0.0186)	<.0001	0.0466 (0.0211)	0.0273	0.0526 (0.0229)	0.0218	0.0363 (0.0305)	0.2339
site: rural	β_2	0.3316 (0.0454)	<.0001	0.2340 (0.0510)	<.0001	0.2493 (0.0553)	<.0001	0.2797 (0.0724)	<.0001
age	β_3	-0.0548 (0.0028	<.0001	-0.0261 (0.0031)	<.0001	-0.0144 (0.0034)	<.0001	-0.0048 (0.0046)	0.2967
month†	β_4	-0.3209 (0.0057)	<.0001	-0.0268 (0.0072)	0.0002	-0.0044 (0.0082)	0.5915	-0.0452 (0.0137)	0.0010
$y_{i,j-1}$	λ_1			-2.6139 (0.0383)	<.0001	-2.3758 (0.0481)	<.0001	-2.2262 (0.0646)	<.0001
$y_{i,j-2}$	λ_2					-0.5858 (0.0413)	<.0001	-0.8044 (0.0575)	<.0001
$y_{i,j-5}$	λ_5							-0.1742 (0.0492)	0.0004

† Month of follow up. F = female children

After seeing the performance of Model (6.7), it was noted that the evolution over time may be gender dependent. Hence, a sex-by-month interaction was also explored, and was found to be significant. The interaction investigated whether the malnutrition status improved differently over time for male and female children. We corrected the incompleteness by the direct likelihood approach. These results are displayed in Table ??. Examining the table, it is observed that there is a significant difference in malnutrition improvement over time between male and female children. This is evidenced by a significant sex-by-month interaction, where the decrease in cumulative odds of higher malnutrition is stronger for male children compared to female children. That is, for male children, the cumulative odds decreases by a factor of $\exp(-0.3294) = 0.7194$ compared to a decrease of $\exp(0.0115 - 0.3293) = 0.7277$ for female children. But with this interaction term, sex effect is insignificant. When the first lag is introduced, the interaction becomes insignificant, but sex effect becomes significant. When dependence of the current response is extended to lag 2, the interaction becomes significant again, but it makes month insignificant, although it was initially significant. When dependence now goes to lag 5, the interaction remains significant, month becomes significant again but age is rendered insignificant.

Generally, the dependence of the current outcomes was evident with significant effects on lags 1, 2, and 5. The dependence on lags 1 and 2 are intuitively clear but the significance of lag 5 is not immediately clear and may require further research and subject matter input.

6.5 Discussion and conclusion

The exploratory analysis of the RSCM dataset suggested the use of an autoregressive transition model. After investigation, cumulative logit models of order 1, 2 and 5 were fitted using a SAS procedure GENMOD. In this way, the parameters were obtained by maximizing the ordinal likelihood function. To account for the incomplete nature of the outcome variable, direct likelihood and multiple imputation approaches were used. The direct likelihood approach is simple and straightforward for it does not require any extra analyst's adjustments on the missing values, i.e., neither deletion of the incomplete cases nor imputation. Both direct likelihood and multiple imputation are valid for an ignorable, missing at random missing data process. These two methods will in most cases produce similar results provided that data are missing on the outcome and the same information is used for both models. In this paper, similar results were obtained from the two methods.

Table 6.4: Parameter estimates, standard errors (StdErr) and P-values obtained from fitting a model with month-by-sex interaction. Incompleteness handled by the direct likelihood approach.

Effect	Par	Est (StdErr)	Pr> $\lvert Z\rvert$	Est (StdErr)	Pr> $\lvert Z\rvert$	Est (StdErr)	Pr> $\lvert Z\rvert$	Est (StdErr)	Pr> $\lvert Z\rvert$
Intercept1	α_1	1.0933 (0.0457)	<.0001	3.4068 (0.0676)	<.0001	3.6414 (0.0801	<.0001	4.2353 (0.1420)	<.0001
Intercept2	α_2	2.5163 (0.0507)	<.0001	6.1625 (0.0907))	<.0001	6.5645 (0.1040)	<.0001	7.3151 (0.1689)	<.0001
Intercept3	α_3	4.0807 (0.0595)	<.0001	9.2191 (0.1210)	<.0001	9.8377 (0.1382)	<.0001	10.7094 (0.2093)	<.0001
sex(F)	β_1	0.0578 (0.0311)	0.0631	0.0915 (0.0398)	0.0215	0.1569 (0.0488)	0.0013	0.3241 (0.1009)	0.0013
site: rural	β_2	0.3715 (0.0443)	<.0001	0.2514 (0.0511)	<.0001	0.2708 (0.0554)	<.0001	0.3125 (0.0729)	<.0001
age	β_3	-0.0552 (0.0027)	<.0001	-0.0265 (0.0032)	<.0001	-0.0146 (0.0034)	<.0001	-0.0045 (0.0046)	0.3312
month[†]	β_4	-0.3294 (0.0057)	<.0001	-0.0285 (0.0072)	<.0001	-0.0057 (0.0082)	0.4891	-0.0512 (0.0137)	0.0002
month*sex(F)	β_5	0.0115 (0.0051)	0.0247	-0.0064 (0.0063)	0.3071	-0.0158 (0.0073)	0.0300	-0.0344 (0.0121)	0.0046
$y_{i,j-1}$	λ_1			-2.6026 (0.0370)	<.0001	-2.3674 (0.0460)	<.0001	-2.2053 (0.0607)	<.0001
$y_{i,j-2}$	λ_2					-0.5940 (0.0403)	<.0001	-0.8295 (0.0568)	<.0001
$y_{i,j-5}$	λ_5							-0.1916 (0.0490)	<.0001
AIC		25622.0260		16198.6752		13693.9800		7776.5542	

[†] Month of follow up. F = female children

The focus of the paper was to investigate the influence of history on the current response. For the RSCM dataset, it was found that dependence on past responses had an effect on the current outcome, and that the influence diminished with time from the current response. Another important observation of the study was that the cumulative odds of malnutrition against better nutrition levels decreased for a child with a healthier history. This observation showed that the intervention that was in place was working and children were improving with time.

Although as per the scope of this paper, we endeavoured to achieve precision by the inclusion of the previous responses as part of the predictor variables, an advanced transition model that includes random effects can be considered for future work. This random effects structure will then capture the variation in the transition probabilities across subjects. In fact, as reported by Uranga and Molenberghs (2014), including random effects corrects for a bias that would have been introduced by the classical transition model. The random effects add precision to the profile estimates. Nonetheless, it should be noted that when there is a component of serial correlation, analyst should be careful in including random effects other than the random intercepts. This is because of a competition between the two sources of variation. Also, this transition-random effects model is recommended when there are long sequences of longitudinal data (Aitkin and Alf, 2003). Ghahroodi et al. (2009) present a transition model that can be used for a long sequence of longitudinal data. Further, one may also consider investigating the effects of interactions between past outcomes and other explanatory variables. We believe the work done in this paper will contribute to the knowledge about the methodology and application of transition models for longitudinal discrete outcomes where incomplete outcome sequences are present.

Chapter 7

Conclusion and Further Research

This research aimed at analyzing longitudinal data, with key emphasis on incomplete discrete longitudinal data. It entailed both simulation and real data applications. Real data applications involved two datasets; the Arthritis dataset – a secondary ordinal outcome dataset made available in Pawitan (2001), and the recovery from severe childhood malnutrition (RSCM) dataset – a clinical trial dataset provided by KEMRI/Wellcome Trust Research Programme, Kilifi, Kenya. These datasets contained missing values, both monotone and non-monotone. The longitudinal nature of the datasets in this thesis (both simulated and real) influenced the use of each one of three modelling frameworks: the marginal, random effects and transition models.

Analyzing incomplete longitudinal data, both of a Gaussian as well as non-Gaussian nature can be done under the somewhat less strict missing at random (MAR) assumption using standard statistical software. Likelihood based methods like the linear mixed models and generalized linear mixed models can be used for Gaussian and non-Gaussian data respectively. Alternatively, weighted generalized estimating equations can be used. Weighting makes these valid under MAR. Further, other methods can be used that do not need an explicit model for the missing data process to be used jointly with the substantive analysis model, like the expectation-maximization algorithm and multiple imputation (MI) and its extension, MI-GEE. These methods can be carried out in 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. GLIMMIX in addition to the NLMIXED procedure can be used for generalized linear mixed modelling. The MIXED procedure is suitable for linear mixed models. With all these powerful tools and sensible strategies, it then leaves almost no reason to still use the highly restrictive ad hoc methods, such as the complete case analysis, last observation carried forward, baseline observation carried forward, and single imputation techniques – although simple and easy to implement. This thesis investigated and compared analysis methods for incomplete correlated data, with primary emphasis on discrete longitudinal data.

Although the main focus of this research was to deal with incomplete discrete outcome data, this thesis started, in Chapter 2, with a continuous outcome data case. Using a simulation study and a real data application, the thesis compared multiple imputation (MI), direct likelihood (DL) analysis and inverse probability weighted generalized estimating equations (IPW-GEE) method. In correcting the incomplete nature of the data, MI and DL approaches were found to perform similar. Although the IPW-GEE method could be used for continuous outcome data, it was observed to be rather different with DL and MI, in some cases giving estimates of opposite sign than expected.

In Chapter 3, using a simulation study, the thesis compared the performance of weighted generalized estimating equations (WGEE) and generalized estimating equations after multiple imputations (MI-GEE). We provided theoretical considerations, as well as a simulated illustration of the two extensions of generalized estimating equations. We simulated count outcome data, first complete then afterwards caused dropouts. The research found that, under different dropout rates and sample sizes, MI-GEE was preferable compared to WGEE. Since the generalized linear mixed model (GLMM) analysis is valid under MAR, it was used as a basis against which the GEE methods were contrasted. The study found that in most cases the GLMM analysis performed similar to MI-GEE. However, it should be noted here that this comparison (between GLMM and GEE extensions) was done after some adjustments to make the two comparable. A difference exists with respect to the interpretation of the fixed effects, β. In random-effects models, the difference between the conditional mean and the marginal mean of an individual is the random effect. The fixed effects under the random-effects model, say β^R, and the marginal model, say β^M, are therefore different from each other in the sense that when the random effects model is considered, the marginal mean profile can be derived, but parameters should be interpreted conditional upon the individuals' heterogeneities. Care should be taken in the interpretation of the fixed effects under these two model families. But it should be noted that for the Poisson model this just applies to the intercept. For more on the relationships between the two model families, see for example, Lee and Nelder (2004); Ritz and Spiegelman (2004); Molenberghs and Verbeke (2005); Mitchell et al. (2013).

In Chapter 4, the thesis evaluated the performance of multiple imputation strategies, namely, fully conditional specification and multivariate normal imputation. The latter relies on the assumption of normally distributed variables whereas the former does not. When applied to ordinal outcome data, we found that the two methods were equally appropriate when faced with missingness in ordinal variables. These methods were assessed via a simulation study and a real data application. In the simulation study, datasets of different sample sizes ($N = 100, 200, 500$) were generated with different levels ($C = 3, 4, \text{and } 5$) of the ordinal outcome. The real application involved the Arthritis dataset. Similar results to those found in this chapter were reported by Lee and Carlin (2010).

Chapter 5 assessed the comparative performance of the direct maximum likelihood analysis, complete case analysis and three multiple imputation strategies, namely fully conditional specification, multivariate normal imputation and the ordinal imputation method (OIM). The methods were applied to ordinal outcome variables. Investigation involved both a simulation study and a real data application. The Kilifi malnutrition dataset was used. Under different dropout rates, the research found that the complete case analysis was generally a poor method. It was also observed that multiple imputation strategies slightly outperformed the direct maximum likelihood method in the simulation study. But, with an 8% non-monotone missing data patterns on the real application dataset, direct maximum likelihood and multiple imputation were found to perform very similarly. This therefore made it difficult to decide which of the two approaches is stronger when the missingness is fairly limited, at least with the settings of the dataset in that chapter. But, generally both approaches were presented as plausible methods that can be satisfactorily used for ordinal longitudinal outcome data.

In Chapter 6, the thesis dealt with the issue of transition modelling. A transitional likelihood missing at random model was built and the effects of conditioning on response history investigated, in addition to estimating the effects of measured covariates. By following this route, we in effect dealt with the dependence which under GEE would be handled using the so-called working correlation matrix together with the sandwich estimator under a marginal likelihood based model. When the latter route is used, there is a correction needed for incomplete data: weighted GEE or MI-GEE.

The research used the Kilifi malnutrition dataset and observed that dependence on history (past responses) had an effect on the current response, and that the influence diminished with distance from the current response. It was also observed that the cumulative odds of higher malnutrition levels against lower malnutrition levels decreased for a child with a healthier history. The missing data was handled by the direct (ignorable) likelihood analysis and multiple imputation. The research found that both methods can be used on equal footing as a way of correcting incompleteness. Similar results were obtained from the two methods.

Overall, the research gave an insight into the methods that may be considered when faced with missing data in longitudinal studies. Although, one may argue that much of what has been discussed in this thesis is well known in literature, these ideas are not always put into proper practice. We demonstrated the mastery of these ideas. For example, the cumulative logistic transition model (discussed in 6), is not commonly used in longitudinal data - especially in the medical field. Also, the relatively poor performance of IPW as demonstrated in the simulation studies of incomplete continuous and categorical data has not been widely reported elsewhere. These areas beyond the scope of this thesis can be further examined. However, as a general note, researchers must plan their studies very well from data collection. Proper data collection procedures must be followed so as to minimize missing data; although there is no agreed upon amount of missing data that can be acceptable. In case of missing data, knowing the reasons leading to the incompleteness plays a crucial role in determining the appropriate statistical procedure to analyze the available data. In fact, there exists no universal procedure for handling all missing data situations. However, proper design of the study and understanding the missing data process can help to a considerable extent.

In this research, we dealt mainly with missingness on the outcome variable, when in actual sense it is most probably that data may miss on the outcome when a covariate misses. These assumptions must be investigated so as to see if they replicate what has been observed in this thesis. However, it should be realized that these replications may depend on the type of study. For example, we may expect DL to be preferable or perform similar to MI when data are missing on the response but the results may not be replicated when data are missing for both covariates and responses.

The research also noted that, to use the OIM method for non-monotone missing data, the missing cases have to be monotonozed first before proceeding to carry out the OIM method. We view this as a double effort and more research on OIM is recommended to avoid the monotization step in its implementation.

Regarding missingness mechanisms, we assumed the MAR process. It is quite agreeable that in practical sense, MAR is the most likely mechanism that is expected to occur. This is due to the fact that an individual's probabilities of response may be related only to their own set of measured items. This set can change from one individual to another. But it is beneficial to make this assumption for analytic simplifications (Schafer and Graham, 2002). MAR also forms the general condition under which valid inferences can be obtained without having to model the missingness process explicitly, given that inferences are likelihood based or Bayesian (Beunckens, Molenberghs and Kenward, 2005; Kenward and Carpenter, 2007), and given the technical *separable parameters* assumption holds; meaning the parameters governing the missing data process are distinct from the measurement model parameters. However, the MAR process can only be fully substantiated in preplanned missingness designs e.g., in simulation studies. Otherwise, in real life situations, like clinical trials, the MAR assumption cannot be fully substantiated from the data, and that MAR and MNAR cannot be distinguished on formal statistical grounds. One only suspects that the data are not MAR but nothing from the data will indicate whether or not that is true. Arguably, for such, beyond the scope of this thesis, we recommend enhancing the common ignorable analyses with an appropriate class of sensitivity analyses and examine how MNAR mechanisms could jeopardize the MAR results. In this case, different sensitivity analysis models could be investigated. We acknowledge that, in terms of non-ignorable missingness, a completely sufficient data analysis that can be used in the process is not readily viable. Standard statistical models can result into very biased results. This is because the available observed measurements cannot provide sufficient information to confirm or refute ignorability. Researchers have proposed inclusion of the missingness in the modelling process. They suggested modelling the missingness process jointly with the measurement process, and then proceed to apply likelihood-based approaches like the maximum likelihood or consider a Bayesian inference. Two principal modelling frameworks that can be specified from the joint distribution of the measurement and missingness processes have been proposed; selection modelling and pattern mixture modelling. In this thesis we were mainly dealing with the ignorable missingness type, while at the same time acknowledging that it is possible to have the non-ignorable cases in real life applications. But, because of the constrains of time we would not cover everything in the scope of missing data problem, we therefore suggested the area of selection and pattern mixture modelling as an area for further research outside the scope of this thesis.

Finally, we acknowledge that we did not cover every aspect of how to deal with missingness in longitudinal studies. Apart from lack of handling the MNAR mechanism and non-ignorable missingness, these were some other more limitations of the study. For instance, proper inclusion of correlation structures for ordinal outcomes was one of the limitations in the study that need further investigation. The random effects model for ordinal outcomes was not handled in the current research and could also be an area of further research. Also, in this study, the number of samples simulated was chosen arbitrarily, some of which may be deemed small (e.g., for 300 datasets, especially when N is small). We recommend further investigations with sufficiently larger samples.

Bibliography

[1] Aerts, M., Molenberghs, G., Ryan, L.M., and Geys, H. (Eds.). (2002). *Topics in modelling of clustered data.* CRC Press.

[2] Agresti, A. (1989). Tutorial on modeling ordered categorical response data. *Psychological Bulletin*, 105(2), 290-301.

[3] Agresti, A. (2007). *Categorical data analysis.* Second Edition. John Wiley and Sons.

[4] Agresti, A. (2010). *Analysis of ordinal categorical data* (Vol. 656). John Wiley and Sons.

[5] Aitchison, J., and Silvey, S. D. (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika*, 44(1/2), 131-140.

[6] Aitkin, M., and Alfo, M. (2003). Longitudinal analysis of repeated binary data using autoregressive and random effect modelling. *Statistical Modelling*, 3(4), 291-303.

[7] Albert, P. S., and Follmann, D. (2009). Shared-parameter models. *Longitudinal data analysis*, 433–452.

[8] Ali, M. W., and Talukder, E. (2005). Analysis of longitudinal binary data with missing data due to dropouts. *Journal of Biopharmaceutical Statistics*, 15(6), 993-1007.

[9] Allison P. D. (2000). Multiple imputation for missing data: a cautionary tale. *Sociological Methods and Research*, 28(3), 301-309.

[10] Allison, P. D. (2001). *Missing data*, Thousand Oaks, CA: SAGE.

[11] Allison, P. D. (2005). Imputation of categorical variables with PROC MI. In *SAS Users Group International, 30th meeting (SUGI 30)*, 113(30), 1-14.

[12] Allison, P. D. (2012). Handling missing data by maximum likelihood. In *SAS global forum* (Vol. 312).

[13] Allison P. D. (2014). Sensitivity analysis for missing not at random. In *Statistical Horizons* 2014, September 25; Accessed on 18^{th} August 2016 at http//www.Statisticalhorizons.com/sensitivity-analysis.

[14] Ananth, C. V. and Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications, *International Journal of epidemiology* 26, 1323-1333.

[15] Armstrong, B. G., and Sloan, M. (1989). Ordinal regression models for epidemiologic data. *American Journal of Epidemiology*, 129(1), 191-204.

[16] Baker, S. G. (1992). 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.

[17] Baraldi, A. N., and Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1), 5-37.

[18] Barnard, J., and Meng, X. L. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research*, 8(1), 17-36.

[19] Bauer, D. J., and Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological methods*, 16(4), 373-390.

[20] Bender, R., and Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology*, 51(10), 809-816.

[21] Berkley, J., Mwangi, I., Griffiths, K., Ahmed, I., Mithwani, S., English, M., Newton, C., and Maitland, K. (2005). Assessment of severe malnutrition among hospitalized children in rural Kenya: comparison of weight for height and mid upper arm circumference. *Journal of the American Medical Association*, 294(5), 591-597.

[22] Berkley, J. A., Ngari, M., Thitiri, J., Mwalekwa, L., Timbwa, M., Hamid, F., Ali, R., Shangala, J., Mturi, M., Jones, K. D. J., Alphan, H., Mutai, B., Bandika, V., Hemed, T., Awuondo, K., Morpeth, S., Kariuki, S., and Fegan, G. (2016). Daily co-trimoxazole prophylaxis to prevent mortality in children with complicated severe acute malnutrition: a multicentre, double-blind, randomised placebo-controlled trial. *The Lancet Global Health*, 4(7), 464-473.

[23] Beunckens, C., Molenberghs, G., and Kenward, M. G. (2005). Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical Trials*, 2(5), 379-386.

[24] Beunckens, C., Sotto, C., and Molenberghs, G. (2008). A simulation study comparing weighted estimating equations and multiple imputation based generalized estimating equations for binary longitudinal data. *Computational Statistics & Data Analysis*, 52(2), 1533-1548.

[25] Birhanu T., Molenberghs, G., Sotto, C., Clark, C.J., and Kenward, M.G. (2011). Doubly robust and multiple imputation-based generalized estimating equations. *Journal of Biopharmaceutical Statistics*, 21(2), 202-225.

[26] Bolker, B. M., Brooks, M.E., Clark, C.J., Geanange, S.W., Poulsen, J.R., Stevens, M.H.H., and White, J.S.S. (2009). Generalized linear models: a practical guide for ecology on evolution. *Trends in Ecology and Evolution*, 24(3), 127-135.

[27] Brand, J. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets.* Dissertation, Erasmus University, Rotterdam.

[28] Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 88(421), 9-25.

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

[30] Carpenter, J., and Kenward, M. (2013). *Multiple imputation and its application.* John Wiley and Sons, Chichester.

[31] Carpenter, J. R., Kenward, M. G., and White, I. R. (2007). Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research*, 16(3), 259-275.

[32] Celeux, G., and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2(1), 73-82.

[33] Chambers, R. (2001). Evaluation Criteria for Statistical Editing and Imputation, National Statistics Methodological Series No. 28. *University of Southampton.*

[34] Chen, L., Toma-Drane, M., Valois, R. F., and Drane, J. W. (2005). Multiple imputation for missing ordinal data. *Journal of Modern Applied Statistical Methods*, 4(1), 288-299.

[35] Choi, K. H., Hoff, C., Gregorich, S. E., Grinstead, O., Gomez, C., and Hussey, W. (2008). 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.

[36] Clayton, D., Spiegelhalter, D., Dunn, G., and Pickles, A. (1998). Analysis of longitudinal binary data from multi-phase sampling. *Journal of Royal Statistical Society. Series B*, 60(1), 71-87.

[37] Collins, L. M., Schafer, J. L., and Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330-351.

[38] Cox, C. (1995). Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Statistics in Medicine*, 14(11), 1191-1203.

[39] Cox, D. R. (1972). The analysis of multivariate binary data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 21(2), 113-120.

[40] Cox, D. R., and Snell, E. J. (1989). *Analysis of binary data* (Vol. 32). CRC Press.

[41] Daniel, R. M., and Kenward, M. G. (2012). A method for increasing the robustness of multiple imputation. *Computational Statistics and Data Analysis*, 56(6), 1624-1643.

[42] Das, S., and Rahman, R. M. (2011). Application of ordinal logistic regression analysis in determining risk factors of child malnutrition in Bangladesh. *Nutritional Journal*, 10(1), 124.

[43] Demirtas, H., and Hedeker, D. (2008). An imputation strategy for incomplete longitudinal ordinal data. *Statistics in Medicine*, 27(20), 4086-4093.

[44] Demirtas, H., Freels, S. A., and Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation*, 78(1), 69-84.

[45] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1-38.

[46] Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics*, 959-971.

[47] Diggle, P., and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 43(1), 49-93.

[48] Diggle, P., Heagerty, P., Liang, K. Y., and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

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

[50] Eekhout, I., de Boer, R. M., Twisk, J. W., de Vet, H. C., and Heymans, M. W. (2012). Missing data: a systematic review of how they are reported and handled. *Epidemiology*, 23(5), 729-732.

[51] Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Publications.

[52] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (3rd ed). New York: John Wiley.

[53] Fielding, S., Fayers, P. M., and Ramsay, C. R. (2009). Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches. *Health and Quality of Life Outcomes*, 7(1), 57-66.

[54] Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, 309-317.

[55] Fitzmaurice, G. M, Davidian, M., Verbeke, G., and Molenberghs, G. (Eds.). (2008). *Longitudinal Data Analysis*. CRC Press.

[56] Fitzmaurice, G. M., Molenberghs, G., and Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society, Series B*, 57, 691-704.

[57] Gameroff, M. J. (2005, April). Using the proportional odds model for health-related outcomes: Why, when, and how with various SAS procedures. In *Proceedings of the Thirtieth Annual SAS Users Group International Conference: April 10-13*, 205-230.

[58] Ganjali, M. (2010). Fitting transition models to longitudinal ordinal response data using available software. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8), Ljubljana*, Slovenia. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/ iase/publications.php [2010 ISI/IASE]

[59] Ghahroodi, Z. R., Ganjali, M., and Berridge, D. (2009). A transition model for ordinal response data with random dropout: an application to the fluvoxamine data. *Journal of Biopharmaceutical Statistics*, 19(4), 658-671.

[60] Glynn, R. J., Laird, N. M., and Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association,* 88(423), 984-993.

[61] Goncalves, M. H., Cabral, M. S., and Azzalin, A. (2012). The r package bild for the analysis of binary longitudinal data. *Journal of Statistical Software*, 46(i09), 46-62.

[62] Gosho, M. (2014). Criteria to Select a Working Correlation Structure for the Generalized Estimating Equations Method in SAS. *Journal of Statistical Software*, 57(CS 1), 1-10.

[63] Gosho, M., Hamada, C., and Yoshimura, I. (2011). Criterion for the selection of a working correlation structure in the generalized estimating equation approach for longitudinal balanced data. *Communications in Statistics-Theory and Methods*, 40(21), 3839-3856.

[64] Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine*, 13(16), 1665-1677.

[65] Greenland, S., and Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, 142(12), 1255-1264.

[66] Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4*, NBER, Cambridge, 475-492.

[67] Hin, L. Y., Carey, V. J., and Wang, Y. G. (2007). Criteria for working correlation structure selection in GEE: Assessment via simulation. *The American Statistician*, 61(4), 360-364.

[68] Hin, L. Y., and Wang, Y. G. (2009). Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine*, 28(4), 642-658.

[69] Hogan, J. W., Roy, J., and Korkontzelou, C. (2004). Tutorial in biostatistics: Handling dropout in longitudinal studies. *Statistics in Medicine*, 23, 1455-1497.

[70] Horton, N. J., and Lipsitz, S. R. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, 55, 244-254.

[71] Horton, N.J., Lipsitz, S.R., and Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician*, 57(4), 229-232.

[72] Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

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

[74] Kadengye, D. T., Cools, W., Ceulemans, E., and Van den Noortgate, W. (2012). Simple imputation methods versus direct likelihood analysis for missing item scores in multilevel educational data. *Behavior Research Methods*, 44(2), 516-531.

[75] Kalton, G., and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19(2), 81-97.

[76] Kenward, M. G. (1998). Selection models for repeated measurements with nonrandom dropout: an illustration of sensitivity. *Statistics in Medicine*, 17(23), 2723-2732.

[77] Kenward, M. G., and Carpenter, J. (2007). Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, 16(3), 199-218.

[78] Kiernan, K., Tao, J., and Gibbs, P. (2012). Tips and strategies for mixed modeling with SAS/STAT procedures. In *SAS Global Forum*, Vol. 2012, Paper 332-2012.

[79] Kombo, A. Y., Mwambi, H., and Molenberghs, G. (2016). Multiple imputation for ordinal longitudinal data with monotone missing data patterns. *Journal of Applied Statistics*, 1-18.

[80] Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 79-86.

[81] Lago, L. P., and Clark, R. G. (2015). Imputation of household survey data using linear mixed models. *Australian and New Zealand Journal of Statistics*, 57(2), 169-187.

[82] Laird, N. M., and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963-974.

[83] Lee, J. (1992). Cumulative logit modelling for ordinal response variables: applications to biomedical research. *Computer Applications in the Biosciences: CABIOS*, 8(6), 555-562.

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

[85] Lee, K. J., Galati, J. C., Simpson, J. A., and Carlin, J. B. (2012). Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of nonlinear effects in a large cohort study. *Statistics in Medicine*, 31(30), 4164-4174.

[86] Lee, Y., and Nelder, J. A. (2004). Conditional and marginal models: another view. *Statistical Science*, 19(2), 219-238.

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

[88] Lin, G., and Rodriguez, R. N (2015). Weighted methods for analyzing missing data with the GEE procedure. *SAS Institute Inc*

[89] Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association*, 94(448), 1147-1160.

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

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

[92] Little, R. J., and Rubin, D. B. (2014). *Statistical analysis with missing data.* New York: Wiley.

[93] Liu, I., and Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. *Test*, 14(1), 1-73.

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

[95] Mallinckrodt, C. H., Clark, W. S., Carroll, R. J., and Molenberghs, G. (2003). Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *Journal of Biopharmaceutical Statistics*, 13(2), 179-190.

[96] Mallinckrodt, C. H., Sanger, T. M., Dube, S., DeBrota, D. J., Molenberghs, G., Carroll, R. J., Potter, W. Z., and Tollefson, G. D. (2003). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry*, 53(8), 754-760.

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

[98] McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models* (Vol. 2). London: Chapman and Hall.

[99] McLachlan, G., and Krishnan, T. (2007). *The EM algorithm and extensions* (Vol. 382). New York: Wiley.

[100] Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 538-558.

[101] Mitchell, S., Ozonoff, A., Zaslavsky, A. M., Hedt-Gauthier, B., Lum, K., and Coull, B. A. (2013). A comparison of marginal and conditional models for capture–recapture data with application to human rights violations data. *Biometrics*, 69(4), 1022-1032.

[102] Molenberghs, G., Beunckens, C., Jansen, I., Thijs. H., Van Steen, K., Verbeke, G., and Kenward, M.G. (2010). The Analysis of Incomplete Data, In: *Pharmaceutical Statistics With SAS.*

[103] Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2), 371-388.

[104] Molenberghs, G., Fitzmaurice, G., Kemward, M.G., Tsiatis, A., and Verbeke, G. (2014). *Handbook of missing data methodology.* CRC Press

[105] Molenberghs, G., and Kenward, M. (2007). *Missing Data in Clinical Studies* (Vol. 61). New York: John Wiley & Sons.

[106] Molenberghs, G., Thijs, H., Kenward, M. G., and Verbeke, G. (2003). Sensitivity analysis of continuous incomplete longitudinal outcomes. *Statistica Neerlandica*, 57(1), 112-135.

[107] Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

[108] Moodie, E. E. M., Delaney, J. A. C., LeFebvre, G., and Platt, R. W. (2008). Missing confounding data in marginal structural models: A comparison of inverse probability weighting and multiple imputation. *The International Journal of Biostatistics*, 4, 1-13.

[109] Nakai, M., and Ke, W. (2011). Review of the Methods for Handling Missing Data in Longitudinal Data Analysis. *International Journal of Mathematical Analysis*, 5(1), 1-13.

[110] National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*, Panel on Handling Missing Data in Clinical Trials, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, Washington, DC: National Academies Press.

[111] Newsom, J., Jones, R. N., and Hofer, S. M. (Eds.). (2012). *Longitudinal Data Analysis: A Practical Guide for Researchers in Aging, Health, and Social Sciences.* Routledge, New York.

[112] Noorian, S., and Ganjali, M. (2012). Bayesian Analysis of Transition Model for Longitudinal Ordinal Response Data: Application to Insomnia Data. *International Journal of Statistics in Medical Research*, 1(2), 148.

[113] Paik, M.C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association*, 92(440), 1320-1329.

[114] Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1), 120-125.

[115] Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, Oxford.

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

[117] Preisser, J. S. (2013). Weighting and imputation methods for missing data in longitudinal studies. *Bios* 767, Spring 2013.

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

[119] Ratitch, B., Lipkovich, I., and O'Kelly, M. (2013). Combining analysis results from multiply imputed categorical data. *PharmaSUG 2013-Paper SP03*, 1-10.

[120] Rhoads, C. H. (2012). Problems with tests of the missingness mechanism in quantitative policy studies. *Statistics, Politics, and Policy*, 3(1).

[121] Ritz, J., and Spiegelman, D. (2004). Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research*, 13(4), 309-323.

[122] Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008). Shared parameter models under random effects misspecification. *Biometrika*, 95(1), 63-74.

[123] Robins, J. M., and Gill, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine*, 16(1), 39-56.

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

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

[126] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106-121.

[127] Rodriguez, R. N., and Stokes, M. (2014). SAS/STAT 13.1 Round-Up. Paper SAS181-2014, Paper presented at 35th Institute in Research and Statistics. *SAS Institute Inc.*

[128] Rotnitzky, A., and Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77(3), 485-497.

[129] Royston, P. (2004). Multiple imputation of missing values. *Stata Journal*, 4(3), 227-241.

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

[131] Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72(359), 538-543.

[132] Rubin, D. B. (1978a). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, Vol. 1, 20-34. American Statistical Association.

[133] Rubin, D. B. (1978b). Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. In: *Imputation and Editing of Faulty or Missing Survey Data*. Washington, DC: U.S. Department of Commerce.

[134] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley and Sons.

[135] Rubin, D. B. (1991). EM and beyond. *Psychometrika*, 56(2), 241-254.

[136] Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434), 473-489.

[137] Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57(1), 3-18.

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

[139] Rubin, D. B., and Schenker, N. (1991). Multiple imputation in healthcare databases: An overview and some applications. *Statistics in Medicine, 10*(4), 585-598.

[140] Satty, A., and Mwambi, H. (2013). Selection and pattern mixture models for modelling longitudinal data with dropout: An application study. *SORT-Statistics and Operations Research Transactions*, 1(2), 131-152.

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

[142] Schabenberger, O. (2005). Introducing the GLIMMIX procedure for generalized linear mixed models. *SUGI 30 Proceedings*, 196-30.

[143] Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* New York: Chapman and Hall.

[144] Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8, 3-15.

[145] Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 57(1), 19-35.

[146] Schafer, J. L., and Graham, J. W. (2002). Missing data: Our view of the state of the art.*Psychological Methods*, 7(2), 147-177.

[147] Schafer, J. L., and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analysts perspective. *Multivariate Behavioral Research*, 33, 545-571.

[148] Schafer, J. L., and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, 11(2), 437-457.

[149] Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable dropout using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*, 94, 1096-1146.

[150] Seaman, S. R., Bartlett, J. W., and White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Medical Research Methodology*, 12(1), 46-58.

[151] Seaman, S. R., and White, L. R. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 1-18.

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

[153] Smith, T., and Smith, B. (2006). PROC GENMOD with GEE to analyze correlated outcomes data using SAS. *San Diego, CA: Department of Defence Center for Deployment Health Research, Naval Health research Center.*

[154] Soley-Bori, M. (2013). *Dealing with missing data: Key assumptions and methods for applied analysis* (No. 4). Technical Report.

[155] Song, X., Davidian, M., and Tsiatis, A. A. (2002). A Semiparametric Likelihood Approach to Joint Modeling of Longitudinal and Time-to-Event Data. *Biometrics*, 58(4), 742-753.

[156] Stokes, M. E., Davis, C. S., and Koch, G. G. (2012). *Categorical data analysis using SAS.* SAS institute Inc., Cary, NC, USA.

[157] Sutradhar, B. C., and Das, K. (2000). On the accuracy of efficiency of estimating equation approach. *Biometrics*, 56(2), 622-625.

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

[159] Tao, J., Kiernan, K., and Gibbs, P. (2015). Advanced Techniques for Fitting Mixed Models Using SAS/STAT Software. *Paper SAS*, 1919-2015. SAS Institute Inc., Cary, NC, USA.

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

[161] Tipa, M. E. V., Murphy, S. A., and McLaughlin, D. K. (1996). *Ignorable dropout in longitudinal studies.* Tech. rep. 96-04, The Pennsylvania State University, University Park, PA.

[162] Tsiatis, A. A., and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 809-834.

[163] Tsonaka, R., Verbeke, G., and Lesaffre, E. (2009). A Semi-Parametric Shared Parameter Model to Handle Nonmonotone Nonignorable Missingness. *Biometrics*, 65(1), 81-87.

[164] UCLA: Statistical Consulting Group, *Statistical computing seminars. Multiple imputation in STATA, Part1*,
URL http://www.ats.ucl.edu/stat/stata/seminars/missing_data/mi_in_stat_pt1.htm (accessed on November 12, 2015).

[165] Uranga, R., and Molenberghs, G. (2014). Longitudinal conditional models with intermittent missingness: SAS code and applications. *Journal of Statistical Computation and Simulation*, 84(4), 753-780.

[166] Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3), 219-242.

[167] Van Buuren, S. (2012). *Flexible Imputation of Missing Data.* CRC press, Boca Raton, FL.

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

[169] Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G. M., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049-1064.

[170] Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software* 45(3).

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

[172] Verbeke, G., and Molenberghs, G. (2009). *Linear Mixed Models for Longitudinal Data.* Springer Science and Business Media.

[173] Verbeke, G., Molenberghs, G., and Rizopoulos, D. (2010). Random effects models for longitudinal data. In *Longitudinal research with latent variables*, 37 - 96. Springer-Verlag Berlin Heidelberg.

[174] Wang, Y. G., and Carey, V. (2003). Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. *Biometrika*, 90(1), 29-41.

[175] Wang, Y., and Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, 96(455), 895-905.

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

[177] Wolfinger, R. D. (1999). Fitting nonlinear mixed models with the new NLMIXED procedure. Paper 287, SUGI Proceedings 1999, SAS Institute Inc., Cary, NC.

[178] Wolfinger, R., and O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48(3-4), 233-243.

[179] Wood, A. M., White, I. R., Hillsdon, M., and Carpenter, J. (2005). 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.

[180] World Health Organization, and Unicef (2009). WHO child growth standards and the identification of severe acute malnutrition in infants and children: joint statement.
URL http://www.who.int/nutrition/publications/severemalnutrition/97892415981 63/en/ Software at: http://www.who.int/childgrowth/software/en/

[181] Wu, M. C., and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 175-188.

[182] Yi, G. Y., and Cook, R. J. (2002a). Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American Statistical Association*, 97, 1071-1080.

[183] Yi, G. Y., and Cook, R. J. (2002b). Second order estimating equations for clustered longitudinal binary data with missing observations. In *Recent Advances in Statistical Methods*, Y. P. Chaubey (ed), 352-366. London: World Scientific Publishing Company, Inc.

[184] Yi, G.Y., He, W., and Liang, H. (2011). Semiparametric marginal and association regression methods for clustered binary data. *Annals of the Institute of Statistical Mathematics*, 63(3), 511-533.

[185] Yi, G.Y., Zeng, L., and Cook, R.J. (2011). A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters. *Canadian Journal of Statistics*, 39(1), 34-51.

[186] Yuan, Y. C. (2010). Multiple imputation for missing data: Concepts and new development (Version 9.0). *SAS Institute Inc, Rockville, MD*, 49.

[187] Yucel, R. M., and Zaslavsky, A. (2004). Practical suggestions on rounding in multiple imputation. *JSM Proceedings, Survey Research Methods Section*, 4679-4683.

[188] Yu, L. M., Burton, A., and Rivero-Arias, O. (2007). Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research*, 16(3), 243-258.

[189] Zeger, S. L., liang, K. Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equations approach. *Biometrics*, 44(4), 1049-1060.

[190] Zeger, S. L., and Karim, M. L. (1991). Generalized linear models with random effects: A Giggs sampling approach. *Journal of the American Statistical Association* 86, 79-86.

[191] Zeger, S. L., and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42(1), 121-130.

[192] Zeger, S. L., and Liang, K. Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, 11(14 - 15), 1825-1839.