Analysis of clustered data when the cluster size is informative

Menelaos Pavlou

Department of Statistical Science

University College London

A thesis submitted for the degree of

Doctor of Philosophy

February 2012

Statement of originality

I, Menelaos Pavlou, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Clustered data arise in many scenarios. We may wish to fit a marginal regression model relating outcome measurements to covariates for cluster members. Often the cluster size, the number of members, varies. Informative cluster size (ICS) has been defined to arise when the outcome depends on the cluster size conditional on covariates. If the clusters are considered complete then the population of all cluster members and the population of typical cluster members have been proposed as suitable targets for inference, which will differ between these populations under ICS. However if the variation in cluster size arises from missing data then the clusters are considered incomplete and we seek inference for the population of all members of all complete clusters.

We define informative covariate structure to arise when for a particular member the outcome is related to the covariates for other members in the cluster, conditional on the covariates for that member and the cluster size. In this case the proposed populations for inference may be inappropriate and, just as under ICS, standard estimation methods are unsuitable. We propose two further populations and weighted independence estimating equations (WIEE) for estimation.

An adaptation of GEE was proposed to provide inference for the population of typical cluster members and increase efficiency, relative to WIEE, by incorporating the intra-cluster correlation. We propose an alternative adaptation which can provide superior efficiency. For each adaptation we explain how bias can arise. This bias was not clearly described when the first adaptation was originally proposed.

Several authors have vaguely related ICS to the violation of the 'missing completely at random' assumption. We investigate which missing data mechanisms can cause ICS, which might lead to similar inference for the populations of typical cluster members and all members of all complete clusters, and we discuss implications for estimation.

Acknowledgements

This research project would not have been possible without the support of many people. I express my gratitude towards MRC-CTU for supporting financially the first three years of my study through a grant. I am thankful to Andrew Copas for his guidance, encouragement and supervision. I thank Shaun Seaman for his supervision and invaluable assistance. Deepest gratitude are also due to the members of the grant committee Vernon Farewell, Caroline Sabin, Abdel Babiker who offered their insight and ideas throughout the duration of the project. The UCL Centre for Sexual health and HIV Research, and in particular Richard Gilson and Andrew Copas who gave me the opportunity to work part-time as a Statistician and Data manager whilst working towards the completion of this thesis. Thanks to the Delta trial steering committee for making available the data. I am grateful to the Department of Statistical Science, UCL, for giving me the opportunity to develop my skills by participating in the departmental teaching activities. I thank my family and friends for their constant support, motivation and encouragement.

To my parents, Andreas and Elita

List of Figures

List of Tables

List of Abbreviations

ARC	AIDS Related Conditions
CGEE	Conditional Generalised Estimating Equations
CL	Conditional Likelihood
DWGEE	Doubly Weighted Generalised Estimating Equations
GEE	Generalised Estimating Equations
GLM	Generalised Linear Model
GLMM	Generalised Linear Mixed Model
GLMMBW	Generalised Linear Mixed Model separating Between- and Within-cluster effects
ICOVST	Informative Covariate Structure
ICS	Informative Cluster Size
IPW	Inverse Probability Weighting
LMM	Linear Mixed Model
LOCF	Last Observation Carried forward
MAR	Missing At Random
MCAR	Missing Completely At Random
MDM	Missing Data Mechanism
MI	Multiple Imputation

MNAR	Missing Not At Random
MWCR	Multiple Within-Cluster Resampling
OC	Oral Candidiasis
PMM	Pattern-Mixture Model
WBGEE	Weighted Estimating Equations with Block-Diagonal working correlation
WCR	Within-Cluster Resampling
WIEE	Weighted Independence Estimating Equations
WRGEE	Weighted Estimating Equations with Realistic working correlation

Contents

Chapter 1

Introduction

Clustered data arise in many fields of research. In longitudinal studies, a response may be measured repeatedly on the same person at different times; a person is a cluster. In toxicology, a response maybe measured on pups in a litter; a litter is a cluster. In educational studies, data are recorded on pupils in a school; a school is a cluster. Members which belong to the same cluster are likely to be more similar than members from different clusters because of genetic factors, persistent environmental characteristics or other determinants. So, clustered data are likely to exhibit correlation between outcome measurements for members in the same cluster. This feature renders standard statistical methods for univariate responses inappropriate for the analysis of clustered data.

Statistical methods for clustered data based on extensions of methods for univariate responses take into account the association between responses in the same cluster. For univariate data, interest lies in modelling the population average (or marginal mean) in terms of a set of regression covariates. For clustered data there is a wider choice of inferences, depending on the way the intracluster association is accounted for. Three main classes of models for clustered data have been considered: marginal models, random effects models and conditional models. Marginal models provide population average inference and the association between measurements is captured using measures of association, such as correlations and odds ratios, and a set of association parameters. In random effects models, the responses in a cluster are assumed to be independent given a set of cluster-specific parameters (random effects). Random effects models provide inference specific to each cluster. In conditional models, the clustered responses are modelled conditional on other responses (on previously observed responses in transition models).

Clusters may vary in size due to missing data. So, the observed clusters are regarded as incomplete and we seek inference for the population of all members of all complete clusters. Several authors have examined the performance of GEE and random effects models, when the variation in cluster size has arisen because of missing data and the aim is to estimate parameters of the distribution of the complete data. In the present work, we, on the other hand, are more concerned with the situation where the variability in cluster size is considered to be an inherent feature of the data and not due to missing data, i.e. the observed data are complete, and our interest is in parameters of the distribution of the observed data.

1.1 Informative cluster size and structure

When cluster size varies, cluster size is said to be informative if, for a given outcome variable of interest and set of covariates, the conditional expected value of the outcome given the covariates and the cluster size depends on the cluster size. That is, the relation between covariates and outcome is different in clusters of different size. Formally, if we denote the outcome for a cluster member by Y, the corresponding vector of covariates by $\boldsymbol{X}$ and the size of the cluster to which the member belongs by N, then cluster size is said to be informative if $E(Y|\boldsymbol{X}, N) \neq E(Y|\boldsymbol{X})$. For example, in studies of factors associated with periodontal disease (Williamson et al., 2003), a cluster corresponds to a person's mouth and members to the teeth. The disease status of the teeth may be associated with the number of teeth in the mouth, even conditional on covariates, because it is likely that genetic and environmental factors causing periodontal disease to also lead to tooth loss.

1.1.1 Methods for Marginal inference

Williamson et al. (2003) suggest that there are two marginal analyses of interest when clusters are complete: one for the population of all cluster members (population M) and one for a typical member of a typical cluster. In the first, larger clusters contribute more to inference than smaller ones; in the second all clusters contribute equally. We view the latter as inference for the population of typical cluster members (population C1), which is a subpopulation of population M, formed by selecting one member at random from each cluster. Therefore if $E(.)$ denotes expectation in population M (we

shall also use $E^M(.)$ to denote this) and $E^{C1}(.)$ denotes expectation in population C1, then $E(Y|\boldsymbol{X}) = E_{N|\boldsymbol{X}} E_{Y|\boldsymbol{X},N}(Y)$ and $E^{C1}(Y|\boldsymbol{X}) = E_{N|\boldsymbol{X}}[\frac{1}{N} E_{Y|\boldsymbol{X},N}(Y)]/E_{N|\boldsymbol{X}}(\frac{1}{N})$.

Williamson et al. (2003) provide a guide to the analyst as to which population should be selected for inference according to the objective of the analysis. In an economic assessment of how many, and which, teeth among patients seen at a dental clinic require a costly intervention, the population of all members (teeth) might be preferred, as clustering by patient may not be of direct relevance. Conversely, in a study of patient factors linked to the disease status of teeth, the population of typical cluster members (typical teeth for patients) might be of more interest.

Inference for population M can be obtained by applying the standard GEE with independence working correlation. For population C1 two inference methods were initially proposed: the computationally-intensive within-cluster resampling method (WCR - Hoffman et al., 2001) and the simpler inversely-weighted-by-cluster-size GEE with independence working correlation (Williamson et al., 2003; Benhin et al., 2005), abbreviated as WIEE. Williamson et al. (2003) proved that the two methods are asymptotically equivalent and showed through simulations that WIEE may perform better than WCR in terms of bias when the number of clusters is small.

When the covariates in the regression model are cluster-varying another type of informativeness may arise. We define informative covariate structure to arise when for a particular member, the outcome is related to the covariates for other members in the cluster, conditional on the covariates for that member and the cluster size. We say that covariates are size-balanced if their distribution is independent of the cluster size; otherwise they are non-size-balanced. When cluster size is informative, informative covariate structure may arise if the covariates are cluster-varying and non-size-balanced. Informative covariate structure may also arise when cluster size is non-informative and even when the cluster size is not varying. When the covariate structure is informative and the covariates are categorical cluster-varying we introduce populations for inference, additional to the ones previously proposed. We present estimation methods which are modifications of the WIEE method.

1.1.2 Efficient estimation for marginal inference

To provide inference for population C1 a potentially more efficient method (MWCR) was proposed by Chiang and Lee (2008), based on the WCR method. When the minimum cluster size, m, is greater than one, the authors propose randomly sampling m members from each cluster and then applying the GEE with a realistic working correlation to each resampled dataset. As the intracluster correlation is accounted for, efficiency may be gained.

Previous authors focused primarily on simple cases of informative cluster size, in the sense that the covariates of interest were either cluster-constant or size balanced. These authors have also focused on scenarios in which the expected value of the outcome depends on cluster size and covariates but not on interactions between the two. In this work we consider more general scenarios where the covariates involved are cluster-varying and non-size balanced. We explain why MWCR may lead to biased inference in these cases, a fact that was not mentioned in the original presentation of the method (Chiang and Lee, 2008). Furthermore, bias in MWCR can arise from realistic choices of the working correlation.

We derive an alternative estimator that is suitable in certain situations and which has the potential to be more efficient than WIEE. We call this method WRGEE because it may be used with a realistic working correlation, rather than requiring the use of the independence working assumption. We compare the performance of WRGEE to MWCR for scenarios where they are both unbiased and also show how WRGEE can give unbiased inference with moderate efficiency gains relative to WIEE in certain scenarios where MWCR is biased.

1.2 Informative cluster size versus missing data

When the variation in cluster size has arisen because of missing data, the observed clusters are incomplete; all clusters may be in fact of the same size, but not all of their members have been observed. We want to make inference for the population of all members of all complete clusters based on the sample of observed (incomplete) clusters in the dataset.

Methods for missing data are well known by statisticians; methods for informative cluster size are less well known. Previous authors referred to the relation between in-

formative cluster size and missing data mechanisms but this relation has not been made clear. We clarify the relation between the two and, having surveyed the methods available for inference about observed clusters and complete clusters, we provide intuition as to why different methods are needed for the two. We identify scenarios and special missing data mechanisms where some of the populations might be equivalent and we discuss implications for estimation methods.

1.3 Data example

To illustrate the methodology we use data from the Delta trial which compares three antiretroviral therapies. Zidovudine (AZT) in HIV-infected individuals was found to have a small and not long lasting effect. The Delta trial (Aber et al., 1996) was a three arm international randomised controlled trial designed to test whether combinations of AZT with zalcitabine (ddC) or AZT with didanosine (ddI) were more effective than AZT alone in extending survival and delaying disease progression for HIV infected patients. Full blood count and immunology subsets were measured on patients at all visits during follow-up. The primary endpoints were death in those with AIDS at entry, and AIDS and death in those without AIDS at entry.

There were 3207 individuals who took part in the Delta Trial; 1055 in AZT arm, 1080 in the AZT+ddC arm and 1072 in the AZT+ddI arm. The median follow-up time was 30 months and during this interval 699(22%) participants died. Of the 2765 participants without AIDS at entry, 936 (34%) developed AIDS or died. The number of months spent on allocated therapy was greatest in the AZT+ddC arm, the median (IQR) being 19 (8 to 29) compared to 17 (5 to 28) in the AZT+ddI and 18 (11 to 27) in the AZT arm.

The primary analysis indicated that for participants who had not had AZT before, initiation of AZT+ddI and AZT+ddC combination regimens had substantial benefits in prolonging survival compared to AZT alone. In particular, there was a significant relative reduction in mortality for both AZT+ddI (33%; $p < 0.001$) and AZT+ddC (21%; $p = 0008$) compared to AZT alone. For participants who had been treated with AZT before, the addition of ddI improved survival but there was no direct evidence of benefit from the addition of ddC. The benefit in terms of disease progression was seen mainly in patients not previously treated with AZT.

For each patient, any adverse events during follow-up were recorded. Apart from acknowledging the benefit of different treatment regimens in extending survival, it is also of interest to clinicians and researchers to establish whether the treatment regimens and other factors are associated with the types of adverse events experienced. In our illustrations we consider patients with at least one AIDS Related Conditions (ARC) event. Each cluster is composed of all the ARC events reported during a patient's follow-up. We identified 979 patients with sufficient information and at least one ARC event. The median number of events was 2; the range 1-15. There were roughly equal numbers of patients with 1 event, 2-3 events and more than 4 events. The most prevalent ARC event type was Oral Candidiasis (OC). The proportion of events that are OC decreases from 27% in patients with 1 event to 22% in patients with 2-3 events and to 15% in patients with more than 3 events.

In secondary analysis to illustrate the methods in Chapter 3, amongst all ARC events recorded, we examine the relation between whether or not the event is of type OC (binary outcome) and the covariates randomisation arm, CD4 count (most recent to the event) and time of the event since entry in the trial. As the total number of ARC events experienced by the patient increases, the percentage of events that are Oral Candidiasis decreases, suggesting that the cluster size might be informative.

It is also of interest to investigate how the immune status of a patient (of which CD4 count is an indicator), at times where ARC events are experienced, changes over time and whether it differs between the treatment arms. In Chapter 4 we examine the relation between CD4 count (continuous outcome) and randomisation arm and time since entry in the study. We discuss the application of the methods considered in Chapter 4 in the context of this example.

1.4 Structure of the thesis

The thesis is organised as follows. Chapter 2 introduces the main methods for analysing clustered data. The model families and the corresponding estimation methods are presented. The method of generalised estimating equations, modifications of which feature as main estimation methods in the subsequent chapters, is considered in detail. We also review methods for missing data. In Chapter 3, we introduce informative cluster size, populations for inference and relevant estimation methods for marginal and

cluster-specific inference. We define informative structure and two populations for inference, additional to the ones previously considered. We develop estimation methods for marginal inference for the additional populations. The application of the methods is illustrated using data from the Delta Trial. Chapter 4 deals with efficient methods for marginal inference. An existing efficient method is presented and its limitations are noted. Also, an alternative efficient method is proposed. We again use the Delta Trial example to demonstrate the application of the methods. In Chapter 5, we attempt to bridge the gap between informative cluster size and missing data. Chapter 6 concludes with a discussion about the proposed methodology and its limitations and areas for further work.

Chapter 2

Approaches to the analysis of clustered data

2.1 Introduction

In univariate statistics, each experimental unit gives rise to a single outcome variable and a vector of explanatory variables. In multivariate statistics each unit provides a number of different response variables. For example, the blood pressure and heart rate for each patient may be measured simultaneously. We concentrate on settings where responses emerging from each unit measure the same physical quantity and naturally form clusters.

Clustered data of this type arise in many situations. The response may be measured on different members of a group. For example, in toxicology measurements are obtained on the offsprings within a litter (cluster). Alternatively, the response may be repeatedly measured on each subject at several time occasions (repeated measurements). For example, in a clinical trial aiming to assess the effectiveness of a particular treatment compared to another, a measure of health outcome is recorded for each patient (cluster) at each follow-up time, giving rise to a vector of responses with natural time ordering among the measurements. In the latter scenario, if time is at least partly under scientific investigation, we specifically refer to *longitudinal data*.

The most important merit of longitudinal studies is enabling the direct study of change. As each subject is measured repeatedly over time, the researcher can study temporal changes within subjects (age effects) and factors that influence change. For example, longitudinal studies may aid understanding of how chronic diseases evolve

over time. Also, age effects can be separated from cohort effects (differences between subjects at baseline). In a cross-sectional study only cohort effects can be estimated. Additionally, in longitudinal studies each subject can serve as a control of himself/herself since the age effects can be estimated by comparing the individual's response at different times. This feature is particularly useful in crossover treatment studies where the experimental condition for a patient may change from control to alternative treatment, or vice-versa. Finally, longitudinal studies economise on the subjects, since fewer subjects are required in a longitudinal study, to achieve the same power as in a cross-sectional one.

Clustered and longitudinal data have important advantages, but also pose challenges to analysts. The main feature of such data is that the within-cluster responses tend to be correlated. Extensions of generalised linear models for univariate responses have been developed to account for this correlation. Three broad classes of models have been proposed: marginal, random effects and conditional models. The estimation methods for analysing clustered data and the interpretation of regression estimates tend to be more complicated than the ones for univariate responses. Also, the presence of time-varying covariates often complicates estimation. Finally, when dealing with studies with clustered or longitudinal data, a frequently encountered issue is missing data. Missing data arise when the outcome and/or covariates are not recorded at all of the intended measurement occasions. Appropriate handling of missing data requires development of more sophisticated methods, the validity of which often relies on untestable assumptions about the missing data mechanism.

In the next section, we define the notation which will be used throughout the thesis. In Sections 2.3-2.7, we introduce the three broad classes of models for analysing clustered data, present estimation methods for each and contrast the three approaches. As the proposed methodology in the Chapters 3 and 4 mainly relates to methods for marginal inference, more attention is paid to estimation methods for marginal models. In Section 2.8, we discuss the separation of covariate effects into the between- and within-cluster components. We also review methods for estimating the within-cluster covariate effects in the presence of cluster-confounding. Finally, we discuss missing data in Section 2.9. We introduce the missing data mechanisms which set a framework for analysing missing data, and we present the main estimation approaches for marginal

and cluster-specific inference.

2.2 Notation

We use capital letters to denote random variables, while lowercase letters are used for the realisation of each variable as a specific observation. The normal type is used to denote scalar quantities and the bold type is used to denote vectors and matrices. We use Greek letters to denote fixed effects parameters and Latin letters to denote random effects. Letter 'T' as a superscript is used to denote matrix transposition.

We now introduce the notation for clustered data. Suppose that K clusters represent a random sample from a population of clusters. The values of an outcome Y and a $q \times 1$ vector of covariates $\boldsymbol{X}$ are recorded for each member of each of these clusters. Let N denote the number of members in a cluster. We use subscripts i and j for the cluster and the member, respectively.

So, N_i is the number of members in cluster i, and Y_{ij} and $\boldsymbol{X}_{ij} = (X_{ij1}, \ldots, X_{ijq})^T$ are the values of Y and $\boldsymbol{X}$ for member j of cluster i ($i = 1, \ldots, K$; $j = 1, \ldots, N_i$). Let $\boldsymbol{Y}_i^* = (Y_{i1}, \ldots, Y_{iN_i})^T$ be the $N_i \times 1$ vector of responses and $\boldsymbol{X}_i^* = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iN_i})^T$ the $N_i \times q$ matrix of covariate values; the jth row of $\boldsymbol{X}_i^*$ corresponds to the vector of covariates for the jth member in cluster i. Let $\mu(\boldsymbol{X}) = E(Y \mid \boldsymbol{X})$, $\mu_{ij} = \mu(\boldsymbol{X}_{ij})$ and $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iN_i})^T$.

Often, a regression model which associates the expected response with covariates is assumed. For example, a linear regression model is specified as:

$$\mu_{ij} = E(Y_{ij} \mid \boldsymbol{X}_{ij}) = \beta_0 + \beta_1 X_{ij1} + \ldots + \beta_q X_{ijq}. \tag{2.1}$$

We denote $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_q)^T$ and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$ is a $(q+1) \times 1$ vector of regression parameters. We shall refer to β_0 as 'the intercept term' and to $\boldsymbol{\beta}_1$ as 'the effect of $\boldsymbol{X}$' or simply 'the covariate effects'. Model (2.1) is often expressed as $\mu_{ij} = E(Y_{ij} \mid \boldsymbol{X}_{ij}) = \beta_0 + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_1$ or in matrix form as $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i^* \mid \boldsymbol{X}_i^*) = \beta_0 \mathbf{1} + \boldsymbol{X}_i^* \boldsymbol{\beta}_1$, where $\mathbf{1} = (1, \ldots, 1)^T$ denotes a $N \times 1$ vector of units. To simplify calculations in some of the sections to follow we also define $\underline{\boldsymbol{X}}_i^*$ to be the matrix of covariates for cluster i with the first column being an N-dimensional vector of units (to account for the intercept term). So the model can be written as $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i^* \mid \boldsymbol{X}_i^*) = \underline{\boldsymbol{X}}_i^* \boldsymbol{\beta}$.

2.3 Model families for the analysis of clustered data

Generalised Linear Models (GLMs - McCullogh and Nelder, 1989) are used for analysing univariate continuous and discrete outcomes. They have a two-part specification. Firstly, the relationship between the expected outcome and covariates is specified: $\mu(\boldsymbol{X}_i) = E(Y_i \mid \boldsymbol{X}_i) = h^{-1}(\beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta}_1)$, $i = 1, \ldots, K$, through a monotone and differentiable function h, the *link function*. Secondly, the variance of the response is assumed to depend on the mean: $\text{var}(Y_i \mid \boldsymbol{X}_i) = v(\mu_i)\phi$, where $v(\mu)$ is the *variance function* and ϕ is a scale parameter. We write $v(\mu) = \frac{dh^{-1}(\theta)}{d\theta}\Big|_{\theta=h(\mu)}$. Examples are: for a continuous outcome, $h^{-1}(\theta) = \theta$ and $v(\mu) = 1$; for a binary outcome, $h^{-1}(\theta) = e^{\theta}/(1 + e^{\theta})$, $v(\mu) = \mu(1 - \mu)$ and $\phi = 1$; and for a count outcome, $h^{-1}(\theta) = e^{\theta}$, $v(\mu) = \mu$ and $\phi = 1$.

In univariate response settings, attention is unavoidably restricted in modelling the *population average* of Y or otherwise the *marginal mean* in terms of regression covariates. Dealing with clustered data is more complicated because there are two levels of replication; clusters and repeated measurements within clusters. The structure of clustered data implies more than one sources of variability: differences between clusters and differences between members within clusters. An additional complexity is that measurements which belong to the same cluster tend to be correlated.

Extensions of generalised linear models to handle clustered data have been developed. Three broad classes of models have been used: (i) marginal or population-averaged models (ii) random effects or cluster-specific models and (iii) conditional or transition models. These classes of models differ in the way they handle the dependence among the clustered responses. In marginal models, a model is specified for the within-cluster associations between responses and these within-cluster associations are modelled separately from the marginal mean. In random effects models, the within-cluster associations are accounted for by the inclusion of random effects, specific to each cluster. In conditional models, the dependence between clustered responses is handled by directly conditioning on other responses (previous responses for transition models).

The three model classes are distinct in several ways. Apart from the different way they handle the within-cluster dependence, the targets for inference are different in each approach. This is reflected by the difference in the interpretation of the regression

parameters for each class of models. Also, each modelling approach requires different estimation methods. In the following sections, we separately consider each model class and the corresponding estimation methods. We do not discuss missing data and their implications in estimation methods and inference until Section 2.9. In this chapter, we assume that the cluster size and covariate structure are non-informative (and even if some settings imply informative cluster size or informative covariate structure we shall not discuss these issues until Chapter 3).

2.4 Marginal models

The first approach in analysing clustered data is using a marginal model. The marginal expectation, $\mu_{ij} = E(Y_{ij} \mid \boldsymbol{X}_{ij})$, of the responses at each occasion is modelled in terms of explanatory variables, as in cross-sectional studies. The term 'marginal expectation' refers to the average response of a sub-population that shares common values of the explanatory variables. These models are useful when interest lies on inference about the population-average.

The additional complexity compared to GLMs is that the repeated measurements are not independent and this association must be accounted for. The specification of marginal models for clustered data consists of three components:

1. The marginal mean is related to the covariates through a link function, $h(.)$:

$$\mu_{ij} = E(Y_{ij} \mid \boldsymbol{X}_{ij}) = h^{-1}(\beta_0 + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_1). \tag{2.2}$$

2. The variance of each response is assumed to depend on the mean through the variance function: $\text{var}(Y_{ij} \mid \boldsymbol{X}_{ij}) = v(\mu_{ij})\phi$.

3. The within-cluster association between the responses is assumed to be a function of association parameters.

The first two components are analogous to the ones for GLMs. The third component reflects the lack of independence between the responses within the same cluster. The use of the term 'correlation' was deliberately avoided in the third component. For continuous responses, correlation is a natural measure of dependence between the responses. For discrete ones, such as binary, correlation is not the most natural measure of within-cluster association because is constrained by the marginal means; instead the

odds ratio may be used. In the next sections we present estimation methods for continuous and discrete outcomes. We initially consider maximum likelihood estimation (MLE). We note the considerable difficulties implementing MLE for discrete outcomes, mainly due to the complex form of the likelihood function for the clustered responses. We then present generalised estimating equations (GEE - Liang and Zeger, 1986) which simplify estimation by avoiding to specify the full distribution of the responses for each cluster. GEE can be used in fitting marginal regression models for continuous and discrete outcomes and extensions of them are used as key methodological tools in this work.

2.4.1 MLE: continuous responses

For continuous clustered responses, a linear regression model is specified to describe the relationship between the response and covariates at each occasion:
$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \ldots + \beta_q X_{ijq} + \epsilon_{ij}$, where ϵ_{ij}'s are zero-mean error terms associated with each response and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_q)^T$ is a $(q+1) \times 1$ vector of marginal regression parameters. The error terms in the standard linear model for univariate data are assumed to be mutually independent, identically distributed variables; for clustered data they tend to be correlated within the clusters.

It is often reasonable to treat the vector of responses (or a transformation of them) for cluster i, as a realisation of a multivariate Gaussian vector $\boldsymbol{Y}_i^*$. So the general linear model is

$$\boldsymbol{Y}_i^* \mid \boldsymbol{X}_i^* \sim \text{MVN}\left(\underline{\boldsymbol{X}}_i^* \boldsymbol{\beta}, \sigma^2 \boldsymbol{V}_i(\boldsymbol{\rho})\right), \tag{2.3}$$

where $\boldsymbol{V}_i(\boldsymbol{\rho})$ is the variance matrix of the responses for each cluster as a function of variance parameters, $\boldsymbol{\rho}$, while σ^2 is the residual variance. The vector of parameters to be estimated consists of one set for the regression parameters, $\boldsymbol{\beta}$, and two more sets for the variance parameters, $\boldsymbol{\rho}$ and σ^2.

Assuming that responses from different clusters are independent, the density function of the multivariate normal distribution for each cluster $f(\boldsymbol{Y}_i^* \mid \boldsymbol{X}_i^*, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho})$ is used to obtain the density function for the joint distribution of the responses, $f(\boldsymbol{Y}_1^*, \ldots, \boldsymbol{Y}_K^*) = \prod_{i=1}^K f(\boldsymbol{Y}_i^* \mid \boldsymbol{X}_i^*, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho})$. MLE maximises the likelihood func-

tion (expressed as a function of the unknown parameters given the data):

$$L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}) = \prod_{i=1}^{K} \left\{ (2\pi\sigma^2)^{-N_i/2} |\boldsymbol{V}_i(\boldsymbol{\rho})|^{-1/2} \, \mathrm{e}^{\left(-\frac{\sigma^2}{2}(\boldsymbol{Y}_i^* - \underline{\boldsymbol{X}}_i^* \boldsymbol{\beta})^T V_i(\boldsymbol{\rho})^{-1} (\boldsymbol{Y}_i^* - \underline{\boldsymbol{X}}_i^* \boldsymbol{\beta})\right)} \right\}. \tag{2.4}$$

Assuming that $\boldsymbol{\rho}$ is known, the MLE estimator for $\boldsymbol{\beta}$ (by maximising the likelihood conditional on $\boldsymbol{\rho}$) does not depend on σ^2 (Laird and Ware, 1982) and is given by

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\rho}) = \left(\sum_{i=1}^{K} \underline{\boldsymbol{X}}_i^{*T} \boldsymbol{V}_i(\boldsymbol{\rho})^{-1} \underline{\boldsymbol{X}}_i^* \right)^{-1} \sum_{i=1}^{K} \underline{\boldsymbol{X}}_i^{*T} \boldsymbol{V}_i(\boldsymbol{\rho})^{-1} \boldsymbol{Y}_i^*. \tag{2.5}$$

Assuming model (2.3), and conditionally on $\boldsymbol{\rho}$, $\boldsymbol{\beta}$ is normally distributed with mean as in equation (2.5) and variance

$$\left(\sum_{i=1}^{K} \underline{\boldsymbol{X}}_i^{*T} \boldsymbol{V}_i(\boldsymbol{\rho})^{-1} \underline{\boldsymbol{X}}_i^* \right)^{-1}. \tag{2.6}$$

In practice, $\boldsymbol{\rho}$ is generally unknown. So, the maximum likelihood function must be firstly expressed in terms of $\boldsymbol{\rho}$ and once an estimate for $\boldsymbol{\rho}$ is obtained, $\hat{\boldsymbol{\beta}}$ can be obtained by substitution in equation (2.5).

The procedure is as follows. The maximum likelihood function in terms of σ^2, and $\boldsymbol{\rho}$ is obtained by substituting $\hat{\boldsymbol{\beta}}(\boldsymbol{\rho})$ in equation (2.4) which becomes:

$$L(\sigma^2, \boldsymbol{\rho}) = \prod_{i=1}^{K} \left\{ (2\pi\sigma^2)^{-N_i/2} |\boldsymbol{V}_i(\boldsymbol{\rho})|^{-1/2} \, \mathrm{e}^{\left(-\frac{\sigma^2}{2}\mathrm{RSS}[\boldsymbol{V}_i(\boldsymbol{\rho})]\right)} \right\}, \tag{2.7}$$

where $\mathrm{RSS}[\boldsymbol{V}_i(\boldsymbol{\rho})] = \left(\boldsymbol{Y}_i^* - \underline{\boldsymbol{X}}_i^* \hat{\boldsymbol{\beta}}(\boldsymbol{\rho})\right)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\rho}) \left(\boldsymbol{Y}_i^* - \underline{\boldsymbol{X}}_i^* \hat{\boldsymbol{\beta}}(\boldsymbol{\rho})\right)$. Expression (2.7) is then maximised with respect to σ^2 to give the MLE estimate for σ^2 in terms of $\boldsymbol{\rho}$. Substituting $\hat{\sigma}^2(\boldsymbol{\rho})$ in equation (2.7) gives an expression for the likelihood in terms of $\boldsymbol{\rho}$ only:

$$L(\boldsymbol{\rho}) = \prod_{i=1}^{K} \left\{ [2\pi\hat{\sigma}^2(\boldsymbol{\rho})]^{-N_i/2} |\boldsymbol{V}_i(\boldsymbol{\rho})|^{-1/2} \mathrm{e}^{\left(-\frac{\hat{\sigma}^2(\boldsymbol{\rho})}{2}\mathrm{RSS}[\boldsymbol{V}_i(\boldsymbol{\rho})]\right)} \right\}. \tag{2.8}$$

Maximisation of expression (2.8) to obtain estimates for $\boldsymbol{\rho}$ usually requires numerical optimisation methods. When an estimate for $\boldsymbol{\rho}$ is obtained, backwards substitution in the corresponding expressions provides maximum likelihood estimates of σ^2 and $\boldsymbol{\beta}$.

Restricted Maximum Likelihood

When it comes to estimation of variance components, MLE is consistent (i.e. asymptotically unbiased and with variance that tends to zero as the sample size tends to infinity)

but can be biased in finite samples because it does not adjust for the degrees of freedom lost by estimating the regression coefficients (Diggle et al., 2002, pg. 65). Especially in cases where the number of clusters in the sample is small relative to the number of regression parameters, MLE can be heavily biased in estimating variance components.

Restricted maximum likelihood estimation (REML) was proposed by Patterson and Thompson (1971) to correct for this finite-sample bias. The underlying principle in REML is that estimation of variance components should not require the estimation of regression parameters first.

Let $\boldsymbol{Y}_a$ be the vector of all responses, $\boldsymbol{Y}_a = (\boldsymbol{Y}_1^{*T}, \ldots, \boldsymbol{Y}_K^{*T})^T$, so model (2.3) can be expressed as

$$\boldsymbol{Y}_a \sim \text{MVN}\left(\underline{\boldsymbol{X}}_a\boldsymbol{\beta}, \sigma^2\boldsymbol{V}_a(\boldsymbol{\rho})\right), \tag{2.9}$$

where $\underline{\boldsymbol{X}}_a$ is the combined matrix of the covariate matrices $\underline{\boldsymbol{X}}_i^*$ stacked below each other. $\boldsymbol{V}_a(\boldsymbol{\rho})$ is a block-diagonal matrix, each block being the covariance matrix $\boldsymbol{V}_i(\boldsymbol{\rho})$. REML is a maximum likelihood estimator for $\boldsymbol{\rho}$ after a linear transformation, defined by a matrix $\boldsymbol{A}$, is applied to the data, $\boldsymbol{U}_a = \boldsymbol{A}^T\boldsymbol{Y}_a$. Matrix $\boldsymbol{A}$ can be any matrix which results in the distribution of $\boldsymbol{U}$ being independent of $\boldsymbol{\beta}$. Let $M = \sum_{i=1}^{K} N_i$. Verbeke and Molenberghs (2000, pg. 45) state that $\boldsymbol{A}$ could be any $M \times (M - q)$ matrix with the property that its $(M - q)$ columns are orthogonal to the columns of matrix $\boldsymbol{X}_a$.

The transformed responses, $\boldsymbol{U}_a$, then follow a zero mean multivariate normal distribution (which does not depend on $\boldsymbol{\beta}$ anymore) with covariance matrix $\boldsymbol{A}^T\boldsymbol{V}_a(\boldsymbol{\rho})\boldsymbol{A}$. Patterson and Thompson (1971) proved that inferences for $\boldsymbol{\rho}$, based on $\boldsymbol{U}_a$ rather than $\boldsymbol{Y}_a$, provide consistent estimation for $\boldsymbol{\rho}$. Importantly, no information on $\boldsymbol{\rho}$ is lost in the absence of information on $\boldsymbol{\beta}$. REML does not itself provide estimates for $\boldsymbol{\beta}$. However, when the variance components have been estimated using REML, estimates for $\boldsymbol{\beta}$ can be obtained by substitution in the corresponding MLE expression for $\hat{\boldsymbol{\beta}}$.

Parametric modelling of the covariance structure

In the estimation procedures described so far, no parametric form was assumed for the covariance matrix $\boldsymbol{V}_i(\boldsymbol{\rho})$ (non-parametric approach). All the elements of $\boldsymbol{V}_i(\boldsymbol{\rho})$ are explicitly estimated using MLE or REML. For example, for balanced datasets ($N_i = N$ $\forall\, i$) the number of parameters in $\boldsymbol{V}_i$ adds up to $\frac{1}{2}N(N-1)$.

The non-parametric approach has important limitations. The number of variance

parameters to be estimated and the computational burden increases with the cluster size. Moreover, the true covariance matrix often depends on much fewer parameters which can be estimated more efficiently than the $\frac{1}{2}N(N-1)$ parameters. Also, since replication across clusters is used to estimate $\boldsymbol{\rho}$, the non-parametric approach is more suitable when the dataset consists of a large number of equally sized small clusters. For unbalanced data the non-parametric approach might cause problems. If, for example, only a few large clusters exist in the dataset, the statistical power to estimate the subset of variance parameters which correspond to these large clusters might be low.

For these reasons explicitly modelling the covariance matrix is not recommended except for balanced datasets consisting of long sequences of small clusters. In more general scenarios, parametric modelling of the covariance matrix can be considered. In the parametric approach, the covariance matrix of the responses is assumed to be a function of a small number of unknown parameters. Popular choices for the form of the covariance matrix are the exchangeable and autoregressive correlation models under which the assumed form of the covariance matrix depends only on two parameters; ρ and σ^2.

2.4.2 MLE: discrete responses

Under Gaussian assumptions for the responses, MLE is a feasible and relatively straightforward procedure to apply, mainly due to the elegant properties of the multivariate Normal distribution. When the clustered responses are discrete (e.g. binary, categorical or counts), MLE for model (2.2) tends to be more complicated in theory and more cumbersome in computation.

Although likelihood based inference for marginal models with discrete outcomes is not the focus of this work, we consider some probability models for the joint distribution of binary responses. This will facilitate direct comparisons with GEE which are simpler to apply and can be used for fitting marginal regression models for discrete and continuous outcomes, without the need of specifying the joint distribution of the responses in each cluster.

The Bahadur representation

Bahadur (1961) proposed a parameterisation which uses marginal means, while second- and higher-order moments are described in terms of correlations. The Bahadur repre-

sentation of the probability model for the binary clustered responses is

$$P(\boldsymbol{Y}_i^* = \boldsymbol{y}_i^* \mid \boldsymbol{X}_i^*, \boldsymbol{\beta}) = \left(\prod_{j=1}^{N_i} \mu_{ij}^{y_{ij}} (1-\mu_{ij})^{1-y_{ij}} \right) \times \left(1 + \sum_{j_1<j_2} \rho_{ij_1j_2} r_{ij_1} r_{ij_2} + \right.$$

$$\left. + \sum_{j_1<j_2<j_3} \rho_{ij_1j_2j_3} r_{ij_1} r_{ij_2} r_{ij_3} + \ldots + \rho_{i123\ldots N_i} r_{i1} r_{i2} r_{in_i} \right), \tag{2.10}$$

where $\mu_{ij} = P(Y_{ij} = 1 \mid \boldsymbol{X}_{ij}, \boldsymbol{\beta})$ and $r_{ij} = \frac{Y_{ij}-\mu_{ij}}{\sqrt{\mu_{ij}(1-\mu_{ij})}}$ are the standardised Pearson residuals. The second- and higher-order correlations are defined as $\rho_{ijk} = E(r_{ij} r_{ik})$ and $\rho_{i123\ldots N_i} = E(r_{i1} r_{i2} \ldots r_{iN_i})$ respectively.

The Bahadur representation of the joint probability function is straightforward and the likelihood comes in an attractive closed form and involves marginal probabilities and correlations which are familiar concepts from the analysis of continuous outcomes. However, it suffers from restrictions in the parameter space of second- and higher-order correlations, since the correlations between the responses are constrained by the marginal means. As a result, the Bahadur parameterisation requires a set of complicated constraints on the model parameters, which make maximisation of the likelihood very difficult. Also, the computational work increases as the cluster size increases. Except in settings with small cluster sizes, the Bahadur parameterisation has not been widely adopted for the analysis of clustered data with binary responses.

Other parameterisations

The log-linear model (Bishop et al., 2000) is another popular example of a probability model for multivariate binary responses. Assuming that the all clusters are of equal size, i.e. $N_i = N \; \forall \; i$, the joint probability distribution of a vector of binary responses is

$$P(\boldsymbol{Y}_i^* = \boldsymbol{y}_i^*) = c(\boldsymbol{\theta}) \exp \left(\sum_{j=1}^{N} \theta_j^{(1)} y_{ij} + \sum_{j_1<j_2} \theta_{j_1,j_2}^{(2)} y_{ij_1} y_{ij_2} + \ldots + \theta_{1,2,\ldots,N}^{(N)} y_{i1} y_{i2} \ldots y_{iN} \right), \tag{2.11}$$

where $\boldsymbol{\theta} = (\theta_1^{(1)}, \ldots, \theta_N^{(1)}, \theta_{1,2}^{(2)}, \ldots, \theta_{N-1,N}^{(2)}, \ldots, \theta_{1,2,\ldots,n}^{(N)})$ is the vector of $2^N - 1$ canonical parameters, and $c(\boldsymbol{\theta})$ is the normalising constant for the probability distribution to add up to one. These parameters have interpretations in terms of conditional probabilities. Consider, for example, the quadratic exponential model, where third- and higher-order terms in equation (2.11) are set equal to zero:

$$\theta_j^{(1)} = \text{logit}\{P(Y_{ij} = 1 | Y_{ik} = 0, \; k \neq j)\}, \quad j = 1, \ldots, N,$$

$$\theta_{jk}^{(2)} = \log \text{OR}(Y_{ij}, Y_{ik}|Y_{il} = 0,\ l \neq j, k), \quad j < k = 1, \ldots, N.$$

As discussed in Diggle et al. (2002, pg.142-144), canonical parameters are useful for the calculation of cell probabilities, but not convenient when the target is to describe the joint distribution of a vector of binary responses as a function of explanatory variables, i.e. when the parameters $\boldsymbol{\theta}$ depend on explanatory variables, $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{x})$. Another limitation of the canonical parameters is the sensitivity of interpretation to changes in the number of observations in the cluster. If N changes, the value and the interpretation of the canonical parameters change as well.

Given the limitations of the canonical parameters $\boldsymbol{\theta}$, alternative parameterisations, stemming from the original log-linear model, have been proposed. These parameterisations start with the N marginal parameters $\mu_{ij} = P(Y_{ij} = 1)$, $i = 1, \ldots, K$, $j = 1, \ldots, N$, as the building block for the model. The remaining $2^N - N - 1$ parameters maybe specified in various ways.

Fitzmaurice and Laird (1993) proposed a parameterisation which uses marginal means but the dependence between the responses is modelled as a function of conditional odds ratios, as in the original log-linear model. The interpretation of the conditional association parameters depends on the cluster size, so is more attractive for balanced datasets. The marginal means are routinely modelled using a regression model, $\mu_{ij} = \text{logit}^{-1}(\beta_0 + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_1)$ and, as noted by Fitzmaurice and Laird (1993), the likelihood equations for $\boldsymbol{\beta}$ have the form of a GEE-type score equation.

Liang et al. (1992) parameterised the association structure between the responses using marginal odds ratios which do not have interpretations conditional on other responses, do not depend on the cluster size and have weaker constraints than correlations. The second-order associations using the marginal odds ratio are $\text{OR}(Y_{ij}, Y_{ik}) = \gamma_{ijk} = \frac{P(Y_{ij}=1,Y_{ik}=1)P(Y_{ij}=0,Y_{ik}=0)}{P(Y_{ij}=1,Y_{ik}=0)P(Y_{ij}=0,Y_{ik}=1)}$, $j < k = 1, \ldots, N$. The parameterisation of the full distribution can be completed by expressing higher-order associations in terms of contrasts of marginal odds ratios. The constraints in the parameters' space when $N > 2$ (weaker though from the Bahadur model) and the computational burden which becomes larger as the cluster size increases limit the applicability of this parameterisation.

To summarise, several parameterisations have been proposed for the joint distribution of binary responses. The second- and higher-order associations can be described

in terms of correlations, marginal odds ratios or conditional odds ratios. Each of the parameterisations has certain limitations. Even when the joint distribution is fully defined, the likelihood can be complicated and its maximisation difficult, except in settings with constant and small cluster size. An alternative is to consider methods based on the quasi-likelihood function which is calculated from the first two moments only (Wedderburn, 1974). GEE are a quasi-likelihood method and are presented next.

2.4.3 Generalised estimating equations

GEE (Liang and Zeger, 1986; Zeger and Liang, 1986) emerge as a multivariate extension of the quasi-likelihood principle (Wedderburn, 1974) for GLMs, and can be viewed as an extension of GLMs for correlated data. A basic feature of GEE is that only the univariate distribution for each response needs to be specified, not the joint distribution of responses in each cluster. Due to the ease of implementation, the method has become popular for the analysis of categorical and count clustered responses but can be used as well for analysis of clustered continuous outcomes.

The Quasi-Likelihood principle

In contrast to the classical maximum likelihood estimation where the actual form of the distribution of the dependent variable needs to be specified, quasi-likelihood estimation requires fewer distributional assumptions about the response variable. Only the relationship between the mean and the covariates and the relationship between the mean and variance need to be specified.

We initially consider quasi-likelihood for univariate response settings, i.e. $N_i = 1$ $\forall\ i$. So, $\boldsymbol{Y}_i^* = Y_i$ is the ith univariate response and $\boldsymbol{X}_{ij} = \boldsymbol{X}_i$ is the q-dimensional vector of covariates for the ith univariate response; $i = 1, \ldots, K$. Also, $\mu_i = E(Y_i \mid \boldsymbol{X}_i)$. A generalised linear regression model to describe the relationship between the expected response and covariates is specified:

$$\mu_i = h^{-1}(\beta_0 + \boldsymbol{X}_i^T \boldsymbol{\beta}_1),$$

where $h(.)$ is a link function and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$ is a $(q+1)$-dimensional vector of regression parameters.

The variance of the response variable is assumed to be a function of the mean i.e. var$(Y_i|\boldsymbol{X}_i) = \phi v(\mu_i)$, where ϕ is a scale parameter and $v(\mu_i)$ is called the variance

function. The quasi-likelihood function for each univariate response was defined by Wedderburn (1974) as $U_i = U(Y_i, \mu_i) = \frac{\partial Q_i(Y_i,\mu_i)}{\partial \mu_i} = \frac{Y_i - \mu_i}{v(\mu_i)}$. It can be shown to have similar important properties to the derivative of a log-likelihood:

$$\text{(i) } E(U_i) = 0, \quad \text{(ii) var}(U_i) = \frac{1}{\phi v(\mu_i)} \text{ and (iii) } -E\left(\frac{\partial U_i}{\partial \mu_i}\right) = \frac{1}{\phi v(\mu_i)}. \quad (2.12)$$

When the responses are independent and their variance is constant, the log-quasi-likelihood can be expressed as $Q = Q(\mu_1, \ldots, \mu_K; Y_1, \ldots, Y_K) = \sum_{i=1}^{K} Q_i(\mu_i, Y_i)$ and has similar properties to the actual log-likelihood.

Estimates for the regression parameters are obtained by maximizing the log-quasi-likelihood, i.e. setting its derivative with respect to the parameters β_k ($k = 0, 1, \ldots, q$), equal to zero:

$$\begin{aligned} \frac{\partial Q}{\partial \beta_k} &= \frac{\partial \left(\sum_{i=1}^{K} Q_i(Y_i, \mu_i) \right)}{\partial \beta_k} = \sum_{i=1}^{K} \frac{\partial Q_i(Y_i, \mu_i)}{\partial \beta_k} \\ &= \sum_{i=1}^{K} \frac{\partial Q_i(Y_i, \mu_i)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_k} = \sum_{i=1}^{K} U(Y_i, \mu_i) \frac{\partial \mu_i}{\partial \beta_k} = \sum_{i=1}^{K} \frac{y_i - \mu_i}{v(\mu_i)} \frac{\partial \mu_i}{\partial \beta_k} = 0. \end{aligned}$$

The above estimating equations can be written in matrix form as

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{K} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} v(\mu_i)^{-1} (Y_i - \mu_i) = \mathbf{0}. \quad (2.13)$$

GEE

GEE are an extension of the quasi-likelihood approach to clustered data. The method requires specification of (a) a model for the marginal mean (b) the variance of each measurement in terms of the mean and (c) a model for the pairwise association between responses in the cluster.

Returning to the setting of clustered data recall that $\mu(\boldsymbol{X}_{ij}) = E(Y_{ij} \mid \boldsymbol{X}_{ij})$, $\mu_{ij} = \mu(\boldsymbol{X}_{ij})$ and $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iN_i})^T$. A marginal regression model $\mu(\boldsymbol{X}_{ij}) = h^{-1}(\beta_0 + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_1)$ is specified, where $h(.)$ is a known link function and $\boldsymbol{\beta}$ a $(q+1)$-dimensional vector of unknown parameters of interest. A working correlation structure is also chosen. Depending on this choice, the actual working correlation may involve unknown parameters $\boldsymbol{\rho}$ that need to be estimated. Let $\boldsymbol{R}_i(\boldsymbol{\rho})$ denote the working correlation matrix for cluster i.

If the marginal model is correctly specified, then under regularity conditions the solution $\hat{\boldsymbol{\beta}}$ to the following GEE is a consistent and asymptotically normally distributed estimator of $\boldsymbol{\beta}$:

$$\sum_{i=1}^{K} \boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{\rho}; \boldsymbol{Y}_i^*, \boldsymbol{X}_i^*) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1}(\boldsymbol{\rho})(\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{2.14}$$

where $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\rho}) \boldsymbol{A}_i^{1/2} \phi$ is the working covariance matrix for cluster i and $\boldsymbol{A}_i$ is the $N_i \times N_i$ diagonal matrix whose jth diagonal element is $v(\mu_{ij})$. Robustness to misspecification of the working correlation is one of the most important features of GEE; even if the working correlation assumption is false, the GEE estimator provides consistent estimation of $\boldsymbol{\beta}$. Pepe and Anderson (1994) raise a note of caution when $\boldsymbol{X}_i^*$ includes cluster-varying covariates. They show that equations (2.14) are consistent provided that the condition:

$$E(Y_{ij} \mid \boldsymbol{X}_{ij}) = E(Y_{ij} \mid \boldsymbol{X}_i^*) \; \forall \; j \tag{2.15}$$

is satisfied. When condition (2.15) is unlikely to hold, equations (2.14) should be applied with independence working correlation for consistent estimation.

The variance of $\hat{\boldsymbol{\beta}}$ is consistently estimated by the *sandwich estimator*

$$\begin{aligned} &\left(\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i(\boldsymbol{\rho})^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T} \right)^{-1} \left(\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i(\boldsymbol{\rho})^{-1} \text{var}(\boldsymbol{Y}_i^*) \boldsymbol{V}_i(\boldsymbol{\rho})^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T} \right) \\ \times \quad &\left(\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i(\boldsymbol{\rho})^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T} \right)^{-1}, \end{aligned} \tag{2.16}$$

where an estimate for $\text{var}(\boldsymbol{Y}_i^*)$ is given by $(\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i)^T$ and all quantities are evaluated at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\rho}}$. The variance estimator is also known as *robust variance estimator* because it provides consistent variance estimation even when the working variance assumption is false, provided the number of clusters is large. If the working correlation matrix is the true one, then $\text{var}(\boldsymbol{Y}_i) = \boldsymbol{V}_i(\boldsymbol{\rho})$ and the sandwich variance estimator reduces to the model-based estimator, $\left(\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} V_i(\boldsymbol{\rho})^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T} \right)^{-1}$.

Liang and Zeger (1986) suggested using moment estimates for $\boldsymbol{\rho}$ and ϕ, and a modified Fisher scoring iterative algorithm to solve the GEE:

1. Obtain an initial estimate for $\hat{\boldsymbol{\beta}}^{(0)}$, from the fit of a GLM ignoring the dependence between repeated measurements.

2. Obtain estimates of ϕ and $\boldsymbol{\rho}$ and $\boldsymbol{R}_i(\boldsymbol{\rho})$, $i = 1, \ldots, K$ (see below).

3. Use the current estimates for $\boldsymbol{\beta}$, ϕ and $\boldsymbol{\rho}$ to obtain an updated estimate for $\boldsymbol{\beta}$ (see below).

4. Return to Step 2 and continue iterating until convergence.

Step 3 uses the iterative procedure

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} - \left\{ \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T} \right\}^{-1} \left\{ \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \right\}, m = 1, \ldots \tag{2.17}$$

and current estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ are substituted in the right hand side of equation (2.17) to update $\boldsymbol{\beta}$.

In Step 2, moment estimates for ϕ and $\boldsymbol{\rho}$ are obtained. Firstly, ϕ is estimated: $\hat{\phi} = \frac{1}{\sum_{i=1}^{K} N_i - p} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \hat{r}_{ij}^2$ where $r_{ij} = \frac{Y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}$ and $\hat{r}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{v(\hat{\mu}_{ij})}}$ are the theoretical and observed standardised Pearson residuals respectively.

The estimate of ϕ is then used to estimate $\boldsymbol{\rho}$. Although estimation for $\boldsymbol{\beta}$ is robust to misspecification of the working correlation, an appropriate specification of the working correlation can increase the efficiency of the regression parameter estimates. The choice of the working correlation structure should generally be consistent with the observed correlations and prior knowledge. The off-diagonal elements of the $(N_i \times N_i)$ working correlation matrix $\boldsymbol{R}_i(\boldsymbol{\rho})$ can be expressed in terms of the correlation parameters $\boldsymbol{\rho}$.

Commonly used working correlation structures are independence, exchangeable, auto-regressive, unstructured and fixed. Independence working correlation ($\boldsymbol{R}_i = \boldsymbol{I}_{N_i}$) assumes that the observations within a cluster are independent and does not require estimation of any correlation parameters. Exchangeable working correlation ($\boldsymbol{R}_i(j,k) = \rho, \ j \neq k$) assumes that all observations within a cluster are equicorrelated. The moment estimate for ρ using all available pairs is: $\hat{\rho} = \frac{1}{(K^* - q)\phi} \sum_{i=1}^{N} \sum_{j \neq k} \hat{r}_{ij} \hat{r}_{ik}$, where $K^* = \sum_{i=1}^{K} N_i(1 - N_i)$. Auto-regressive-1 working correlation ($\boldsymbol{R}_i(j,k) = \rho^{|j-k|}, \ j \neq k$) assumes that the correlation decreases as the distance between observations increases. A moment estimate of ρ using adjacent pairs is: $\hat{\rho} = \frac{1}{(K^{**} - q)\phi} \sum_{i=1}^{K} \sum_{j \leq N_i - 1} \hat{r}_{ij} \hat{r}_{ik}$, where $K^{**} = \sum_{i=1}^{K} (N_i - 1)$. For longitudinal data, auto-regressive-t correlation structure might be defined as $\boldsymbol{R}_i(j,k) = \rho^{|t_j - t_k|}, \ j \neq k$,

where t_j and t_k indicate the times of measurements j and k. For unstructured correlation structure ($\boldsymbol{R}_i(j,k) = \rho_{jk},\ j < k$) the working correlation structure is left completely unspecified. The number of correlation parameters increases with the maximum cluster size. For balanced datasets ($N_i = N\ \forall\ i$), $\frac{1}{2}N(N-1)$ parameters need to be estimated. The moment estimates are $\hat{\rho}_{jk} = \frac{1}{(K-p)\phi}\sum_{i=1}^{N}\hat{r}_{ij}\hat{r}_{ik}$. Estimation may be inefficient for unbalanced datasets with few large cluster sizes.

Alternative GEE formulations

In their seminal article, Liang and Zeger (1986) emphasise that GEE are proposed as an estimation method for marginal inference when the principal interest is in the associations between the expected outcome and covariates. The association structure between the responses is treated as nuisance. Valid inference for the regression parameters is obtained even if the association structure (expressed through correlations) is misspecified. The GEE proposed by Liang and Zeger (1986) were subsequently termed 'GEE1'. When the scientific question regards the association structure of the responses as well as the estimation of regression parameters, moment estimation generally performs poorly for the estimation of correlation parameters and several extensions have been proposed to address this issue.

Prentice (1988) amended the GEE1, by adding a separate set of estimating equations for the parameters corresponding to pairwise correlations. For the additional set of estimating equations the components that carry information about the correlation between responses Y_{ij} and Y_{ik} are defined as $Z_{ijk} = \frac{(Y_{ij}-\mu_{ij})(Y_{ik}-\mu_{ik})}{\sqrt{v(\mu_{ij})}\sqrt{v(\mu_{ik})}}, i = 1,\ldots,K,\quad j,k = 1,\ldots,N_i$. $\boldsymbol{Z}_i$ consists of $\binom{N_i}{2} + N_i$ elements, $\boldsymbol{Z}_i = (Z_{i12}, Z_{i13}, \ldots, Z_{iN_i-1N_i}, Z_{i11}, Z_{i22}, \ldots, Z_{iN_iN_i})$.

Let $\eta_{ijk} = E(Z_{ijk})$. The additional set of estimating equations for the correlation parameters are:

$$U_{\boldsymbol{\rho}}(\boldsymbol{\beta},\boldsymbol{\rho}) \;=\; \sum_{i=1}^{K}\frac{\partial\boldsymbol{\eta}_i}{\partial\boldsymbol{\rho}}\boldsymbol{H}_i^{-1}(\boldsymbol{Z}_i - \boldsymbol{\eta}_i) = \boldsymbol{0}, \tag{2.18}$$

where $\boldsymbol{H}_i$ is a working covariance matrix for $\boldsymbol{Z}_i$. If the primary interest lies in modelling the marginal mean, Prentice (1988) suggests using an identity or diagonal matrix for $\boldsymbol{H}_i$. Although these choices are not optimal (Godambe, 1960), they result in simple computation and minimally affect the efficiency of $\hat{\boldsymbol{\beta}}$. Comparative studies on the issue of using a diagonal matrix $\boldsymbol{H}_i$ as opposed to alternative forms are given by Hall and

Severini (1998).

The solution, $\hat{\boldsymbol{\rho}}$, to equations (2.18) is obtained iteratively:

$$\hat{\boldsymbol{\rho}}^{(m+1)} = \hat{\boldsymbol{\rho}}^{(m)} - \left\{ \sum_{i=1}^{K} \frac{\partial \boldsymbol{\eta}_i^T}{\partial \boldsymbol{\rho}} \boldsymbol{H}_i^{-1} \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\rho}^T} \right\}^{-1} \left\{ \sum_{i=1}^{K} \frac{\partial \boldsymbol{\eta}_i^T}{\partial \boldsymbol{\rho}} \boldsymbol{H}_i^{-1} (\boldsymbol{Z}_i - \boldsymbol{\eta}_i) \right\}, m = 1, \ldots . \tag{2.19}$$

An initial value for $\boldsymbol{\rho}$ is needed to start the iterative procedure in (2.19). Estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ are obtained by iteration between equations (2.14) and (2.18). If the model for the marginal mean is correctly specified, the solution to equations (2.14) is a consistent estimate of $\boldsymbol{\beta}$ even when the model for the pairwise correlation is not correctly specified.

The estimating equations for $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$ can be written jointly as:

$$\begin{aligned} U_{GEEP}(\boldsymbol{\beta}, \boldsymbol{\rho}) &= \sum_{i=1}^{K} \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \boldsymbol{\eta}_i^T}{\partial \boldsymbol{\rho}} \end{pmatrix} \begin{pmatrix} \text{cov}(\boldsymbol{Y}_i) & \mathbf{0} \\ \mathbf{0} & \text{cov}(\boldsymbol{Z}_i) \end{pmatrix} \begin{pmatrix} \boldsymbol{Y}_i - \boldsymbol{\mu}_i \\ \boldsymbol{Z}_i - \boldsymbol{\eta}_i \end{pmatrix} \\ &= \sum_{i=1}^{K} \boldsymbol{C}_i \boldsymbol{B}_i \boldsymbol{S}_i = \mathbf{0}. \end{aligned} \tag{2.20}$$

Under regularity assumptions the joint distribution of the solution $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\rho}})$ to equations (2.20) is asymptotically Normally distributed with mean $(\boldsymbol{\beta}, \boldsymbol{\rho})$ and variance consistently estimated by $\left(\sum_{i=1}^{K} \boldsymbol{C}_i^T \boldsymbol{B}_i \boldsymbol{C}_i\right)^{-1} \left(\sum_{i=1}^{K} \boldsymbol{C}_i^T \boldsymbol{B}_i \boldsymbol{S}_i \boldsymbol{S}_i^T \boldsymbol{B}_i \boldsymbol{C}_i\right) \left(\sum_{i=1}^{K} \boldsymbol{C}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{B}_i\right)^{-1}$; all quantities are evaluated at $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\rho}})$.

In Prentice's GEE1, the estimating equations for $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ are considered to be independent (this is reflected by the diagonal matrices $\boldsymbol{C}_i$ and $\boldsymbol{B}_i$ in equation (2.20)). This might have some cost in terms of efficiency in the estimation of the regression coefficients but has the advantage of retaining consistency under misspecification of the working correlation structure. The advantage over the GEE1 of Liang and Zeger (1986) is that formal inferences can be made about the correlation parameters. Prentice (1988) mentioned that even in cases where scientific interest is in the marginal means rather than in the correlations, careful modelling of the association structure might improve the efficiency of $\hat{\boldsymbol{\beta}}$.

Several authors considered extensions of the Prentice's GEE where estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ is performed simultaneously. These methods are often termed 'GEE2'. One

motivation for such extensions is to increase the efficiency of the parameter estimates, especially for $\boldsymbol{\rho}$. Another motivation is that, certain choices for the working correlation structure as opposed to the 'true' one, might cause breakdown of the asymptotic properties of GEE1 (Crowder, 1995).

Preserving notation from Prentice's GEE1, the following single estimating equation was proposed by Prentice and Zhao (1991) for joint estimation of the vector of parameters of interest $(\boldsymbol{\beta}, \boldsymbol{\rho})$;

$$\begin{aligned} U_{GEE2}(\boldsymbol{\beta}, \boldsymbol{\rho}) &= \sum_{i=1}^{K} \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} & \mathbf{0} \\ \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} & \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\rho}} \end{pmatrix}^T \begin{pmatrix} \text{cov}(\boldsymbol{Y}_i^*) & \text{cov}(\boldsymbol{Y}_i^*, \boldsymbol{Z}_i) \\ \text{cov}(\boldsymbol{Z}_i, \boldsymbol{Y}_i^*) & \text{cov}(\boldsymbol{Z}_i) \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{Y}_i^* - \boldsymbol{\mu}_i \\ \boldsymbol{Z}_i - \boldsymbol{\eta}_i \end{pmatrix} \\ &= \sum_{i=1}^{K} \boldsymbol{C}_i \boldsymbol{B}_i \boldsymbol{S}_i = \mathbf{0}. \end{aligned} \tag{2.21}$$

Contrary to GEE1, in GEE2 the matrix of covariances $\boldsymbol{B}_i$, and more importantly the matrix of derivatives $\boldsymbol{C}_i$, are non-diagonal. The implication of the first being non-diagonal is that consistent estimation of $\boldsymbol{\beta}$ requires that, apart from the model for the marginal mean to be correctly specified, the model for the pairwise associations to also be correctly specified. An additional complication in GEE2 comes with the specification of the working covariance matrix $\boldsymbol{B}_i$. In GEE2 (as well as in GEE1), three-way and higher-order correlations are set equal to zero. Although $\text{cov}(\boldsymbol{Y}_i^*)$ is fully specified by the marginal means and pairwise correlations, specification of the other covariance matrices requires third- and fourth-order moments of $\boldsymbol{Y}_i^*$. Calculation of these moments becomes computationally cumbersome as the cluster size becomes larger.

When the model for the marginal mean and the model for the pairwise correlations are correctly specified, equations (2.21) yield consistent estimates for $\boldsymbol{\beta}$. Under regularity conditions the solution $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\rho}})$ to equations (2.21) is asymptotically normally distributed with mean $(\boldsymbol{\beta}, \boldsymbol{\rho})$ and variance which can be consistently estimated by $\left(\sum_{i=1}^{K} \boldsymbol{C}_i^T \boldsymbol{B}_i \boldsymbol{C}_i\right)^{-1} \left(\sum_{i=1}^{K} \boldsymbol{C}_i^T \boldsymbol{B}_i \boldsymbol{S}_i \boldsymbol{S}_i^T \boldsymbol{B}_i \boldsymbol{C}_i\right) \left(\sum_{i=1}^{K} \boldsymbol{C}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{B}_i\right)^{-1}$ and evaluated at $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\rho}})$.

Liang et al. (1992) also developed GEE2 for clustered binary responses using odds ratios, instead of correlations, as a measure of within-cluster associations between the responses. The choice between GEE1 and GEE2 should be based on the bias-precision trade-off and also depending on the scientific question to be answered. In GEE2, if

the model for the marginal mean is true, consistency of $\hat{\boldsymbol{\beta}}$ depends also on the correct specification of the model for the pairwise associations. If both models are correctly specified, GEE2 provide consistent estimates for $\boldsymbol{\beta}$ with higher efficiency compared to GEE1. GEE1 might provide slightly less efficient estimates for $\boldsymbol{\beta}$ but still consistent, even if the model for pairwise associations is not correctly specified. Simulation studies (Liang et al., 1992) indicate that GEE1 can be highly efficient for the estimation of $\boldsymbol{\beta}$ but could be extremely inefficient for the estimation of $\boldsymbol{\rho}$. Therefore, GEE1 should be the preferred estimation method when interest is in modelling the marginal mean; if modelling the marginal pairwise associations is equally important, GEE2 could be the considered.

Finally, we draw attention to another extension of GEE for binary responses proposed by Carey et al. (1993) and termed *alternating logistic regressions (ALR)*. Through an elegant combination of marginal and conditional specifications to capture the pairwise associations between the responses, addressing the third- and fourth-order moments is completely avoided, which is different to setting them to zero (Molenberghs and Verbeke, 2006). The estimating equations for $\boldsymbol{\beta}$ remain the same as in (2.14). ALR is an attractive procedure since it retains the computational ease of GEE1 and robustness against misspecification of the model for the pairwise associations, but also enables highly efficient estimation of association parameters with precision estimates, as in GEE2. The computational burden is much less than in GEE2 and consequently ALR can be used in datasets where the cluster sizes are moderate or large.

2.5 Random effects models

2.5.1 Formulation

In marginal models, the regression coefficients are considered to be constants and acquire population-average interpretations. Random effects models (Laird and Ware, 1982) are differentiated from marginal models by the inclusion of regression coefficients specific to each cluster. The variability in regression coefficients between clusters represents the natural heterogeneity between clusters due to unmeasured factors, so the regression parameters measure the direct influence of covariates on the expected outcome for specific clusters.

The typical specification of a random effects model has two components:

1. A regression model for the expected outcome, conditional on random effects is specified

$$\mu_{ij} = E(Y_{ij} \mid \boldsymbol{X}_{ij}, \boldsymbol{b}_i) = h^{-1}(\beta_0 + b_{0i} + \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1 + \boldsymbol{D}_{ij}^T\boldsymbol{b}_{1i}),\ \forall\ i,\ j \qquad (2.22)$$

   where $\boldsymbol{b}_i = (b_{0i}, \boldsymbol{b}_{1i}^T)^T$ is a $(q_b + 1) \times 1$ vector of random effects and $\boldsymbol{D}_{ij}$ is q_b-dimensional subset of $\boldsymbol{X}_{ij}$. Thus, a subset of regression coefficients is assumed to vary between clusters. Conditional on $\boldsymbol{b}_i$, the clustered responses, Y_{ij}, are assumed to be independent and have densities which belong to the exponential family of distributions.

2. The random effects are assumed to share a common underlying multivariate distribution with zero mean and density function $f(\boldsymbol{b}_i)$. Also, the random effects are assumed to be independent of covariates, i.e. $\boldsymbol{X}_i^* \perp \boldsymbol{b}_i\ \forall\ i$.

Model (2.22) is known as the *Generalised Linear Mixed Model* (GLMM). In the most general scenario $\boldsymbol{D}_{ij} = \boldsymbol{X}_{ij}$, which means that each cluster has different regression coefficients for all covariates. Usually the vector of the random effects, $\boldsymbol{b}_i$, is assumed to follow a multivariate Normal distribution, $\boldsymbol{b}_i \sim \text{MVN}(\boldsymbol{0}, \boldsymbol{G})$, where $\boldsymbol{G}$ is a $(q_b + 1) \times (q_b + 1)$ variance-covariance matrix with elements to be estimated. Contrary to marginal models where the marginal mean is modelled separately from the within-cluster correlation, in random effects models the correlation among the responses in the same cluster arises from the shared random effects.

A frequently used random effects model is the random intercepts regression model

$$Y_{ij} = \beta_0 + b_i + \boldsymbol{\beta}_1\boldsymbol{X}_{ij} + \epsilon_{ij},\ i = 1, \ldots, K,\ j = 1, \ldots, N_i, \qquad (2.23)$$

where $b_i \sim \text{N}(0, \sigma_b^2)$ and $\epsilon_{ij} \sim \text{N}(0, \sigma_\epsilon^2)$. The random effects, $b_i,\ i = 1, \ldots, K$ are independent between each other and independent from the error terms ϵ_{ij}. The fixed effect β_0 corresponds to the population average when $\boldsymbol{X}_{ij} = \boldsymbol{0}$; b_i represents the deviation of the cluster-specific average from the overall mean. Usually σ_ϵ^2 is known as the 'within-cluster variance' and σ_b^2 as the 'between-cluster variance'.

It is of interest to consider the covariance matrix and subsequently the correlation matrix of the responses in cluster i for the random intercepts model. The diagonal and

off-diagonal elements of the covariance matrix are, respectively:

$$\text{var}(Y_{ij}) = \text{var}(b_i + \epsilon_{ij}) = \text{var}(\epsilon_{ij}) + \text{var}(b_i) = \sigma_\epsilon^2 + \sigma_b^2,$$
$$\text{cov}(Y_{ij}, Y_{ik}) = \text{cov}(b_i + \epsilon_{ij}, b_i + \epsilon_{ik}) = \text{var}(b_i) = \sigma_b^2.$$

The correlation matrix has common off-diagonal elements, $\rho = \frac{\sigma_b^2}{\sigma_\epsilon^2 + \sigma_b^2}$. This observation serves as a justification of the fact that the exchangeable correlation model (often used when modelling the covariance structure in marginal linear models for clustered data) arises from a random intercepts model.

Random effects models are also known as 'multilevel' or 'hierarchical' models because of the underlying hierarchical structure of the model. For example, one can think of the linear random intercepts model (2.23) as a model formed in two levels or stages. The first stage, which refers to the measurement level is $Y_{ij} = \beta_i + \boldsymbol{\beta}_1 \boldsymbol{X}_{ij} + \epsilon_{ij}$; in the second stage, which corresponds to the cluster level, the cluster-specific intercept is partitioned into a fixed part and random part, $\beta_{0i} = \beta_0 + b_i$, where β_0 is the overall intercept and $b_i \sim N(0, \sigma_b^2)$. Terms ϵ_{ij} and b_i are known as level-1 and level-2 residuals respectively.

2.5.2 Estimation

Considering the general random effects model in equation (2.22), the parameters of main interest are the vector of fixed regression coefficients and the variance parameters in the distribution of random effects. Interest might also lie in estimating the cluster-specific parameters $\boldsymbol{b}_i$. These can be often used to detect groups of clusters that evolve differently over time, to detect special cluster profiles (for example outlying clusters) or to make predictions about cluster-specific evolutions.

There are three approaches to obtain inference for random effects models. The first, 'Conditional likelihood', is used when interest lies only in regression coefficients that do not have a random part. The underlying principle in conditional likelihood estimation is that the random effects $\boldsymbol{b}_i$ are regarded as nuisance effects and only the part of the data which does not contain information about $\boldsymbol{b}_i$ is used to make inference about the subset of regression coefficients which only have a fixed part. Such inference is called 'conditional inference'. The second, maximum likelihood estimation, can be used to estimate the parameters of main interest, while estimates for the random effects can be obtained using special methods such as 'Empirical Bayes'. The third approach

is to implement a fully Bayesian approach where prior distributions are specified for the parameters of main interest and the random effects.

Conditional inference

In the conditional likelihood approach (McCullogh and Nelder, 1989, Section 7.2), $\boldsymbol{b}_i$ are considered 'nuisance' or 'incidental' parameters that need to be conditioned out of the problem. They may be considered as nuisance in the sense that, although they are necessary for the assumed model to make sense, their value is not of main interest to the researcher. The likelihood of the responses, $(\boldsymbol{Y}_1^*, \ldots, \boldsymbol{Y}_k^*)$, is maximised conditional on the sufficient statistics of $\boldsymbol{b}_i$. Conditional likelihood has two important advantages: first, no distributional assumptions are needed for $\boldsymbol{b}_i$ and second, it does not require the condition of independence between the random effects and covariates to be true.

Although conditional inference simplifies the estimation procedure by suppressing the complexity that arises from defining the distribution of $\boldsymbol{b}_i$, it also has some important limitations. All information about $\boldsymbol{b}_i$ is lost. More importantly, the covariate effects of any cluster-constant covariates cannot be estimated. Also, it is not possible to estimate the fixed part of regression coefficients which have a random part. Consider for example, a random intercepts model with a single cluster-varying covariate X and a vector of *cluster-constant covariates* $\boldsymbol{S}$

$$E(Y_{ij}|X_{ij}, b_i) = h^{-1}(\beta_0 + b_i + \beta_1 X_{ij} + \boldsymbol{\beta}_2 \boldsymbol{S}_i).$$

Inference using conditional likelihood (which in this case conditions on the sufficient statistics $(\sum_{j=1}^{N} Y_{ij})$ for b_i) only allows estimation of β_1 which is the within-cluster effect of X. By conditioning out of the problem the random effects b_i, all information on β_0 and $\boldsymbol{\beta}_2$ is also lost (see, for example, Goetgeluk and Vansteelandt, 2008).

An application of conditional inference, conditional logistic regression, is frequently encountered in case-control studies. Consider for example a case-control study of 100 matched pairs (200 subjects) and a single risk factor of interest. The number of parameters to be estimated adds up to 101, since there exist 99 dummy variables for each matched pair, plus the intercept term and one parameter for the risk factor of interest. Ordinary logistic regression is unsuitable in this case because of the large number of parameters compared to observations. As often estimating or modelling the baseline risk is not of direct scientific interest, the problem can be alleviated through

conditional inference. The pair-specific parameters are considered to be nuisance and are integrated out of the problem by conditioning on their sufficient statistics; only the effect of the risk factor is estimated.

Several authors (see, for example, Neuhaus and Kalbfleisch, 1998; Neuhaus and MCculloch, 2006) studied the connection between conditional likelihood and estimation of within-cluster effects in the context of cluster confounding. This relation is further discussed in Section 2.8.

Maximum likelihood estimation

MLE can be used when $\boldsymbol{b}_i$ is also of interest, apart from the fixed part of the regression coefficients. Contrary to the conditional inference approach where the within-cluster effect of cluster-varying covariates can be estimated by utilising only longitudinal information (i.e. information from clusters where the covariates vary), in the maximum likelihood approach also cross-sectional information (i.e. information from clusters where $\boldsymbol{X}$ does not vary within-clusters) is used to make inferences about $\boldsymbol{\beta}$ in model. The underlying idea is that knowledge about a cluster's regression coefficients can be informed by the variability of the regression coefficients across the population. So, the terms $\boldsymbol{b}_i$ are assumed to be a sample of unobservable variables from a distribution.

The distribution of random effects is usually assumed to be a multivariate Normal distribution with mean zero, covariance matrix $\boldsymbol{G}$ and density $f(\boldsymbol{b}_i \mid \boldsymbol{G})$. The vector of unknown parameters to be estimated is denoted by $\boldsymbol{\delta}$ and includes the fixed effects $\boldsymbol{\beta}$ and the variance elements in $\boldsymbol{G}$.

The likelihood contribution of cluster i (suppressing dependence on covariates) is obtained by integrating over the distribution of the random effects, $\boldsymbol{b}_i$:

$$f(\boldsymbol{Y}_i^* \mid \boldsymbol{\delta}) = \int \prod_{j=1}^{N_i} f(Y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\beta}) f(\boldsymbol{b}_i \mid \boldsymbol{G}) \, \mathrm{d}\boldsymbol{b}_i, \quad i = 1, \ldots, K. \tag{2.24}$$

So, the likelihood is:

$$L(\boldsymbol{\delta} \mid \boldsymbol{Y}_1^*, \ldots, \boldsymbol{Y}_K^*) = \prod_{i=1}^{K} f(\boldsymbol{Y}_i^* \mid \boldsymbol{\delta}) = \prod_{i=1}^{K} \int \prod_{j=1}^{N_i} f(Y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\beta}) f(\boldsymbol{b}_i \mid \boldsymbol{G}) \, \mathrm{d}\boldsymbol{b}_i. \tag{2.25}$$

A key issue in the evaluation of the likelihood function in expression (2.25) is the calculation of the K integrals of the form (2.24). Analytic expressions for the likelihood contribution of cluster i exist only in limited cases. In a linear mixed model, for example, the contribution of each cluster in the likelihood is the density of a multivariate

Normal distribution with mean $\underline{\boldsymbol{X}}_i^*\boldsymbol{\beta}$ and variance $\boldsymbol{V}_i = \underline{\boldsymbol{D}}_i^{*T}\boldsymbol{G}\underline{\boldsymbol{D}}_i^* + \boldsymbol{\Sigma}_i$, where matrices $\underline{\boldsymbol{X}}_i^*$ and $\underline{\boldsymbol{D}}_i^*$ are the $N_i \times (q+1)$ and $N_i \times (q+1)$ design matrices for the $(q+1)$- and (q_b+1)-dimensional vectors $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$, respectively (recall that the first column of $\underline{\boldsymbol{X}}_i^*$ and $\underline{\boldsymbol{D}}_i^*$ is a column of units). Also, matrix $\boldsymbol{\Sigma}_i$ is the covariance matrix of the error terms in the linear regression model and usually $\boldsymbol{\Sigma}_i = \sigma^2\boldsymbol{I}_{N_i}$. Estimation can then be completed by maximising the likelihood or the restricted likelihood.

In general, for discrete responses no analytic expressions exist for the integrals in equation (2.24) and numerical approximations are required for the evaluation of the K integrals involved. Popular methods for the approximation of the integral in (2.24) are the Gaussian Quadrature and Adaptive Gaussian Quadrature. We also note that the original formulation of REML, as it was described in Section 2.4.1, only applies to LMMs; the extension to GLMMs is not straightforward. Alternative REML-type estimators have been used for GLMMs (Liao and Lipsitz, 2002; Bellio and Brazzale, 2011).

Once the issue of the evaluation of the likelihood function has been addressed, the resulting likelihood function needs to be maximised to obtain maximum likelihood estimates. By setting the derivative of the log-likelihood equal to zero, score equations for $\boldsymbol{\delta}$ are obtained. The Expectation-Maximisation (EM) algorithm, popularised by Dempster et al. (1977) as a computing algorithm for incomplete data settings, finds an application in solving the score equations for $\boldsymbol{\delta}$. In this case, there are no missing data but it is assumed that the 'complete data' for a cluster comprise of the observed responses $\boldsymbol{Y}_i$ and the unobserved random effects $\boldsymbol{b}_i$. Following Diggle et al. (2002, pg. 173-175), the complete-data score functions for $\boldsymbol{\beta}$ and $\boldsymbol{G}$ have a simple form and are functions of the unobserved quantities $\boldsymbol{b}_i$. The 'observed-data' score functions can be obtained by taking the expectation of the complete data score equations with respect to the unobserved random effects $\boldsymbol{b}_i$. The EM algorithm iterates between two steps. The 'Expectation' step involves calculation of the expectations in the score functions using the current parameter estimates. In the 'Maximisation' step the score equations are solved to obtain updated parameter estimates. The expectation step involves integration to obtain the expectation conditional on $\boldsymbol{b}_i$. When the dimension of $\boldsymbol{b}_i$ is one or two, the integration can be carried out using numerical techniques. For higher dimension of $\boldsymbol{b}_i$ Monte Carlo integration methods are required.

Apart from the estimation of fixed effects and variance components it might also be of interest to obtain estimates for the cluster-specific parameters $\boldsymbol{b}_i$. Since these are assumed to be random variables, Bayesian techniques are usually deployed for such purposes. The target is to calculate the 'posterior' distribution of $\boldsymbol{b}_i$ conditional on the responses, where the unknown parameters ($\boldsymbol{\beta}$ and $\boldsymbol{G}$) have been replaced by their ML or REML estimates. In the Bayesian framework, the distribution of the random effects, $\boldsymbol{b}_i$, is the prior distribution with density $f(\boldsymbol{b}_i \mid \boldsymbol{G})$ and $f(\boldsymbol{Y}_i^* \mid \boldsymbol{b}_i)$ is the density function of the distribution of $\boldsymbol{Y}_i^*$ conditional on $\boldsymbol{b}_i$. The posterior density of $\boldsymbol{b}_i$ is given by

$$f(\boldsymbol{b}_i \mid \boldsymbol{Y}_i^*, \boldsymbol{\beta}, \boldsymbol{G}) = \frac{f(\boldsymbol{Y}_i^* \mid \boldsymbol{b}_i)\ f(\boldsymbol{b}_i \mid \boldsymbol{G})}{\int f(\boldsymbol{Y}_i^* \mid \boldsymbol{b}_i)\ f(\boldsymbol{b}_i \mid \boldsymbol{G})\ \mathrm{d}\boldsymbol{b}_i}. \tag{2.26}$$

In the special case of linear responses with Gaussian random effects, the density in (2.26) is a Normal density and the posterior mean of that Normal distribution is used as a point estimate for $\boldsymbol{b}_i$. More generally, for non-Gaussian responses the density in (2.26) is not Normal and the mode of the posterior distribution is used as a point estimate of $\boldsymbol{b}_i$. The obtained estimates, $\hat{\boldsymbol{b}_i}$ are called 'Empirical Bayes (EB)' estimates.

Bayesian approach

Although maximum likelihood estimation is convenient in the case of linear mixed models because of the conjugation of the assumed distribution of the responses and the distribution of random effects, numerical integration is necessary to obtain likelihoods when the responses are non-Gaussian. The hierarchical structure of model (2.22) makes a Bayesian formulation (Zeger and Karim, 1991) for alternative estimation of parameters very appealing. Prior distributions are firstly introduced for the fixed effects $\boldsymbol{\beta}$ and the variance components $\boldsymbol{G}$. Commonly used priors for $\boldsymbol{\beta}$ are Normal or flat distributions and for $\boldsymbol{G}$ non-informative priors. Given the specification of the priors, the joint posterior distribution is

$$f(\boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_K \mid \boldsymbol{Y}_1^*, \ldots, \boldsymbol{Y}_K^*) \propto \prod_{i=1}^{K} \prod_{j=1}^{N_i} f(Y_{ij} \mid \boldsymbol{\beta}, \boldsymbol{b}_i) \prod_{i=1}^{K} f(\boldsymbol{b}_i \mid \boldsymbol{G}) f(\boldsymbol{G}) f(\boldsymbol{\beta}).$$

The Gibbs sampler is used for estimating the desired posterior distributions. Estimates for $\boldsymbol{\beta}$, $\boldsymbol{G}$ and also $\boldsymbol{b}_i$ are obtained by drawing samples from their posterior distributions. The Bayesian approach is easy to implement (although computationally intensive) and is flexible in changes in the dimension and the distribution of random effects.

2.6 Conditional models

Conditional models are used to characterise the conditional expectation of the outcome Y_{ij} in terms of subsets of other outcomes and covariates of interest. A special case of conditional models arises in longitudinal studies where the conditional distribution of each response Y_{ij} is described in terms of previous responses. Such models are known as *transition models*. The dependence among the repeated responses arises through the past responses which are viewed as influencing the one being modelled.

The specification of a conditional model depends on the researcher's beliefs as to how the current response might be associated with other responses. A general specification of a conditional regression model for the conditional expectation of the current outcome Y_{ij} in terms of all other outcomes and covariates is:

$$E(Y_{ij}|\{Y_{ik},\ k \neq j\}, \boldsymbol{X}_{ij}) = h^{-1}(\beta_0 + \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1 + \boldsymbol{Y}_{ij'}^T\,\boldsymbol{\gamma}), \quad i = 1, \ldots, K,\ \ j = 1, \ldots, N_i, \tag{2.27}$$

where $\boldsymbol{Y}_{ij'}$ denotes the vector of all outcomes except the one being modeled.

Due to the sequential nature of repeated responses, transition models have found application in longitudinal studies. The number of previous responses upon the current response is assumed to be associated with is called the *order* of the model. A transition model of order s is expressed as

$$E(Y_{ij}|Y_{ij-1}, \ldots, Y_{ij-s}, \boldsymbol{X}_{ij}) = h^{-1}\{\beta_{0s} + \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{1s} + \sum_{r=1}^{s}\gamma_r f_r(\boldsymbol{H}_{ij})\}, \tag{2.28}$$

where $\boldsymbol{H}_{ij} = (Y_{i1}, \ldots, Y_{ij-1})$ denotes the history of observed responses and $f_r(\boldsymbol{H}_{ij})$ denotes functions (often non-linear) of the history of previously observed responses. The notation β_{0s} and $\boldsymbol{\beta}_{1s}$ in equation (2.28) is used to emphasise that, in general, the value and interpretation of the regression coefficients depends on s.

For example, in a simple first-order stationary transition model the current response is modelled only in terms of the previous response
$(\sum_{r=1}^{s}\gamma_r f_r(\boldsymbol{H}_{ij}) = \gamma_1 f_1(Y_{ij-1}))$ and covariates. So, model (2.28) is simplified to

$$E(Y_{ij}|Y_{ij-1}, \boldsymbol{X}_{ij}) = h^{-1}(\beta_0 + \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1 + \gamma\, Y_{i,j-1}), \quad i = 1, \ldots, K,\ \ j = 1, \ldots, N_i.$$

Such a transition model is known as 'auto-regressive model of order 1'.

In fitting transition models such as the ones described by (2.28), past responses can be treated as additional explanatory variables. If terms $f_r(\boldsymbol{H}_{ij})$ do not depend

on $\boldsymbol{\beta}$, estimation simply proceeds as in standard GLMs for independent responses. Otherwise, estimation is feasible using a re-weighted least squares algorithm for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (Diggle et al., 2002, pg. 193). As conditional and transition models are not the main focus in this work, no further details are provided regarding the estimation methods.

There are certain limitations in the use of transition models and their application is somewhat limited compared to marginal and random effects models. In general, transition models have been developed for equally spaced responses; their generalisation in scenarios with non-equidistant responses or in scenarios with missing data is not straightforward (Fitzmaurice et al., 2009, pg. 21). Additionally, the regression coefficients are sensitive to assumptions about the time dependence (Fitzmaurice et al., 2009, pg. 21) and also their interpretation depends on the order of the serial dependence. A final concern is that, if a vector of covariates is known to be associated with the expected response at all times, then by conditioning on past responses the effect of covariates might be attenuated.

2.7 Comparison of approaches

The choice between one of the three modeling approaches depends on the scientific question to be answered. In marginal models, regression parameters acquire population-average interpretation, for random effects models interpretation specific to each cluster (i.e. conditional on random effects), while for conditional models interpretation conditional on other responses. Arguably, the choice between a marginal and a random effects approach is subtle (see, for example, Carriere and Bouyer, 2002). On the other hand, if interest lies in the direct effect of previous responses on the current one, a transition model would be the obvious choice. We next consider special cases where marginal regression parameters coincide with the corresponding ones from random effects or transition models.

When $h(.)$ is the identity function, it is noted (Diggle et al., 2002, pg. 132; Begg and Parides, 2003) that fixed effects parameters in random effects models also have a marginal interpretation. We write $E(Y_{ij} \mid \boldsymbol{X}_{ij}) = \beta_0^M + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_1^M$ and $E(Y_{ij} \mid \boldsymbol{X}_{ij}, \boldsymbol{b}_i) = \beta_0^{RE} + b_{0i} + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_1^{RE} + \boldsymbol{D}_{ij}^T \boldsymbol{b}_{1i}$ for the marginal and random effects models, respectively. Integrating over the distribution of $\boldsymbol{b}_i$ for the random effects model, $E(Y_{ij}|\boldsymbol{X}_{ij}) = E\{E(Y_{ij} \mid \boldsymbol{b}_i)\} = E(\beta_0 + b_{0i} + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_1^{RE} + \boldsymbol{D}_{ij}^T \boldsymbol{b}_{1i}) = \beta_0 + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_1^{RE}$ and it follows

directly that $\boldsymbol{\beta}^M = \boldsymbol{\beta}^{RE}$.

Similar equivalence in parameter interpretation between marginal and transition models for the link function other than the identity one does not exist in general (Diggle et al., 2002, pg. 134). Nevertheless, Diggle et al. (2002, pg. 133) showed that for a certain formulation of a transition model, the transition model regression parameters have the same interpretation as the marginal ones. They considered a model with an auto-regressive error structure:

$$Y_{ij} = \beta_0^{CO} + \beta_1^{CO} X_{ij} + \epsilon_{ij} \tag{2.29}$$

and

$$\epsilon_{ij} = \alpha \epsilon_{ij-1} + Z_{ij}, \quad Z_{ij} \sim N(0, \sigma^2). \tag{2.30}$$

Model (2.29) can be expressed as $Y_{ij} = \beta_0^{CO} + \beta_1^{CO} X_{ij} + \alpha(Y_{ij-1} - \beta_0^{CO} - \beta_1^{CO} X_{ij-1}) + Z_{ij}$, which describes a transition model with the current response depending on the previous one. From equations (2.29) and (2.30) it follows that $E(Y_{ij}|X_{ij}) = \beta_0^{CO} + \beta_1^{CO} X_{ij}$ and therefore the parameters of the particular transition model have a marginal interpretation as well.

For discrete responses, regression parameters in marginal, random effects and transition models are generally different. Consider, for example, the marginal and random intercept models, respectively, for logistic regression:

$$E(Y_{ij}|\boldsymbol{X}_{ij}) = \text{logit}^{-1}(\beta_0^M + \boldsymbol{\beta}_1^M \boldsymbol{X}_{ij}) \tag{2.31}$$

and

$$E(Y_{ij}|\boldsymbol{X}_{ij}, b_i) = \text{logit}^{-1}(\beta_0^{RE} + b_i + \boldsymbol{\beta}_1^{RE} X_{ij}), \quad b_i \sim \text{N}(0, \sigma_b^2). \tag{2.32}$$

The marginal expectation $E(Y_{ij}|\boldsymbol{X}_{ij})$ in the random intercepts model can be obtained by integrating over the distribution of the random effects: $E\{E(Y_{ij}|\boldsymbol{X}_{ij}, b_i)\} = E\{\text{logit}^{-1}(\beta_0^{RE} + b_i + \boldsymbol{\beta}_1^{RE} \boldsymbol{X}_{ij})\} \neq \text{logit}^{-1}(\beta_0^M + \beta_1^M \boldsymbol{X}_{ij})$. From the last expression, it follows that there is no equivalence between the marginal regression parameters and the fixed regression coefficients in a random intercepts model. The interpretation of the parameters in marginal and random effects logistic models differs as well: the first one is used to describe the ratio of population odds, while the second to describe the ratio of the odds for a specific cluster.

In the case of logistic regression with random intercepts, the relation between the parameters $\boldsymbol{\beta}^{RE}$ in model (2.32) and $\boldsymbol{\beta}^{\mathrm{M}}$ in model (2.31) has been established. Neuhaus et al. (1991) showed that $|\beta_k^{\mathrm{M}}| \leq |\beta_k^{\mathrm{RE}}|, k = 0, \ldots, q$. The equality holds only if a parameter's value is zero. Also, they proved that the difference between β_k^{M} and β_k^{RE} increases as the between-cluster variability (defined by σ_b^2) increases. Zeger et al. (1988) proved that the approximate relationship between $\boldsymbol{\beta}^{M}$ and $\boldsymbol{\beta}^{RE}$ is
$\boldsymbol{\beta}^{M} \approx (c^2\sigma_b^2 + 1)^{-1/2}(\boldsymbol{\beta}^{RE}), \;\; c = 16\sqrt{3}/15\pi.$

In models for count data (log-linear regression), marginal and random effects parameters can be equivalent in magnitude and interpretation in some cases (Diggle et al., 2002, pg. 137). The most important one is when a Poisson random intercepts regression model is considered: $E(Y_{ij}|\boldsymbol{X}_{ij}, b_i) = \mathrm{e}^{\beta_0 + b_i + \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1^{RE}}$, $b_i \sim \mathrm{N}(0, \sigma_b^2)$. The marginal expectation is obtained by integrating over the distribution of b_i:
$E(Y_{ij}|\boldsymbol{X}_{ij}) = \int \mathrm{e}^{\beta_0 + \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1^{RE}}\mathrm{e}^{b_i}\ \mathrm{d}b_i = \mathrm{e}^{\beta_0 + \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1^{RE}} \int \mathrm{e}^{b_i}\ \mathrm{d}b_i = \mathrm{e}^{\beta_0^{RE} + c + \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1^{RE}}$, where c is an additive constant coming from the integration. Comparing with the marginal Poisson model, $E(Y_{ij}|\boldsymbol{X}_{ij}) = \mathrm{e}^{\beta_0^{M} + \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1^{M}}$, it is inferred that apart from the intercept term, the other components of $\boldsymbol{\beta}^{M}$ and $\boldsymbol{\beta}^{RE}$ are equal.

2.8 Between- and within-cluster effects

An advantage, as well as a challenge in the analysis of clustered data, is the presence of cluster-varying (also termed within-cluster) covariates. In estimating the effect of a cluster-varying covariate, confounding due to cluster-level (observed or unobserved) characteristics might cause problems for the analysis. Standard methods for regression analysis imply models that relate the response with the covariates assuming (explicitly or implicitly) that the between- and within-cluster effects are equal (Neuhaus and Kalbfleisch, 1998; Mancl et al., 2000).

We initially consider models whose only random effects are random intercepts, i.e.

$$E(Y_{ij} \mid \boldsymbol{X}_{ij}, b_i) = h^{-1}(\beta_0 + b_i + \boldsymbol{\beta}_1^T\boldsymbol{X}_{ij}). \tag{2.33}$$

Under cluster-level confounding, a subset of $\boldsymbol{X}$ may be correlated with the random intercepts, violating a basic assumption of random effects models which states that covariates must be independent of the random effects. This violation is likely to jeopardise the validity of GLMMs. Research has focused on assessing the validity of

parameters estimates obtained from GLMMs under such violation and also on developing methods which are robust to this violation.

Neuhaus and Kalbfleisch (1998) proposed GLMMs which incorporate between- and within-cluster covariate effects by partitioning covariate $\boldsymbol{X}$ into the between-cluster component ($\bar{\boldsymbol{X}}_i = N_i^{-1}\sum_{j=1}^{N_i}\boldsymbol{X}_{ij}$) and the within-cluster component ($\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i$):

$$E(Y_{ij} \mid \boldsymbol{X}_{ij}, \bar{\boldsymbol{X}}_i, b_i) = h^{-1}\{\beta_0 + b_i + \boldsymbol{\beta}_B^T\bar{X}_i + \boldsymbol{\beta}_W^T(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)\}, \;\; b_i \sim N(0, \sigma_b^2). \quad (2.34)$$

The between-cluster effect, $\boldsymbol{\beta}_B$, corresponds to the difference in the expected response between two clusters whose average covariate value ($\bar{\boldsymbol{X}}_i$) differs by one unit. The within-cluster effect, $\boldsymbol{\beta}_W$, indicates that for a given cluster, the expected response increases by β_W units, for each unit increase in the deviation from the covariate mean value ($\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i$) within that cluster. Note that the partitioning of the effect of $\boldsymbol{X}$ into the between- and within-cluster components can also be applied in marginal models, leading to marginal rather than cluster specific inference.

Neuhaus and Kalbfleisch (1998) noted that fitting the standard random effects model (2.33) is equivalent to assuming that the between-cluster effects are the same as the within-cluster effects, i.e. $\boldsymbol{\beta}_B = \boldsymbol{\beta}_W$. When this assumption does not hold, it is known (Scott and Hold, 1982; Neuhaus and Kalbfleisch, 1998; Palta and Yao, 1991) that the regression coefficient $\boldsymbol{\beta}_1$ in (2.33) is a weighted average of the between- and within-cluster coefficients in (2.34). Therefore, when the between- and within-cluster effects are different, models which assume that these are the same, estimate neither the between-cluster nor the within-cluster coefficients.

There are two exceptions to the rule above. First, for a cluster-level covariate, since $\boldsymbol{X}_{ij} = \boldsymbol{X}_{ik} \;\forall\; j, k$ and $\bar{\boldsymbol{X}}_i = \boldsymbol{X}_{ij} \;\forall\; j$, $\boldsymbol{\beta}_1$ in model (2.33) corresponds to the between-cluster effect of $\boldsymbol{X}$. Second, if $\boldsymbol{X}$ is a 'designed' within-cluster covariate (i.e. a cluster-varying covariate where $\bar{\boldsymbol{X}}_i$ is the same for all clusters), $\boldsymbol{\beta}_1$ consistently estimates the within-cluster effect of $\boldsymbol{X}$.

Other methods have also been used for estimating the within-cluster effect of $\boldsymbol{X}$ under cluster-confounding. For canonical link functions Neuhaus and MCculloch (2006) suggested using the conditional likelihood method (conditioning on $\left(\sum_{j=1}^{N_i} Y_{ij}\right)$ to 'eliminate' the random intercepts). Conditional likelihood remains valid when the

covariates are correlated with the random intercepts, and provides consistent estimation of the within-cluster effect of $\boldsymbol{X}$ under the effect of observed or unobserved cluster-level confounders. Neuhaus and MCculloch (2006) describe the method of Neuhaus and Kalbfleisch (1998) as the 'poor man's' method, when compared to conditional likelihood. They note that the poor man's method provides estimates for the within-cluster effect of $\boldsymbol{X}$ with little or no bias when the link function is the identity or the logit link one. They say the poor man's method can be used as an alternative to conditional likelihood when the second cannot be used (e.g. when the link function is non-canonical). Note that conditional likelihood only uses within-cluster information to estimate the within-cluster effect of $\boldsymbol{X}$, therefore it eliminates the effects of any cluster-constant covariates in the model along with the with the effect of cluster-constant confounders. Neuhaus and MCculloch (2006) also investigate the performance of standard GLMMs when some covariates are not independent of the random intercepts. For the case of a single covariate ($\boldsymbol{X} = X$), they show (theoretically, for the identity link function and through simulations for the logit link function) that when X is correlated with the random intercepts, a GLMM provides inconsistent inference for the intercept term and the effect of X. They explain that this bias can be seen as arising because of misspecification of the distribution of random effects.

More recent work includes the conditional GEE (CGEE - Goetgeluk and Vansteelandt, 2008) approach. Goetgeluk and Vansteelandt (2008) propose a semi-parametric model of the form $E(Y_{ij} \mid \boldsymbol{X}_{ij}, \boldsymbol{V}_{ij}, \boldsymbol{S}_i, b_i) = h(\beta_0 + \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1 + \boldsymbol{V}_{ij}^T\boldsymbol{\gamma} + \boldsymbol{S}_i^T\boldsymbol{\delta} + b_i)$. The covariate of main interest is $\boldsymbol{X}$, while $\boldsymbol{V}$ and $\boldsymbol{S}$ are observed cluster-varying and cluster-constant confounders, respectively. The method can be used to remove cluster confounding due to observed confounders by including them in the model but also due to unobserved cluster-level confounders, b_i. Simply viewed, CGEE removes confounding due to unmeasured cluster-level characteristics by making within-cluster comparisons (as conditional likelihood) and fits the model of interest (at least for the identity link function) by using regression of change in the outcome to change in covariates. Goetgeluk and Vansteelandt (2008) showed that the poor man's method consistently estimates the within-cluster effect of $\boldsymbol{X}$ for the identity but not for log link function. CGEE, on the other hand, provides consistent estimation for the within-cluster effect of $\boldsymbol{X}$ under the effect of cluster-constant (observed or unobserved) and cluster-varying

observed confounders. CGEE provides consistent estimation of the effect of $\boldsymbol{X}$ for the identity and log link functions but cannot be applied when the link function is the logit one. Although the CGEE estimator solves GEE-type estimating equations, it provides a cluster-specific rather than marginal inference.

Brumback et al. (2010) provide a useful review of methods for the within-cluster effect of $\boldsymbol{X}$ when the responses are binary. They verify the findings of Neuhaus and MCculloch (2006) by showing through simulations that bias from the use of the poor man's method is small in the case of logit function. So, the method remains useful due to its ease of implementation but also because it allows for estimation of between-cluster effects which is not possible under the conditional likelihood approach.

In summary, CL and CGEE are alternatives to using the poor man's method when we wish to estimate the within-cluster effect of $\boldsymbol{X}$ in the presence of cluster-confounding. Both methods assume underlying random effect models in which the effect of $\boldsymbol{X}$ is the same in all clusters, i.e. the effect of $\boldsymbol{X}$ does not have a random part. For later reference we introduce the concept of a *homogeneous effect of* $\boldsymbol{X}$.

Definition 2.1 *The effect of* $\boldsymbol{X}$ *is said to be* ***homogeneous****, if* $\boldsymbol{\beta}_1$ *is the same in every cluster. Otherwise, the effect of* $\boldsymbol{X}$ *is said to be non-homogeneous.*

The GLMM in equation (2.33) assumes a homogeneous effect of $\boldsymbol{X}$. An example of a random effects model where the effect of $\boldsymbol{X}$ is non-homogeneous is:

$$E(Y_{ij} \mid \boldsymbol{X}_{ij}) = h^{-1}(\boldsymbol{\beta}_0 + b_{0i} + \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1 + \boldsymbol{X}_{ij}^T\boldsymbol{b}_{1i}). \tag{2.35}$$

In scenarios where the assumption of homogeneous effect of $\boldsymbol{X}$ is not true, CL and CGEE methods do not consistently estimate the within-cluster effect of $\boldsymbol{X}$. We discuss this issue further in Section 3.12.

2.9 Missing Data

Missing data arise in a repeated measurements context when for one or more experimental units, intended measurements have not been recorded for whatever reason. As a result, some sequences of measurements are incomplete. Missing data are a commonly encountered issue in repeated measures studies. Specifically, the term *dropout* is used to describe the scenario where the sequence of intended measurements for an

experimental unit is terminated prematurely. Dropouts are often referred to as *monotone missingness*. When intermediate intended measurements are missing, the terms *intermittently missing data* or *non-monotone missingness* are used. In the rest of this section, we exclusively concentrate on missingness in the response.

Methods of fitting regression models for clustered data were discussed in previous sections but no reference was made regarding their robustness to missing data. In fact, the presence of missing observations is often a decisive factor in selecting a method for analysis. This is because certain missing data mechanisms render some methods inappropriate. In the next section, we introduce the notation for missing data scenarios and a unified missing data framework, as it was outlined by Rubin (1976) and Little and Rubin (1987).

2.9.1 Notation and Definitions

A simple random sample of K clusters is drawn from a population of clusters. We assume that all complete clusters are of the same size, which we denote by N_{comp}. Whenever the size of a sampled cluster, N, is less than N_{comp} then there are further $N_{\text{comp}} - N$ members who are missing, i.e. on whom Y and $\boldsymbol{X}$ are not observed. Depending on the application it may be possible to index the members of each complete cluster; when not possible we imagine an arbitrary indexing is applied.

Let Y_j and $\boldsymbol{X}_j$ denote the outcome and covariate vector for member j in the complete cluster. Let $\widetilde{\boldsymbol{Y}}^* = (Y_1, \ldots, Y_{N_{\text{comp}}})^T$ and $\widetilde{\boldsymbol{X}}^* = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{N_{\text{comp}}})$. We shall additionally use subscript i where necessary to denote the cluster $(i = 1, \ldots, K)$.

The random process that determines which of the N_{comp} members are missing is called the missing data mechanism (MDM). For each cluster let $\boldsymbol{R} = (R_1, \ldots, R_{N_{\text{comp}}})^T$, where $R_j = 1$ if member j is observed and $R_j = 0$ otherwise. Let the observed and missing parts of $\widetilde{\boldsymbol{Y}}^*$ and $\widetilde{\boldsymbol{X}}^*$ be denoted by $\widetilde{\boldsymbol{Y}}^*_{(\boldsymbol{R})}$, $\widetilde{\boldsymbol{Y}}^*_{(\bar{\boldsymbol{R}})}$, and $\widetilde{\boldsymbol{X}}^*_{(\boldsymbol{R})}$, $\widetilde{\boldsymbol{X}}^*_{(\bar{\boldsymbol{R}})}$ respectively.

The classification of missing data mechanisms by Rubin (1976) and Little and Rubin (1987) is fundamental to the analysis of incomplete data because is independent of the statistical framework used to analyse the data. The data are said to be missing completely at random (MCAR) if $P(\boldsymbol{R} = \boldsymbol{r} \mid \widetilde{\boldsymbol{X}}^*, \widetilde{\boldsymbol{Y}}^*) = P(\boldsymbol{R} = \boldsymbol{r}) \ \forall \ \boldsymbol{r}$; they are covariate-dependent MCAR if $P(\boldsymbol{R} = \boldsymbol{r} \mid \widetilde{\boldsymbol{X}}^*, \widetilde{\boldsymbol{Y}}^*) = P(\boldsymbol{R} = \boldsymbol{r} \mid \widetilde{\boldsymbol{X}}^*) \ \forall \ \boldsymbol{r}$.

They are missing at random (MAR) if $P(\boldsymbol{R} = \boldsymbol{r} \mid \widetilde{\boldsymbol{X}}^*, \widetilde{\boldsymbol{Y}}^*) = \pi(\boldsymbol{r}, \widetilde{\boldsymbol{X}}^*_{(\boldsymbol{r})}, \widetilde{\boldsymbol{Y}}^*_{(\boldsymbol{r})}) \ \forall \ \boldsymbol{r}$ for some function $\pi(.)$; they are covariate-dependent MAR if $P(\boldsymbol{R} = \boldsymbol{r} \mid \widetilde{\boldsymbol{X}}^*, \widetilde{\boldsymbol{Y}}^*) = \pi(\boldsymbol{r}, \widetilde{\boldsymbol{X}}^*, \widetilde{\boldsymbol{Y}}^*_{(\boldsymbol{r})}) \ \forall \ \boldsymbol{r}$. Note that MCAR is a special case of MAR, and MAR is a special case of covariate-dependent MAR. If the data are not MAR, then they are said to be missing not at random (MNAR).

The property of making valid inferences (by using an estimation method) about the measurements process, without explicitly dealing with the missingness process, is called *ignorability*. Whether the missing data structure is ignorable or not, depends on the chosen method of analysis. In the next sections, we discuss methods for analysing datasets with missing data.

We may wish to investigate how various covariates are related to the outcome. A marginal model can be specified

$$E(Y_{ij} \mid \boldsymbol{X}_{ij}) = h^{-1}(\beta_0 + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_1) \tag{2.36}$$

where $h(.)$ is a link function and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$ is a vector of parameters we wish to estimate. Alternatively, a cluster-specific model can be specified

$$E(Y_{ij} \mid \boldsymbol{X}_{ij}, \boldsymbol{b}_i) = h^{-1}(\beta_0 + b_{0i} + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_1 + \boldsymbol{D}_{ij}^T \boldsymbol{b}_{1i}) \tag{2.37}$$

where $\boldsymbol{b}_i$ is a vector of random effects following a zero mean multivariate distribution with variance matrix $\boldsymbol{G}$ and $\boldsymbol{D}_{ij}$ is a q_b-dimensional subset of $\boldsymbol{X}_{ij}$. The parameters of main interest are the vector of regression coefficients, $\boldsymbol{\beta}$ and the variance matrix, $\boldsymbol{G}$.

2.9.2 Simple methods for missing data

In this section, we briefly review a few simple approaches, the majority of which is suitable for cases where the measurement and missingness process are independent and their parameters are separated. These methods are usually easy to implement and can be used when data are MCAR. Although simple, they suffer from serious drawbacks and wide use of them is considered bad practice.

The easiest approach to deal with clustered incomplete data is to retain those clusters for which all intended measurements have been obtained and discard all clusters with missing data. This approach is termed *Complete Case Analysis* (CCA). The method is easy to implement as the resulting dataset has the same data structure as the

hypothetical complete one and therefore any statistical software can be used for analysis. However, CCA suffers from important drawbacks. An important disadvantage is the potentially substantial efficiency loss, since a lot of useful information is discarded along with all the incomplete clusters. In the best case, if the missingness process is unrelated to the measurement process then a complete case analysis will simply be a waste of data (ethical issues may arise as well). In the worst, if the missingness mechanism is not MCAR, this approach yields inconsistent parameter estimates.

The second category involves *imputation methods* which, instead of deleting incomplete clusters, fill in the missing values to produce a 'completed' dataset. One can distinguish two types of imputation: single and multiple. In single imputation each missing value is filled in once and a single completed dataset is constructed. In multiple imputation each missing value is filled in more than once to produce a number of imputed datasets. Parameter estimates are then obtained by combining the estimates from each dataset. In this section, we concentrate on single imputation approaches.

The simplest imputation-based technique is *Last Observation Carried Forward* (LOCF) and is more suited for monotone missingness. It suggests extrapolating the last observed measurement for a cluster until filling in the intended number of measurements. A possible improvement would be to estimate the 'trend' for each cluster and then complete the cluster not by extrapolating a constant value, but by using the predicted value from the estimated trend. Another simple imputation technique is *unconditional mean imputation* where a missing value of a variable in a given cluster is filled in by the average of the observed values of the same variable from other clusters. It does not use information on a given cluster to complete its missing values.

Other single imputations methods also exist. For example 'hot deck' imputation uses 'matching' to fill in the missing values in a cluster by using values from a similar complete one. Most of the single imputation methods are only valid under the strong assumption that data are MCAR. Also, even if they provide consistent point estimates, they routinely fail to provide consistent precision estimators. In fact, the variance of the estimates is usually underestimated because the artificially filled-in values are treated as if they were observed.

2.9.3 Methods for MAR

Here we provide an overview of the methods which provide valid inference when data are MAR. We distinguish three main categories: likelihood-based methods, imputation-based methods and modifications of the GEE, focusing on inverse probability weighting (IPW) methods. Most of the methods subsequently presented are more suited to monotone missingness. Whenever a method, or a modification of it, might also be applicable in scenarios of non-monotone missingness, this will be indicated. Methods specifically designed for non-monotone missingness, irregularly spaced observations and 'outcome dependent follow-up' have also been developed. They are beyond the scope of this thesis though and are not further considered.

Likelihood-based methods

Inference for complete data using likelihood based estimation only requires a correctly specified model for the joint distribution of the responses. For incomplete data, a model for the joint distribution of $\widetilde{\boldsymbol{Y}}^*$ and $\boldsymbol{R}$ is typically required. One possible factorisation for the likelihood contribution of the ith cluster is

$$f(\widetilde{\boldsymbol{Y}}_i^*, \boldsymbol{R}_i \mid \widetilde{\boldsymbol{X}}_i^*, \boldsymbol{\gamma}, \boldsymbol{\phi}) = f(\widetilde{\boldsymbol{Y}}_i^* \mid \widetilde{\boldsymbol{X}}_i^*, \boldsymbol{\gamma}) f(\boldsymbol{R}_i | \widetilde{\boldsymbol{Y}}_i^*, \widetilde{\boldsymbol{X}}_i^*, \boldsymbol{\phi}). \tag{2.38}$$

Often, little is known about the process leading to missing values. Inferences about $\boldsymbol{\gamma}$, which are the parameters of substantive interest, are generally sensitive to the assumptions in the model for $\boldsymbol{R}$, even if the joint distribution of $\widetilde{\boldsymbol{Y}}^*$ is correctly specified.

In likelihood-based inference the missingness is *ignorable* under a MAR missing mechanism, i.e. $f(\boldsymbol{R}_i \mid \widetilde{\boldsymbol{Y}}_i^*, \widetilde{\boldsymbol{X}}_i^*, \boldsymbol{\phi}) = f(\boldsymbol{R}_i \mid \widetilde{\boldsymbol{Y}}_{(\boldsymbol{R})i}^*, \widetilde{\boldsymbol{X}}_{(\boldsymbol{R})i}^*, \boldsymbol{\phi})$. Then the joint distribution of the responses for the observed data in cluster i can be expressed as

$$f(\widetilde{\boldsymbol{Y}}_{(\boldsymbol{R})i}^*, \boldsymbol{R}_i \mid \widetilde{\boldsymbol{X}}_{(\boldsymbol{R})i}^*, \boldsymbol{\gamma}, \boldsymbol{\phi}) = f(\widetilde{\boldsymbol{Y}}_{(\boldsymbol{R})i}^* \mid \widetilde{\boldsymbol{X}}_{(\boldsymbol{R})i}^*, \boldsymbol{\gamma}) f(\boldsymbol{R}_i | \widetilde{\boldsymbol{Y}}_{(\boldsymbol{R})i}^*, \widetilde{\boldsymbol{X}}_{(\boldsymbol{R})i}^*, \boldsymbol{\phi}). \tag{2.39}$$

Inferences can be based only on the first component, $f(\widetilde{\boldsymbol{Y}}_{(\boldsymbol{R})i}^* \mid \widetilde{\boldsymbol{X}}_{(\boldsymbol{R})i}^*, \boldsymbol{\gamma})$, of the right hand side of (2.39). So, in practise, when the missingness is ignorable, maximum likelihood estimates can be obtained by maximising the likelihood function corresponding to the distribution of the observed data.

Apart from the assumption that data are MAR, it is also implicitly assumed that the parameters, $\boldsymbol{\gamma}$, of the measurement process and, $\boldsymbol{\phi}$, of the missingness process are separately parameterised (separability). Fitzmaurice et al. (2009, pg. 411) argue

that the assumption of separability is less important (than the MAR assumption) and often reasonable in applications. Also, Diggle et al. (2002, pg. 284) note that if the parameters are functionally dependent, ignoring the second term in the right hand side of equation (2.39) might lead to an efficiency loss.

Ignorability under MAR in likelihood-based inference is an important result. All likelihood-based methods for marginal models and random effects models result in valid inference, provided that the model for $f(\widetilde{\boldsymbol{Y}}^*_{(\boldsymbol{R})i} \mid \widetilde{\boldsymbol{X}}^*_{(\boldsymbol{R})i}, \boldsymbol{\gamma})$ is correctly specified; no modelling of the missingness process is required. Likelihood-based methods can be applied in scenarios of monotone and non-monotone missingness.

As it was seen in Section (2.4.2), maximum likelihood estimation is generally unattractive for marginal inference when dealing with non-Gaussian outcomes because of the complexities in formulating the joint distribution of the responses in each cluster. To circumvent this limitation, methods which do not require specification of the full distribution of the responses have been developed to provide valid inference under MAR. Some of these methods are reviewed below.

Multiple imputation

In multiple imputation, each missing value is imputed M times (instead of 1) to reflect the uncertainty in the value. Multiple imputation was introduced by Rubin (1987) and consists of three main steps. First, an imputation model is assumed and each missing value is replaced by M predicted ones by drawing samples from the predictive distribution of the missing data given the observed. So, M imputed datasets are constructed. Second, each imputed dataset is analysed using standard statistical methods (e.g. random effect models, GEE), as if the data were complete. Thus, each dataset gives rise to a set of point and precision estimates. Third, the estimates from the M datasets are appropriately combined using Rubin's rules (Rubin, 1987) into a single estimate for the parameters of interest and corresponding standard errors. If the imputation model is correctly specified, multiple imputation is valid (i.e. provides consistent parameter estimation) when the data are MAR (implementations of MI also exist under MNAR assumptions but these are not frequently used).

The most important part of the imputation procedure is the first step, i.e. producing the imputed responses. The underlying idea is to impute the unobserved responses from the conditional distribution of the missing responses given the observed ones us-

ing an imputation model. This model can be quite flexible, in the sense that it may include covariates which are not included in the main analysis model (auxiliary covariates). As far as the number of imputation covariates does not affect the stability of the estimates, all covariates which are envisaged making the MAR assumption plausible can be included in the imputation model. Two main methods have been proposed for imputing the unobserved responses: the multivariate Normal approach (MVNI) and the full conditional specification, often termed the 'Chained Equations' approach.

Schafer (1997) suggested joint imputation models for $P(\widetilde{\boldsymbol{Y}}^*, \widetilde{\boldsymbol{X}}^*)$ where the variables in the model are assumed to follow a multivariate Normal distribution. MVNI comes with an elegant theoretical justification and can be applied for monotone and non-monotone missing patterns. Clearly, the assumption of multivariate normality is not always plausible. However, Schafer (1997) seems to suggest that even if the multivariate Normal assumption is clearly false, MVNI can still provide valid inference. It is not clear whether this statement is true since there exist mixed results suggesting that MVNI may or may not provide valid inferences when the assumption of multivariate normality does not hold (e.g. van Buuren et al., 2006; Yu et al., 2007; Lee and Carlin, 2010).

Chained equations (van Buuren, 2000, 2007) impute the missing responses variable-by variable, by assuming a conditionally specified imputation model for the variable being imputed conditional on all other variables. For longitudinal data, chained equations can be applied by considering the responses at different (discrete) time points as different variables to be imputed. For example, let Y_j be the response at time point j. Y_1 is first regressed on all responses at other time points and covariates (including auxiliary ones). The missing values in Y_1 are filled by sampling from the posterior predictive distribution of Y_1. Then Y_2 is imputed by specifying a suitable model for a Y_2 in terms of all other variables and using the imputed values for Y_1. This sequential imputation of variables continues until missing values for all variables are imputed. This procedure consists a 'cycle' and is repeated several times (say 10 to 20) to stabilize the results. Chained equations can be applied for monotone and non-monotone missing data, but imputing multilevel data is generally not straightforward (White et al., 2011).

The most attractive feature of chained equations is that there is no need to specify multivariate density for $P(\widetilde{\boldsymbol{Y}}^*, \widetilde{\boldsymbol{X}}^*)$. Instead, a separate imputation model is specified

according to the nature of the variable to be imputed (e.g. linear model for a continuous variable, logistic model for a binary variable etc). The method has been criticised of lack of solid theoretical background (Goldstein et al., 2009). In particular, the conditional distributions can be incompatible in the sense that no multivariate distribution exists which yields the specified conditional distributions. Nevertheless, the method has been seen to perform well in terms of bias and coverage for a number of occasions through simulations and also to perform adequately in applications when compared to MVNI (e.g. Lee and Carlin, 2010).

Modified GEE, IPW and Doubly Robust methods

While GEE are known to provide valid inferences under the stringent MCAR assumption, the robustness of the method under the weaker MAR is questionable. Research has focused on extensions of GEE where the objective is the validity of GEE under MAR assumptions.

The 'modified' GEE approach, is essentially the standard GEE approach complemented by careful modelling of the working correlation structure. It is known that moment estimation of correlation parameters based on available pairs is generally inefficient. For binary data, Lipsitz et al. (2000) and Fitzmaurice et al. (2001) use 'Gaussian estimation' and 'conditional residuals' respectively for the estimation of the correlation parameters. Both methods perform equally well and the authors' epitome is that provided that the correlation structure is correctly specified and the correlation parameters are correctly estimated, the modified GEE result in negligible bias. More recently, Copas and Seaman (2010) report that the modified GEE can be substantially biased in certain scenarios and therefore should be used with caution. Also, Seaman and Copas (2009) report that the GEE using Gaussian estimation and conditional residuals did not perform any better (in their data illustration) in terms of bias, coverage and mean square error compared to GEE using available pairs. Often, the missingness process might depend on past response history but also on the measurements of an auxiliary process. A limitation of the modified GEE is that such auxiliary information cannot be readily utilised.

A more flexible modification of the GEE involves inverse probability of weighting methods (IPW). Robins et al. (1995) proposed Weighted GEE (WGEE) which use inverse probability of observation weights to provide valid inference for random

dropouts. In WGEE, each subject is inversely weighted by the probability of being observed at that particular measurement occasion. The underlying idea is to up-weight measurements with a small probability of being observed, so as to compensate for the missing measurements from similar individuals who actually dropped out.

Similarly to equations (2.14), the WGEE are

$$\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1} \boldsymbol{\Delta}_i(\boldsymbol{\alpha})(\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{2.40}$$

where $\boldsymbol{\Delta}_i(\boldsymbol{\alpha}) = \text{diag}(R_{i1}/\psi_{i1}, \ldots, R_{i\text{Nmax}}/\psi_{iN_{\text{comp}}})$ and ψ_{ij} is the probability of the response measurement on individual i being observed at time j. Let $\boldsymbol{H}_{i,j-1}$ denote the complete history up to time $j-1$. $\boldsymbol{H}_{i,j-1}$ may include observed responses, observed regression covariates and observed auxiliary variables which are not included in the regression model for outcome; all up to time $j-1$. The probability of the response measurement on individual i being observed at time j, $\psi_{ij} = \psi_{ij}(\boldsymbol{\alpha}) = p(R_{ij} = 1 \mid \boldsymbol{H}_{i,j-1})$ is estimated from an assumed dropout model:

$$\lambda_{ij}(\boldsymbol{\alpha}) = P(R_{ij} = 0 | R_{i,j-1} = 1, \boldsymbol{H}_{i,j-1}) \tag{2.41}$$

and $\psi_{ij} = (1 - \lambda_{i1}) \times \ldots \times (1 - \lambda_{ij})$. The parameters, $\boldsymbol{\alpha}$, of the dropout model are estimated using logistic regressions. When the dropout model is correctly specified and data are MAR, the solution of equations (2.40) is a consistent estimate of $\boldsymbol{\beta}$.

Doubly robust approaches have also been proposed (Bang and Robins, 2005; Seaman and Copas, 2009). These methods combine WGEE and imputation methods and require specification of both a dropout model and an imputation model. They are called doubly robust or doubly protected because they provide consistent estimation if at least one of the dropout and imputation models is correctly specified.

Recently, Diggle et al. (2007) (see also Farewell, 2010) proposed a challenging new framework of analysis of longitudinal data subject to dropout. They proposed methods where the objective of analysis is to make inferences for the longitudinal features of the hypothetical population had nobody dropped out. This method is beyond the scope of this work and is not further considered.

2.9.4 Methods for MNAR

When data are MNAR, the missingness process is non-ignorable in the sense that the missingness structure must not be ignored when making inferences about the measure-

ment process. Under MNAR, methods which are designed to provide valid inference under MAR assumptions can yield severely biased results. Since the measurement and missingness process are not independent their joint distribution is needed. One difficulty is that the observed data alone cannot verify or reject assumptions about the missing data. Nevertheless, there is an active course of research towards the development of methods for non-ignorable missingness.

Two main categories of models have been used; selection models and pattern mixture models. These approaches mainly differ in the factorisation used to specify the joint distribution of $\widetilde{\boldsymbol{Y}}_i^*$ and $\boldsymbol{R}_i$. Also, shared parameters models have been proposed and these may be used in a selection or pattern-mixture framework. We overview these approaches in the next few paragraphs.

Selection Models

Selection models use the factorisation

$$f(\widetilde{\boldsymbol{Y}}_i^*, \boldsymbol{R}_i \mid \widetilde{\boldsymbol{X}}_i^*, \boldsymbol{\gamma}, \boldsymbol{\phi}) = f(\widetilde{\boldsymbol{Y}}_i^* \mid \widetilde{\boldsymbol{X}}_i^*, \boldsymbol{\gamma}) f(\boldsymbol{R}_i | \widetilde{\boldsymbol{Y}}_i^*, \widetilde{\boldsymbol{X}}_i^*, \boldsymbol{\phi}). \tag{2.42}$$

This is a natural way of factorising the joint distribution of $\widetilde{\boldsymbol{Y}}_i^*$ and $\boldsymbol{R}_i$ into two components. The first, $f(\widetilde{\boldsymbol{Y}}_i^* \mid \widetilde{\boldsymbol{X}}_i^*, \boldsymbol{\gamma})$, is the model for $\widetilde{\boldsymbol{Y}}_i^*$ if data were not missing. The second, $f(\boldsymbol{R}_i \mid \widetilde{\boldsymbol{Y}}_i^*, \widetilde{\boldsymbol{X}}_i^*, \boldsymbol{\phi})$, is the model for the missing data which determines which parts of $\widetilde{\boldsymbol{Y}}_i^*$ are missing. The model for the missing data also imposes an assumption on the missing data mechanism.

The joint likelihood of $\widetilde{\boldsymbol{Y}}_i^*$ and $\boldsymbol{R}_i$ is specified and the selection model is fitted using an iterative procedure, such as the EM algorithm. Term $\boldsymbol{\gamma}$ which usually contains the parameters of main interest is obtained directly from the selection model analysis. Considering the case of dropout, if MNAR is assumed, the dropout model for the probability of a subject dropping out at a given occasion given that was in the study in the previous occasion specifies dependence on the *current* and previous outcomes (and possibly on other covariates). If MAR is assumed instead, then the dropout model specifies dependence only on previously observed outcomes (and possibly covariates). Then selection modelling leads directly to an ignorable likelihood analysis since inference can be based solely on the first component of equation (2.42).

A source of criticism for selection models is that distributional assumptions about the missing data cannot be verified (Little, 1995; Little and Rubin, 2002). This issue

is addressed to some extent by Kenward (1998) who suggested performing sensitivity analysis by varying the assumptions about the missingness model.

Pattern-mixture models

Pattern-mixture models (Little, 1993, 1994) use the factorisation

$$f(\widetilde{\boldsymbol{Y}}_i^*, \boldsymbol{R}_i \mid \widetilde{\boldsymbol{X}}_i^*, \boldsymbol{\delta}, \boldsymbol{\nu},) = f(\boldsymbol{R}_i \mid \widetilde{\boldsymbol{X}}_i^*, \boldsymbol{\delta}) f(\widetilde{\boldsymbol{Y}}_i^* \mid \widetilde{\boldsymbol{X}}_i^*, \boldsymbol{R}_i, \boldsymbol{\nu}). \tag{2.43}$$

This factorisation of the joint likelihood makes the application of pattern-mixture models appealing under MNAR assumptions. By their construction, pattern-mixture models are under-identified (or over-specified). Therefore, assumptions or restrictions should be made to ensure identifiability. Although this feature could be considered as a drawback of the method, it could also be considered as a requirement to make the assumptions made about missingness explicit.

Molenberghs and Verbeke (2006, Chapters 18 and 20) suggest three possible strategies to deal with the under-identifiability of the model. The first strategy is to set identifiability restrictions from the beginning and these can be seen as analogous to the assumptions on the missing data process, when using selection models. Three special cases are briefly considered. The simplest identification used is the Complete Case Missing Values (CCMV), where unavailable information is always borrowed from the completers. In the second identification, unavailable information is obtained from the nearest identified pattern and is termed Neighbouring Case Missing Values (NCMV). The third case is Available Case Missing Values (ACMV), which is equivalent (Molenberghs and Verbeke, 2006, pg. 334) to the MAR assumption used in selection modelling.

The second and third strategies are simpler. The subjects are firstly divided into patterns according to their time of dropout (or their missing data pattern more generally). In Strategy 2, model simplification is achieved by fitting a separate model in each pattern. In Strategy 3, the pattern indicator is included in the model as an additional covariate. We note that in Strategies 2 and 3, untestable assumptions are made about missingness. In Strategy 1, the assumptions about missingness (as identifiability restrictions) are made clear from the beginning. Technical details on implementing these strategies in practice and choosing between them can be found in Molenberghs and Verbeke (2006, Chapter 20); these are not discussed further in the context of this work.

Shared parameters models

Shared parameters models is another category of methods used for MNAR missingness. These combine the usual random effects regression model with a model for missing data. The two models are linked through shared random effects (Hogan and Laird, 1997; De Gruttola and Tu, 1994; Wu and Carrol, 1988). Suppressing the dependence on covariates and parameters, shared random effects models use the factorisation $f(\widetilde{\boldsymbol{Y}}_i^*, \boldsymbol{R}_i \mid \boldsymbol{b}_i) = f(\widetilde{\boldsymbol{Y}}_i^* \mid \boldsymbol{b}_i) f(\boldsymbol{R}_i \mid \boldsymbol{b}_i) f(\boldsymbol{b}_i)$, where $\boldsymbol{b}_i$ is a vector of shared random effects. Conditional on the shared random effects, the models for the main outcome and missingness are assumed to be independent. This factorisation is analogous to the one used in selection modelling. The joint distribution of $\widetilde{\boldsymbol{Y}}_i^*$ and $\boldsymbol{R}_i$ is obtained by integrating over the distribution of the shared random effects. Parameter estimates are obtained by maximising the log-likelihood function using the EM algorithm.

Similar principles can be applied in pattern-mixture models, using the corresponding factorisation for the joint likelihood. For pattern-mixture models, over-parameterisation issue can also be circumvented by considering the pattern-specific parameters as nuisance and treating them as random (Guo et al., 2004).

2.10 Discussion

In this chapter we have introduced marginal, random effects and transition models for the analysis of clustered data. We have discussed estimation methods and interpretation of regression parameters in each approach.

One important issue in the analysis of clustered data is missing data. It has been made clear that missing data, and in particular the assumptions about the missingness mechanism, are decisive factors regarding the method of analysis to be chosen.

In the next two chapters, we discuss informative cluster size and informative covariate structure. We shall often refer to specific sections of the current chapter when discussing adaptations of existing methods to deal with informative cluster size and informative covariate structure. We shall discuss missing data and related estimation methods again in Chapter 5 where we contrast informative cluster size to missing data.

Chapter 3

Informative cluster size and covariate structure

3.1 Introduction

In many medical research and audit studies, the data are clustered; for example repeated measurements of health status obtained on patients, each one relating to a clinical episode. For each member (clinical episode) of a cluster (patient) an outcome (health status) and a set of covariates (e.g. treatment) are measured. We may wish to investigate how the various covariates are related to the outcome by using a marginal regression model. The cluster size (number of clinical episodes) may vary between clusters. When the cluster size varies, we might be interested in two types of inference. In the first, the variability in size is an inherent feature of the data, rather than arising because some of the data are missing. So, the observed clusters are complete and we seek to make inference for the 'observed clusters'. In the second, the variation in cluster size is viewed as arising through missing data, i.e. some clusters are smaller because some of their members were not observed. Therefore, the observed clusters are viewed as incomplete and we seek inference for the 'complete clusters'.

In this chapter, we are concerned about inference for the observed clusters. When the cluster size varies, informative cluster size has been defined to arise when the outcome is conditionally associated with the cluster size given the covariates, i.e. $E(Y \mid \boldsymbol{X}, N) \neq E(Y \mid \boldsymbol{X})$. We see that informative cluster size can only arise when 1) for scientific reasons the cluster size is not included as a covariate in the regression model, and 2) the link between the outcome and cluster size does not only

arise because both are linked to the covariates. We view $\boldsymbol{X}$ as the vector of covariates of main interest (primary predictors), while N is regarded as a nuisance variable.

When the cluster size is informative Hoffman et al. (2001) and Williamson et al. (2003) suggest that there are two marginal analyses of possible interest. Williamson et al. (2003) describe this choice as being between inference for the population of all cluster members and inference for a typical member of a typical cluster. For example, in an audit of clinic consultations to examine resource use, inference for the population of all cluster members (i.e. consultations) might be preferred, as clustering by patient may not be of inherent interest. Conversely in a study of disease progression, inference for a typical member of a typical cluster (i.e. a typical episode for a typical patient) might be of more interest.

We define *informative covariate structure* to arise when the conditional expectation of the outcome for a particular member given the covariate values of that member and the cluster size depends on the values of the covariates of other members of the same cluster. Informative covariate structure is a generalisation to variable cluster size of the situation studied by Pepe and Anderson (1994). When cluster size is constant, they recommended the use of GEE with independence working correlation in this situation. In the current work, the type of informative covariate structure we shall most closely consider, is when the expected outcome for a member in a cluster is associated with the number of members in that cluster where the covariates take certain values.

The method of generalised estimating equations (Liang and Zeger, 1986) is widely used for the marginal regression analysis of clustered data, due to its robustness against the misspecification of correlation structure and its relative ease of use. Under informative cluster size, inference for the population of all cluster members may be obtained through the application of standard GEE if independence is selected as the working correlation. To provide inference for a typical member of a typical cluster two methods were first proposed: the within cluster resampling method termed WCR (Hoffman et al., 2001) and the inversely weighted by cluster size GEE (CWGEE-Williamson et al., 2003) also termed WIEE, since it uses the independence working correlation. Recently, another potentially more efficient method was proposed by Chiang and Lee (2008), based on an extension of the WCR method. However, these papers only considered in detail simple cases of informative cluster size, in the sense that the outcome

depends only on cluster size and covariates and not on the interaction between these. These papers also primarily considered covariates with simple distributions e.g. cluster constant.

Other authors (Dunson et al., 2003; Gueorguieva, 2005; Chen et al., 2011) have developed approaches based on jointly modelling the cluster size and the outcome measurements. These methods are more complex and do not address marginal regression. Alternatively, joint models for the outcome measurements and the cluster size can be fitted in the special case where it is known when episodes could in principle occur and covariate values at these times are available (e.g. Su et al., 2009). In this chapter, though, we consider more general settings. We also assume that interest does not lie in modelling the cluster size itself.

In the next section, we introduce the notation for this chapter and provide important definitions. In Section 3.3 we present popular examples of informative cluster size problems encountered in the literature. In Section 3.4 we explain why informative cluster size causes problems to analysis and clarify that simple methods such as including cluster size alongside the primary predictors in a regression model for the expected outcome are not appropriate in general. We also explain the limitations of an alternative approach which assumes a model for the expected outcome in terms of $\boldsymbol{X}$ and N but then obtains the marginal effect of $\boldsymbol{X}$ by marginalising over the distribution of N. We next present and discuss the current methodology for marginal and cluster-specific inference under informative cluster size. In Section 3.6 we formally define informative covariate structure, additional populations for inference and provide guidance concerning the choice of population for inference. We also propose simple adaptations of the WIEE to provide unbiased inference for these populations. We comment on possible strategies for selecting an analysis method when informative cluster size or covariate structure are possible in Section 3.7. In Section 3.8 we highlight differences between populations and in Section 3.9 we present simulation studies to assess the performance of the WIEE. In Section 3.10 we apply the WIEE method to data on AIDS related conditions from the Delta trial of HIV treatment. In Section 3.11 we bring into attention another recent approach (Huang and Leroux, 2011), which relates to informative covariate structure. We discuss issues on the practical application of this approach and identify areas for future work in Section 3.12. Finally we discuss our findings.

3.2 Informative cluster size: notation and definitions

In this section we briefly remind the reader of the notation used in settings with clustered data. We also formally define concepts that will appear frequently in the current and the chapters to follow.

Suppose that an independent random sample of clusters is drawn from a population of clusters. On each member of each cluster an outcome Y and a q-dimensional covariate vector $\boldsymbol{X}$ are measured. For each cluster, let N denote the number of members in the cluster, $\boldsymbol{Y}^*$ the vector whose jth element is the jth member's value of Y, and $\boldsymbol{X}^*$ the $N \times q$ matrix of covariate values.

When the cluster size varies two populations of members were initially proposed by Williamson et al. (2003).

Definition 3.1

- *The **population of all members (M)** consists of all members of all clusters in the population.*
- *The **population of typical cluster members 1 (C1)** is the subpopulation of population M in which each cluster contributes a single member at random. Thus, the probability that each member is contributed to the population is N^{-1}. This population provides inference for a typical member of a typical cluster (Williamson et al., 2003).*

Conditional and unconditional expectations of Y may differ across the two populations above. In this work, the expectation notation $E^p(.)$ refers to the population p (p =M, C1), and $E(.) = E^M(.)$ refers to the population of all members. For population M we write $E(Y \mid \boldsymbol{X}) = E_{N|\boldsymbol{X}} E_{Y|\boldsymbol{X},N}(Y)$ and for population C1, $E^{C1}(Y \mid \boldsymbol{X}) = E_{N|\boldsymbol{X}}[\frac{1}{N} E_{Y|\boldsymbol{X},N}(Y)]/E_{N|\boldsymbol{X}}(\frac{1}{N})$.

Definition 3.2 *Cluster size is non-informative if*

$$E(Y \mid \boldsymbol{X} = \boldsymbol{x}, N = n) = E(Y \mid \boldsymbol{X} = \boldsymbol{x}) \; \forall \; \boldsymbol{x}, n. \tag{3.1}$$

Otherwise cluster size is informative.

This definition was given by previous authors (Hoffman et al., 2001; Williamson et al., 2003; Benhin et al., 2005; Chiang and Lee, 2008). We define $\mu^p(\boldsymbol{x}) = E^p(Y|\boldsymbol{X} = \boldsymbol{x})$ where $p =$ M, C1 according to which population is considered.

Another important concept in which we shall often refer to relates to whether or not the distribution of covariates is associated with the distribution of the cluster sizes.

Definition 3.3 *$\boldsymbol{X}$ is said to be cluster-size balanced (or size-balanced) if*

$$f(\boldsymbol{X}_{J_1} \mid N = n_1) = f(\boldsymbol{X}_{J_2} \mid N = n_2) \tag{3.2}$$

for all n_1 and n_2 such that $P(N = n_1) > 0$ and $P(N = n_2) > 0$, where J_1 and J_2 are independent uniform random variables on $\{1, \ldots, n_1\}$ and $\{1, \ldots, n_2\}$ and $f(A)$ denotes the probability density function for a generic random variable A. Otherwise, $\boldsymbol{X}$ is said to be non-cluster-size balanced (or non-size-balanced).

3.3 Examples of informative cluster size

The issue of informative cluster size is more clearly understood through some examples from dental, reproductive toxicology and pregnancy studies. Also, informative cluster size may arise in scenarios of clustering by clinical episodes experienced by patients. This is the scenario used to illustrate the methodology in the current and the subsequent chapter.

Dental studies

In dental studies of periodontal (i.e. gum) disease, interest often lies in exploring the associations between factors (age, dental hygiene, smoking habits, plaque etc.) and the disease status of the teeth. The teeth (members) in a patient's mouth consist a cluster. Individuals with fewer teeth are likely to have worse dental health than individuals with more teeth because factors associated to deteriorated dental health might also lead to tooth loss, leading to informative cluster size. Williamson et al. (2003) note that the particular scenario can also be viewed as a missing data problem: as the number of teeth in a complete cluster (mouth) is known (excluding supernumerary teeth) any individuals with fewer teeth can be regarded as clusters with missing teeth.

Reproductive toxicology studies

Reproductive toxicology studies often assess the effect of a toxicant on pups within litters (e.g. Dunson et al., 2003). A litter (mother) is a cluster and a pup is a member. It is likely that litters exposed to the effect of a toxicant to produce fewer pups than unexposed litters. This is because more pups experience foetal resorptions (death of the foetus at any stage after the completion of organogenesis) under the effect of the

toxicant, thus reducing the size of the litter. As the maximum number of pups a litter can produce is unknown, the variability in litter sizes is regarded as an inherent feature of the data.

Pregnancy studies

Hoffman et al. (2001) mentioned that informative cluster size might arise when investigating how the pregnancy outcome (successful or abortion) is associated with environmental factors, maternal characteristics, ethnic origin and other characteristics. Each pregnancy is a member and the set of a mothers's pregnancies is the cluster. Mothers with higher risk of abortion tend to have more pregnancy-attempts until they reach the desired family size suggesting that the number of pregnancies might be informative.

Illustration: Secondary analysis of the Delta trial

Informative cluster size may also arise in clinical settings. For example, in a study of episodes (experienced by patients) of clinical care at the hospital, at each episode a biological health outcome is obtained alongside information on treatment and other patient characteristics. We might be interested in the marginal association between covariates and the measure of health outcome. Cluster size might be informative if the number of episodes experienced is related to the health outcome measured.

In our data illustration we perform a secondary analysis on a dataset of adverse events from the Delta trial of HIV therapy. As described in Chapter 1, a cluster consists of the ARC events experienced by a patient. We are interested in modelling whether or not an ARC event is of type Oral Candidiasis, in terms of randomisation arm, CD4 count at the time of episode and time since entry in the trial. As the proportion of the events that are Oral Candidiasis decreases as the number of ARC events per patient increases, the cluster size might be informative. More details are provided in Section 3.10 where we illustrate the proposed methodology of this chapter.

3.4 Why informative cluster size might cause problems in analysis

In this section we firstly attempt to explain why the apparently naive solution of adjusting for cluster size in the regression model is not, in general, good practice. We then explain that adjusting for cluster size and then integrating over the distribution of N to

derive the marginal effect of $\boldsymbol{X}$ (i.e. not conditional on cluster size) on the expected outcome can be useful in certain scenarios but has important limitations. From our experience these two issues seem to cause confusion to analysts and to our knowledge, have not been discussed by earlier authors. From the considerations in this section it will become apparent that special methods for informative cluster size problems are required and these are introduced in Section 3.5.

3.4.1 Adjusting for cluster size

A possible strategy to deal with informative cluster size problems is to include cluster size alongside the predictors of main interest in a marginal regression model for the expected outcome

$$E(Y \mid \boldsymbol{X}, N) = h^{-1}(\gamma_0 + \boldsymbol{X}^T\boldsymbol{\gamma}_1 + \gamma_2 N). \tag{3.3}$$

This approach is generally not desirable for a number of reasons, if interest lies in the marginal effect (i.e. not conditional on cluster size) of $\boldsymbol{X}$ on the expected outcome.

If $\boldsymbol{X}$ is non-size-balanced, the effect of $\boldsymbol{X}$ conditional on N as in equation 3.3 will not be equal to its marginal effect. For example, it is likely that unmeasured cluster-level factors which are linked to Y and N, are also linked to $\boldsymbol{X}$. We assume it is not scientifically meaningful to adjust that N, i.e. include N as part of $\boldsymbol{X}$.

Even if $\boldsymbol{X}$ is size-balanced, adjusting for N does not result in parameter estimates for the model of interest (which is for $E^p(Y \mid \boldsymbol{X})$) in any population, if the effect of $\boldsymbol{X}$ is not the same for all cluster sizes. The speculated cluster-size-adjusted model becomes:

$$E(Y \mid \boldsymbol{X}, N) = h^{-1}(\gamma_0 + \boldsymbol{X}^T\boldsymbol{\gamma}_1 + \gamma_2 N + \boldsymbol{X}^T\boldsymbol{\gamma}_3 N). \tag{3.4}$$

This model provides an estimate of the effect of $\boldsymbol{X}$ for each group of cluster sizes. Such an estimand is not generally useful in applications.

However, there are at least two circumstances in which adjustment for N would be appropriate. The first arises in volume-outcome studies (Panageas et al., 2007). This type of studies is used to evaluate whether surgeons who treat a higher number of patients for a specific condition, demonstrate better outcomes than those who treat fewer patients. Each patient is seen as a member and a cluster is formed by the patients treated by a surgeon. The uniqueness of this setting is that the 'volume' (number of

patients) is both a primary predictor and the cluster size. As the 'volume' is a main predictor, for scientific reasons it is appropriate to include cluster size alongside the other covariates in a regression model.

The second arises when $h(.)$ is the identity link function, $\boldsymbol{X}$ is size-balanced and the effect of $\boldsymbol{X}$ is the same for all cluster sizes (for example when the effect of $\boldsymbol{X}$ is homogeneous). In this case it can be easily seen that $\boldsymbol{\gamma}_1$ in model (3.3), which is the effect of $\boldsymbol{X}$ conditional on N, coincides with the marginal effect of $\boldsymbol{X}$. If - as it often happens in practice - the intercept term is not of direct interest, fitting model (3.3) can be used to obtain consistent estimates for the marginal the effect of $\boldsymbol{X}$ but not for the intercept term.

3.4.2 Integrating over the distribution of N

We have established when it may or may not be possible to directly estimate the marginal parameters of interest from a model that adjusts for cluster size. We now consider the possibility of deriving the marginal effect of $\boldsymbol{X}$ on the expected outcome by first considering a model for $E(Y \mid \boldsymbol{X}, N)$ and then marginalising over the conditional distribution of N given $\boldsymbol{X}$. So,

$$\begin{aligned}
\mu^{C1}(\boldsymbol{x}) &= E^{C1}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = E^{C1}_{N|\boldsymbol{X}}[E(Y \mid \boldsymbol{X} = \boldsymbol{x}, N)] \\
&= \sum_n E(Y \mid \boldsymbol{X} = \boldsymbol{x}, N = n) P^{C1}(N = n \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= \sum_n E(Y \mid \boldsymbol{X} = \boldsymbol{x}, N = n) \frac{f(\boldsymbol{X} = \boldsymbol{x} \mid N = n) P(N = n)}{\sum_j E(Y \mid \boldsymbol{X} = \boldsymbol{x}, N = j) f(N = j)} \quad (3.5) \\
&\quad \text{for all values of } \boldsymbol{X} = \boldsymbol{x}.
\end{aligned}$$

Apart from a model for $E(Y \mid \boldsymbol{X}, N)$ we also need to specify a model for $f(\boldsymbol{X} \mid N)$. The main limitation of this approach is that it relies on the correct specification of $E(Y \mid \boldsymbol{X}, N)$ and $f(\boldsymbol{X} \mid N)$. Important, but perhaps less of a concern, are computational difficulties that might arise when a multivariate distribution needs to be specified for $f(\boldsymbol{X} = \boldsymbol{x} \mid N = n)$. Finally, even if the 'true' model for $E(Y \mid \boldsymbol{X}, N)$ is linear in the predictors (for example, as in model (3.3)), $E^{C1}(Y \mid \boldsymbol{X})$ will not be of the form $E^{C1}(Y \mid \boldsymbol{X}) = \beta_0 + \boldsymbol{X}^T \boldsymbol{\beta}_1$, in general.

In special cases where the model for $E^{C1}(Y \mid \boldsymbol{X})$ is saturated, for example when $\boldsymbol{X} = X$ is a single binary covariate, models for $E(Y \mid \boldsymbol{X}, N)$ and $f(\boldsymbol{X} \mid N)$ are not required. Instead, estimates for the conditional expectations and probabilities in

expression (3.5) can be obtained by calculating the corresponding observed averages and proportions from the sample. When X is binary, the assumed model of interest is $E^{C1}(Y \mid X) = \beta_0 + \beta_1 X$. So, $\hat{\mu}^{C1}(x)$ is computed for $X = 0, 1$ and then, $\hat{\beta}_0 = \hat{\mu}(0)$ and $\hat{\beta}_1 = \hat{\mu}(1) - \hat{\mu}(0)$. In case of concerns about model uncertainty due to rarely observed values of X, a model for $f(X \mid N)$ can be specified but the model we end up with might not then be of the from $E^{C1}(Y \mid X) = \beta_0 + \beta_1 X$.

3.5 Methods for informative cluster size: current methodology

3.5.1 Marginal inference

We now present existing estimation methods for marginal inference for population C1. We use subscripts i and j to denote the cluster and the member, respectively. We assume a marginal regression model for population C1

$$\mu^{C1}(\boldsymbol{X}_{ij}) = \beta_0 + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_1,$$

where $h(.)$ is a link function and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$ is a $(q+1)$-dimensional vector of unknown parameters to be estimated.

Within-cluster resampling (WCR)

Hoffman et al. (2001) introduced the concepts of measurement-based sampling and cluster-based sampling and these relate to the populations for inference defined earlier (see Definition 3.1, pg. 67). Measurement-based sampling is the one implicitly assumed in a generalised estimating equations approach with independence working correlation matrix and corresponds to population M. Cluster-based sampling corresponds to sampling one observation from each cluster and leads to an inference about a randomly chosen member of a randomly chosen cluster (population C1). They noted that when the cluster size is informative, inferences under these two sampling schemes are generally different. Inference under cluster-based sampling can be obtained using their proposed Within Cluster Resampling method (WCR).

In WCR a new dataset is created from the original dataset by sampling at random (with replacement) one member from each of the K clusters. This is done repeatedly (say Q times), so that Q datasets are created, each containing K members. Since the

K observations are independent, a generalised linear model is used to estimate $\boldsymbol{\beta}$ and var($\boldsymbol{\beta}$) for each of these Q datasets. Let $\hat{\boldsymbol{\beta}}^{(k)}$ and var($\hat{\boldsymbol{\beta}}^{(k)}$) denote the estimates for $\boldsymbol{\beta}$ and var($\boldsymbol{\beta}$) in the kth resampled dataset ($k = 1, \ldots, Q$). Then these Q estimates are averaged:

$$\hat{\boldsymbol{\beta}}_{WCR} = \frac{1}{Q}\sum_{l=1}^{k}\hat{\boldsymbol{\beta}}^{(k)}$$

and

$$\text{var}\sqrt{N}(\hat{\boldsymbol{\beta}}_{WCR} - \boldsymbol{\beta}) = \frac{1}{Q}\sum_{k=1}^{Q}\text{var}(\hat{\boldsymbol{\beta}}^{(k)}) - \frac{1}{Q}\sum_{k=1}^{Q}(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}_{WCR})(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}_{WCR})^T.$$

As each cluster contributes one member to each estimate regardless of its size, it is apparent that the parameter estimated is that for the population of typical cluster members 1. Provided the marginal regression model $\mu^{C1}(\boldsymbol{X}) = h^{-1}(\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}_1)$ is correctly specified, the WCR estimator is a consistent estimator of $\boldsymbol{\beta}$. This is in contrast with a standard application of GEE, where large clusters are considered to be more important in the estimation of regression parameters. So, population of typical cluster members 1 corresponds to an inference for a randomly chosen measurement from a randomly chosen cluster.

Inversely weighted by the cluster size GEE (CWGEE)

In the same spirit, when the cluster size is informative, Williamson et al. (2003) suggested that there are two marginal analyses of possible interest: one for the population of all members and one for the population of typical cluster members 1. Compared to the definitions of Hoffman et al. (2001), population of all members corresponds to an observation-based sampling scheme, whereas population of typical cluster members 1 to cluster-based sampling.

WCR method is computationally intensive. CWGEE (or weighted independence estimating equations-WIEE) proposed by Williamson et al. (2003), provides an estimator that is asymptotically equivalent to WCR (as $K, Q \to \infty$) but avoids the Monte Carlo element of WCR. The CWGEE are

$$\sum_{i=1}^{K}\left(\frac{\partial\boldsymbol{\mu}_i^{C1}}{\partial\boldsymbol{\beta}}\right)^T \frac{1}{N_i}(\boldsymbol{V}_i^{\mathrm{I}})^{-1}(\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i^{C1}) = \mathbf{0}. \tag{3.6}$$

Note that $\boldsymbol{V}_i^{\mathrm{I}}$ is the working covariance matrix based on an independence working correlation and the inclusion of the term N^{-1} means that clusters are inversely weighted

by their size. If the marginal model $\mu^{C1}(\boldsymbol{X}) = h^{-1}(\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}_1)$ is correctly specified, the solution to equations (3.6) is a consistent estimator of $\boldsymbol{\beta}$.

If, instead, marginal model $\mu(\boldsymbol{X}) = h^{-1}(\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}_1)$ is correctly specified, inference can be made for population M by deleting N_i^{-1} from equations (3.6) and replacing $\boldsymbol{\mu}_i^{C1}$ by $\boldsymbol{\mu}_i$. Doing this gives the standard GEE (equations (2.14)) with independence working correlation. Note that using a non-independence working correlation in equations (3.6) will not give unbiased inference for population M when cluster size is informative.

Parameter estimates, $\hat{\boldsymbol{\beta}}$, are iteratively estimated, as in (2.17). Under mild regularity conditions, it is proved that $\hat{\boldsymbol{\beta}}$ follows a multivariate Normal distribution with mean $\boldsymbol{\beta}$ and variance consistently estimated by

$$\left(\sum_{i=1}^{K}\left(\frac{\partial\boldsymbol{\mu}_i^{C1}}{\partial\boldsymbol{\beta}}\right)^T(\boldsymbol{V}_i^{\mathrm{I}})^{-1}\frac{\partial\boldsymbol{\mu}_i^{C1\,T}}{\partial\boldsymbol{\beta}}\right)^{-1}\left(\sum_{i=1}^{K}\left(\frac{\partial\boldsymbol{\mu}_i^{C1}}{\partial\boldsymbol{\beta}}\right)^T(\boldsymbol{V}_i^{\mathrm{I}})^{-1}\mathrm{var}(\boldsymbol{Y}_i^*)(\boldsymbol{V}_i^{\mathrm{I}})^{-1}\frac{\partial\boldsymbol{\mu}_i^{C1}}{\partial\boldsymbol{\beta}^T}\right)$$
$$\times\quad\left(\sum_{i=1}^{K}\left(\frac{\partial\boldsymbol{\mu}_i^{C1}}{\partial\boldsymbol{\beta}}\right)^T(\boldsymbol{V}_i^{\mathrm{I}})^{-1}\frac{\partial\boldsymbol{\mu}_i^{C1}}{\partial\boldsymbol{\beta}^T}\right)^{-1}.$$

Williamson et al. (2003) carried out simulation studies to compare the finite sample properties of WCR and WIEE. They concluded that for large sample sizes both methods perform equally well in terms of bias and efficiency. For small samples sizes WCR provides parameter estimates with small but noticeable bias, whereas WIEE exhibits negligible bias.

Williamson et al. (2003) also extended the approach of Prentice (1988) for obtaining estimates for the correlation parameters, $\boldsymbol{\rho}$, when the cluster size is informative and the target of inference is population C1. Compared to equation (2.18), in Williamson's proposed estimating equations for $\boldsymbol{\rho}$, the contribution of each cluster is inversely weighted by the number of distinct pairs in the cluster. Notation is retained as in equations (2.18). The Williamson's estimating equations for $\boldsymbol{\rho}$ are:

$$U_{\boldsymbol{\rho}}(\boldsymbol{\beta},\boldsymbol{\rho}) = \sum_{i=1}^{K}\frac{1}{\binom{n_i}{2}}\frac{\partial\boldsymbol{\eta}_i^T}{\partial\boldsymbol{\rho}}\boldsymbol{H}_i^{-1}(\boldsymbol{Z}_i - \boldsymbol{\eta}_i) = \boldsymbol{0}. \tag{3.7}$$

Estimates for $\boldsymbol{\rho}$ are obtained iteratively, as in equation (2.19), but these estimates are not part of the iterative procedure to compute $\hat{\boldsymbol{\beta}}$, since equations (3.6) use independence working correlation.

3.5.2 Comparison of populations: a simple hypothetical example

We now wish to contrast the methods (and populations for inference) discussed in this section in the context of an example from a hypothetical toxicology study. Suppose that we are interested in the effect of a toxin (X) on the average weight (Y) of pups from a typical mother (litter). So, we are interested in inference for population C1. Toxin is a mother-specific covariate ($X = 0$ and 1 for unexposed and exposed litters, respectively).

For simplicity we assume that there are only two 'types' of mothers. The mother's type is unobserved, e.g. it is a genetic characteristic. Mothers of type 1 produce heavy pups and mothers of type 2 produce light pups. We also assume that because of a characteristic inherent to the type of the mother, type-1 mothers are affected by the effect of the toxin; this results in foetal resorptions and therefore reduced litter size. On the other hand, because of a different inherent characteristic, type-2 mothers are not affected by the effect of the toxin. So, the effect of the toxin on the average pup weight can be seen to be different for each type of mother.

It is useful to consider the quantities to be estimated when we seek inference for population M and C1, respectively. In the absence of the toxin, both types of mothers contribute approximately equally in the estimation for both populations, so $E^M(Y \mid X = 0)$ is very similar to $E^{C1}(Y \mid X = 0)$. However, when the toxin is present $E^M(Y \mid X = 1)$ is predominantly estimated from mothers of type 2 because mothers of type 2 produce more pups than mothers of type 1 in the presence of the toxin. Therefore, $E^M(Y \mid X = 1)$ can be seen to be lower than $E^{C1}(Y \mid X = 1)$ and consequently biased since we are interested in inference for a typical pup of a typical mother.

In relation to Section 3.4.1, adjusting for the cluster size, i.e. the number of pups in the litter (using, for example, model (3.4)) provides an estimate of the effect of toxin conditional on the number of pups. From an epidemiological point of view, such an inference might not be of scientific interest.

3.5.3 Equivalence of covariate effects in populations M and C1 in special cases

In their simulation studies, Hoffman et al. (2001) considered scenarios with a binary outcome and a binary cluster-constant and size-balanced covariate (exposure). As it

can be seen in their Table 3, estimates for the parameters in population M and C1 only differ for the intercept term; estimates for the exposure effects are very similar.

This similarity of the exposure effects for populations M and C1, need not be the case in general. For example, in the simulation study of Williamson et al. (2003) for binary outcomes, covariate X is also cluster-constant but non-size balanced; exposed clusters are more likely to have a smaller cluster size. Their Table 1 shows that not only the intercept term, but also the exposure effect differs between populations M and C1. Also, the real-data analyses of Hoffman et al. (2001) and Williamson et al. (2003) show differences between the estimates in the slope terms for populations M and C1 (although these differences are not statistically significant).

Neuhaus and MCculloch (2011) briefly comment on the simulations results of Hoffman et al. (2001) and Williamson et al. (2003) and they seem to infer that only the intercept term and not the exposure effects will differ between the two populations. In our view, this is not true in general (see, for example, Section 3.5.2).

It helps understanding the rational behind the results in the simulations and illustrations of Hoffman et al. (2001) and Williamson et al. (2003), and also the comment of Neuhaus and MCculloch (2011), if we start by considering a scenario as follows. Suppose that cluster size is informative, the link function is the identity one and $\boldsymbol{X}$ is size-balanced. Also, suppose that the effect of $\boldsymbol{X}$ is the same for all clusters and therefore for all cluster sizes. We assume that a model for $E(Y \mid \boldsymbol{X}, N)$ of the form (suppressing dependence on i and j):

$$E(Y \mid \boldsymbol{X}, N) = \gamma_0 + \boldsymbol{X}^T\boldsymbol{\gamma}_1 + \gamma_N, \tag{3.8}$$

is true. Each cluster size is allowed to have a different intercept. We instead postulate the 'pragmatic model' of interest $E^p(Y \mid \boldsymbol{X}) = \beta_0^p + \boldsymbol{X}^T\boldsymbol{\beta}_1^p$.

For population M we write:

$$\begin{aligned} E^M(Y \mid \boldsymbol{X}) &= E_{N|\boldsymbol{X}}[E_{Y|\boldsymbol{X},N}(Y)] = E_N[E_{Y|\boldsymbol{X},N}(Y)], \text{ because } N \perp \boldsymbol{X} \\ &= E_N(\gamma_0 + \boldsymbol{X}^T\boldsymbol{\gamma}_1 + \gamma_N) = \underbrace{\gamma_0 + E(\gamma_N)}_{\beta_0^M} + \boldsymbol{X}^T \underbrace{\boldsymbol{\gamma}_1}_{\boldsymbol{\beta}_1^M}. \end{aligned} \tag{3.9}$$

For population C1 we write:

$$
\begin{aligned}
E^{C1}(Y \mid \boldsymbol{X}) &= \frac{E_{N|\boldsymbol{X}}\left[\frac{1}{N}E_{Y|\boldsymbol{X},N}(Y)\right]}{E_N\left[\frac{1}{N}\right]} = E_{N|\boldsymbol{X}}\frac{\left[\frac{1}{N}E_{Y|\boldsymbol{X},N}(Y)\right]}{E_{N|\boldsymbol{X}}\left[\frac{1}{N}\right]}, \text{ because } N \perp \boldsymbol{X} \\
&= \frac{E_N\left[\frac{1}{N}(\gamma_0 + \boldsymbol{X}^T\boldsymbol{\gamma}_1 + \gamma_N)\right]}{E_N\left(\frac{1}{N}\right)} = \underbrace{\gamma_0 + \frac{E_N\left[\frac{1}{N}\gamma_N\right]}{E_N\left[\frac{1}{N}\right]}}_{\beta_0^{C1}} + \boldsymbol{X}^T\underbrace{\boldsymbol{\gamma}_1}_{\boldsymbol{\beta}_1^{C1}} \quad (3.10)
\end{aligned}
$$

To summarize, it follows from expressions (3.9) and (3.10) that when the conditions:

A.1. The link function is the identity one.

A.2. The effect of $\boldsymbol{X}$ is the same in all clusters,

A.3. $\boldsymbol{X}$ is size-balanced,

are satisfied, the intercept parameters will differ between populations M and C1 but the exposure effects will be the same, i.e. $\beta_0^M \neq \beta_0^{C1}$ but $\boldsymbol{\beta}_1^M = \boldsymbol{\beta}_1^{C1}$.

If Condition (A.1) is not satisfied, then the models for $E^p(Y \mid \boldsymbol{X})$ might not be in the same form as in model (3.8). This is known as the problem of non-collapsibility. Nevertheless, it is worth noting that for the pragmatic model $E^p(Y \mid \boldsymbol{X}) = h^{-1}(\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}_1)$, it has been seen in simulation studies (Hoffman et al., 2001; Benhin et al., 2005) that if Condition (A.1) does not hold (when the link function is the logit one) but A.2 and A.3 are true, the estimates for $\boldsymbol{\beta}_1^M$ and $\boldsymbol{\beta}_1^{C1}$ are very similar.

Condition (A.2) is not satisfied, if there is an interaction between N and $\boldsymbol{X}$ in the model for $E(Y \mid \boldsymbol{X}, N)$ i.e. the assumed true model for $E(Y \mid \boldsymbol{X}, N)$ is of the form $E(Y \mid \boldsymbol{X}, N) = \gamma_0 + \boldsymbol{X}^T\boldsymbol{\gamma}_1 + \gamma_{0N} + \boldsymbol{X}^T\boldsymbol{\gamma}_{1N}$. It can be easily seen that not only the intercept terms but also the other regression coefficients will be different for the parameters in populations M and C1. This is so even if $\boldsymbol{X}$ is cluster-constant and Conditions (A.1) and (A.3) are true.

If Condition (A.3) is not satisfied, $E^M(Y \mid \boldsymbol{X} = \boldsymbol{x})$ and $E^{C1}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ will not be equal for all values of $\boldsymbol{x}$, in general. In this case, as we explain in detail in Section 3.6, inference for population C1 does not provide useful parameter estimates. Instead, for scenarios where $\boldsymbol{X}$ is categorical cluster-varying and non-size balanced we define further populations for inference for typical cluster members (see Definition 3.4). We

also propose corresponding estimators which result in more interpretable parameter estimates than the ones in population C1.

It is important to emphasise that when conditions (A.1-A.3) are true and the cluster size is non-informative, CWGEE will be less efficient than IEE in estimating $\boldsymbol{\beta}_1$ by fitting a model for $E^{C1}(Y \mid \boldsymbol{X})$. This is because in using CWGEE all clusters receive the same total weight in the estimation even if larger clusters are naturally more informative than the smaller ones.

3.5.4 Cluster-specific inference and the joint modelling approach

Dunson et al. (2003) developed a Bayesian approach based on joint modelling the cluster size and the main outcome with shared random effects to provide cluster-specific inference when the cluster size is informative.

Joint models are specified for Y and N with a continuation ratio (CR) probit model for the latter. For simplicity we assume random intercepts models:

$$E(Y_{ij} \mid \boldsymbol{X}_{ij}, b_i) = h^{-1}(\beta_0 + \boldsymbol{X}_{ij}^T\beta + \lambda_1 b_i) \tag{3.11}$$

and

$$P(N_i = k \mid \boldsymbol{X}_i^*, b_i) = F(\delta_k - \boldsymbol{X}_{ik}^T\boldsymbol{\alpha} - \lambda_2 b_i) \prod_{t=1}^{k-1} \left\{1 - F(\delta_t - \boldsymbol{X}_{it}^T\boldsymbol{\alpha} - \lambda_2 b_i)\right\}, \tag{3.12}$$

$k = 1, \ldots, N_{max}-1$, where N_{max} is the maximum cluster size in the sample of clusters, F denotes the cumulative density function of the standard Normal distribution and the terms $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{N_{max}-1})$ represent cut-points. The random effects b_i follow a zero mean Normal distribution. Dependence between cluster size and main outcome is accommodated through the shared random effects, b_i.

In their illustration, Dunson et al. (2003) consider a toxicology study where pups are clustered within litters. The target is mainly to estimate the effect of the toxin on a typical pup's weight. Gueorguieva (2005) also considers the joint modelling approach from a frequentist viewpoint, using maximum likelihood estimation, where the random effects for the two models are correlated rather than shared.

More recently, Chen et al. (2011) investigated the robustness of such joint models to misspecification of the cluster size model, assuming that the distributions of random effects and errors terms are correctly specified. They consider two 'types' of misspecification of the cluster size model: (1) misspecification of the distribution for the model

for N and (2) misspecification of the functional form of random effects in the model for N. The distribution of N in the cluster size model is misspecified when, for example, a Poisson or Negative Binomial model instead of a CR model when the CR model is the correct one. They found that using an incorrect distribution for the cluster size model may lead to small or moderate biases. Misspecification of the functional form of the random effect in the cluster size model (for example misspecification of the polynomial form) results in nearly unbiased estimation. The efficiency loss, compared to a cluster size model that correctly specifies the functional form of the random effects, was found to be small.

Neuhaus and MCculloch (2011) examined the performance of GLMMs when the cluster size is informative. They assumed that the random effects, shared by outcome and cluster size model, are independent of the covariates and also that the response is independent of the cluster size conditional on the shared random effects. They focused on the performance of maximum likelihood estimation for a GLMM when ignoring the informative cluster sizes. They showed theoretically that for linear mixed models (LMMs), ignoring informative cluster size results in consistent estimates for the effects of covariates uncorrelated to the random effects, but biased estimates for the effects of covariates correlated with the random effects. Similar results hold for GLMMs in general. Based on simulation results, Neuhaus and MCculloch (2011) suggested that the effects of covariates uncorrelated with the random effects are estimated with little or no bias for regression parameters corresponding to covariates independent of the random effects but with bias otherwise. For example, in models with random intercepts only, fitting a random intercepts model ignoring informative cluster sizes, may provide a biased estimate for the intercept term but nearly unbiased estimates for the other regression parameters. Finally, they demonstrated how the bias observed in GLMMs when ignoring informative cluster sizes can be seen as arising because of misspecification of the distribution of random effects.

There are two additional issues regarding the use of the joint modelling approach of Dunson et al. (2003) worth noting. First, contrary to the case of marginal inference under informative cluster size, inference for the populations for typical and all cluster members has not been distinguished. This is because when $\boldsymbol{X} \perp \boldsymbol{b}$ and the assumption of conditional independence of Y and N given $\boldsymbol{b}$ holds, then conditional on $\boldsymbol{b}$ the

cluster size is not informative and consequently $\mu(\boldsymbol{x}, \boldsymbol{b}) = E(Y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{b})$ is the same in populations M and C1 $\forall$ $\boldsymbol{x}$. Second, although the method of Dunson et al. (2003) has been proposed as a method for observed cluster inference, it can be seen to have similarities with shared parameters models for the outcome and missingness process under MNAR assumptions (see Chapter 2, Section 2.9.4) and hence provides cluster-specific inference for the complete clusters, in general. Under certain assumptions about the missingness process, the complete-cluster inference coincides with the observed-cluster one. We elaborate on these two issues in Chapter 5, Section 5.4.

Finally, we remind the reader that when the random effects are not independent of the covariates, estimation of covariate effects can be biased (see Section 2.8 and Neuhaus and MCculloch, 2006). This issue is further discussed in Section 3.12.

For the rest of this chapter we primarily focus on methods for marginal inference. We shall not consider cluster-specific inference under informative cluster size until Chapter 5. In the next section we define *informative covariate structure* and demonstrate how this may cause problems in marginal regression.

3.6 Informative covariate structure and new methodology

3.6.1 Additional notation and further definitions

We retain the notation from Section 3.2 whilst introducing additional quantities. For simplicity we omit the indicator, i, for the cluster. Denote the cluster-varying components of $\boldsymbol{X}$ (i.e. the components of $\boldsymbol{X}$ that can vary between members of the same cluster) as $\boldsymbol{X}^{(1)}$ and the cluster-constant elements as $\boldsymbol{X}^{(2)}$. So, $\boldsymbol{X} = (\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$. Let Λ, $\Lambda^{(1)}$ and $\Lambda^{(2)}$ denote the sample spaces of $\boldsymbol{X}$, $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, respectively, i.e. the sets of possible values of $\boldsymbol{X}$, $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$. For each cluster, let $B = 1$ if the columns of $\boldsymbol{X}^*$ include all the elements of Λ, i.e. if all possible values of $\boldsymbol{X}$ are realised in the cluster; $B = 0$ otherwise. For each member of a cluster, Z denotes the total number of members in that cluster who share the same value of $\boldsymbol{X}$ as the member in question. Finally, although N, B and $\boldsymbol{X}^*$ are cluster-level variables, we also use these symbols to refer to member-level variables: the values of N, B and $\boldsymbol{X}^*$ for a member are equal to the values of N, B and $\boldsymbol{X}^*$ of the cluster to which it belongs.

We define two additional populations of members that may be of interest depending on the research objective.

Definition 3.4

- *The **population of typical cluster members 2 (C2)** is the subpopulation in which each cluster contributes at random, for each distinct value of $\boldsymbol{X}$ represented in the cluster, a single member with that value of $\boldsymbol{X}$. Thus, the probability that each member is contributed is Z^{-1}.*

- *The **population of typical cluster members 3 (C3)** is the subpopulation in which each cluster with $B = 1$ contributes a single member at random for each distinct covariate value present in that cluster. Thus, the probability that each member is contributed is BZ^{-1}. This population is only defined if $\Lambda^{(1)}$ is finite.*

Note that populations C2 and C3 depend on the choice of covariates $\boldsymbol{X}$ whereas M and C1 do not.

In line with the definitions in Section 3.2 the expectation notation $E^p(.)$ refers to the population p (p =M, C1, C2, C3), and $E(.) = E^M(.)$. For a given member with $\boldsymbol{X} = \boldsymbol{x}$ in a given cluster, let $Z_{\boldsymbol{x}}$ denote the total number of members in that cluster with $\boldsymbol{X} = \boldsymbol{x}$. Formally, for population C2 we write

$$E^{C2}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{E_{N,Z_{\boldsymbol{x}}|\boldsymbol{X},Z_{\boldsymbol{x}}\geq 1}\left[\frac{1}{Z_{\boldsymbol{x}}}E_{Y|\boldsymbol{X},Z_{\boldsymbol{x}},N}(Y)\right]}{E_{N,Z_{\boldsymbol{x}}|\boldsymbol{X},Z_{\boldsymbol{x}}\geq 1}\left(\frac{1}{Z_{\boldsymbol{x}}}\right)}. \tag{3.13}$$

$E^{C3}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ is defined analogously, by replacing the condition $Z_{\boldsymbol{x}} \geq 1$ in equation (3.13) with the condition $B = 1$.

Definition 3.5 *The covariate structure is non-informative if*

$$E(Y \mid \boldsymbol{X}, \boldsymbol{X}^*, N) = E(Y \mid \boldsymbol{X}, N). \tag{3.14}$$

Otherwise, the covariate structure is informative.

Note that informative covariate structure may occur even if all clusters are of the same size, and also that the combination of informative covariate structure and informative cluster size is possible.

We define $\mu^p(\boldsymbol{x}) = E^p(Y \mid \boldsymbol{X} = \boldsymbol{x})$ where p = M, C1, C2, or C3 according to which population is considered. If (3.1) and (3.14) hold, then $\mu^p(\boldsymbol{x})$ will be the same for all four populations $\forall\ \boldsymbol{x}$. Where either (3.1) or (3.14) does not hold, the four populations may differ, but there are some scenarios where either some populations are equivalent or where populations differ but $\mu^p(\boldsymbol{x})$ would be equal:

- If $E(Y|\boldsymbol{X}, N) = E(Y|\boldsymbol{X})$, i.e. cluster size is non-informative, $\mu^M(\boldsymbol{x}) = \mu^{C1}(\boldsymbol{x})\ \forall\ \boldsymbol{x}$.
- If $E(Y|\boldsymbol{X}, Z, N) = E(Y|\boldsymbol{X})$, then $\mu^M(\boldsymbol{x}) = \mu^{C1}(\boldsymbol{x}) = \mu^{C2}(\boldsymbol{x})\ \forall\ \boldsymbol{x}$.
- If $E(Y|\boldsymbol{X}, Z, B, N) = E(Y|\boldsymbol{X}, Z, N)$, then $\mu^{C2}(\boldsymbol{x}) = \mu^{C3}(\boldsymbol{x})\ \forall\ \boldsymbol{x}$.
- As a consequence of the three items above, if $E(Y|\boldsymbol{X}, Z, B, N) = E(Y|\boldsymbol{X})$, then $\mu^p(\boldsymbol{x})$ is the same for all four populations $\forall\ \boldsymbol{x}$.
- If $Z = 1$ for all members, for example if $\boldsymbol{X}$ is continuous, then populations M and C2 are the same.
- If $B = 1$ for all members, then populations C2 and C3 are the same.
- If $\boldsymbol{X}$ is cluster-constant, then $Z = N$ for all members, and populations C1 and C2 are the same.

3.6.2 Selection of population for inference - a hypothetical example

We provide guidance on the choice of which of the four populations is most suitable according to the research objectives, building on earlier work by Williamson et al. (2003) on the choice between the populations M and C1. We focus on the hypothetical example of clustering by patient, each member corresponding to a clinical episode of care in hospital for a certain medical condition whilst Y represents a measure of health outcome from the episode, such as length of stay in hospital or a biological measure of health at discharge from hospital. Suppose that the higher is the value of Y, the worse is the outcome. The binary covariate of interest, X, represents a characteristic of the episode such as whether a treatment had been received for the episode before admission to hospital, or whether the condition appeared severe at admission. Let $X = 1$ denote severe and $X = 0$ non-severe. The distribution of X at a given episode

may be associated with the total number of episodes. For example, if X measures severity at admission, then it may be that patients with initially severe episodes also have more episodes due to common underlying factors. Similarly, the relationship between Y and X may differ according to the total number of episodes so that there is informative cluster size. The relationship between Y and X may differ according to the proportion of episodes that are severe at admission due to common underlying factors, leading to informative covariate structure.

We first consider the simple regression of Y on the binary indicator X. For population M, $\mu^M(x)$ is the average of Y over all the episodes for which $X = x$. Each episode is viewed as equally important, and so patients with a greater number of episodes with treatment x are considered more important in estimating $\mu^M(x)$. For population C2, $\mu^{C2}(x)$ is the average of Y over patients who experienced $X = x$ treating patients equally, as only one episode with $X = x$ is included per patient, so that $\mu^{C2}(x)$ can be thought to correspond to the conditional expectation for a typical patient. In particular, $\mu^{C2}(x)$ is not unduly influenced by the possibly small number of patients with many episodes and poor health outcomes as would be $\mu^M(x)$. Inference for population M allows statements such as "On average better outcomes were seen for episodes with treatment before admission". This population and this type of statement will be of interest should Y represent the length of stay in the hospital and X indicates pre-admission treatment and the analysis is conducted by health economists to examine patterns of resource use across the health service. The experience of individual patients is not of direct interest as costs are modelled at the aggregate level. Inference for population C2 allows statements such as "The outcome for an episode rated severe at admission in a typical patient experiencing such an episode was worse on average than the outcome for an episode rated non-severe in a typical patient experiencing a non-severe episode". This population and this type of statement will be of interest, should Y represent a measure of health at discharge and X the severity of disease at admission and the analysis is performed to examine how factors are linked to health at discharge for the typical patient.

Population C3 is restricted to patients who experience all values of X, which here means they experience both episodes rated severe and non-severe at admission. $\mu^{C2}(x)$ and $\mu^{C3}(x)$ may differ if patients who experience one or more episodes rated severe at

admission more often have other underlying health problems which persist over time. In this scenario whilst $\mu^{C2}(1)$ and $\mu^{C3}(1)$ might be similar, $\mu^{C2}(0)$ might be lower (meaning better health at discharge) than $\mu^{C3}(0)$, as $\mu^{C3}(0)$ is estimated only from patients with other underlying problems. Population C3 would be an unnatural choice if our interest lies separately in $\mu^p(0)$ and $\mu^p(1)$ but if we are interested in the effect of X, e.g. $\mu_p(1) - \mu_p(0)$, then this population could be selected to remove 'cluster confounding' (Neuhaus and Kalbfleisch, 1998; Ten Have et al., 1995) through estimating what can be seen as a within-cluster effect of X, conditional on experiencing episodes with both $X = 0$ and $X = 1$. In our example, confounding by patient characteristics that do not vary with time, such as other persistent underlying problems, is removed through the 'matching' of episodes with $X = 0$ and $X = 1$ by patient.

For population C1, $\mu^{C1}(x)$ is the average of Y over episodes with $X = x$ but with weighting according to the proportion of the episodes experienced by the patient for which $X = x$. In our view this is somewhat unnatural, and we cannot identify any scenarios when such inference is likely to be of direct interest, except when $\mu^{C1}(x) = \mu^p(x)$ for another population p.

We have considered only the simplest regression model with a single binary factor X. In realistic scenarios we will generally wish to remove confounding by measured factors by including them alongside the factor(s) of primary interest in the covariate vector $\boldsymbol{X}$ and fitting a regression model for $\mu^p(\boldsymbol{x})$. If the confounders are patient characteristics that do not vary with time and inference for a population of typical episodes experienced by a typical patient is required, then the analyst has a choice of two ways of removing the confounding. Either the analyst fits a model for $\mu^{C2}(\boldsymbol{x})$, or $\boldsymbol{X}^{(2)}$ is excluded from $\boldsymbol{X}$ and the analyst fits a model for $\mu^{C3}(\boldsymbol{x}^{(1)})$, a model conditional only on $\boldsymbol{X}^{(1)}$, which provides a within-cluster effect of $\boldsymbol{X}^{(1)}$ for patients who experience events with all values $\boldsymbol{x}^{(1)} \in \Lambda^{(1)}$. The latter approach removes cluster confounding by measured or unmeasured time-invariant patient characteristics but involves discarding all data from patients who do not experience all possible values of $\boldsymbol{X}^{(1)}$. If a large proportion of data is discarded to provide inference for population C3 and unmeasured patient confounders are not of great concern, then C2 will typically be preferred.

Populations C2 and C3 depend on the choice of $\boldsymbol{X}$, and in some scenarios these populations may be equivalent to other populations or be empty. In the most complex

scenarios where either the covariates of interest or confounders are continuous cluster-varying then populations C2 and M are equivalent and population C3 is empty; Population C1 then is the only typical cluster member population distinct from population M.

3.6.3 Estimation through weighted independence estimating equations

We now present estimation methods for the proposed populations for inference. We choose one of the four populations and assume that for this population, p,

$$\mu^p(\boldsymbol{X}) = h^{-1}(\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}_1), \tag{3.15}$$

where h is a known link function and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$ is an unknown $(q+1)$-dimensional parameter. We write $\mu_{ij} = h^{-1}(\beta_0 + \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1)$ and $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iN_i})^T$. The weighted estimating equations can be written in matrix form, dropping the superscript p for notational simplicity, as

$$\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \boldsymbol{W}_i \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i) = \boldsymbol{0}, \tag{3.16}$$

where K is the number of clusters in the sample and $\boldsymbol{V}_i$ is the diagonal matrix with jth element equal to the assumed variance function $v(\mu_{ij}) = \text{var}(Y_{ij} \mid \boldsymbol{X}_{ij})$. $\boldsymbol{W}_i$ is a $N_i \times N_i$ diagonal matrix whose jth element provides a weight for the jth measurement in cluster i. It depends on the target population of inference. As mentioned earlier, for μ^M, i.e. for estimation in population M, one should use $\boldsymbol{W}_i = \boldsymbol{I}_{N_i}$, whereas for μ^{C1}, i.e. for population C1, $\boldsymbol{W}_i = \text{diag}\ (N_i^{-1}, \ldots, N_i^{-1})$. We add to this that for μ^{C2}, $\boldsymbol{W}_i = \text{diag}(Z_{i1}^{-1}, \ldots, Z_{iN_i}^{-1})$ and for μ^{C3}, $\boldsymbol{W}_i = \text{diag}(B_i Z_{i1}^{-1}, \ldots, B_i Z_{iN_i}^{-1})$.

To illustrate the differences between these $\boldsymbol{W}$ matrices, consider again the example of patients, hospital episodes and a single binary covariate X. Suppose that two patients have $N = 5$ and $N = 3$ episodes, respectively, and the covariate values for these episodes are $\boldsymbol{X}^* = (0, 0, 0, 0, 1)$ and $\boldsymbol{X}^* = (0, 0, 0)$, respectively. To estimate the effect of X in population M such patients would have their episodes weighted by (1,1,1,1,1) and (1,1,1), respectively. For population C1 the weights are (1/5,1/5,1/5,1/5,1/5) and (1/3,1/3,1/3), respectively. For population C2 they are (1/4,1/4,1/4,1/4,1) and (1/3,1/3,1/3); and for population C3, (1/4,1/4,1/4,1/4,1) and (0,0,0).

Assuming equation (3.15) is correctly specified and standard regularity conditions apply, $\hat{\boldsymbol{\beta}}$ is asymptotically Normally distributed with mean $\boldsymbol{\beta}$ and with variance consistently estimated by

$$\begin{aligned} \boldsymbol{V}_R &= \left(\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \boldsymbol{W}_i \boldsymbol{V}_i^{-1} \boldsymbol{W}_i \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}\right)^{-1} \left(\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \boldsymbol{W}_i \boldsymbol{V}_i^{-1} \text{var}(\boldsymbol{Y}_i^*) \boldsymbol{V}_i^{-1} \boldsymbol{W}_i \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}\right) \\ &\times \left(\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \boldsymbol{W}_i \boldsymbol{V}_i^{-1} \boldsymbol{W}_i \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}\right)^{-1}, \end{aligned} \tag{3.17}$$

where $\text{var}(\boldsymbol{Y}_i^*) = (\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i)^T$. This is proven in Appendix A.1. Consistent estimates of $\boldsymbol{\beta}$ are iteratively obtained using a Fisher scoring algorithm, as in equation (2.17). The sandwich or empirical estimator of $\boldsymbol{V}_R$ is obtained by substituting a consistent estimate for $\boldsymbol{\beta}$ in (3.17). This is the variance estimator we use in our simulations and data analysis.

We used the software R. The function *geese* from the package *geepack* was used, which allows for weights that vary within clusters. In STATA, weights that vary within clusters cannot be incorporated in the GEE function *xtgee*. However, to obtain inference for populations C2 and C3 the survey analysis functions (*svy* prefix commands) can be used instead. These use a different (but also valid) variance estimator based on a jacknife procedure.

3.7 Strategies for Practical Implementation

If the cluster size and covariate structure are non-informative, then methods such as the standard GEE with a realistic working correlation matrix can be used, and would be expected to have greater efficiency than WIEE. Consequently, the analyst may wish to use standard methods if the cluster size or covariate structure are thought unlikely to be informative *a priori*, and otherwise use WIEE.

In practise, the analyst would have to make a decision of whether the cluster size and/or covariate structure are informative and then use the appropriate method of analysis. We propose a number of simple exploratory strategies can be of use to help the applied statistician decide whether the cluster size and/or the covariate structure are informative.

In an initial exploratory analysis, the mean of the response can be calculated for each cluster size in the sample. If the distribution of the cluster size is skewed and

therefore some cluster sizes appear infrequently, the cluster sizes can be grouped into a small number of equally sized groups (e.g. small, medium, large) and the mean response calculated for each cluster size group. An association of smaller (larger) cluster sizes with lower (higher) outcomes on average can be a first indication that the cluster size is informative. As a second step, we propose carrying out an initial regression where the cluster size, N, is included alongside the covariates of interest $\boldsymbol{X}$ in a marginal model for the expectation of Y. If the regression model of interest is $\mu^p(\boldsymbol{X}) = h^{-1}(\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}_1)$ for a known link function h and selected population p we propose first to fit the model

$$E(Y \mid \boldsymbol{X}, N) = h^{-1}(\gamma_0 + \boldsymbol{X}^T\boldsymbol{\gamma}_1 + \gamma_2 N) \tag{3.18}$$

using unweighted independence estimating equations, i.e. the WIEE in (3.16) with $\boldsymbol{W} = \boldsymbol{I}_N$. Assuming (3.18) is of the correct form then $\gamma_2 \neq 0$ corresponds to informative cluster size. A Wald test of $\gamma_2 = 0$ may be performed. Interactions between $\boldsymbol{X}$ and N can be tested to conclude whether the effect of $\boldsymbol{X}$ is likely to be different in clusters of different sizes.

When the covariate of interest (exposure) is cluster-varying then the covariate structure can also be informative, whether the cluster sizes vary or not. As for informative cluster size, simple strategies can be used to help the analyst decide whether the covariate structure is informative. One way to test for informative covariate structure would be to include the cluster-mean of the exposure (and also the cluster size, if it varies) alongside the exposure and other covariates in a model for the expected outcome. A significant effect of the term corresponding to the cluster mean of the exposure would be an indication for non-informative covariate structure.

A final approach would be to apply the WIEE with weighting for different populations, as suggested by Benhin et al. (2005) to investigate whether cluster size and/or covariate structure are informative. If the parameter estimates are all similar, then this provides some evidence that the cluster size and covariate structure are non-informative, and therefore standard GEE methods could be used.

3.8 Comparison of populations

Next we compare regression parameters for different populations across a range of scenarios. We examine the performance of our proposed estimation method through a simulation study in the next section. We consider binary and normally distributed Y and a binary cluster-varying covariate X. We choose to induce informative cluster size and/or covariate structure through an underlying 'susceptibility' that does not vary within the cluster. Data for each cluster are generated independently. For cluster i:

1. The underlying susceptibility U_i is generated from $U_i \sim N(0, 0.5^2)$.

2. The cluster size N_i depends on U_i, $N_i|U_i \sim \text{Poisson}\{\exp(\alpha_0 + \alpha_1 U_i)\} + 1$.

3. X is a cluster-varying covariate and $X_{i1}, \ldots, X_{i_{N_i}}$ are independently generated from $X_{ij}|U_i \sim \text{Bernoulli}\{\lambda_0 + \lambda_1 \text{logit}^{-1}(U_i)\}$. If $\lambda_1 = 0$ then X is size balanced, while if $0 < \lambda_1 \leq 1$ it is non-size balanced.

4. $Y_{i1}, \ldots, Y_{iN_i}$ are independently generated from $Y_{ij} \mid U_i, X_{ij} \sim \text{Bernoulli}\{\text{logit}^{-1}(\eta_{ij})\}$ for binary responses and $N(\eta_{ij}, 1)$ for continuous responses, where $\eta_{ij} = \gamma_0 + \gamma_1 X_{ij} + \gamma_2 U_i + \gamma_3 U_i X_{ij}$. Parameter γ_2 controls the degree of association between the susceptibility (and consequently cluster size) and the outcome when $X = 0$, while γ_3 controls the magnitude of the interaction between the susceptibility and the covariate.

The parameters for the cluster size model are selected to be $\alpha_0 = \alpha_1 = 1$ and these result in a mean cluster size of approximately 4. As $\alpha_1 \neq 0$, when $\gamma_2 \neq 0$ or $\gamma_3 \neq 0$ the cluster size is informative. Note that, when the cluster size is informative, the variance in the underlying susceptibility terms U governs the 'degree' of informativeness: the higher it gets the more informative the cluster size becomes. This variance was kept constant (0.25) for all scenarios considered in this section. We select $\gamma_0 = 1$ and $\gamma_1 = 1$, and γ_2 and γ_3 are varied across scenarios. For size balanced X we select $\lambda_0 = 0.4$ and $\lambda_1 = 0$, while for non-size balanced $\lambda_0 = 0$ and $\lambda_1 = 1$. When X is non-size balanced, the covariate structure is informative if either $\gamma_2 \neq 0$ or $\gamma_3 \neq 0$.

For each population, p, the analysis model is of the form $\mu^p(X) = h^{-1}(\beta_0 + \beta_1 X)$, with h being the logit link for binary Y and the identity link function for continuous Y. The true values of β_0 and β_1 for the four populations for inference are calculated

using numerical integration. Details are provided in Appendix A.2 and an example of the numerical integration using R-code is provided in Appendix A.3.

For $\gamma_3 = 0$ and non-size balanced X, Figure 3.1 shows the true values of β_0, β_1 and β_0+ β_1 for each population as γ_2 is varied and Y is either Normal or binary (top and bottom graphs respectively).

For Normal Y (top graphs), $E^p(Y|X = x) = \gamma_0 + \gamma_1 X + E^p(U|X = x)\gamma_2$, and the graph of $E^p(Y|X = x)$ against γ_2 is a straight line with slope $E^p(U|X = x)$. Let PM(0) (and PM(1)) denote the subpopulation of population M composed of the members with $X = 0$ (and $X = 1$). Similarly, let PC1(0), PC2(0) and PC3(0) (and PC1(1), PC2(1) and PC3(1)) denote the subpopulations of C1, C2 and C3 with $X = 0$ (and $X = 1$). Although U is a cluster-level variable, we also use U below to denote a member-level variable: the value of U for a member is equal to the value of U of the cluster to which it belongs. We consider $E^p(Y|X = 0)$ and $E^p(Y|X = 1)$ in turn.

$E^p(Y|X = 0)$: It is not immediately clear whether $E^M(U|X = 0)$ should be greater or less than zero. Smaller clusters tend to have smaller values of U and have a higher proportion of members with $X = 0$, while larger clusters have smaller proportion of members with $X = 0$ but of course contain more members. In this scenario the contribution of members with $X = 0$ from larger clusters outweighs the corresponding contribution from smaller clusters, so $E^M(U|X = 0) > 0$ and the slope in the graph of $E^M(Y|X = 0)$ against γ_2 is positive. Each cluster contributes only one member to PC1(0). As smaller clusters have a higher proportion of members with $X = 0$, relative to PM(0), PC1(0) contains more members from smaller clusters. Since smaller clusters tend to have smaller values of U, $E^{C1}(U|X = 0) \leq E^M(U|X = 0)$. We would also expect $E^{C2}(U|X = 0) \leq E^M(U|X = 0)$. We would however expect $E^{C2}(U|X = 0) \geq E^{C1}(U|X = 0)$ because larger clusters have proportionately fewer measurements with $X = 0$, and so are less likely than smaller clusters to contribute a measurement with $X = 0$ to PC1(0), whereas they are very likely to contribute one to PC2(0). Thus PC2(0) contains proportionately more measurements from big clusters than PC1(0), and big clusters tend to have higher values of U. PC3(0) is derived only from clusters that contain both $X = 0$ and $X = 1$. Such clusters are rare among smaller clusters (with smaller U) because the size restricts the likelihood of both values of X occurring and because for small U the proportion of members with $X = 0$ is high, so

that many clusters have no member with $X = 1$. Indeed, in our scenarios PC3(0) is dominated by members from large clusters, with consequently high values of U, such that the expectation $E^{C3}(U|X = 0)$ is higher than in any other population, and so this population has the steepest slope of $E^p(Y|X = 0)$ against γ_2.

$E^p(Y|X = 1)$: It is clear that $E^M(U|X = 1) > 0$, as bigger clusters and those with a greater proportion of members with $X = 1$ (both of which tend to have a higher value of U) contribute more measurements. For the same reasons we would expect $E^{C1}(U|X = 1) \leq E^M(U|X = 1)$ and $E^{C2}(U|X = 1) \leq E^M(U|X = 1)$. Indeed, this results in a very large positive slope in the graph of $E^M(Y|X = 1)$ against γ_2, and less steep slopes for populations C1 and C2. We would expect $E^{C1}(U|X = 1) \geq E^{C2}(U|X = 1)$ because clusters with low values of U tend to have a lower proportion of measurements with $X = 1$. They will therefore tend not to contribute to PC1(1), but will contribute to PC2(1) unless they have no measurements with $X = 1$. As with $X = 0$, it is not entirely clear how $E^{C3}(U|X = 1)$ will compare to other populations. In this scenario, it is comparable to $E^{C1}(U|X = 1)$.

When $\gamma_3 \neq 0$ (graphs not shown), $E^p(Y|X = 0)$ remains unchanged, whereas $E^p(Y|X = 1) = \gamma_0 + \gamma_1 + E^p(U|X = 1)\gamma_3 + E^p(U|X = 1)\gamma_2$ and so the slope of the graph of $E^p(Y|X = 1)$ against γ_2 is the same as for $\gamma_3 = 0$, but the expectation is increased by $E^p(U|X = x)\gamma_3$.

When the responses are binary, the relationship between Y and U is not linear and this introduces additional complexities to the explanations above. In particular, $E^p(Y|X = x) = E^p\left\{\text{logit}^{-1}(\gamma_0 + \gamma_1 x + \gamma_2 U) \mid X = x\right\}$. Therefore, the whole distribution of $U|X = x$, rather than simply its mean, affects the expectation of Y for each population for inference. As can be seen in the bottom three graphs in Figure 3.1, the patterns observed are analogous to those of the top graphs in Figure 3.1, but the relationship between $E^p(Y|X)$ and γ_2 is now non-linear, and indeed the graphs exhibit stationary points.

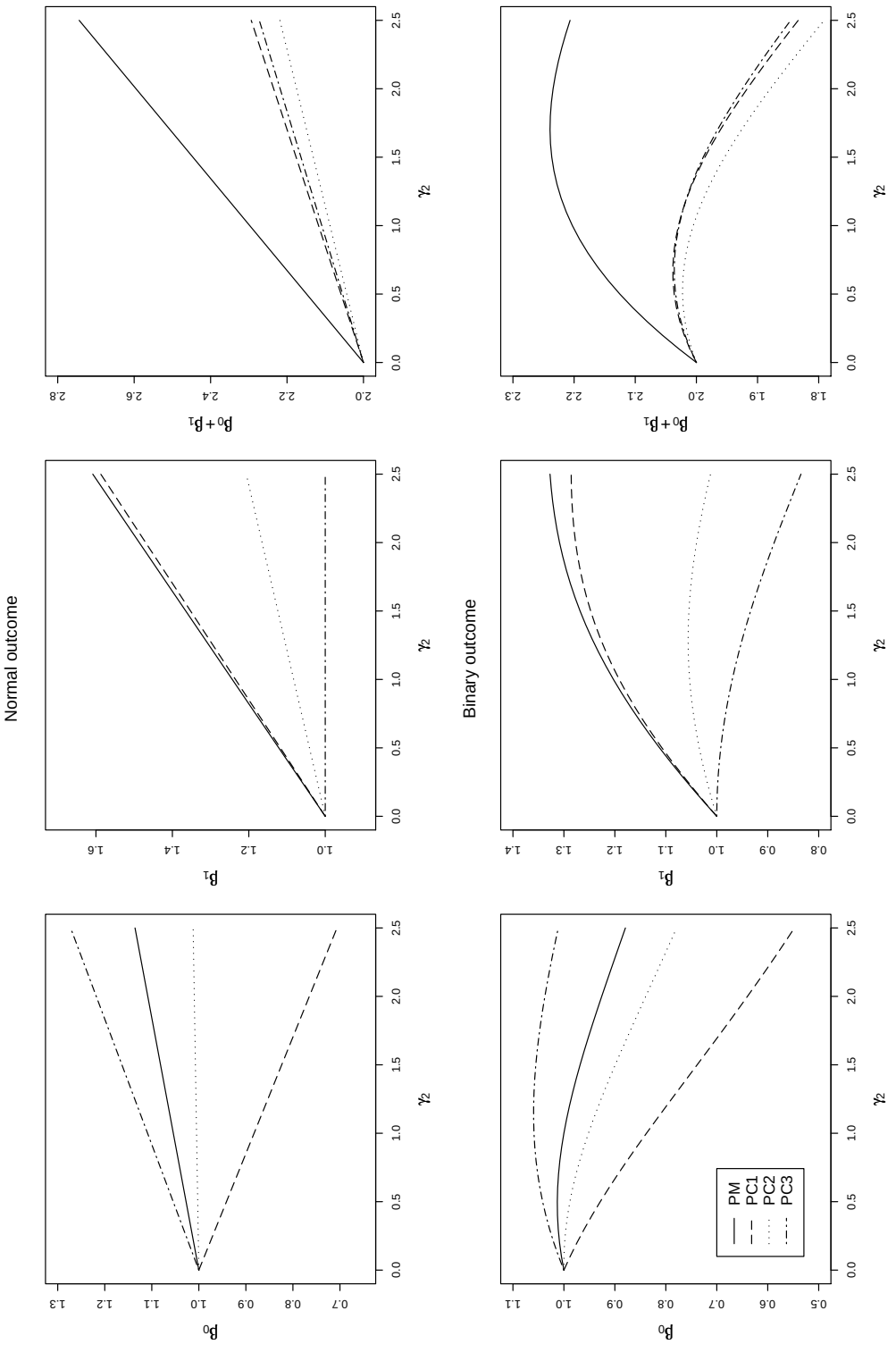

Figure 3.1: Non-size balanced X: True values for β_0 and β_1 as γ_2 varies and $\gamma_3 = 0$.

3.9 Simulation studies

To investigate the performance of our proposed estimation methods in terms of bias and coverage we conducted a simulation study using the same data generating mechanism as described in the previous section, for moderate sample sizes. Each simulated dataset contained 200 clusters and for each different scenario we generated S=2000 simulated datasets. We used the estimation methods described in Section 3.6.3 to fit the models and obtain 4 sets of parameter estimates of $\boldsymbol{\beta}$ in each simulation. We considered two sets of scenarios. The first set corresponds to a scenario where both the cluster size ($\alpha_0 = \alpha_1 = 1$) and covariate structure ($\lambda_0 = 0; \lambda_1 = 1$) are informative. In the second scenario, the cluster size is non-informative ($\alpha_0 = 1.5; \alpha_1 = 0$) and the covariate structure is informative. Note that informative covariate structure might also arise when the cluster size is constant.

In Tables 3.1 and 3.2 we present mean estimated values of the parameters from the estimation methods over the 2000 simulated datasets and their empirical standard errors (ese), i.e. the square root of the variance of the 2000 estimates, for the case of binary responses. Specifically,

$$\hat{\beta}_k = \frac{1}{S}\sum_{s=1}^{S}\hat{\beta}_{k,s}, \quad \text{ese}(\hat{\beta}_k) = \left(\frac{1}{S(S-1)}\sum_{s=1}^{S}(\hat{\beta}_{k,s}-\hat{\beta}_k)^2\right)^{1/2}, \quad k \in 0,1$$

where $\hat{\beta}_{k,s}$ is the estimated parameter value in the sth simulated dataset. Average standard errors based on the proposed variance estimators can be calculated using

$$\text{se}(\hat{\beta}_k) = \frac{1}{S}\sum_{s=1}^{S}[\text{var}(\hat{\beta}_{k,s})]^{1/2}, \quad k \in 0,1,$$

where var$(\hat{\beta}_{k,s})$ is the variance of the parameter estimate in the sth simulated dataset, using the proposed variance estimator. Coverage probabilities for a 95% confidence interval are given by

$$\text{cover}(\beta_k) = \frac{1}{S}\sum_{s=1}^{S} I(\beta_k \in [\hat{\beta}_{k,s} - z_{0.95}\ \text{se}(\hat{\beta}_{k,s}),\ \hat{\beta}_{k,s} + z_{0.95}\ \text{se}(\hat{\beta}_{k,s})]),$$

where $z_{0.95}$ is the 95^{th} percentile of the Normal distribution and β_k is the true value of the parameter for a given population for inference.

The mean estimates in both scenarios are in agreement with the corresponding true value computed using numerical integration, demonstrating that our estimation

		$\gamma_2 = 1,\ \gamma_3 = 1$		
p	TRUE(β_0, β_1)	$\hat{\beta}_0$(ese$(\hat{\beta}_0)$)	$\hat{\beta}_1$(ese$(\hat{\beta}_1)$)	cover($\hat{\beta}_0$,$\hat{\beta}_1$)
M	(1.000,1.234)	1.00(0.12)	1.24(0.21)	(0.943,0.945)
C1	(0.838,1.079)	0.84(0.14)	1.09(0.25)	(0.945,0.940)
C2	(0.952,0.927)	0.96(0.13)	0.93(0.23)	(0.948,0.944)
C3	(1.059,0.866)	1.07(0.16)	0.88(0.25)	(0.953,0.953)
		$\gamma_2 = 1,\ \gamma_3 = 0$		
M	(1.000,1.203)	1.00(0.12)	1.22(0.20)	(0.943,0.948)
C1	(0.838,1.191)	0.84(0.14)	1.21(0.24)	(0.945,0.938)
C2	(0.952,1.053)	0.96(0.14)	1.07(0.23)	(0.948,0.945)
C3	(1.059,0.970)	1.07(0.16)	0.98(0.26)	(0.953,0.945)
		$\gamma_2 = 0,\ \gamma_3 = 1$		
M	(1.000,1.202)	1.01(0.12)	1.21(0.21)	(0.951,0.951)
C1	(1.000,1.029)	1.01(0.14)	1.04(0.24)	(0.950,0.950)
C2	(1.000,1.005)	1.01(0.13)	1.02(0.24)	(0.950,0.949)
C3	(1.001,1.028)	1.01(0.15)	1.04(0.26)	(0.953,0.946)

Table 3.1: Binary responses: Parameter estimates, empirical standard errors and coverage for the four populations for inference when cluster size is informative ($\alpha_0 = \alpha_1 = 1$) and covariate structure is informative ($\lambda_0 = 0; \lambda_1 = 1$ i.e. X is non-size balanced).

		$\gamma_2 = 1,\ \gamma_3 = 1$		
p	TRUE(β_0, β_1)	$\hat{\beta}_0$(ese$(\hat{\beta}_0)$)	$\hat{\beta}_1$(ese$(\hat{\beta}_1)$)	cover($\hat{\beta}_0$,$\hat{\beta}_1$)
M	(0.838,1.079)	0.84(0.10)	1.09(0.16)	(0.952,0.953)
C1	(0.838,1.079)	0.84(0.11)	1.09(0.17)	(0.949,0.958)
C2	(0.924,0.825)	0.93(0.11)	0.83(0.18)	(0.948,0.948)
C3	(0.951,0.759)	0.95(0.12)	0.77(0.19)	(0.947,0.946)
		$\gamma_2 = 1,\ \gamma_3 = 0$		
M	(0.838,1.079)	0.84(0.10)	1.08(0.16)	(0.952,0.953)
C1	(0.838,1.079)	0.84(0.11)	1.09(0.19)	(0.949,0.958)
C2	(0.924,0.825)	0.93(0.11)	0.83(0.18)	(0.948,0.948)
C3	(0.951,0.759)	0.95(0.12)	0.77(0.19)	(0.947,0.945)
		$\gamma_2 = 0,\ \gamma_3 = 1$		
M	(1.000,1.029)	1.00(0.10)	1.03(0.17)	(0.942,0.943)
C1	(1.000,1.029)	1.00(0.11)	1.04(0.18)	(0.944,0.951)
C2	(1.000,0.939)	1.00(0.11)	0.94(0.19)	(0.951,0.953)
C3	(1.000,0.916)	1.00(0.12)	0.92(0.19)	(0.952,0.949)

Table 3.2: Binary responses: Parameter estimates, empirical standard errors and coverage for the four populations for inference when cluster size is non-informative ($\alpha_1 = 0$) and covariate structure is informative ($\lambda_0 = 0; \lambda_1 = 1$).

methods are approximately unbiased. We also present estimated coverage probabilities for 95% confidence intervals based on our sandwich variance estimator. These are in the range 94-96% across all scenarios. For set 2, note that since cluster size is non-informative, inference would be for populations M, C2 and C3 depending on the objectives of the analysis. Since cluster size is non-informative inference for population C1 coincides with inference for population M. As shown in Table 3.2, parameter estimates for population C1 are approximately equal to those for population M. The empirical standard errors though, are in all cases higher when inference is for population C1.

3.10 Illustration: secondary analysis of the Delta trial

To illustrate the proposed methodology we examine the relation amongst all AIDS related condition (ARC) events recorded between whether or not the event is one of the most prevalent ARC event types (Oral Candidiasis) and the covariates randomisation arm (RA), CD4 count (most recent to the event) and time of the event since entry in the trial. Each cluster is composed of all the ARC events reported during a patient's follow-up. There were 979 clusters, i.e. patients with sufficient information and at least 1 ARC event during follow-up. The median number of events was 2, the range 1-15.

For exploratory analysis we categorised cluster size into small (1 event), medium (2-3) and large (4-15), where each of these groups contains roughly equal number of clusters. As the total number of ARC events experienced by the patient increases the percentage of events that are Oral Candidiasis decreases, from 27% in patients with 1 event, to 22% in patients with 2 or 3 events, and to 15% in those with 4 or more. This is a first indication that the cluster size might be informative.

Of our covariates, randomisation arm is cluster-constant, CD4 count and time are cluster-varying. The average CD4 count at the time of event for small, medium and large clusters was found to be 165, 121 and 67 respectively, which suggests that CD4 count is non-size balanced. The corresponding averages for time (days) were 402, 465 and 489. To illustrate our methods we categorised CD4 as low [0, 20], medium (20, 80] or high (80+), these categories being chosen to make roughly equal number of events fall into each category. Similarly time was categorised as Start ([0, 12] months), Middle ((12, 24] months) and End (24+ months), where the categories contain roughly equal numbers of events. Since the cluster-varying covariates CD4 and Time appear to be

Set	Covariates	Step 1 Test for N		Step 2 Test for interactions with N
		β_N	p-value	p-value
M1	RA	-0.092	<0.001	0.39
M2	CD4	-0.044	0.041	0.52
M3	Time	-0.051	<0.001	<0.001
M4	RA, CD4	-0.076	0.042	0.33
M5	RA, Time	-0.086	<0.001	<0.001
M6	RA, CD4, Time	-0.086	<0.001	0.024

Table 3.3: Testing for informative cluster size in the data example from Delta Trial. Step 1: Test for the coefficient of N. Step 2: Joint Wald test for all interaction terms between N and covariates.

non-size-balanced it is likely that the covariate structure is also informative.

Let N_i denote the number of events experienced by patient i, and let j index these N_i events. Let $Y_{ij} = 1$ if the i^{th} patient's j^{th} event was Oral Candidiasis, and $Y_{ij} = 0$ otherwise. We fit models including each covariate alone, each adjusted for randomisation arm, and finally a 'fully adjusted' regression model

$$\text{logit}\{E^p(Y)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6, \qquad (3.19)$$

where X_1 and X_2 are the indicator variables of randomisation to the drug combinations AZT+ddl and AZT+ddC respectively, X_3 and X_4 are indicators that the CD4 count is medium or high respectively and finally X_5 and X_6 are indicators that the time is middle or end respectively.

For each set of covariates in the second column of Table 3.3 we tested whether cluster size is informative using the strategy described in Section 3.7. That is, we firstly fitted a model including cluster size, N, alongside the covariates in each model and performed a Wald test for the coefficient of cluster size. For example, for the set of covariates M1 the model considered was
$\text{logit}\{E^p(Y)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_N N$ and it was fitted using unweighted GEE with independence working correlation. For each of the models considered for the set of covariates M1-M6 we present the term corresponding to cluster size and the

associated p-values in the third and fourth columns of Table 3.3, respectively. The main effect of cluster size (β_N) was found significant (p < 0.05) and negative in every case, suggesting that, adjusting for covariates, the prevalence of Oral Candidiasis is lower in smaller clusters. Secondly, in each of the models considered in the first step, we added all interaction terms between cluster size and the covariates at once and tested them jointly using a Wald χ^2 test. For example, for the set of covariates M1, the model considered was logit$\{E^p(Y)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{N0} N + \beta_{N1} N X_1 + \beta_{N2} N X_2$, and similarly for the other sets of covariates. The p-values from the joint Wald tests for the interaction terms are presented in the fifth column of Table 3.3. The interaction between cluster size and time was found significant. These results indicate that the cluster size is probably informative.

Popul.	Covariate	Group	Prevalence of Oral Candid.	Unadjusted OR (95%CI)	RA Adjusted OR (95% CI)	Fully Adjusted OR (95%CI)
PM	RA	AZT	0.169	1.00	-	1.00
		AZT+ddC	0.178	1.06 (0.80,1.42)	-	1.02(0.76,1.37)
		AZT+ddI	0.182	1.09 (0.81,1.47)	-	0.98(0.72,1.32)
	CD4	Low	0.074	1.00	1.00	1.00
		Medium	0.169	2.53(1.83,3.50)	2.54(1.83,3.51)	2.25(1.62,3.13)
		High	0.285	4.97(3.64,6.79)	4.99(3.66,6.80)	4.28(3.11,5.89)
	Time	Start	0.244	1.00	1.00	1.00
		Middle	0.152	0.55(0.45,0.69)	0.56(0.45,0.69)	0.72(0.57,0.90)
		End	0.101	0.34(0.25,0.47)	0.35(0.25,0.47)	0.49(0.35,0.68)
PC1	RA	AZT	0.199	1.00	-	1.00
		AZT+ddC	0.194	0.97(0.71,1.32)	-	0.96(0.70,1.32)
		AZT+ddI	0.232	1.21(0.88,1.67)	-	1.13(0.82,1.56)
	CD4	Low	0.082	1.00	1.00	1.00
		Medium	0.178	2.43(1.69,3.50)	2.41(1.67,3.46)	2.34(1.63,3.37)
		High	0.300	4.81(3.40,6.80)	4.77(3.37,6.73)	4.58(3.22,6.53)
	Time	Start	0.244	1.00	1.00	1.00
		Middle	0.196	0.76(0.58,0.98)	0.76(0.58,0.99)	0.95(0.72,1.25)
		End	0.156	0.57(0.40,0.83)	0.57(0.40,0.83)	0.74(0.50,1.07)
...	...	...	... Continued in the next page ...		...	...

Population	Covariate	Group	Prevalence of Oral Candid.	Unadjusted OR (95%CI)	RA Adjusted OR (95% CI)	Fully Adjusted OR (95%CI)
PC2	RA	AZT	0.199	1.00	-	1.00
		AZT+ddC	0.194	0.97(0.71,1.32)	-	1.02(0.76,1.37)
		AZT+ddI	0.232	1.21(0.88,1.67)	-	1.09(0.80,1.47)
	CD4	Low	0.081	1.00	1.00	1.00
		Medium	0.198	2.80(1.97,3.95)	2.78(1.96,3.94)	2.20(1.57,3.09)
		High	0.291	4.66(3.33,6.51)	4.64(3.32,6.48)	4.05(2.90,5.64)
	Time	Start	0.254	1.00	1.00	1.00
		Middle	0.185	0.67(0.53,0.84)	0.67(0.53,0.84)	0.83(0.66,1.05)
		End	0.141	0.48(0.35,0.67)	0.48(0.35,0.67)	0.63(0.45,0.89)
GEE(EX)	RA	AZT	-	1.00	-	1.00
		AZT+ddC	-	1.01(0.76,1.33)	-	0.99(0.75,1.32)
		AZT+ddI	-	1.13(0.85,1.50)	-	1.02(0.76,1.36)
	CD4	Low	-	1.00	1.00	1.00
		Medium	-	2.50(1.86,3.34)	2.50(1.86,3.34)	2.19(1.64,2.93)
		High	-	4.77(3.60,6.31)	4.77(3.60,6.30)	3.98(2.99,5.31)
	Time	Start	-	1.00	1.00	1.00
		Middle	-	0.56(0.46,0.69)	0.57(0.46,0.69)	0.72(0.58,0.89)
		End	-	0.37(0.28,0.49)	0.37(0.28,0.49)	0.52(0.39,0.70)

Table 3.4: Data example from Delta Trial: Modelling the prevalence of adverse event Oral Candidiasis as a function of randomisation arm (RA), CD4 count and time. Odds ratios and confidence intervals are presented for the 3 populations for inference and for the application of standard GEE with exchangeable working correlation (GEE(EX)).

In Table 3.4 we present odds ratios and confidence intervals for three populations for inference, obtained by fitting the models described above by the estimation methods described in Sections 3.5 and 3.6.3. In addition, for comparison we fit the models using the standard GEE [GEE(EX)] with exchangeable working correlation. We present an estimate of the prevalence of Oral Candidiasis among ARC events in each subcategory of each covariate and for each population. For population M this prevalence is simply the proportion of Oral Candidiasis events. For population C2 our estimates of the prevalence of Oral Candidiasis according to each covariate are calculated by considering that covariate alone, i.e. using weights as if fitting a model including only that covariate. We do not include population C3 in our analysis because there are no events in the final time interval with high CD4, and so there are no data to estimate the parameters of the fully adjusted model.

The estimated odds ratios for the three populations are broadly similar in magnitude for randomisation arm and CD4, but different for the effect of time, the factor seen to interact with cluster size in our initial exploration. Note that for the randomisation arm adjusted inference for populations C1 and C2 coincides since randomisation arm is a cluster-constant variable (see Section 3.6). The standard GEE provide inference similar to that for the population of all members. This is because the correlation parameter was estimated to be approximately 0.15 for all the models. The GEE with exchangeable working correlation will provide similar parameter estimates to the GEE with independence working correlation when the correlation parameter value is modest, at least for the range of cluster sizes considered here.

In all populations we conclude that the prevalence of Oral Candidiasis among ARC events is not linked to randomisation arm but seems higher in events with a higher CD4 count and earlier follow-up time. This association with time is weaker for a typical event of a typical patient than among all events, a feature that remains after adjustment for randomisation arm and CD4. We have learnt that whilst Oral Candidiasis is a very common event at early follow-up times, at the 'aggregate level', i.e. considering all events, its prevalence declines sharply and at later times it represents only a small proportion of all the events diagnosed and so requiring a clinical response. However, considering the experience of a typical patient, the decline in prevalence over time is less marked, and it remains a relatively common event diagnosis. As revealed by our

initial analysis exploring the interaction between cluster size and time, the difference in findings across populations arises because at later times there are some patients with several events where Oral Candidiasis is not diagnosed, and these patients disproportionately affect inference for all events.

3.11 A related recent approach

We have documented the work in Sections 3.6-3.10 in reports to the Department of Statistical Science, UCL. We have also presented part of it at the International Society of Clinical Biostatistics (ISCB) conference held in Prague in August 2009 (Pavlou et al., 2009) and at an open seminar at MRC-CTU. It later became evident that Huang and Leroux (2011) were independently developing similar methods.

They use slightly different terminology. They define *informative covariate distribution* to arise when the covariate of interest is categorical cluster-varying and the expected outcome is related to the probability of a member having a certain covariate level. We interpret non-informative covariate distribution to mean that

$$f(Y \mid \boldsymbol{X}, \boldsymbol{X}^*, N) = f(Y \mid \boldsymbol{X}, N).$$

Otherwise, we say that the covariate distribution is informative.

This definition can be seen to be similar to our definition of informative covariate structure (see Definition 3.2, pg. 67) which is defined in terms of expectations rather than probability distributions. Huang and Leroux (2011) clarify that the covariate of interest might be (i) a characteristic inherent to the cluster under study or (ii) a treatment applied to it. They present their exposition for scenarios where the covariate of interest is binary and is termed 'exposure'. They clarify that the extension to the case of a categorical covariate is straightforward. They consider methods for *non-manipulable* and *manipulable exposure*. We briefly describe these below.

3.11.1 Non-manipulable exposure

When the exposure is non-manipulable, the covariate of interest is assumed to be an inherent characteristic of the cluster under study. Huang and Leroux (2011) proposed two estimators, DWGEE1 and DWGEE2, to provide inference for a further population of typical cluster members where each cluster contributes one member at random for each value of X.

Firstly, when all values of the exposure are realised in every cluster they propose the DWGEE1 estimator to provide consistent estimation of the effect of the exposure. For member j of cluster i we define Z_{ij} to denote the number of members in cluster i with the same exposure level as member j. DWGEE1 are weighted independence estimating equations where each cluster member is inversely weighted by Z_{ij}. The DWGEE1 estimator can be seen to be identical to the WIEE-C3 (and also WIEE-C2) estimator we proposed in Section 3.6.3.

Secondly, when not all clusters include all exposure levels, Huang and Leroux (2011) propose DWGEE2, which requires parametric modelling of the number of members in each cluster with certain exposure levels. We let $Z_{i,x}$ denote the number of members in cluster i with $X = x$ and $Z_{iX_{ij}}$ (or Z_{ij} as above for simplicity) the number of members in cluster i with exposure level equal to X_{ij}. Let $\boldsymbol{L}$ denote a vector of observed *cluster-level* covariates which are assumed to describe the frequency distribution of $Z_{i,x}$. DWGEE2 is analogous to the DWGEE1 estimator, but member j in cluster i is inversely weighted by the *expected* (rather than the observed) value of Z_{ij}. The expected value of Z_{ij} is obtained by fitting (separately for each level of exposure) a model for $Z_{i,x}$ in terms of $\boldsymbol{L}$. If there exists a natural maximum value for $Z_{i,x}$, then $Z_{i,x}$ is modelled using binomial regression; otherwise Poisson regression can be used. The necessary condition for the validity of DWGEE2 is that conditional on $\boldsymbol{L}$ the covariate distribution is not informative, i.e. $f(Y \mid \boldsymbol{X}, \boldsymbol{X}^*, N, \boldsymbol{L}) = f(Y \mid \boldsymbol{X}, N)$.

So, the DWGEE2 are

$$\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \boldsymbol{\Pi}_i \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{3.20}$$

where K is the number of clusters in the sample and $\boldsymbol{V}_i$ is the diagonal matrix with jth element equal to the assumed variance function $v(\mu_{ij}) = \text{var}(Y_{ij} \mid \boldsymbol{X}_{ij})$. $\boldsymbol{\Pi}_i$ is a $N_i \times N_i$ diagonal matrix whose jth diagonal element, π_{ij}, is the inverse expected value of Z_{ij}. Consistent estimates, $\hat{\pi}_{ij}$, for π_{ij}, are substituted in $\boldsymbol{\Pi}_i$. If the model for π_{ij} is correctly specified, equations (3.20) provide consistent parameter estimation for a further population for inference which does not coincide with populations C1, C2 or C3, in general. We discuss further this population for inference in Section 3.12.1.

3.11.2 Manipulable exposure

Huang and Leroux (2011) also considered the scenario where the covariate of inter-

est (previously termed exposure) is not a characteristic inherent to the cluster, but a treatment applied to it. The expected outcome may be related to the probability of a member of a cluster receiving the treatment and ignoring this dependence may lead to biased estimation of the treatment effect. For example, when patients are clustered within clinics, factors may relate to the probability of a patient receiving the treatment and the potential health outcome for the patient. This issue is known as 'informative treatment propensity' and use of inverse probability of treatment has been used for the type of informativeness (Robins et al., 2000). Huang and Leroux (2011) extended the inverse probability of treatment approach for scenarios where not only the treatment propensity, but also the cluster size might be informative. Such a scenario would arise if the size of the clinic to which a patient belongs, is related to the patient's potential outcome.

Let Y denote the outcome for a member and T the treatment indicator. For $t = 0, 1$ let also $D(t)$ denote the binary indicator of receiving treatment t and $Y(t)$ the potential outcome given treatment t. $\boldsymbol{W}$ denotes a vector of variables that define treatment allocation so that $D(t) \perp\!\!\!\perp Y(t) \mid \boldsymbol{W}$. Also, let $\boldsymbol{S}$ be a subset of $\boldsymbol{W}$ which is included in the speculated regression model. The target is to estimate the *potential effect* of the treatment on outcome, adjusting for other covariates of interest, $\boldsymbol{S}$, for a typical member of a typical cluster. The relevant regression model is:

$$E(Y \mid T, \boldsymbol{S}) = h^{-1}(T\beta_T + \boldsymbol{S}^T\boldsymbol{\beta}_S)$$

and the parameters of interest are β_T and $\boldsymbol{\beta}_S$.

Let $\theta_{ij} = \theta(T_{ij}, \boldsymbol{W}_{ij}) = P(T_{ij} = t \mid \boldsymbol{W}_{ij})$ denote the probability that a member receives treatment t. The proposed estimating equations, DWGEE3, are

$$\sum_{i=1}^{K} \frac{1}{N_i} \frac{\partial \boldsymbol{\mu}_i^T}{\partial(\beta_T, \boldsymbol{\beta}_S)} \boldsymbol{\Theta}_i \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{3.21}$$

where $\boldsymbol{\Theta}_i$ is a $N_i \times N_i$ diagonal matrix whose jth element, θ_{ij}, is the probability that the jth member receives the treatment T_{ij}. Consistent estimates, $\hat{\theta}_{ij}$, for θ_{ij}, are substituted in $\boldsymbol{\Theta}_i$. If the model for θ_{ij} is correctly specified, equations (3.21) provide consistent parameter estimation of the potential treatment effect for a typical member of a typical cluster. Huang and Leroux (2011) noted that in the case where all clusters have both treatments, Z_{ij} is a consistent estimate of the member specific weight $N_i\ \theta(T_{ij}, \boldsymbol{W})$, so the DWGEE1 method remains valid, although potentially less efficient than DWGEE3.

3.12 Use of DWGEE1/2 and future work

3.12.1 DWGEE2: use and limitations

Both DWGEE1 and DWGEE2 aim to estimate the within-cluster effect of the exposure. Huang and Leroux (2011) state that when not all exposure levels are present in all clusters there are two possible populations for inference. The first is the one where attention is restricted to clusters which contain all levels of exposure and for the analysis the rest of the clusters are discarded. The DWGEE1 method provides inference for this population.

The DWGEE2 method seeks inference for the second population which is not well defined by Huang and Leroux (2011). We interpret inference using DWGEE2, as one which aims to answer the question: "What would have been the within-cluster effect of the exposure if in clusters with one level of exposure only, the other level had been realised as well?". Clearly, it is assumed that the exposure levels which are not present could have been observed but they are actually missing. So, DWGEE2 can be regarded as a method for complete-cluster inference where the observed-cluster members are up-weighted to represent both themselves and missing members from clusters with the same value of $\boldsymbol{L}$ but also down-weighted as if there existed only one member with each exposure level in every cluster. Introducing the weights, precisely serves the purpose of creating a pseudopopulation of complete clusters where all clusters experience all levels of exposure. Importantly, this pseudopopulation depends on the choice of $\boldsymbol{L}$.

3.12.2 Practical application

In real scenarios, the applicability of DWGEE2 can be limited, mainly because of the restrictive nature of the covariates in the model for $Z_{i,x}$. In the absence of suitable cluster-constant auxiliary variables application of DWGEE2 is not feasible.

Huang and Leroux (2011) neither provide guidance for the choice of suitable $\boldsymbol{L}$ nor comment on methods for assessing the appropriateness of the model for $Z_{i,x}$. In their data illustration, gender and cluster size are used as predictors in the model for $Z_{i,x}$, but this choice is not explained. Arguably, the choice of the set of covariates in $\boldsymbol{L}$, can be based on exploratory analysis. To examine whether a covariate is suitable to be included in $\boldsymbol{L}$, a reasonable strategy would be to examine the distribution of each candidate auxiliary variable at the different levels of the exposure. Any imbalances would indicate

that such a covariate is a potential confounder of the effect of the exposure and is suitable to be included in $\boldsymbol{L}$. For practical application, we propose (as Huang and Leroux (2011) did in their data application) that N is included as a predictor in the model for $Z_{i,x}$ since is inherently related to $Z_{i,x}$ (e.g. $Z_{i,x} \leq N_i$).

Another issue which may cause problems in the application of DWGEE2 is the presence of unusually small/large weights or variable weights which may lead to unstable parameter estimates (Joffe et al., 2004). One way to protect against the first is to exclude clusters with unusually large or small weights or 'trim' the extreme weights.

In the application of DWGEE2, it is not clear whether $\boldsymbol{L}$ is allowed to involve covariates from the regression model for Y. There is a greater issue of whether to include cluster-constant covariates in the main regression model as part of $\boldsymbol{X}$, or use them to construct the weights, i.e. as part of $\boldsymbol{L}$. This issue is not discussed by Huang and Leroux (2011). In analogy with marginal structural models, one reason for using such covariates in $\boldsymbol{L}$ rather than including them in $\boldsymbol{X}$, is to aid variable selection of $\boldsymbol{X}$ (structural part of the model) and also to produce a more meaningful model which only includes covariates, $\boldsymbol{X}$, of main interest and not nuisance covariates. For the case of non-manipulable exposure, it is of interest to investigate whether cluster-varying auxiliary variables can contribute to $\boldsymbol{L}$ (e.g. through a cluster-level summary) as this may broaden the range of scenarios where DWGEE2 can be applied.

The extension of DWGEE1, DWGEE2 and DWGEE3 to scenarios where $\boldsymbol{X}$ is continuous remains an open question. For practical application we proposed categorising the continuous covariate, to make the application of DWGEE1 feasible. Such an approach was used in our illustration in Section 3.10. Two issues arise from using this approach. First, the selection of number of categories and the 'cut-off' points is subjective. Second, collapsing the values of a continuous covariate into a categorical one involves discarding useful information, thus reducing the precision of the parameter estimates. In response to the second issue, a possible solution would be to use a sensible categorisation to obtain the weights but use the original values of the continuous covariate when applying the WIEE. In this way, the information in $\boldsymbol{X}$ is fully utilised and the members receive suitable weights. The underlying assumption is that for the subcluster formed by members of a cluster in the same category, the covariate structure is non-informative. Whether this assumption is satisfied depends on the selection of

categories.

3.12.3 Choice of method

When all levels of exposure are realised in all clusters, a recommended approach would be to use WIEE-C3 (DWGEE1). When not all exposure levels are present in all clusters the analyst has two possible choices of method: WIEE-C3 (DWGEE1) or DWGEE2. We aim to estimate the within-cluster effect of the exposure in all clusters and the choice of method can be based on the efficiency/bias trade-off.

We first assume that the effect of the exposure is homogeneous (i.e. the effect of the exposure is the same in all clusters, see also Definition 2.1, pg. 52). If the proportion of clusters with missing exposure levels is small, we suggest using WIEE-C3 since it does not require any assumptions about the missingness mechanism and can remove confounding due to measured or unmeasured cluster-level confounders. If a large proportion of clusters have missing exposure levels, application of DWGEE1 can be inefficient. If the analyst is prepared to make assumptions about the missingness of exposure levels, DWGEE2 may then be considered (subject to availability of suitable cluster-level auxiliary variables).

If the effect of the exposure is non-homogeneous, WIEE-C3 will only provide consistent estimation if all clusters contain all levels of exposure; otherwise, estimation using WIEE-C3 will generally be inconsistent. When cluster-level auxiliary variables are available, the analyst may use DWGEE2 where $\boldsymbol{L}$ must be chosen by thinking of what can predict missingness of exposure levels.

If confounding is due to cluster-level confounders, these are observed and we wish to estimate the effect of the exposure adjusting for these confounders in a regression model, a further possibility would be to include these confounders in $\boldsymbol{X}$ and seek inference for population C2. If the effect of the exposure is assumed to vary between different clusters, we may fit a regression model for the expected outcome in population C2, adjusting for the cluster-level confounders and their interactions with exposure. Arguably, application of WIEE-C2 would lead to more complex models for the expected outcome compared to the application of DWGEE2, in which the observed confounders are only used to obtain the weights and are not part of $\boldsymbol{X}$.

3.12.4 Relation of DWGEE1/2 to methods estimating the within-cluster effect of the exposure

Informative covariate distribution can be seen to be strongly related to cluster-confounding (Palta and Yao, 1991; Neuhaus and Kalbfleisch, 1998; Neuhaus and MCculloch, 2006; Goetgeluk and Vansteelandt, 2008; Brumback et al., 2010). DWGEE1/2 are naturally viewed as methods to estimate the within-cluster effect of the exposure. However, as discussed in Section 2.8, there are alternative methods to estimate the effect of the exposure under cluster-confounding by using only within-cluster comparisons. These are:

1. GLMMs which separate between- and within-cluster effects (GLMMBW-Neuhaus and Kalbfleisch, 1998),
2. Conditional GEE (CGEE-Goetgeluk and Vansteelandt, 2008),
3. Conditional likelihood (CL-Neuhaus and MCculloch, 2006).

There are two important issues to consider when choosing a method: whether the effect of the exposure is homogeneous and whether the cluster size is informative.

DWGEE1 provides marginal inference and is valid when the cluster size is informative. If all values of the exposure are present in every cluster, it does not require the condition of homogeneous exposure effect to be true. When the within-cluster effect is different for every cluster, it estimates what can be seen as an 'average' within-cluster-effect, where each cluster contributes equally to the estimation. It requires the exposure to be categorical. If not all values of the exposure are present in each cluster, then DWGEE2 can be used instead, subject to availability of auxiliary variables which adequately predict missingness of exposure levels.

In the special case where the link function is the identity one, the exposure effect is homogeneous, and the exposure is categorical and all clusters contain all values of the exposure, all four methods (DWGEE1, GLMMBW, CGEE and CL) can be used to estimate the effect of the exposure. However, DWGEE1 is expected to be less efficient than the other methods, because in DWGEE1 all clusters are equally weighted in the estimation even if some are smaller or less informative than others.

GLMMBW, CGEE and CL provide cluster-specific inference under cluster-confounding and do not require the exposure to be categorical, as does DWGEE1/2.

Nevertheless, GLMMBW, CGEE and CL can be used to estimate the within-cluster effect of the exposure under the condition of homogeneous exposure effects. When the effect of the exposure is not homogeneous, we do not expect any of these methods to consistently estimate the within-cluster effect of the exposure. Adaptation of these methods for such scenarios requires further investigation. As part of future work, it is of interest to investigate the performance of GLMMBW, CGEE and CL under cluster-confounding and also informative cluster size. We would expect these methods to provide consistent estimation for the within-cluster effect of the exposure when the cluster size is informative but only affects intercept term (see Section 3.5.3, pg. 75).

3.13 Discussion

In this chapter we have investigated existing methods for informative cluster size and also considered informative cluster size in more general scenarios than previous authors. We have defined informative covariate structure and additional populations for inference with corresponding estimation methods.

We initially explained that adjusting for cluster size in a marginal regression model for the expected outcome is not in general a good approach to deal with informative cluster size, except in scenarios where cluster size is a predictor of scientific interest such as in volume-outcome studies. We have also investigated an approach where a model for the expected outcome conditional on cluster size and covariates is assumed, but then the 'marginal effect' of covariates can be obtained by marginalising over the distribution of the cluster size. We identified limitations in using such an approach but also noted that it can be useful in certain scenarios and therefore can be considered for further investigation. From these two considerations it becomes apparent that special methods are needed to deal with informative cluster size.

We have examined informative cluster size in more general scenarios than previous authors (Hoffman et al., 2001; Williamson et al., 2003; Benhin et al., 2005). We questioned the suitability of the population of typical cluster members 1 proposed, clarifying that it can only be useful in simple cases where the covariates are cluster-constant or cluster-varying but size-balanced. For these cases, we identified scenarios where the covariate effects can be equal in inference for populations M and C1, and listed conditions under which the intercepts as well as the covariate effects can be different between

the two populations.

Importantly, we have extended the field of research beyond informative cluster size to informative covariate structure. We defined additional populations (C2 and C3), inference for which is more interpretable than inference for population C1. Under informative covariate structure standard methods such as GEE with a realistic working correlation do not provide consistent estimation for any well-defined population for inference. Whilst informative cluster size is known to occur widely in medical research, by contrast we are unaware of how commonly the problem of informative covariate structure occurs, and how commonly it is considered when clustered data are analysed. Our proposed method of estimation by weighted independence estimating equations (WIEE) is easy to implement in standard statistical software. We hope to see our proposed methodology implemented across a range of study types, in particular so that the relative merits of the possible populations for inference in different practical situations can be better understood.

We emphasise that the choice between populations C2 and C3 depends on the choice of which covariates to investigate. In scenarios where it is wished to remove confounding by adjusting for many covariates, and particularly cluster-varying covariates, population C2 will become very similar to population M and population C3 may be (nearly) empty. In such scenarios with many confounders, performing a 'cluster-level' analysis can be problematic. Categorising continuous covariates may allow a cluster-level analysis but will clearly involve some loss of information.

Huang and Leroux (2011) independently identified the issue of informative covariate structure (in their terminology, informative covariate distribution). They have proposed an estimator (DWGEE1) which is identical to our WIEE for inference for population C3. They also considered further estimators (DWGEE2, DWGEE3) which require models for the frequency distribution of the covariate of interest (exposure). The validity of DWGEE2 and DWGEE3 relies on the correct specification of the frequency distribution of the exposure. In our view, DWGEE2 is a method for complete-cluster inference where the target is to estimate the within-cluster effect of the exposure and all complete clusters contain all levels of exposure. For scenarios where the all clusters contain all values of the exposure, DWGEE1 consistently estimates the within-cluster effect of the exposure, regardless of whether the effect of the exposure is homogeneous

or not. When not all clusters contain all values of the exposure and the effect of the exposure is not homogeneous DWGEE1 does not consistently estimate the within-cluster effect of the exposure. In this scenario, if the analysts are prepared to make assumptions about the missingness of exposure levels, then they may consider use of DWGEE2/3 (subject to availability of suitable auxiliary variables).

In this chapter, we have focused on applications involving outcome measurements concerning repeated episodes, in contrast to previous authors of papers regarding informative cluster size. We acknowledge that for such longitudinal data there may alternatively be interest in directly modelling the risk of episodes of different types. Take for example our illustrative application to data from an HIV trial; we are aware that the risk of particular adverse events could be modelled directly through using hazard models (Prentice et al., 1978). By applying our method here, we are assessing the characteristics of adverse events given that they have happened, and such a model could be complemented by a model for adverse events of any type (Hachen, 1988; Ghilagaber, 1998). More generally we acknowledge that there is often interest in formulating a model for the cluster size, as forms part of those approaches that jointly model size and outcome measures (Dunson et al., 2003; Gueorguieva, 2005), or in answering 'causal' questions concerning what might happen were clinical practice to change.

In many longitudinal studies concerning episodic data, as well as cross-sectional studies such as in dental research, informative cluster size and/or covariate structure may arise and standard methods may be appreciably biased. The methods we present provide easily implementable ways to perform unbiased regression analysis for well-defined populations of interest.

A.1 Proof of consistency and asymptotic Normality

Here we prove that the estimator that solves equations (3.16) is consistent and asymptotically Normally distributed when equation (3.15) holds and standard regularity conditions are satisfied. We also show that the variance estimator given by equation (3.17) is a consistent estimator of the variance of this parameter estimator. We assume that the maximum cluster size, $N_{\max}$, is finite. This proof is for population C2, but it can be adapted for the other populations. For population C2 the true value of $\boldsymbol{\beta}$ is the solution to

$$E_{IJ}[\frac{\partial \mu_{IJ}}{\partial \boldsymbol{\beta}^T}\boldsymbol{V}_{IJ}^{-1}(Y_{IJ}-\mu_{IJ})]=0, \tag{A-1}$$

where $\mu_{ij}=h^{-1}(\beta_0+\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1)$, Y_{ij}, $\boldsymbol{X}_{ij}$, and $\boldsymbol{V}_{ij}$ are the observed values of Y, $\boldsymbol{X}$, and the assumed variance for the jth measurement in cluster i, and (I,J) is a randomly chosen measurement in population C2. Since the probability that a particular measurement is chosen (when population C2 is formed from the overall population of all members - see Section 3.6) is inversely proportional to Z, then (A-1) can be written

$$E_I[\sum_{j=1}^{N_I}\frac{1}{Z_{Ij}}\frac{\partial \mu_{Ij}}{\partial \boldsymbol{\beta}^T}\boldsymbol{V}_{Ij}^{-1}(Y_{Ij}-\mu_{Ij})]=0. \tag{A-2}$$

Now (A-2) can be written

$$E_I[\sum_{j=1}^{N_{\max}}\frac{R_{Ij}}{\widetilde{Z}_{Ij}}\frac{\partial \widetilde{\mu}_{Ij}}{\partial \boldsymbol{\beta}^T}\widetilde{\boldsymbol{V}}_{Ij}^{-1}(\widetilde{Y}_{Ij}-\widetilde{\mu}_{Ij})]=0, \tag{A-3}$$

where $R_{ij}=1$ if $N_i\geq j$ and $R_{ij}=0$ otherwise, and $\widetilde{Z}_{ij}$, $\widetilde{\mu}_{ij}$, $\widetilde{\boldsymbol{V}}_{ij}$ and $\widetilde{Y}_{ij}$ are equal to Z_{ij}, μ_{ij}, $\boldsymbol{V}_{ij}$ and Y_{ij} if $R_{ij}=1$ and are equal to any non-zero value if $R_{ij}=0$.

Then (A-3) can be written

$$E_I[\widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{X}}_I,\boldsymbol{\beta})(\widetilde{\boldsymbol{Y}}_I-\widetilde{\boldsymbol{\mu}}_I(\widetilde{\boldsymbol{X}}_I,\boldsymbol{\beta}))]=0,$$

where $\tilde{\mathbf{Y}}_i=(\widetilde{Y}_{i1},...,\widetilde{Y}_{iN_{\max}})$, $\widetilde{\boldsymbol{\mu}}_i=(\widetilde{\mu}_{i1},...,\widetilde{\mu}_{iN_{\max}})$, $\widetilde{\mathbf{X}}_i=(\widetilde{\boldsymbol{X}}_{i1},...,\widetilde{\boldsymbol{X}}_{iN_{\max}})$, and

$$\widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{X}}_I,\boldsymbol{\beta})=\frac{\partial \widetilde{\boldsymbol{\mu}}_I}{\partial \boldsymbol{\beta}^T}\widetilde{\boldsymbol{W}}_I\widetilde{\boldsymbol{V}}_I^{-1}$$

where $\tilde{\boldsymbol{V}}_i=\text{diag}(\widetilde{\boldsymbol{V}}_{i1},...,\widetilde{\boldsymbol{V}}_{iN_{\max}})$ and $\widetilde{\boldsymbol{W}}_I=\text{diag}(R_{I1}\widetilde{Z}_{I1}^{-1},...,R_{IN_{\max}}\widetilde{Z}_{IN_{\max}}^{-1})$.

It follows from Tsiatis (2006, pg. 54-57), that the solution $\widehat{\boldsymbol{\beta}}$ to the estimating equations, where K is the number of clusters in the sample,

$$\sum_{i=1}^{K}\widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{X}}_i,\boldsymbol{\beta})[\widetilde{\boldsymbol{Y}}_i-\widetilde{\boldsymbol{\mu}}_i(\widetilde{\boldsymbol{X}}_i,\boldsymbol{\beta})]=\mathbf{0} \tag{A-4}$$

is asymptotically Normally distributed with mean equal to the true value of $\boldsymbol{\beta}$ and variance

$$E(\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}})^{-1}E(\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{A}}^T)E(\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}})^{-1^T}, \tag{A-5}$$

where

$$\widetilde{\boldsymbol{D}}(\widetilde{\boldsymbol{X}}, \boldsymbol{\beta}) = \frac{\partial \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{X}}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}.$$

Finally, we show equivalence between these estimating equations and the weighted independence equations we proposed in Section 3.6.3. It is easy to see that

$$\widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{X}}, \boldsymbol{\beta}) = \begin{bmatrix} \boldsymbol{A}(\boldsymbol{X}_i, \boldsymbol{\beta}) & \mathbf{0}_{K\times(N_{\max}-N_i)} \\ \mathbf{0}_{(N_{\max}-N_i)\times N_i} & \mathbf{0}_{(N_{\max}-N_i)\times(N_{\max}-N_i)} \end{bmatrix} \tag{A-6}$$

where $\boldsymbol{A}(\boldsymbol{X}_i, \boldsymbol{\beta}) = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}\boldsymbol{W}_i\boldsymbol{V}_i^{-1}$, $\boldsymbol{W}_i = \text{diag}(Z_{i1}^{-1}, \ldots, Z_{iN_i}^{-1})$, and $\mathbf{0}_{a\times b}$ denotes an $(a \times b)$ matrix of zeros. Hence (A-4) can be written

$$\sum_{i=1}^{K} \boldsymbol{A}(\boldsymbol{X}_i, \boldsymbol{\beta})[\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i(\boldsymbol{X}_i, \boldsymbol{\beta})] = \mathbf{0}$$

which is identical to (3.16).

Tsiatis (2006, pg. 56-57) also states that the variance in (A-5) can be estimated by a sandwich variance estimator, and by examining $\widetilde{\boldsymbol{D}}$ and $\widetilde{\boldsymbol{V}}$ as we examined $\widetilde{\boldsymbol{A}}$ in (A-6) it can be shown that indeed the sandwich estimator we proposed in (3.17) is also identical to this estimator.

A.2 Computation of true regression parameter values

Here we outline the procedure we used to compute the true parameter values for the scenarios examined in Sections 3.8 and 3.9 for the population of all members and the three populations of typical cluster members. Let $f_A(.)$ denote the probability density (or distribution) function for a generic random variable A. Then, for population C1,

$$\begin{aligned} E^{C1}(Y \mid X = x) &= \int h^{-1}[E(Y \mid X = x, U = u)] f_U(u \mid X = x)\, du \\ &= \frac{\int h^{-1}[E(Y|X = x, u)] f_X(x \mid U = u) f_U(u)\, du}{\int f_X(x \mid U = u) f_U(u)\, du}. \end{aligned} \tag{A-7}$$

For population M we define Z_x to be the number of members with $X = x$ in the cluster. Now,

$$E^M(Y|X = x) = \frac{\int E(Z_x \mid U = u) h^{-1}[E(Y \mid X = x, U = u)] f_U(u)\, du}{\int E(Z_x \mid U = u) f_U(u)\, du}, \tag{A-8}$$

and $E(Z_x \mid U = u) = E(N \mid U = u) f_X(x \mid U = u)$. The integrals with respect to U in equations (A-7) and (A-8) were approximated using numerical quadrature with 200 quadrature points.

For populations C2 and C3 we further define probabilities for inclusion of a measurement in a population given the value, u, of U. These probabilities depend on the size of a cluster with $U = u$. For population C2 we let $p_x^{in}(u)$ denote the probability that a cluster with $U = u$ includes at least one measurement with $X = x$ and $p_x^{in}(u) = \sum_N f_N(n|u)[1 - P(X = x|U = u)^n]$. For population C3 at least one measurement with $X = 0$ and one with $X = 1$ has to be included in the cluster and the probability of this occurring is denoted by $p_{01}^{in}(u) = \sum_N f_N(n|u)[1 - (1 - P(X = 0|U = u))^n - (1 - P(X = 1|U = u))^n]$. Then,

$$E^{C2}(Y|X = x) = \frac{\int p_x^{in}(u) h^{-1}[E(Y|X = x, U = u)] f(u)\, du}{\int p_x^{in}(u) f(u)\, du},\ x = 0, 1 \qquad \text{(A-9)}$$

and

$$E^{C3}(Y|X = x) = \frac{\int p_{01}^{in}(u) h^{-1}[E(Y|X = x, U = u)] f(u)\, du}{\int p_{01}^{in}(u) f(u)\, du},\ x = 0, 1. \qquad \text{(A-10)}$$

Again, the integrals with respect to U in equations (A-9) and (A-10) were approximated using numerical quadrature with 200 quadrature points. For each quadrature point, $p_x^{in}(u)$ and $p_{01}^{in}(u)$ were computed using Monte Carlo integration over the distribution of N by simulating 10000 cluster sizes for each value of u.

A.3 R-code for the computation of true regression parameter values

We here present an R-software implementation of the general procedure described in Appendix A.2. The code presented is an example of the code used for the computation of the true parameter values in Section 3.8. In particular it was used to compute of true parameter values of the 'analysis model' in each population for inference, assuming that the data were generated under the procedure described in page 88 for Gaussian responses and non-size-balanced X.

```
# u is a vector of 200 quadrature points
u <- qnorm(seq(0.0025, 0.9975, 0.005)) * 0.5
px1 <- expit(u)
```

```
pin0 <- rep(0, length(u))
pin1 <- rep(0, length(u))
pin01 <- rep(0, length(u))
exp.numx0 <- rep(0, length(u))
exp.numx1 <- rep(0, length(u))

# Nclust is the number of the simulated clusters for
# the Monte Carlo integration at each quadrature point
Nclust <- 1e4
for (i in 1:length(u))
{n <- rpois(Nclust, exp(a0+a1*u[i]))+1
pin0[i] <- mean(1 - px1[i]^n)
pin1[i] <- mean(1 - (1-px1[i])^n)
pin01[i] <- mean( 1 - px1[i]^n - (1-px1[i])^n )
exp.numx0[i] <- (1-px1[i]) * mean(n)
exp.numx1[i] <- px1[i] * mean(n)
}

# beta0.x and beta1.x are the true values of beta_0 and
# beta_1 for populations M,C1,C2 and C3 respectively

beta0.m <- mean(exp.numx0*(g0+g2*u))/mean(exp.numx0)
beta0.c1<- mean((g0+g2*u)*(1-px1))/mean(1-px1)
beta0.c2<- mean((g0+g2*u)*pin0)/mean(pin0)
beta0.c3<- mean((g0+g2*u)*pin01)/mean(pin01)

beta1.m <- mean(exp.numx1*(g0+g1+g2*u+g3*u))/mean(exp.numx1)
           -beta0.m
beta1.c1<- mean((g0+g1+g2*u+g3*u)*px1)/mean(px1)-beta0.c1
beta1.c2<- mean((g0+g1+g2*u+g3*u)*pin1)/mean(pin1)-beta0.c2
beta1.c3<- mean((g0+g1+g2*u+g3*u)*pin01)/ mean(pin01)-beta0.c3
```

Chapter 4

Efficient estimation methods when the cluster size is informative

4.1 Introduction

In the previous chapter we introduced informative cluster size, as this was defined by previous authors. When the cluster size is informative, inference might be for the population of all members or the population of typical members 1. We also defined informative covariate structure and two populations for inference, additional to the ones previously considered. Marginal inference for the population of all members and the three populations of typical cluster members can be obtained using weighted independence estimating equations.

Williamson et al. (2003) provide a useful discussion regarding the use of a WIEE or IEE with a realistic correlation structure when the cluster size is informative. Application of GEE with a non-diagonal correlation matrix may increase the efficiency of the parameter estimates compared to independence, by recognising the correlation among the members of each cluster. They note that when the cluster size is informative, use of non-diagonal correlation structures generally causes bias because the weight attributed to each member is altered, the total weight attributed to the cluster is altered and the result is inference for none population for inference.

To provide inference for population C1 a potentially more efficient method (MWCR) was proposed by Chiang and Lee (2008), based on the WCR method. When the minimum cluster size, m, is greater than 1, the authors proposed randomly sampling m members from each cluster and then applying the GEE with a realistic working cor-

relation to each resampled dataset. As the intracluster correlation is accounted for, efficiency may be gained.

Previous authors focused primarily on simple cases of informative cluster size, in the sense that the covariates of interest were either cluster-constant or cluster-size balanced (see Definition 3.3, pg. 68). These authors also focused on scenarios in which the expected value of the outcome depends on cluster size and covariates but not on interactions between the two. In this work we consider more general scenarios where the covariates involved are cluster-varying and non-size balanced. We explain why MWCR may lead to biased inference in these cases, a fact that is not clear in the original presentation of the method (Chiang and Lee, 2008). Furthermore, bias in MWCR can arise from realistic choices of the working correlation.

In addition, we derive an alternative estimator that is suitable in certain situations and which has the potential to be more efficient than WIEE. We call this method WRGEE because it may be used with a realistic(R) working correlation, rather than requiring the use of the independence working assumption. We compare the performance of WRGEE to MWCR for scenarios where they are both unbiased and also show how WRGEE can give unbiased inference with moderate efficiency gains relative to WIEE in certain scenarios where the MWCR method is biased.

In the next section we introduce the standard notation used in GEE, present the MWCR method and explain why bias can occur with MWCR in some scenarios. In Section 4.3 we present the WRGEE method. In Section 4.4 we use simulation studies to compare the performance of the methods in terms of bias and relative efficiency. We apply the methods to AIDS related conditions data from the Delta Trial of HIV treatment in Section 4.5. Finally, we discuss our results and possible future extensions of the methodology.

4.2 Existing methods of estimation

4.2.1 Standard GEE and notation

In this chapter we use a slightly different notation from the one used when introducing GEE in Section 2.4.3. We start by introducing the additional notation. As before, a marginal regression model $\mu(\boldsymbol{X}) = h^{-1}(\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}_1)$ is specified, where h is a known canonical link function and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$ is a $(q+1)$-dimensional vector of unknown

parameters of interest.

We consider the GEE1 of Liang and Zeger (1986) as described by equations (2.14). In this chapter, the working correlation structure is assumed to be either independence, exchangeable, auto-regressive or fixed. Depending on this choice, the actual working correlation may involve unknown parameters $\boldsymbol{\rho}$ that need to be estimated. Let $\hat{\boldsymbol{\rho}}$ denote the estimate of $\boldsymbol{\rho}$. We assume that $\hat{\boldsymbol{\rho}}$ converges to a value $\boldsymbol{\rho}_0$ as $K \to \infty$. Let $\hat{\boldsymbol{R}}_i$ and $\boldsymbol{R}_i$ denote the working correlation matrix for cluster i evaluated at $\hat{\boldsymbol{\rho}}$ and at $\boldsymbol{\rho}_0$, respectively. Note that $\hat{\boldsymbol{R}}_i$ and $\boldsymbol{R}_i$ can depend on observed variables which may or may not be included in $\boldsymbol{X}^*$. For example, for an auto-regressive working correlation structure, they will depend on the times of observation of the members in the clusters, and time may or may not be included as a variable in the analysis model.

If cluster size is non-informative, $E(Y \mid \boldsymbol{X})$ and $E^{C1}(Y \mid \boldsymbol{X})$ are the same, i.e. the expectation of Y given $\boldsymbol{X}$ is the same in the populations of all members and typical members 1. If, furthermore, the marginal model $\mu(\boldsymbol{X}) = h^{-1}(\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}_1)$ is correctly specified, then under regularity conditions the solution $\hat{\boldsymbol{\beta}}$ to the following GEE is a consistent asymptotically Normally distributed estimator of $\boldsymbol{\beta}$:

$$\sum_{i=1}^{K} \boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{Y}_i^*, \boldsymbol{X}_i^*) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i) = \boldsymbol{0}, \tag{4.1}$$

where $\hat{\boldsymbol{V}}_i = \boldsymbol{A}_i^{1/2}\hat{\boldsymbol{R}}_i\boldsymbol{A}_i^{1/2}\phi$ is the working covariance matrix for cluster i and $\boldsymbol{A}_i$ is the $N_i \times N_i$ diagonal matrix whose jth diagonal element is $v(\mu_{ij})$.

Let r_{ilj} denote the $(l,j)^{th}$ element of $\boldsymbol{R}_i^{-1}$, and let $r_{i+j} = \sum_{l=1}^{N} r_{ilj}$ and $r_{i++} = \sum_{j=1}^{N} r_{i+j}$. Let $\hat{r}_{ilj}$, $\hat{r}_{i+j}$ and $\hat{r}_{i++}$ denote the analogous quantities for $\hat{\boldsymbol{R}}_i^{-1}$.

In preparation for Section 4.2.3, it will be useful to consider the special case in which the identity link function $h^{-1}(\theta) = \theta$ is used and there are no covariates. In this case, β_0 is just the population mean of Y. Equation (4.1) then becomes

$$\sum_{i=1}^{K} \boldsymbol{U}(\beta_0; \boldsymbol{Y}_i^*) = \sum_{i=1}^{K} \sum_{j=1}^{N_i} \frac{1}{\phi} \hat{r}_{i+j}(Y_{ij} - \beta_0) = 0,$$

to which the solution is $\hat{\beta}_0 = \sum_{i=1}^{K} \sum_{j=1}^{N_i} \hat{r}_{i+j} Y_{ij} / \sum_{i=1}^{K} \hat{r}_{i++}$. Thus, $\hat{\beta}_0$ is a weighted average of the Y_{ij}'s in which the total weight given to the measurements in cluster i is $\hat{r}_{i++}$.

Recall that when cluster size is informative, estimating equations (4.1) will not, in general, give consistent estimation for either population M or C1. We previously defined $\mu(\boldsymbol{X}) = E(Y \mid \boldsymbol{X})$. This is the expectation of Y in the population of all members and recall that is no longer equal to the expectation in the population of typical members 1. Define, analogously, $\mu^{C1}(\boldsymbol{X}) = E^{C1}(Y \mid \boldsymbol{X})$ for the population of typical cluster members 1.

4.2.2 A more efficient method for the population of typical cluster members

The MWCR method, proposed by Chiang and Lee (2008), is a modification of the WCR method and, like it, provides inference for the population of typical cluster members 1 (population C1). It can be more efficient than WCR when m, the size of the smallest cluster that appears in the dataset, is greater than 1. As it is evident from its asymptotic equivalence to WIEE, WCR effectively uses an independence working correlation. MWCR, on the other hand, allows a non-independence working correlation to be used. We now describe MWCR.

For any subcluster s composed of m elements from cluster i, let $\hat{\boldsymbol{V}}_{i(s)}$, $\boldsymbol{Y}^*_{i(s)}$ and $\boldsymbol{\mu}^{C1}_{i(s)}$ denote, respectively, the submatrix of $\hat{\boldsymbol{V}}_i$ and the subvectors of $\boldsymbol{Y}^*_i$ and $\boldsymbol{\mu}^{C1}_i$ corresponding to those m members. There are two versions of the MWCR method, the first of which is more intuitively understandable but also more computationally intensive. The first resembles WCR; the second, WIEE.

In the first version of MWCR, Q datasets are created from the original dataset by each time sampling at random (and without replacement) m members from each of the K clusters. So, each dataset consists of K clusters each of m members. For each of these Q datasets, $\boldsymbol{\beta}$ is estimated using the standard GEE (equations (4.1)) with $\hat{\boldsymbol{V}}_i$ and $\boldsymbol{Y}^*_i$ replaced by the appropriate submatrix/subvector $\hat{\boldsymbol{V}}_{i(s)}$ and $\boldsymbol{Y}^*_{i(s)}$ and $\boldsymbol{\mu}_i$ replaced by $\boldsymbol{\mu}^{C1}_{i(s)}$. The resulting Q estimates of $\boldsymbol{\beta}$ are then averaged.

Since each cluster contributes m members to each estimate of $\boldsymbol{\beta}$ regardless of its size, the parameter estimated is that for the population of typical cluster members 1. Also, since the intracluster correlation is accounted for, MWCR may give increased efficiency relative to WCR.

The second version of MWCR is asymptotically equivalent to the first version (as

$K, Q \to \infty$), but avoids the Monte Carlo element. In this second version, $\boldsymbol{\beta}$ is estimated as the solution to weighted GEE

$$\sum_{i=1}^{K} \frac{1}{\Delta_i} \sum_{s=1}^{\Delta_i} \left(\frac{\partial \boldsymbol{\mu}_{i(s)}^{C1}}{\partial \boldsymbol{\beta}} \right)^T \hat{\boldsymbol{V}}_{i(s)}^{-1} (\boldsymbol{Y}_{i(s)}^{*} - \boldsymbol{\mu}_{i(s)}^{C1}) = \mathbf{0} \tag{4.2}$$

where $\Delta_i = \begin{pmatrix} N_i \\ m \end{pmatrix}$ denotes the number of subclusters of size m that can be formed from cluster i (no subcluster can contain the same member more than once) and these Δ_i subclusters are indexed $s = 1, \ldots, \Delta_i$. The correlation parameters $\boldsymbol{\rho}$ are estimated using the method outlined in Williamson et al. (2003). The weighted GEE (4.2) can be seen to be the sum of the contributions to standard GEE from each of the subclusters, with each subcluster inversely weighted by the number of subclusters that can be formed from its cluster. This weighting ensures that each of the K clusters contributes equally to the GEE, regardless of its size.

An easily computed variance estimator for $\hat{\boldsymbol{\beta}}$ is given by $\boldsymbol{H}^{-1}\boldsymbol{B}\boldsymbol{H}^{-1}$, where

$$\begin{aligned} \boldsymbol{H} &= \sum_{i=1}^{K} \frac{1}{\Delta_i} \sum_{s=1}^{\Delta_i} \left(\frac{\partial \boldsymbol{\mu}_{i(s)}^{C1}}{\partial \boldsymbol{\beta}} \right)^T \hat{\boldsymbol{V}}_{i(s)}^{-1} \frac{\partial \boldsymbol{\mu}_{i(s)}^{C1}}{\partial \boldsymbol{\beta}^T} \quad \text{and} \\ \boldsymbol{B} &= \sum_{i=1}^{K} \frac{1}{\Delta_i} \sum_{s=1}^{\Delta_i} \left(\frac{\partial \boldsymbol{\mu}_{i(s)}^{C1}}{\partial \boldsymbol{\beta}} \right)^T \hat{\boldsymbol{V}}_{i(s)}^{-1} (\boldsymbol{Y}_{i(s)}^{*} - \boldsymbol{\mu}_{i(s)}^{C1})(\boldsymbol{Y}_{i(s)}^{*} - \boldsymbol{\mu}_{i(s)}^{C1})^T \hat{\boldsymbol{V}}_{i(s)}^{-1} \frac{\partial \boldsymbol{\mu}_{i(s)}^{C1}}{\partial \boldsymbol{\beta}^T} \end{aligned}$$

are evaluated at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\rho}}$. This variance estimator was not clearly described by Chiang and Lee (2008).

4.2.3 Bias in MWCR for general covariate patterns

Chiang and Lee (2008) focus on scenarios where the covariates $\boldsymbol{X}$ are either cluster-constant or cluster-size balanced. In these special cases MWCR gives consistent estimation, but only with certain choices of working correlation. In general, MWCR is biased, a fact which is not evident in their paper. Here we state conditions under which equations (4.2) are consistent estimating equations for $\boldsymbol{\beta}$.

We assume a correctly specified marginal model $\mu^{C1}(\boldsymbol{X}) = h^{-1}(\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}_1)$ for population C1. Consider the following sampling mechanism for members which leads to population C1. A cluster is chosen at random from the population of clusters, a subcluster of size m is sampled at random (without replacement) from this cluster and, finally, a member is chosen at random from this subcluster. Recall that N denotes

the size of the chosen cluster and $\boldsymbol{X}^*$ denotes the covariate values for all members of the cluster. Also Y and $\boldsymbol{X}$ denote the outcome and covariate values for the chosen member of the chosen subcluster, and $\underset{\sim}{\boldsymbol{X}}$ denotes the covariate values for all the other $m-1$ members of the chosen subcluster. Let $\underset{\sim}{\boldsymbol{R}}$ denote the working correlation for the chosen subcluster when $\boldsymbol{\rho} = \boldsymbol{\rho}_0$.

Theorem 1

The solution to the MWCR estimating equations (4.2) is a consistent estimator of $\boldsymbol{\beta}$ for population C1 if the following conditions are satisfied:

1. $Y \perp\!\!\!\perp \underset{\sim}{\boldsymbol{X}} \mid \boldsymbol{X}$.
2. $N \perp\!\!\!\perp \underset{\sim}{\boldsymbol{R}} \mid \boldsymbol{X}, \underset{\sim}{\boldsymbol{X}}$.
3. $Y \perp\!\!\!\perp \underset{\sim}{\boldsymbol{R}} \mid N, \boldsymbol{X}, \underset{\sim}{\boldsymbol{X}}$.

Proofs of all theorems in this chapter are in Appendix B.1. We now discuss Conditions 1–3 of Theorem 1. Condition 1 is closely related to the assumption that $Y \perp\!\!\!\perp \boldsymbol{X}^* \mid \boldsymbol{X}$, a condition identified by Pepe and Anderson (1994) as necessary for consistent estimation when cluster size is constant and GEE are used with a non-independence working correlation. Note that when the cluster size is informative and $\boldsymbol{X}$ involves cluster-varying and non-size-balanced covariates this condition is violated.

Whether Conditions 2 and 3 are satisfied will depend on the choice of working correlation structure. Condition 2 would not be satisfied, for example, if an auto-regressive structure were used and the time intervals between members of larger clusters tended to be longer (or shorter) on average than the intervals between members of smaller clusters, even after taking into account the values of $\boldsymbol{X}$ in the subcluster. Condition 3 is the requirement that the working correlation for a randomly chosen subcluster be conditionally independent of the outcome Y of a randomly chosen member from that subcluster given the size of the cluster to which that member belongs and the covariate values of the members of the subcluster. It would not be satisfied, for example, if an auto-regressive structure were used and a member's Y value tended to be higher (or lower) in subclusters with longer time intervals between members than in subclusters with shorter intervals, even after taking into account the $\boldsymbol{X}$ values in the subcluster and

the size of the cluster from which the subcluster came. However, if an auto-regressive structure were used and time were one of the covariates $\boldsymbol{X}$ in the analysis model, then Conditions 2 and 3 would be satisfied. The independence and exchangeable working correlation structures are guaranteed to satisfy Conditions 2 and 3.

The necessity of Conditions 2 and 3 can be appreciated by considering the special case introduced at the end of Section 4.2.1: $h^{-1}(\theta) = \theta$ and no covariates. In this case, it can be shown, analogously to the result at the end of Section 4.2.1 (see proof of Theorem 1 for full details), that the population mean β_0 is estimated by a weighted average of the Y_{ij}'s in which the total weight given to cluster i is the average, over each of the Δ_i $(m \times m)$ submatrices of $\hat{\boldsymbol{R}}_i$, of the sum of the elements of its inverse matrix. Therefore, if Condition 2 is not satisfied, clusters of different sizes may, on average, be given different total weights. Likewise, if Condition 3 is not satisfied, two clusters of the same size but with different expectations for Y will receive different total weights.

If Conditions 1–3 are not satisfied, MWCR may not give consistent estimation of $\boldsymbol{\beta}$. For this reason, one is restricted in the choice of possible working correlation, a restriction that limits the potential for improving efficiency by using MWCR rather than WIEE. Note that Conditions 1 and 3 would be required for unbiased estimation when using a non-independence matrix even if cluster size were non-informative.

4.3 Weighted GEE for informative cluster size

In this section we develop an alternative efficient method (WRGEE) for certain scenarios where cluster size is informative. WRGEE allows a realistic working correlation to be used and then employs weighting to provide unbiased inference for either population M or C1. The motivation for incorporating the correlation is to increase efficiency. Such an approach is mentioned briefly by Williamson et al. (2007). They assert that there will be little gain in efficiency because the additional scaling weights required to weight each cluster appropriately for the selected population 'cancel out' the effects of the working correlation matrix. We demonstrate later through simulation studies that this is not always true. We present the method in general terms and then demonstrate that it is unbiased in two particular scenarios.

4.3.1 The method

A population p is selected ($p =$ M or C1) and a marginal regression model proposed,

$$\mu^p(\boldsymbol{X}) = h^{-1}(\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}_1), \tag{4.3}$$

where $\mu^M(\boldsymbol{X}) = \mu(\boldsymbol{X})$. The method we propose has three steps:

1. Estimate $\boldsymbol{\beta}$ in the marginal model specified in equation (4.3) by solving the IEE (when $p =$M) or the WIEE (when $p =$C1, see equations 3.16).

2. Choose a working correlation structure. Use the residuals from the model in Step 1 to estimate unknown parameters $\boldsymbol{\rho}$ (if any). If $p =$C1, the estimate, $\hat{\boldsymbol{\rho}}$, of $\boldsymbol{\rho}$ is estimated by applying the method of Williamson et al. (2003); if $p =$M, the method of Prentice (1988) can be used.

3. Solve the weighted GEE (WRGEE)

$$\sum_{i=1}^{K} \boldsymbol{U}_i^W(\boldsymbol{\beta}; \boldsymbol{Y}_i^*, \boldsymbol{X}_i^*) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^{pT}}{\partial \boldsymbol{\beta}} \hat{\boldsymbol{V}}_i^{-1} \frac{s_i^p}{\hat{r}_{i++}} \{\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i^p\} = \boldsymbol{0}, \tag{4.4}$$

   where $s_i^{C1} = 1$ and $s_i^M = N_i$. The value of $\hat{\boldsymbol{\rho}}$ used in equation (4.4) is that obtained at Step 2.

   A sandwich estimator of the variance of $\widehat{\boldsymbol{\beta}}$ is

$$\left(\sum_{i=1}^{K} \frac{s_i^p}{\hat{r}_{i++}} \frac{\partial \boldsymbol{\mu}_i^{pT}}{\partial \boldsymbol{\beta}} \hat{\boldsymbol{V}}_i^{-1} \frac{\partial \boldsymbol{\mu}_i^p}{\partial \boldsymbol{\beta}^T}\right)^{-1} \left(\sum_{i=1}^{K} \frac{(s_i^p)^2}{\hat{r}_{i++}^2} \frac{\partial \boldsymbol{\mu}_i^{pT}}{\partial \boldsymbol{\beta}} \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i^p)(\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i^p)^T \hat{\boldsymbol{V}}_i^{-1} \frac{\partial \boldsymbol{\mu}_i^p}{\partial \boldsymbol{\beta}^T}\right)$$
$$\times \left(\sum_{i=1}^{K} \frac{s_i^p}{\hat{r}_{i++}} \frac{\partial \boldsymbol{\mu}_i^{pT}}{\partial \boldsymbol{\beta}} \hat{\boldsymbol{V}}_i^{-1} \frac{\partial \boldsymbol{\mu}_i^p}{\partial \boldsymbol{\beta}^T}\right)^{-1}.$$

Note that equations (4.4) reduce to equations (3.6) when $p =$C1 and the independence working correlation is used. So, WIEE are a special case of WRGEE.

4.3.2 Unbiased estimation under certain scenarios

Whilst the estimator from the WRGEE method is in general biased, we now show that it is consistent under certain conditions in the case of cluster-constant covariates and in another special case considered by other authors (Chiang and Lee, 2008). In this section we omit the subscript i denoting the cluster.

4.3.2.1 Scenario 1: cluster-constant covariates

Theorem 2.

When $\boldsymbol{X}$ is cluster-constant, equation (4.3) is true and the conditions

1. $\boldsymbol{Y}^* \perp\!\!\!\perp \boldsymbol{R} \mid N, \boldsymbol{X}$ and
2. either (a) $E\{Y_j \mid \boldsymbol{X}, N\} = E\{Y_1 \mid \boldsymbol{X}_1, N\} \ \forall\ j$ or (b) $E_{\boldsymbol{R}}(r_{+j}/r_{++} \mid \boldsymbol{X}_1, N) = N^{-1} \ \forall\ j$.

are satisfied, the solution to WRGEE estimating equations (4.4) is a consistent estimator of $\boldsymbol{\beta}$ for the selected population (p=M or C1).

We now discuss Conditions 1–2 of Theorem 2. Condition 1 is similar to Condition 3 of Theorem 1, but the working correlation and outcome refer to the whole cluster, rather than to a selected subcluster within that cluster. Condition 1 might not be satisfied, for example, if an auto-regressive correlation structure is selected (just as such a structure may violate Condition 3 of Theorem 1 — see Section 4.2.3). As pointed out in Section 4.2.3, Condition 1 is an implicit assumption of the standard GEE. Condition 2(a) is the requirement that the 'position' of a member within a cluster does not affect its expected outcome. It could be violated, for example, in a dental study with plaque as outcome and covariates that take the same value for all teeth in the same mouth, as it may be more likely for molars than incisors to have plaque. Condition 2(b) is satisfied when the selected working correlation is exchangeable. If Condition 2(a) is satisfied, the choice of the working correlation is unrestricted; otherwise, exchangeable is the only safe choice to avoid bias. Note, however, that using an exchangeable working correlation makes WRGEE reduce to WIEE, and so there will be no gain in efficiency.

4.3.2.2 Scenario 2: cluster-size balanced covariates, linear regression, and no interaction between cluster size and covariates

In the simulation scenarios considered by Chiang and Lee (2008) it was assumed that

$$E(\boldsymbol{Y}^* \mid \boldsymbol{X}^*, N) = \theta_0 \mathbf{1} + \boldsymbol{X}^* \boldsymbol{\theta}_1 + \gamma_N \mathbf{1}, \tag{4.5}$$

where θ_0, $\boldsymbol{\theta}_1$ and γ_N are unknown parameters and $\boldsymbol{X}$ is cluster-size balanced (recall that $\mathbf{1}$ denotes an $N \times 1$ vector of units). Note that equation (4.5) implies that the

expected outcome depends on $\boldsymbol{X}$ and N but not on interactions between the two. If equation (4.5) is true, the marginal model $\mu^p(\boldsymbol{X}) = \beta_0^p + \boldsymbol{X}^T\boldsymbol{\beta}_1^p$ is correctly specified for both populations (p =C1 and M), and because the relation between Y and $\boldsymbol{X}$ is the same irrespective of the cluster size N and also because $\boldsymbol{X}$ is cluster-size balanced, $\boldsymbol{\beta}_1^{C1} = \boldsymbol{\beta}_1^M = \boldsymbol{\theta}_1$. For this reason $\boldsymbol{\beta}_1^p$ may be estimated by fitting a standard GEE with a separate intercept term for each cluster size, but we assume here that interest lies in both β_0^p and $\boldsymbol{\beta}_1^p$.

Theorem 3.

When $\boldsymbol{X}$ is cluster-size balanced and $E(\boldsymbol{Y}^* \mid \boldsymbol{X}^*, N) = \theta_0\mathbf{1} + \boldsymbol{X}^*\boldsymbol{\theta}_1 + \gamma_N\mathbf{1}$, the solution to the WRGEE estimating equations (4.4) with identity link function is a consistent estimator of $\boldsymbol{\beta}^p$ for the selected population (p =M or C1) if the following conditions are satisfied:

1. $\boldsymbol{Y}^* \perp\!\!\!\perp \boldsymbol{R} \mid \boldsymbol{X}^*, N$.

2. Either (a) $E(\boldsymbol{X}_j \mid N) = E(\boldsymbol{X}_1 \mid N)\ \forall\ j$ or (b) $E_{\boldsymbol{R}}(r_{+j}/r_{++} \mid N) = N^{-1}\ \forall\ j$.

3. $\boldsymbol{X}^* \perp\!\!\!\perp \boldsymbol{R} \mid N$.

Conditions 1–2 are similar to Conditions 1–2 of Theorem 2. Condition 1 is also analogous to Condition 3 of Theorem 1 and relates to the choice of the working correlation structure, as explained at the end of Section 4.2.3. Condition 2(a) is similar to Condition 2(a) of Theorem 2, with the difference that the expectation in Condition 2(a) of Theorem 3 refers to the covariates rather than the outcome, and is interpreted in a similar manner. Condition 3 is the requirement that the working correlation be conditionally independent of the covariate values of all the members in the cluster given the size of the cluster. It would not be satisfied, for example, if an auto-regressive structure were used, time were one of the covariates in the analysis model, and the times of the members differed between clusters of the same size. However, Condition 3 might be satisfied if time were not one of the covariates in the analysis model.

4.3.3 Adaptation of WRGEE for non-size balanced categorical covariates and informative covariate structure

For non-size-balanced cluster-varying categorical covariates, we propose an adaptation of the WRGEE based on weighting separately for each value of $\boldsymbol{X}$ in the cluster. Steps 1 and 2 of Section 4.3.1 remain the same. However, before implementing Step 3, the working correlation estimated at Step 2 is modified in the following way. For each i and each $j \neq k$, the (j,k)th element of working correlation matrix $\hat{\boldsymbol{R}}_i$ is set to zero whenever $\boldsymbol{X}_{ij} \neq \boldsymbol{X}_{ik}$. Thus, members of cluster i with different values of $\boldsymbol{X}$ are assumed to be independent; members with the same $\boldsymbol{X}$ are assumed to have the correlation estimated at Step 2. Equation (4.4) is also modified, replacing $\hat{r}_{i++}^{-1}$ by $\text{diag}(w(\boldsymbol{X}_{i1}), \ldots, w(\boldsymbol{X}_{iN}))$, where $w(\boldsymbol{X})$ is defined as follows. Let $L_{\boldsymbol{x}}$ denote the number of members in the cluster with $\boldsymbol{X} = \boldsymbol{x}$ and let $\hat{r}_{++}(\boldsymbol{x})$ denote the sum of the elements of the submatrix of $\hat{\boldsymbol{R}}^{-1}$ composed of the rows and columns that correspond to members with $\boldsymbol{X} = \boldsymbol{x}$. Then $w(\boldsymbol{x}) = L_{\boldsymbol{x}}\{N\hat{r}_{++}(\boldsymbol{x})\}^{-1}$. Similarly, efficient estimation for population C2 can be obtained by substituting $s_i^{C2} = 1$ and $w(\boldsymbol{x}) = \{\hat{r}_{++}(\boldsymbol{x})\}^{-1}$ in equation (4.4). This adaption of WRGEE we call WBGEE. As this method, apart from the variance estimation, corresponds to splitting clusters into subclusters based on $\boldsymbol{X}$ and then applying the WRGEE to these subclusters as if they were independent clusters with cluster-constant covariates, it is unbiased by the argument in Section 4.3.2.1. Note that when the working correlation structure is exchangeable, WBGEE reduce to WIEE.

Finally, WBGEE may be adapted for inference when using the DWGEE2 of Huang and Leroux (2011). We term this adaptation DWBGEE2. Since DWGEE2 also refers to categorical covariates, it is straightforward to apply similar ideas as in WBGEE to increase efficiency by using a non-diagonal correlation matrix. The weights should be selected to reflect the expected rather than the observed number of members with $\boldsymbol{X} = \boldsymbol{x}$ (see Section 3.11.1). As with WBGEE, DWBGEE2 method is expected to provide moderate efficiency gains compared to DWGEE2 since the covariates within each stratum defined by the levels of exposure are cluster-invariant.

Some of the R-functions used to fit the models considered in Sections 4.2.24.3 are included in Appendix B.3.

4.3.4 Practical application

If there are several covariates in the model, including cluster-constant, cluster-varying size balanced and cluster-varying non-size balanced, then the block-diagonal method outlined in the previous subsection should be followed, basing the blocks on the cluster varying non-size-balanced covariate alone. This will limit the potential efficiency gain from specifying a realistic working correlation relative to WIEE. When the non-size-balanced covariate is continuous cluster-varying and there are no common values of the particular covariate within a cluster then the correlation matrix will be diagonal and our method reduces to WIEE. If inference for population C2 or C3 is required, we propose first categorising the non-size-balanced covariate. Subsequently we can use either appropriately weighted WIEE for populations C2 or C3, or WBGEE for populations C2 or C3, where the blocks are constructed based on the categorised covariate.

4.4 Simulation study and comparison of methods

We aim to assess the performance of MWCR and WRGEE in terms of bias and efficiency for four sets of scenarios where cluster size and/or covariate structure are informative. We simulated clustered normal responses Y and a binary cluster-varying scalar covariate X. We induced informative cluster size through an underlying 'susceptibility' that did not vary within the cluster. Each simulated dataset contained 100 clusters. Data were generated independently for each cluster as follows. For cluster i:

1. Generate $B_i \sim N(0, 0.5^2)$ to be the underlying susceptibility.

2. Generate $N_i \sim Poisson\{\exp(\alpha_0+\alpha_1 B_i)\}+m$, where m is the minimum cluster size.

3. Generate $X_{ij} \sim Bernoulli\{\lambda_0 + \lambda_1 \text{logit}^{-1}(B_i)\}$ independently for $j = 1, \ldots, N_i$. Note that if $\lambda_1 = 0$ then X is size balanced, while if $0 < \lambda_1 \leq 1$ it is non-size balanced.

4. Calculate the linear predictor $\eta_{ij} = \gamma_0 + \gamma_1 X_{ij} + \gamma_2 B_i + \gamma_3 B_i X_{ij}$, and denote $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iN_i})^T$. Parameter γ_2 determines the association between the underlying susceptibility (and consequently cluster size) and the outcome, while γ_3 determines how this association changes with X.

5. Finally, generate $\boldsymbol{Y}_i^* \sim MVN(\boldsymbol{\eta}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is an exchangeable or AR-1 correlation matrix with parameter (pairwise correlation or autocorrelation) ρ.

We selected $\gamma_0 = \gamma_1 = \gamma_2 = 1$ and either $\gamma_3 = 0$ or $\gamma_3 = 1$. We selected $\rho = 0.2$, 0.5 or 0.8, which correspond to small, medium or high correlation/autocorrelation. For each scenario we generated 2000 simulated datasets. When $\alpha_1 \neq 0$ and either $\gamma_2 \neq 0$ or $\gamma_3 \neq 0$, the cluster size is informative. For size balanced X we selected $\lambda_0 = 0.4$ and $\lambda_1 = 0$ and for non-size balanced $\lambda_0 = 0$ and $\lambda_1 = 1$. For each population p (p =M, C1, C2 or C3) the correctly specified analysis model is of the form $E^p(Y_{ij}) = \beta_0^p + \beta_1^p X_{ij}$. The true values of β_0^p and β_1^p were calculated using numerical integration.

We applied the WIEE, WRGEE, and MWCR methods. For scenarios where WRGEE and/or MWCR were unbiased we calculated their efficiency relative to WIEE. For four sets of scenarios we present the mean estimated values of the parameters over the 2000 simulated datasets and their empirical standard errors (ese), i.e. the square root of the variance of the 2000 estimates. We also present coverage probabilities. For the WRGEE and MWCR methods the working correlation selected was the same as that used to generate the data at step 5 above, though we note this is not generally the correct correlation, because at step 5 the term $\boldsymbol{\eta}_i$ gives $E(\boldsymbol{Y}_i^*|B_i, \boldsymbol{X}_i^*)$ but our regression models condition only on X.

We considered four sets of scenarios. The first set was designed to illustrate bias and relative efficiency of the WRGEE and MWCR methods, primarily focusing on inference for population C1, as this is the only possibility for MWCR. We also considered inference for population M using WRGEE. The second set was chosen to illustrate the WBGEE method, as presented in Section 4.3.3. To demonstrate the possible efficiency gains relative to WIEE from WBGEE, we selected a larger minimum cluster size and AR-1 correlation (there is no efficiency gain under exchangeable correlation) in this second set of scenarios. The third set corresponds to a scenario of informative covariate structure, constant cluster size, and AR-1 (auto-regressive) correlation. We aim to show that standard GEE with AR-1 working correlation are biased (for population M) and WBGEE maybe used to increase efficiency (for populations M, C2 or C3) compared to WIEE. Finally, we considered a set of scenarios where we demonstrated the possible efficiency gains of WRGEE in scenarios with cluster-constant covariates.

Set 1: The correlation structure at step 5 above was exchangeable and the minimum cluster size was $m = 2$. The parameters for the cluster size model were selected to be $\alpha_0 = \alpha_1 = 1$ and these resulted in a mean cluster size of approximately 4. As shown in Table 4.1, MWCR and WRGEE methods were biased when X was non-size balanced. Both MWCR and WRGEE led to unbiased inference with increased efficiency relative to WIEE when X was size-balanced and $\gamma_3 = 0$, which is similar to Scenario 2 in Section 4.3.2.2. In this special case, WRGEE gave greater efficiency gains than MWCR. For the case of size-balanced X and $\gamma_3 = 1$ the only unbiased method with increased efficiency relative to WIEE, was MWCR. Analogous simulation results for population M are presented in Table 4.2. Results are shown from the application of IEE, WRGEE and the standard GEE with exchangeable working correlation matrix (GEE(EX)). When X was size-balanced and $\gamma_3 = 0$, only WRGEE provided unbiased inference with increased efficiency. We note that in this particular scenario GEE(EX) provided unbiased estimates for the slope term. For all other cases WRGEE and GEE(EX) resulted in biased parameter estimates.

Set 2: X was non-size balanced cluster-varying, the parameters for the cluster size model were selected to be $\alpha_0 = \alpha_1 = 1.5$, the correlation structure at step 5 was AR-1, and the minimum cluster size was $m = 5$. As shown in Table 4.3, WBGEE led to unbiased inference with efficiency gains of up to 14% for all populations for inference, relative to WIEE. As it was seen in Set 1, both WRGEE and MWCR are biased when X is non-size balanced and, therefore, results from applying these methods are not presented in Table 4.3.

Set 3: The cluster size was constant $n = 10$, X was cluster-varying and $\lambda_0 = 0$, $\lambda_1 = 1$. So, the distribution of X was associated with the distribution of the outcome Y. The correlation structure at step 5 was AR-1. This is a case of pure informative covariate structure and can be viewed as a violation of the assumptions imposed by Pepe and Anderson (1994), regarding the use of non-diagonal correlation matrices. The populations for inference could be population M, C2 or C3 (inference for population C1 coincides with that for population M since the cluster size is constant). As shown in Table 4.4 the standard GEE with an AR-1 correlation structure (GEE(AR)) were biased for population M, as expected, while WBGEE provided unbiased inference with moderate efficiency gains, for populations M and C2.

Set 4: X was cluster-constant and non-size balanced and the correlation structure, AR-1. The minimum cluster size was $m = 5$ and the parameters for the cluster size model, $\alpha_0 = \alpha_1 = 1$. As shown in Table 4.5, the WRGEE led to unbiased inference with small efficiency gains for populations M and C1.

In summary, we have demonstrated that WRGEE and WBGEE can be used in certain scenarios to give unbiased inference with moderate efficiency gains compared to WIEE and that the variance estimator leads to roughly 95% coverage. We would expect slightly greater efficiency gains, had the working correlations been correctly specified. The variance estimator we have presented for the MWCR method is also seen to have good coverage when that method is unbiased.

Similar simulation studies were carried out also for binary responses with a cluster-constant covariate (as in Set 4) or a cluster-varying categorical non-size balanced covariate (as in Set 2). The results are generally consistent with those (reported above) for the corresponding sets of scenarios for Normal responses. Details for the simulation procedure and tables of results for Set 2 can be found in Appendix B.2.

No Interaction: $\gamma_3 = 0$

		X non-size balanced TRUE $(\beta_0^{C1}, \beta_1^{C1}) = (0.88, 1.23)$			X size balanced TRUE $(\beta_0^{C1}, \beta_1^{C1})$=(1.00,1.00)			
ρ	Method	$\hat{\boldsymbol{\beta}}_0$(ese($\hat{\boldsymbol{\beta}}_0$))	$\hat{\boldsymbol{\beta}}_1$(ese($\hat{\boldsymbol{\beta}}_1$))	CV($\hat{\boldsymbol{\beta}}_0$,$\hat{\boldsymbol{\beta}}_1$)	$\hat{\boldsymbol{\beta}}_0$(ese($\hat{\boldsymbol{\beta}}_0$))	$\hat{\boldsymbol{\beta}}_1$(ese($\hat{\boldsymbol{\beta}}_1$))	CV($\hat{\boldsymbol{\beta}}_0$,$\hat{\boldsymbol{\beta}}_1$)	RE($\hat{\boldsymbol{\beta}}_0$,$\hat{\boldsymbol{\beta}}_1$)
	WIEE	0.88(0.099)	1.23(0.115)	(0.95,0.95)	1.00(0.093)	1.00(0.112)	(0.95,0.94)	–
0.2	WRGEE	0.96(0.092)	1.09(0.089)	–	1.00(0.088)	1.00(0.087)	(0.94,0.95)	(1.12,1.68)
	MWCR	0.92(0.096)	1.15(0.104)	–	1.00(0.090)	1.00(0.101)	(0.95,0.95)	(1.05,1.22)
	WIEE	0.88(0.111)	1.23(0.118)	(0.95,0.95)	1.00(0.105)	1.00(0.113)	(0.96,0.95)	–
0.5	WRGEE	0.98(0.099)	1.04(0.070)	–	1.00(0.098)	1.00(0.069)	(0.95,0.95)	(1.15,2.66)
	MWCR	0.95(0.103)	1.08(0.087)	–	1.00(0.100)	1.00(0.084)	(0.95,0.95)	(1.10,1.80)
	WIEE	0.88(0.122)	1.23(0.122)	(0.95,0.94)	1.00(0.115)	1.00(0.113)	(0.95,0.95)	–
0.8	WRGEE	0.99(0.107)	1.01(0.045)	–	1.00(0.106)	1.00(0.044)	(0.95,0.95)	(1.16,6.88)
	MWCR	0.97(0.110)	1.04(0.058)	–	1.00(0.107)	1.00(0.054)	(0.95,0.95)	(1.15,4.30)

With Interaction: $\gamma_3 = 1$

		X non-size balanced TRUE $(\beta_0^{C1}, \beta_1^{C1}) = (0.88, 1.35)$			X size-balanced TRUE $(\beta_0^{C1}, \beta_1^{C1}) = (1.00, 1.00)$			
ρ	Method	$\hat{\boldsymbol{\beta}}_0$(ese($\hat{\boldsymbol{\beta}}_0$))	$\hat{\boldsymbol{\beta}}_1$(ese($\hat{\boldsymbol{\beta}}_1$))	CV($\hat{\boldsymbol{\beta}}_0$,$\hat{\boldsymbol{\beta}}_1$)	$\hat{\boldsymbol{\beta}}_0$(ese($\hat{\boldsymbol{\beta}}_0$))	$\hat{\boldsymbol{\beta}}_1$(ese($\hat{\boldsymbol{\beta}}_1$))	CV($\hat{\boldsymbol{\beta}}_0$,$\hat{\boldsymbol{\beta}}_1$)	RE($\hat{\boldsymbol{\beta}}_0$,$\hat{\boldsymbol{\beta}}_1$)
	WIEE	0.88(0.099)	1.35(0.139)	(0.95,0.96)	1.00(0.093)	1.01(0.138)	(0.95,0.94)	–
0.2	WRGEE	0.95(0.092)	1.21(0.105)	–	0.94(0.089)	1.14(0.109)	–	–
	MWCR	1.95(0.096)	1.21(0.121)	–	1.00(0.089)	1.00(0.120)	(0.95,0.95)	(1.06,1.32)
	WIEE	0.88(0.111)	1.35(0.142)	(0.95,0.95)	1.00(0.105)	1.01(0.138)	(0.94,0.95)	–
0.5	WRGEE	0.96(0.101)	1.19(0.092)	–	0.93(0.099)	1.16(0.095)	–	–
	MWCR	0.98(0.104)	1.14(0.105)	–	1.00(0.099)	1.00(0.103)	(0.95,0.95)	(1.10,1.77)
	WIEE	0.88(0.122)	1.35(0.144)	(0.95,0.95)	1.00(0.113)	1.01(0.140)	(0.95,0.94)	–
0.8	WRGEE	0.97(0.108)	1.18(0.074)	–	0.93(0.108)	1.18(0.079)	–	–
	MWCR	1.01(0.110)	1.08(0.083)	–	1.00(0.108)	1.00(0.080)	(0.95,0.94)	(1.14,2.95)

Table 4.1: Set 1(a). Application of WIEE, MWCR and WRGEE for population C1. The cluster size is informative and the working correlation is exchangeable.

[a] Relative efficiency compared to WIEE

No Interaction: $\gamma_3 = 0$

		X non-size balanced TRUE $(\beta_0^M, \beta_1^M) = (1.02, 1.24)$			X size balanced TRUE $(\beta_0^M, \beta_1^M) = (1.15, 1.00)$			
ρ	Method	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CV($\hat{\beta}_0$,$\hat{\beta}_1$)	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CV($\hat{\beta}_0$,$\hat{\beta}_1$)	RE($\hat{\beta}_0$,$\hat{\beta}_1$)
	IEE	1.02(0.097)	1.24(0.109)	(0.95,0.95)	1.15(0.096)	1.00(0.105)	(0.96,0.94)	–
0.2	WRGEE	0.96(0.092)	1.07(0.092)	–	1.15(0.094)	1.00(0.092)	(0.95,0.94)	(1.03,1.30)
	GEE(EX)	1.00(0.091)	1.06(0.092)	–	1.02(0.085)	1.00(0.085)	(–,0.95)	(– ,1.49)
	IEE	1.02(0.112)	1.24(0.118)	(0.95,0.95)	1.15(0.111)	1.00(0.106)	(0.96,0.95)	–
0.5	WRGEE	0.98(0.099)	1.03(0.073)	–	1.15(0.108)	1.00(0.075)	(0.95,0.94)	(1.06,1.99)
	GEE(EX)	1.00(0.099)	1.02(0.073)	–	1.01(0.096)	1.00(0.068)	(– ,0.95)	(– ,2.37)
	IEE	1.02(0.121)	1.24(0.121)	(0.95,0.94)	1.15(0.124)	1.00(0.106)	(0.96,0.94)	–
0.8	WRGEE	0.99(0.107)	1.01(0.047)	–	1.15(0.114)	1.00(0.048)	(0.95,0.94)	(1.08,4.86)
	GEE(EX)	1.00(0.106)	1.00(0.047)	–	1.00(0.105)	1.00(0.044)	(– ,0.95)	(– ,5.85)

With Interaction: $\gamma_3 = 1$

		X non-size balanced TRUE $(\beta_0^M, \beta_1^M) = (1.02, 1.51)$			X size-balanced TRUE $(\beta_0^M, \beta_1^M) = (1.15, 1.16)$			
ρ	Method	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CV($\hat{\beta}_0$,$\hat{\beta}_1$)	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CV($\hat{\beta}_0$,$\hat{\beta}_1$)	RE($\hat{\beta}_0$,$\hat{\beta}_1$)
	IEE	1.02(0.098)	1.51(0.145)	(0.94,0.95)	1.15(0.096)	1.16(0.135)	(0.96,0.95)	–
0.2	WRGEE	1.11(0.101)	1.32(0.123)	–	1.09(0.090)	1.31(0.137)	–	–
	GEE(EX)	0.98(0.918)	1.19(0.108)	–	0.96(0.089)	1.18(0.110)	–	–
	IEE	1.02(0.111)	1.51(0.149)	(0.95,0.95)	1.15(0.111)	1.16(0.136)	(0.96,0.95)	–
0.5	WRGEE	1.12(0.113)	1.31(0.111)	–	1.08(0.104)	1.32(0.125)	–	–
	GEE(EX)	0.98(0.100)	1.18(0.093)	–	0.94(0.098)	1.18(0.096)	–	–
	IEE	1.02(0.124)	1.51(0.152)	(0.95,0.95)	1.15(0.124)	1.16(0.136)	(0.96,0.94)	–
0.8	WRGEE	1.13(0.122)	1.30(0.095)	–	1.08(0.116)	1.33(0.110)	–	–
	GEE(EX)	0.97(0.107)	1.16(0.074)	–	0.93(0.107)	1.19(0.079)	–	–

Table 4.2: Set 1(b). Application of IEE, WRGEE and GEE(EX) for the population M. The cluster size is informative and the assumed correlation structure is exchangeable.

[a] Relative efficiency compared to WIEE

		Population C1 TRUE $(\beta_0^{C1}, \beta_1^{C1}) = (0.88, 1.24)$				Population M TRUE $(\beta_0^{M}, \beta_1^{M}) = (1.06, 1.26)$			
ρ	Method	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CV($\hat{\beta}_0$,$\hat{\beta}_1$)	RE($\hat{\beta}_0$,$\hat{\beta}_1$)	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CV($\hat{\beta}_0$,$\hat{\beta}_1$)	RE($\hat{\beta}_0$,$\hat{\beta}_1$)
0.2	WIEE	0.88(0.074)	1.24(0.080)	(0.95,0.94)	-	1.06(0.076)	1.26(0.084)	(0.94,0.95)	-
	WBGEE	0.88(0.074)	1.24(0.079)	(0.95,0.95)	(1.00,1.02)	1.06(0.076)	1.26(0.083)	(0.95,0.94)	(1.00,1.02)
0.5	WIEE	0.88(0.083)	1.24(0.082)	(0.95,0.95)	-	1.06(0.083)	1.26(0.085)	(0.95,0.95)	-
	WBGEE	0.88(0.082)	1.24(0.079)	(0.95,0.95)	(1.02,1.08)	1.06(0.082)	1.26(0.082)	(0.95,0.94)	(1.01,1.08)
0.8	WIEE	0.89(0.099)	1.24(0.085)	(0.95,0.94)	-	1.06(0.098)	1.26(0.089)	(0.94,0.95)	-
	WBGEE	0.89(0.097)	1.24(0.081)	(0.95,0.94)	(1.03,1.11)	1.06(0.096)	1.26(0.085)	(0.95,0.95)	(1.03,1.09)
		Population C2 TRUE $(\beta_0^{C2}, \beta_1^{C2}) = (1.00, 1.02)$				Population C3 TRUE $(\beta_0^{C3}, \beta_1^{C3}) = (1.02, 1.00)$			
ρ	Method	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CV($\hat{\beta}_0$,$\hat{\beta}_1$)	RE($\hat{\beta}_0$,$\hat{\beta}_1$)	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CV($\hat{\beta}_0$,$\hat{\beta}_1$)	RE($\hat{\beta}_0$,$\hat{\beta}_1$)
0.2	WIEE	1.00(0.071)	1.02(0.076)	(0.956,0.926)	-	1.02(0.071)	1.00(0.075)	(0.959,0.933)	-
	WBGEE	1.00(0.071)	1.02(0.074)	(0.954,0.933)	(1.00,1.03)	1.02(0.071)	1.00(0.074)	(0.953,0.932)	(1.00,1.03)
0.5	WGEE	1.00(0.076)	1.02(0.070)	(0.960,0.923)	-	1.02(0.077)	1.00(0.070)	(0.962,0.928)	-
	WBGEE	1.00(0.075)	1.02(0.067)	(0.959,0.924)	(1.01,1.11)	1.02(0.076)	1.00(0.066)	(0.959,0.929)	(1.01,1.11)
0.8	WGEE	1.00(0.088)	1.02(0.056)	(0.965,0.927)	-	1.02(0.089)	1.00(0.054)	(0.962,0.926)	-
	WBGEE	1.00(0.087)	1.02(0.053)	(0.964,0.933)	(1.03,1.14)	1.02(0.087)	1.00(0.050)	(0.958,0.929)	(1.03,1.14)

Table 4.3: Set 2. Application of WIEE and WBGEE for populations M, C1, C2 and C3. The cluster size is informative, X is non-size balanced and the assumed correlation is AR-1.

Population M

TRUE $(\beta_0^M, \beta_1^M) = (0.883, 1.234)$

ρ	Method	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CV($\hat{\beta}_0$,$\hat{\beta}_1$)	RE($\hat{\beta}_0$,$\hat{\beta}_1$)
0.2	IEE	0.88(0.074)	1.23(0.080)	(0.947,0.954)	-
	WBGEE	0.88(0.073)	1.23(0.079)	(0.935,0.949)	(1.02,1.03)
	GEE(AR)	0.97(0.072)	1.06(0.071)	-	-
0.5	IEE	0.88(0.081)	1.23(0.082)	(0.942,0.949)	-
	WBGEE	0.88(0.079)	1.23(0.078)	(0.939,0.947)	(1.05,1.10)
	GEE(AR)	0.99(0.075)	1.04(0.054)	-	-
0.8	IEE	0.88(0.096)	1.23(0.085)	(0.943,0.947)	-
	WBGEE	0.88(0.092)	1.23(0.080)	(0.945,0.942)	(1.08,1.12)
	GEE(AR)	1.00(0.086)	1.01(0.033)	-	-

Population C2

TRUE $(\beta_0^{C2}, \beta_1^{C2}) = (1.00, 1.01)$

ρ	Method	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CV($\hat{\beta}_0$,$\hat{\beta}_1$)	RE($\hat{\beta}_0$,$\hat{\beta}_1$)
0.2	WIEE	1.00(0.074)	1.01(0.072)	(0.943,0.950)	-
	WBGEE	1.00(0.074)	1.01(0.071)	(0.941,0.946)	(1.00,1.03)
0.5	WIEE	1.00(0.080)	1.01(0.067)	(0.952,0.939)	-
	WBGEE	1.00(0.079)	1.01(0.064)	(0.944,0.943)	(1.03,1.11)
0.8	WIEE	1.00(0.093)	1.01(0.053)	(0.950,0.941)	-
	WBGEE	1.00(0.090)	1.01(0.050)	(0.947,0.944)	(1.06,1.11)

Table 4.4: Set 3. Application of WBGEE for populations M and C2. The cluster size is constant, the covariate structure is informative and the assumed correlation structure is AR-1.

Population M					
TRUE $(\beta_0^M, \beta_1^M) = (1.057, 1.262)$					
ρ	Method	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CV($\hat{\beta}_0$,$\hat{\beta}_1$)	RE($\hat{\beta}_0$,$\hat{\beta}_1$)
0.2	WIEE	1.06(0.099)	1.26(0.136)	(0.945,0.946)	-
	WRGEE	1.06(0.099)	1.26(0.135)	(0.946,0.947)	(1.00,1.01)
0.5	WIEE	1.06(0.112)	1.26(0.153)	(0.941,0.939)	-
	WRGEE	1.06(0.110)	1.26(0.150)	(0.936,0.944)	(1.02,1.03)
0.8	WIEE	1.06(0.136)	1.26(0.187)	(0.943,0.939)	-
	WRGEE	1.06(0.132)	1.26(0.180)	(0.942,0.946)	(1.06,1.08)
Population C1					
TRUE $(\beta_0^{C1}, \beta_1^{C1})$ =(0.883,1.235)					
ρ	Method	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CV($\hat{\beta}_0$,$\hat{\beta}_1$)	RE($\hat{\beta}_0$,$\hat{\beta}_1$)
0.2	WIEE	0.88(0.093)	1.24(0.125)	(0.937,0.946)	-
	WRGEE	0.88(0.093)	1.24(0.124)	(0.937,0.945)	(1.00,1.01)
0.5	WIEE	0.88(0.106)	1.24(0.144)	(0.942,0.943)	-
	WRGEE	0.88(0.104)	1.24(0.142)	(0.941,0.946)	(1.01,1.03)
0.2	WIEE	0.88(0.130)	1.24(0.178)	(0.945,0.951)	-
	WRGEE	0.88(0.127)	1.24(0.172)	(0.941,0.948)	(1.04,1.07)

Table 4.5: Set 4. Application of WRGEE for populations M, and C1. The cluster size is informative, X is cluster-constant non-size-balanced and the assumed correlation structure is AR-1.

4.5 Illustration

The Delta trial dataset is used for the illustration and comparison of methods presented in this chapter. In Chapter 3 we modelled whether or not an ARC event was Oral candidiasis, in terms of randomisation arm, CD4 count and the time of the event since entry in the study. In this illustration we investigate how the immune status of a patient (of which CD4 count is an indicator), at times of ARC events, changes over time and whether it differs between the treatment arms.

Since CD4 count had a skewed distribution we modelled its square root, Y. Events are clustered by patient and we let N denote the number of events experienced by a pa-

tient. Let subscript i denote patient and let j index the N_i events experienced by patient i. Let X_1 and X_2 be indicator variables of randomisation to the drug combinations AZT+ddC and AZT+ddI respectively (cluster-constant) and T be the time to the event from entry in the study in units of 60 days. As interactions between T and X_1 and X_2 were found non-significant the model we considered was

$$E(Y_{ij}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 T_{ij} + \beta_4 T_{ij}^2. \tag{4.6}$$

MWCR may only be applied to datasets where the minimum cluster size is 2 or more. Whilst a thorough examination of immune status at the time of events would clearly involve all events, for the purposes of comparison between methods, in our illustration we excluded all patients with one episode. After excluding these clusters of size one, 657 clusters remained; the maximum cluster size was 15 and the median 3. Among the 657 clusters 32% were of size 2 (cluster size group 1), 39% of size 3 or 4 (group 2) and 29% of size 5 to 16 (group 3). The mean square root CD4 count was 9.25, 8.01 and 6.2, for groups 1, 2 and 3 respectively. This is an initial indication that the cluster size might be informative; patients with more episodes tend to have lower CD4 count than patients with fewer episodes.

We fitted a regression model analogous to model (4.6) but for population M and including cluster size alongside the covariates of main interest:
$E(Y_{ij}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 T_{ij} + \beta_4 T_{ij}^2 + \beta_N N_i$. We used independence estimating equations to fit this model. The effect of cluster size was found significant
($\hat{\beta}_N = -0.47$, se$(\hat{\beta}_N) = 0.09$, $p < 0.001$), supporting the initial indication for informative cluster size. We tested for interactions between the cluster size and covariates and these were not found statistically significant.

Model (4.6) was fitted using WIEE and MWCR. WIEE method is known to provide consistent inference. So, although the true value of parameters is unknown, parameter estimates and standard errors from MWCR and WRGEE are compared to the corresponding ones from WIEE to assess the evidence of bias in MWCR and WRGEE and possible efficiency gains. MWCR and WRGEE were applied using either (a) exchangeable working correlation (MWCR(EX),WRGEE(EX)) or (b) an auto-regressive type working correlation with lag 1 we denote AR-1. This choice (b) is the auto-regressive correlation corresponding to treating consecutive events as occurring one

time unit apart, e.g. at times 1, 2, and 3 if cluster size is 3. We considered the application of MWCR with the more conventional auto-regressive correlation structure based on the actual times of episodes, but this was not possible because of computational problems. Specifically, because of the highly irregular times of the episodes, for many clusters the working correlation matrix was non-invertible.

The results from the application of the methods are presented in Table 4.6. Interestingly, CD4 count at ARC events is on average higher for a typical patient who receives the combination treatment AZT+ddI, compared to a typical patient receiving AZT alone. Also, as it would be expected, CD4 count at ARC events for a typical patient decreases over time.

In terms of the performance of MWCR(EX) and MWCR(AR-1), there is some evidence of bias in the estimation of the effects of T and T^2. In particular, for the effect of T^2, the difference between the estimates from MWCR and WIEE is approximately three times the standard error of the estimates when using MWCR(EX) or MWCR(AR-1). The differences between the estimates from WIEE and MWCR of the effects of X_1 and X_2 are negligible. For the intercept term, the difference is small when using MWCR(EX) and negligible for MWCR(AR-1). For the intercept term and the effect of T and T^2, the standard errors of the estimates are considerably smaller for MWCR(EX) and MWCR(AR-1) compared to WIEE.

In our illustration, Condition 2 and 3 of Theorem 1 are satisfied when using MWCR(EX). Condition 2 is not satisfied for MWCR(AR-1) because the correlations specified between members of subclusters will typically be smaller for subclusters from larger clusters than for subclusters from smaller clusters. Condition 3 may not be satisfied for MWCR(AR-1) if the gaps between ARC events are associated with the CD4 at events. Condition 1 is not met for either MWCR(EX) or MWCR(AR-1) because the covariates T and T^2 are not size balanced. The mean time in days from entry to the trial for events in cluster size groups 1, 2 and 3 (see earlier) was 495, 465 and 502 respectively, indicating some deviation from size balance. We view this as the main reason for the probable bias in the application of MWCR seen in our results, as Conditions 2 and 3 were satisfied for MWCR(EX) but the bias seems as large for MWCR(EX) as MWCR(AR-1).

Model: $E(Y_{ij}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 T_{ij} + \beta_4 T_{ij}^2$

Method	$\hat{\boldsymbol{\beta}}_0$(se($\hat{\boldsymbol{\beta}}_0$))	$\hat{\boldsymbol{\beta}}_1$(se($\hat{\boldsymbol{\beta}}_1$))	$\hat{\boldsymbol{\beta}}_2$(se($\hat{\boldsymbol{\beta}}_2$))	$\hat{\boldsymbol{\beta}}_3$(se($\hat{\boldsymbol{\beta}}_3$))	$\hat{\boldsymbol{\beta}}_4$(se($\hat{\boldsymbol{\beta}}_4$))
WIEE	10.96(0.472)	0.71(0.412)	0.97(0.421)	-0.73(0.095)	0.027(0.0053)
MWCR(EX)	11.16(0.353)	0.73(0.414)	0.98(0.426)	-0.63(0.065)	0.016(0.0039)
MWCR(AR-1)	10.95(0.369)	0.70(0.410)	0.96(0.423)	-0.63(0.071)	0.017(0.0042)
WRGEE(EX)	11.48(0.345)	0.72(0.416)	0.97(0.430)	-0.66(0.060)	0.015(0.0036)
WRGEE(AR-1)	11.27(0.348)	0.67(0.414)	0.96(0.422)	-0.65(0.076)	0.018(0.0041)

Table 4.6: Application of WIEE, MWCR and WRGEE using data from the Delta trial.

WRGEE(EX) and WRGEE(AR-1) perform similarly to MWCR(EX) and MWCR(AR-1), i.e. there is some indication of bias in estimating the effects of T and T^2, but negligible bias in the covariate effects of X_1 and X_2 and the intercept term. We now refer to the Conditions of Theorem 3. Arguably, the main reason for bias from the application of WRGEE(EX) and WRGEE(AR-1) is the deviation of size balance for T and T^2; Theorem 3 requires size-balanced covariates. Conditions 1, 2 & 3 of Theorem 3 are satisfied when the working correlation is exchangeable. For the application of WRGEE(AR-1), Conditions 1 and 3 are satisfied since the working correlation does not vary conditional on N but Condition 2(a) is violated by definition given the nature of covariates T and T^2.

From the scenarios considered here, it is evident that application of MWCR and WRGEE might be problematic, at least when dealing with longitudinal data. When the speculated regression model includes a mixture of cluster-constant and cluster-varying covariates these special methods may result in non-negligible bias for the covariate effects corresponding to cluster-varying covariates if these are non-size balanced. Exploratory analysis should be carried out to assess whether conditions are likely to be satisfied. When the conditions imposed for the consistency of each method are not likely to be met, WIEE should be used instead. As noted in Section 4.2.3, some of the conditions in Theorems 1, 2 and 3 are also requirements for the consistency of the standard GEE, even if the cluster size were non-informative. However, especially in longitudinal studies, it may be more likely for these conditions to be violated in datasets with informative cluster size than in more general scenarios where the cluster size is non-informative.

4.6 Discussion

In this chapter we have drawn attention to bias in the MWCR method in scenarios where the covariates are non-size balanced, a bias which was not mentioned by Chiang and Lee (2008). Importantly, even for size-balanced covariates we clarify that an exchangeable working correlation is the only safe choice when using MWCR, and we present a variance estimator which was not clearly described by Chiang and Lee (2008).

We have also proposed weighted GEE (WRGEE), which may be used in certain simple scenarios to increase efficiency. Compared to MWCR, WRGEE is simpler, does not require the minimum cluster size to be greater than one, and allows inference for either the population of all members or that of typical members 1. As with MWCR, some care is required in the choice of working correlation. Our simulation studies have shown that relative to WIEE, WRGEE give small efficiency gains when the covariates are cluster-constant but higher efficiency gains for cluster-varying covariates.

With the aim of increasing efficiency, both MWCR and WRGEE are worthwhile of consideration alongside methods which use independence working correlation. Nevertheless, analysts should be cautious when using either MWCR or WRGEE as their consistency relies upon conditions which relate to the selected working correlation and also the structure of the covariates. We discuss the practical application of MWCR and WRGEE in the next few paragraphs.

Both MWCR and WRGEE require the covariates to be size-balanced. If the cluster size is informative and the cluster varying-covariate $\boldsymbol{X}$ is non-size-balanced, Condition 1 in Theorem 1 is violated. For this reason it is important to consider in advance the likelihood of non-size balanced covariates given the study design and scientific setting. Where size-balance of covariates is assured, either MWCR or WRGEE may be a good choice of analysis method, and where non-size-balanced covariates are likely, WIEE will be a natural choice. In scenarios where non-size-balanced covariates are unlikely, but possible, it may be appropriate to explore whether deviations from size-balance have occurred in the data and then select MWCR/WRGEE or WIEE accordingly. One way to empirically check the size-balance assumption is to plot the cluster size against the cluster mean of each component of the vector $\boldsymbol{X}$. Although the definition of size-balance refers to distributions of $\boldsymbol{X}$ rather than expectations, any clear relationship in the plot would be an indication of departure from size-balance. Alternatively, the

cluster size could be regressed on the cluster mean of $\boldsymbol{X}$. A significant effect of the cluster mean of $\boldsymbol{X}$ is evidence against size-balance.

Even when the 'size-balanced' assumption holds, the consistency of MWCR and WRGEE requires the validity of additional conditions, which relate to the choice of working correlation. For both methods, the relevant conditions are satisfied when the selected working correlation is exchangeable, which would be a natural choice when members are unordered within clusters (e.g. pups in litters). In longitudinal-data settings, an auto-regressive working correlation - although a natural choice - can jeopardise the consistency of the methods, if the relevant conditions are not satisfied. In such longitudinal-data scenarios, MWCR and WRGEE with an auto-regressive correlation structure (MWCR(AR) and WRGEE(AR)) may provide biased estimation. Choosing, instead, to use an exchangeable working correlation, even when the true correlation structure is not exchangeable but rather depends on the times of measurement for the members, guarantees the consistency of the methods, but not increased efficiency of the estimates, compared to WIEE. Exploratory simulation studies have shown that, in scenarios where the relevant conditions are violated, MWCR(AR) and WRGEE(AR) were biased, whereas MWCR(EX) and WRGEE(EX) were more efficient than WIEE while maintaining consistency. The efficiency gain is, however, not guaranteed when the true correlation differs greatly from exchangeable, although other authors (see, for example, Park and Shin (1999)) have found that, when cluster size is not informative, the use of standard GEE with exchangeable correlation can provide more efficient estimates that IEE when the true correlation is auto-regressive.

The efficiency gains from the application of MWCR/WRGEE are expected to be small if the covariates involved are cluster-constant. In fact, if the selected working correlation is exchangeable then WRGEE and MWCR reduce to WIEE, therefore no efficiency gains are possible. Given these considerations, the adoption of either WRGEE or MWCR is not recommended when the covariates are cluster-constant.

When there is a mixture of cluster-constant, cluster-varying size balanced and cluster-varying non-size balanced covariates, MWCR should not be used; WIEE should be used instead. As was seen in the data example, even where cluster-varying covariates (T and T^2) have means that differ modestly across cluster sizes, the bias from the use of MWCR for the effects of these covariates appeared to be appreciable.

When the cluster-varying covariates are categorical non-size-balanced, the adapted WRGEE method, WBGEE, may be suitable, but the possible efficiency gains will be modest, while the consistency conditions which relate to the choice of working correlation are still required. Huang and Leroux (2011) consider further populations for inference when the cluster size is informative. Both the WRGEE and MWCR methods could be extended to these populations. Another area for further work is to identify further scenarios in which the WRGEE method is unbiased or minimally biased.

Though the range of scenarios in which the WRGEE and MWCR methods are unbiased is somewhat limited, and there are restrictions on the choice of working correlation, these methods are simple to implement. Due to the possibility of increased efficiency they are worthy of consideration alongside the more generally applicable methods (WIEE) based on an independence working correlation. The conditions we have specified for consistent estimation from MWCR and WRGEE can form a useful basis for considering whether these methods are appropriate for each specific data example.

B.1 Proofs of Theorems 1, 2 and 3

In preparation for the proofs, note that equation (4.1) can be rewritten as

$$\sum_{i=1}^{K} \boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{Y}_i^*, \boldsymbol{X}_i^*) = \sum_{i=1}^{K} \sum_{j=1}^{N_i} \frac{1}{\phi} \boldsymbol{g}_{ij} (Y_{ij} - \mu_{ij}) = \mathbf{0}, \tag{B-1}$$

where

$$\boldsymbol{g}_{ij} = \sum_{l=1}^{N_i} \frac{\partial \mu_{il}}{\partial \boldsymbol{\beta}} v(\mu_{il})^{-1/2} \hat{r}_{ilj} v(\mu_{ij})^{-1/2} = \sum_{l=1}^{N_i} \boldsymbol{X}_{il}\ v(\mu_{il})^{1/2} \hat{r}_{ilj} v(\mu_{ij})^{-1/2} \tag{B-2}$$

is an implicit weighting for the jth measurement. As mentioned in Section 4.2.1, when $h^{-1}(\theta) = \theta$ and there are no covariates, $\boldsymbol{g}_{ij} = \hat{r}_{i+j}$.

Proof of Theorem 1.

We show that the expectation of the contribution from a single cluster to estimating equations (4.2) evaluated at $\boldsymbol{\rho}_0$ and the true value of $\boldsymbol{\beta}$ equals zero. Hence, estimating equations (4.2) are consistent.

Analogously to equation (B-1), equation (4.2) can be written as

$$\sum_{i=1}^{K} \frac{1}{\Delta_i} \sum_{j=1}^{N_i} \sum_{s \in \Lambda_{ij}} \frac{1}{\phi} \underset{\sim}{\boldsymbol{g}}_{ij(s)} \{Y_{ij} - \mu_{ij}^{C1}\} = \mathbf{0}, \tag{B-3}$$

where $\underset{\sim}{\boldsymbol{g}}_{ij(s)}$, analogously to $\boldsymbol{g}_{ij}$ in equation (B-2), is the implicit weighting for the jth member of cluster i when it is in subcluster s, and Λ_{ij} denotes the set of indices of the $\Delta_i m N_i^{-1}$ subclusters containing the jth member of cluster i. Note that the total weight given to cluster i in equation (B-3) is $\Delta_i^{-1} \sum_{j=1}^{N_i} \sum_{s \in \Lambda_{ij}} \underset{\sim}{\boldsymbol{g}}_{ij(s)}$, i.e. the average of $\Delta_i m$ values of $m\underset{\sim}{\boldsymbol{g}}_{ij(s)}$. In the special case where $h^{-1}(\theta) = \theta$ and there are no covariates, the average value of $m\underset{\sim}{\boldsymbol{g}}_{ij(s)}$ is a scalar and is the average, over each of the Δ_i $(m \times m)$ submatrices of $\hat{\boldsymbol{R}}_i$, of the sum of the elements of its inverse matrix.

For the sampling mechanism described immediately before Theorem 1, let $\underset{\sim}{\boldsymbol{g}}$ denote the resulting implicit weighting for the chosen member when it is in the chosen subcluster (see equation (B-3)). Denote expectations of the distributions of Y, $\boldsymbol{X}$, $\underset{\sim}{\boldsymbol{X}}$ and $\underset{\sim}{\boldsymbol{g}}$ under this sampling mechanism by $E^S(.)$. Note that $E^S(Y|\boldsymbol{X}) = E^{C1}(Y|\boldsymbol{X})$.

It can be seen that the expectation of the contribution of a single cluster to equations (B-3) at $\boldsymbol{\rho} = \boldsymbol{\rho}_0$ and the true value of $\boldsymbol{\beta}$ is $\phi^{-1} m E^S_{Y,\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}},N}[\underset{\sim}{\boldsymbol{g}}\{Y - \mu^{C1}(\boldsymbol{X})\}]$. Now,

$$\begin{aligned} E^S_{Y,\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}},N}[\underset{\sim}{\boldsymbol{g}}\{Y - \mu^{C1}(\boldsymbol{X})\}] &= E^S_{\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}}} E^S_{N|\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}}} E^S_{Y|\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}},N}[\underset{\sim}{\boldsymbol{g}}\{Y - \mu^{C1}(\boldsymbol{X})\}] \\ &= E^S_{\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}}}[\underset{\sim}{\boldsymbol{g}} E^S_{N|\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}}}\{E^S(Y|\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}},N) - \mu^{C1}(\boldsymbol{X})\}]. \end{aligned}$$

From Conditions 2 and 3, respectively, it follows that $N \perp\!\!\!\perp \underset{\sim}{\boldsymbol{g}} \mid \boldsymbol{X}, \underset{\sim}{\boldsymbol{X}}$ and $Y \perp\!\!\!\perp \underset{\sim}{\boldsymbol{g}} \mid N, \boldsymbol{X}, \underset{\sim}{\boldsymbol{X}}$. So,

$$\begin{aligned} E^S_{Y,\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}},N}[\underset{\sim}{\boldsymbol{g}}\{Y - \mu^{C1}(\boldsymbol{X})\}] &= E^S_{\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}}}[\underset{\sim}{\boldsymbol{g}} E^S_{N|\boldsymbol{X},\underset{\sim}{\boldsymbol{X}}}\{E^S(Y|\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},N) - \mu^{C1}(\boldsymbol{X})\}] \\ &= E^S_{\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}}}[\underset{\sim}{\boldsymbol{g}}\{E^S(Y \mid \boldsymbol{X},\underset{\sim}{\boldsymbol{X}}) - \mu^{C1}(\boldsymbol{X})\}]. \end{aligned}$$

So, using Condition 1,

$$\begin{aligned} E^S_{Y,\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}},N}[\underset{\sim}{\boldsymbol{g}}\{Y - \mu^{C1}(\boldsymbol{X})\}] &= E^S_{\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}}}[\underset{\sim}{\boldsymbol{g}}\{E^S(Y \mid \boldsymbol{X}) - \mu^{C1}(\boldsymbol{X})\}] \\ &= E^S_{\boldsymbol{X},\widetilde{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}}}[\underset{\sim}{\boldsymbol{g}}\{E^{C1}(Y \mid \boldsymbol{X}) - \mu^{C1}(\boldsymbol{X})\}] \\ &= E^S_{\boldsymbol{X},\underset{\sim}{\boldsymbol{X}},\underset{\sim}{\boldsymbol{g}}}[\underset{\sim}{\boldsymbol{g}} \times 0] = \mathbf{0}. \end{aligned}$$

Proof of Theorem 2

We shall show that when equation (4.3) is true, the contribution of a cluster to equa-

tions (4.4) evaluated at $\boldsymbol{\rho}_0$ and the true value of $\boldsymbol{\beta}$ has the same expectation as its contribution to the WIEE. Since WIEE are known to be consistent, so will be WRGEE.

Analogously to equation (B-1), we can write the contribution of a cluster to equations (4.4) as

$$\boldsymbol{U}^W(\boldsymbol{\beta}; \boldsymbol{Y}^*, \boldsymbol{X}^*) = \frac{s^p}{\hat{r}_{++}} \sum_{j=1}^{N} \frac{1}{\phi} \boldsymbol{g}_j \{Y_j - \mu^p(\boldsymbol{X}_1)\}.$$

As $\boldsymbol{X}$ is cluster-constant, $\boldsymbol{X}_j = \boldsymbol{X}_1$, $\mu_j^p = \mu_1^p$ and $v(\mu_j^p) = v(\mu_1^p)$ for all j. So, from equation (B-2), $\boldsymbol{g}_j = \boldsymbol{X}_1 \hat{r}_{+j}$. Therefore, assuming Condition 1, we have that at $\boldsymbol{\rho} = \boldsymbol{\rho}_0$ and the true value of $\boldsymbol{\beta}$,

$$\begin{aligned} &E_{\boldsymbol{Y}^*,\boldsymbol{R}}\{U^W(\boldsymbol{\beta}; \boldsymbol{Y}^*, \boldsymbol{X}^*) \mid \boldsymbol{X}_1, N\} = \\ &= s^p \boldsymbol{X}_1 \frac{1}{\phi} \sum_{j=1}^{N} E_{\boldsymbol{R}} \left(\frac{r_{+j}}{r_{++}} \mid \boldsymbol{X}_1, N \right) [E_{Y_j}(Y_j \mid \boldsymbol{X}_1, N) - \mu^p(\boldsymbol{X}_1)]. \end{aligned} \quad \text{(B-4)}$$

Assuming Condition 2, it follows from equation (B-4) that

$$E_{\boldsymbol{Y}^*,\boldsymbol{R}}\{U^W(\boldsymbol{\beta}; \boldsymbol{Y}^*, \boldsymbol{X}^*) \mid \boldsymbol{X}_1, N\} = \frac{s^p}{N} \boldsymbol{X}_1 \frac{1}{\phi} \sum_{j=1}^{N} [E(Y_j \mid \boldsymbol{X}_1, N) - \mu^p(\boldsymbol{X}_1)].$$

This does not depend on the choice of working correlation structure, and hence the WRGEE have the same expectation as the WIEE.

Proof of Theorem 3

For the identity link function, $v(\mu) = 1$. So, from equation (B-2), $\boldsymbol{g}_j = \sum_{l=1}^{N} \boldsymbol{X}_l r_{lj}$. Hence, at $\boldsymbol{\rho} = \boldsymbol{\rho}_0$ and the true value of $\boldsymbol{\beta}^p$,

$$\begin{aligned} &E_{\boldsymbol{Y}^*}\{\boldsymbol{U}^W(\boldsymbol{\beta}^p; \boldsymbol{Y}^*, \boldsymbol{X}^*) \mid \boldsymbol{X}^*, N, \boldsymbol{R}\} = \\ &= E_{\boldsymbol{Y}^*} \left[\frac{s^p}{\phi \, r_{++}} \sum_{j=1}^{N} \sum_{l=1}^{N} \boldsymbol{X}_l r_{lj} \{Y_j - \mu^p(\boldsymbol{X}_j)\} \mid \boldsymbol{X}^*, N, \boldsymbol{R} \right]. \end{aligned}$$

So, assuming Condition 1,

$$\begin{aligned} &E_{\boldsymbol{Y}^*}\{\boldsymbol{U}^W(\boldsymbol{\beta}^p; \boldsymbol{Y}^*, \boldsymbol{X}^*) \mid \boldsymbol{X}^*, N, \boldsymbol{R}\} = \\ &= \frac{s^p}{\phi \, r_{++}} \sum_{j=1}^{N} \sum_{l=1}^{N} \boldsymbol{X}_l r_{lj} \{(\theta_0 + \gamma_N - \beta_0^p) + \boldsymbol{X}_j^T (\boldsymbol{\theta}_1 - \boldsymbol{\beta}_1^p)\} = \\ &= \frac{s^p}{\phi \, r_{++}} \sum_{l=1}^{N} \boldsymbol{X}_l r_{+l} (\theta_0 + \gamma_N - \beta_0^p). \end{aligned} \quad \text{(B-5)}$$

Therefore,

$$E_{\boldsymbol{Y}^*,\boldsymbol{R}}\{\boldsymbol{U}^W(\boldsymbol{\beta}^p;\boldsymbol{Y}^*,\boldsymbol{X}^*) \mid \boldsymbol{X}^*,N\} = (\theta_0+\gamma_N-\beta_0^p)\frac{s^p}{\phi}\sum_{j=1}^{N}\boldsymbol{X}_j E_{\boldsymbol{R}}\left\{\frac{r_{+j}}{r_{++}} \mid \boldsymbol{X}^*,N\right\}. \tag{B-6}$$

Now, assuming Condition 3,

$$\sum_{j=1}^{N}\boldsymbol{X}_j E_{\boldsymbol{R}}\left\{\frac{r_{+j}}{r_{++}} \mid \boldsymbol{X}^*,N\right\} = \sum_{j=1}^{N}\boldsymbol{X}_j E_{\boldsymbol{R}}\left\{\frac{r_{+j}}{r_{++}} \mid N\right\}.$$

So,

$$E_{\boldsymbol{X}^*}\left[\sum_{j=1}^{N}\boldsymbol{X}_j E_{\boldsymbol{R}}\left\{\frac{r_{+j}}{r_{++}} \mid \boldsymbol{X}^*,N\right\} \mid N\right] = \sum_{j=1}^{N}E(\boldsymbol{X}_j \mid N)E_{\boldsymbol{R}}\left\{\frac{r_{+j}}{r_{++}} \mid N\right\}. \tag{B-7}$$

Assuming Condition 2, equation (B-7) implies

$$E_{\boldsymbol{X}^*}\left[\sum_{j=1}^{N}\boldsymbol{X}_j E_{\boldsymbol{R}}\left\{\frac{r_{+j}}{r_{++}} \mid \boldsymbol{X}^*,N\right\} \mid N\right] = \frac{1}{N}\sum_{j=1}^{N}E(\boldsymbol{X}_j \mid N). \tag{B-8}$$

It follows from equations (B-6) and (B-8) that

$$E_{\boldsymbol{Y}^*,\boldsymbol{X}^*,\boldsymbol{R}}\{\boldsymbol{U}^W(\boldsymbol{\beta}^p;\boldsymbol{Y}^*,\boldsymbol{X}^*) \mid N\} = (\theta_0+\gamma_N-\beta_0^p)\frac{s^p}{\phi N}\sum_{j=1}^{N}E(\boldsymbol{X}_j \mid N).$$

This does not depend on the choice of working correlation structure, and hence the WRGEE have the same expectation as the WIEE. Since WIEE are consistent, so are WRGEE.

B.2 Simulations for binary correlated responses with informative cluster size

Here we outline a method for generating binary correlated responses and demonstrate how simulation studies analogous to the ones in Section 4.4 can be carried out for binary responses.

The Bahadur representation (see Section 2.4.2) may be used for generating correlated binary responses but is computationally unattractive for cluster sizes larger than 3. A relatively simple method to generate correlated binary responses with a desired correlation structure is provided by Emrich and Piedmonte (1991). Let $p_{ij} = E(Y_{ij})$ denote the marginal expectation for member j in cluster i. Also let $\rho_{ij} = \text{corr}(Y_{ij}, Y_{ik})$ denote the correlation between responses Y_{ij} and Y_{ik} (note that for binary responses ρ_{ij} is not free to vary over $[-1, +1]$ but is restricted by the marginal means).

The steps to generate binary responses with the desired marginal means and correlations (omitting the indicator for the cluster, i, for simplicity) are:

- Solve the equations:

$$\Phi(z(p_j), z(p_k), \alpha_j) = \rho_j \{p_j(1-p_j)q_j(1-q_j)\}^{1/2} + p_j p_k$$

  where $z(p)$ denotes the pth quantile of a standard normal distribution with distribution function Φ. These equations can be solved for α_j using the bisection method. So, an auxiliary correlation matrix, $\boldsymbol{\Sigma}$, with elements $\hat{\alpha}_j$ is created.

- Generate a multivariate normal variable $\boldsymbol{Z}$ with the correlation matrix $\boldsymbol{\Sigma}$ obtained at the step above.

- Generate $Y_j = \begin{cases} 1 & \text{if } Z_j \leq z(p_j); \\ 0 & \text{otherwise.} \end{cases}$.

The binary responses have the desired properties: $E(Y_j) = p_j$ and $\text{corr}(Y_j, Y_k) = \rho_j$.

In analogy to Section 4.4 we performed simulation studies for binary correlated responses. We here present the results for the Set 2, i.e. for informative cluster size, non-size-balanced X and AR-1 correlation structure. In Step 2 of the simulation procedure (pg. 126), the parameters for the cluster size model were selected to be $\alpha_0 = \alpha_1 = 1$.

The minimum cluster size was $m = 4$. In Step 3 we selected $\lambda_0 = 0$, $\lambda_1 = 1$, so X was non-size-balanced. For the linear predictor in Step 4 we used $\gamma_0 = -0.5$, $\gamma_1 = 0.25$, $\gamma_2 = 2$ and $\gamma_3 = 0$. Finally, in Step 5 we used the method of Emrich and Piedmonte (1991) to induce an AR-1 correlation structure. In each simulation we generated 100 clusters and we repeated this procedure 1000 times.

As shown in Table 4.7, WBGEE led to unbiased inference with efficiency gains of up to 8% for all populations for inference, relative to WIEE.

Population C1					
TRUE $(\beta_0^{CI}, \beta_1^{CI}) = (-0.723, 0.864)$					
ρ	Method	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CV($\hat{\beta}_0$,$\hat{\beta}_1$)	RE($\hat{\beta}_0$,$\hat{\beta}_1$)
	WIEE	-0.72(0.159)	0.86(0.182)	(0.943,0.951)	-
0.2	WBGEE	-0.72(0.158)	0.86(0.180)	(0.945,0.948)	(1.01,1.02)
	WIEE	-0.72(0.171)	0.86(0.192)	(0.946,0.950)	-
0.4	WBGEE	-0.72(0.168)	0.86(0.1185)	(0.947,0.949)	(1.03,1.07)
Population M					
TRUE $(\beta_0^{M}, \beta_1^{M})$ =(-0.432,0.917)					
ρ	Method	$\hat{\beta}_0$(ese($\hat{\beta}_0$))	$\hat{\beta}_1$(ese($\hat{\beta}_1$))	CV($\hat{\beta}_0$,$\hat{\beta}_1$)	RE($\hat{\beta}_0$,$\hat{\beta}_1$)
	WIEE	-0.43(0.155)	0.92(0.188)	(0.952,0.948)	-
0.2	WBGEE	-0.44(0.154)	0.92(0.186)	(0.950,0.948)	(1.00,1.02)
	WIEE	-0.44(0.165)	0.92 (0.200)	(0.941,0.942)	-
0.4	WBGEE	-0.44(0.164)	0.92 (0.192)	(0.943,0.942)	(1.01,1.08)

Table 4.7: Set 2 (binary responses). Application of WIEE and WBGEE for populations M and C1. The cluster size is informative, X is non-size balanced and the assumed correlation is AR-1.

B.3 R functions for the application of the methods

B.3.1 R-function for the application of WRGEE

In the following we present the 'core' function used to fit the WRGEE method for populations M and C1. For simplicity and space-efficiency the code is restricted to the case of linear regression (with the identity link function). It can be easily adapted to the case of logistic or poisson regression. Also, the extension to the application of WBGEE (for the case of cluster-varying categorical exposure) which can provide inference for populations M, C1, C2 or C3 is straightforward.

```
wrgee<-function(formula,dataset,corstr="independence",
                pop="pc",accuracy=0.00005,inrho=NULL,inphi=NULL){
# dataset  : a data-frame which contains the variables in
#            columns and the indicator "id" for the cluster
# formula  : the regression formula
# corstr   : the working correlation structure
#            ("independence", "exchangeable" or "ar1")
# pop      : the population for inference, "M" or "C1"
# accuracy : the accuracy for the convergence criterion
# inrho    : the correlation parameter
# inphi    : the scale parameter

#Model frame
mf<-model.frame(formula,dataset)
#Model matrix
mt<-model.matrix(formula,dataset)

#Obtain the number of members per
#cluster using the function "counts"

counts<-function(id){
clus<-rep(0,length(id))
k0<-0
k1<-1
for(i in 2:length(id)) { i1<-i-1
if(id[i]==id[i1]) {k1<-k1+1
if(i==length(id)) {k0<-k0+1
                   clus[k0]<-k1}}
if(id[i]!=id[i1]) {k0<-k0+1
                   clus[k0]<-k1
                   k1<-1
```

```
if(i==length(id)) {k0<-k0+1
                   clus[k0]<-k1 }}}
clusz<-clus[clus>0]
}
countss<-counts(dataset$id)

#Split the dataset into the clusters.
#Each cluster is defined as a "list"
cluster<-list()
idua<-unique(dataset$id)
for (i in 1:length(idua))
cluster[[i]]<-dataset[dataset$id==idua[i],]

#Fit a GLM to obtain a vector of initial estimates
#for beta (betain)
fit.glm<-glm(formula)
betain<-matrix(unlist(fit.glm$coefficients),ncol=1)

#N is the number of clusters
N<-length(countss)

#p is the number of parameters
p<-length(betain)

#it is the number of iterations
it<-0

#Here starts the main loop until convergence

error<-1

while (error>accuracy){

I0<-0; I1<-0; I2<-0
it<-it+1
X<-list()
Y<-list()
mu<-list()

for (i in 1:N){

#ni is the number of members in the ith cluster
```

```
ni<-countss[i]

#Model frame for each patient
mfi<-model.frame(formula,cluster[[i]])
#X= model matrix for the ith patient
X[[i]]<-matrix(unlist(model.matrix(formula,cluster[[i]])),ncol=p)
#Y= outcome for the ith patient
Y[[i]]<- as.vector(model.response(mfi))

#mu is the linear predictor
mu[[i]]<-as.vector(X[[i]]%*%betain)
D<-X[[i]]

#Working correlation matrix

if (corstr=="independence")
R<-diag(1,ni,ni)

if (corstr=="exchangeable"){
R<-matrix(rho,nrow=ni,ncol=ni); diag(R)<-1}

if (corstr=="ar1"){
R<-matrix(0,ni,ni)
for (j in 1:ni)
    for (k in 1:ni){
        if (j==k) R[j,k]=1
        else R[j,k]=rho^abs(j-k)}}

#The inverse correlation matrix
R<-solve(R)

#Weights for the selected population for inference
w.m<-rep(ni/sum(R),ni)
w.c<-rep(1/sum(R),ni)

#Weighted working correlation matrix
if(pop=="pm") R<-R%*%diag(w.m)
if(pop=="pc") R<-R%*%diag(w.c)

#The working covariance matrix
W<-R*phi
```

```
#Estimate of the true covariance matrix
VY<-(Y[[i]]-mu[[i]])%*%t(Y[[i]]-mu[[i]])

I0in<-t(D)%*%W%*%D
I0<-I0+I0in
I1in<-t(D)%*%W%*%(Y[[i]]-mu[[i]])
I1<-I1+I1in
I2in<-t(D)%*%W%*%VY%*%W%*%D
I2<-I2+I2in
}

#Update the estimates for beta in each iteration

betanew<-betain+solve(I0)%*%I1
error<-sum((betanew-betain)^2)
betain<-betanew
}

#Exit this loop when convergence is achieved

#Compute the robust variance estimator
#at the final values of beta
robust<-solve(I0)%*%(I2)%*%solve(I0)

#Return a list of objects form fitting the model
return<-list()
return$beta<-betanew
return$population<-pop
return$vbeta<-robust
return$iterations<-it
return
}
```

B.3.2 R-function for the application of MWCR

We now present the function used to fit the MWCR method for population C1, for the case of linear regression with the identity link function. The code can be adapted to accommodate logistic or poisson regression.

```
mwcr<-function(formula,dataset,corstr="independence",
               accuracy=0.00005,inrho=NULL,inphi=NULL){
# dataset  : a data-frame which contains the variables in
#            columns and the indicator "id" for the cluster
```

```
# formula  : the regression formula
# corstr   : the working correlation structure
#            ("independence", "exchangeable" or "ar1")
# accuracy : the accuracy for the convergence criterion
# inrho    : the correlation parameter
# inphi    : the scale parameter

#Model frame
mf<-model.frame(formula,dataset)
#Model matrix
mt<-model.matrix(formula,dataset)

#Obtain the number of members per
#cluster using the function "counts"

counts<-function(id){
clus<-rep(0,length(id))
k0<-0
k1<-1
for(i in 2:length(id)) { i1<-i-1
if(id[i]==id[i1]) {k1<-k1+1
if(i==length(id)) {k0<-k0+1
                   clus[k0]<-k1}}
if(id[i]!=id[i1]) {k0<-k0+1
                   clus[k0]<-k1
                   k1<-1
if(i==length(id)) {k0<-k0+1
                   clus[k0]<-k1 }}}
clusz<-clus[clus>0]
}
countss<-counts(dataset$id)

#Split the dataset into the clusters.
#Each cluster is defined as a "list"
cluster<-list()
idua<-unique(dataset$id)
for (i in 1:length(idua))
cluster[[i]]<-dataset[dataset$id==idua[i],]

#Fit a GLM to obtain a vector of initial estimates
#for beta (betain)
fit.glm<-glm(formula)
```

```
betain<-matrix(unlist(fit.glm$coefficients),ncol=1)

#N is the number of clusters
N<-length(countss)

#p is the number of parameters
p<-length(betain)

#it is the number of iterations
it<-0

#Here starts the main loop until convergence
error<-1

while (error2>accuracy){

I0=0; I1<-0; I2<-0
it<-it+1

for (i in 1:N){
# ni is the number of members per cluster
ni<-countss[i]

# m is the minimum cluster size in the dataset
m<-min(countss)

# is a matrix with all the combinations (by m) of members
a<-combn(ni,m)

I0in<-0; I1in<-0; I2in<-0


# subcluster is a list which stores the contribution of
# each subcluster
subcluster<-list()

for(j in 1:ncol(a)){

subcluster[[j]]<-cluster[[i]][a[,j],]

#Model frame for subcluster j
mfj<-model.frame(formula,subcluster[[j]])
```

```
#Xj is the matrix of covariates for the jth subcluster
Xj<-matrix(unlist(model.matrix(formula,subcluster[[j]])),ncol=p)
#Yj is the vector of responses for the jth subcluster
Yj<- as.vector(model.response(mfj))

#eta=the linear predictor
etaj<-as.vector(Xj%*%betain)
#muj=expected value for Yj
muj<-etaj
Dj<-Xj

#R is the working correlation matrix

if (correlation=="independence") R<-diag(m)

if (corstr=="exchangeable"){
R<-matrix(rho,nrow=m,ncol=m); diag(R)<-1}


if(correlation=="ar1"){
R<-matrix(0,m,m)
for (l in 1:m)
      for (k in 1:m){
          if (l==k) R[l,k]=1
          else R[l,k]=rho^abs(l-k)}
}

#Vj is the inverse covariance matrix
Vj<-R*phi
Vj<-solve(Vj)
VYj<-(Yj-muj)%*%t(Yj-muj)

I0inj<-t(Dj)%*%Vj%*%Dj
I1inj<-t(Dj)%*%Vj%*%(Yj-muj)
I2inj<-t(Dj)%*%Vj%*%(Yj-muj)

#Sum the contributions from each subcluster

I0in<-I0in+I0inj
I1in<-I1in+I1inj
I2in<-I2in+I2inj
}
```

```
#Sum the contributions from each cluster

I0<-I0+1/ncol(a)*I0in
I1<-I1+1/ncol(a)*I1in
I2<-I2+1/ncol(a)^2*I2in%*%t(I2in)
}


#Update the estimates for beta in each iteration

betanew<-betain+solve(I0)%*%I1
error<-sum((betanew-betain)^2)
betain<-betanew
}

#Exit this loop when convergence is achieved

#Compute the robust variance estimator
#at the final values of beta
robust<-solve(I0)%*%(I2)%*%solve(I0)

#Return a list of objects form fitting the model
return<-list()
return$beta<-betanew
return$vbeta<-robust
return$iterations<-it
return
}
```

Chapter 5

Contrasting informative cluster size and missing data

5.1 Introduction

When making *marginal inference* for clustered data with varying cluster size, there are several populations of potential interest. First, there is the population of all members of all clusters. Second, there are the populations of typical members of all clusters (see Definition 3.1, pg. 67 and Definition 3.4, pg. 81). Third, it is possible to regard the observed clusters as incomplete and seek inference for the population of all members of all complete clusters. The problem then becomes one of missing data. Alternatively, we may seek *cluster-specific inference* for the observed or the complete clusters by fitting random effects models.

As an example, consider a study in which patients in poor health diagnosed with a certain condition attend for health check-ups. At each check-up various outcomes can be obtained such as health of the patient or cost of consultation. Also, covariates are obtained; patient sex and age, treatments taken etc. Each check-up can be thought of as a *member*, and the check-ups for a patient form a *cluster*. We may wish to investigate how the various covariates are related to the outcome by using a marginal regression model, so that if Y is a measure of health and the covariates determine a vector $\boldsymbol{X}$, then we wish to write down a model such as

$$E(Y \mid \boldsymbol{X}) = h^{-1}(\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}_1) \tag{5.1}$$

for a known link function h and unknown parameters $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$ which we wish to

estimate.

The size of the cluster, N, i.e. the number of cluster members, may vary. In our example the number of check-ups performed for the patient may vary. In this case when considering marginal inference, a population for inference must be specified. The population of all members of all clusters (population M) consists of all members, so here simply all check-ups. By contrast, the population of typical members of all clusters (population C1) can be viewed informally to arise from every cluster contributing exactly one member with equal probability, so that here one check-up is taken per patient. If cluster size is constant, inference for the two populations is the same. Otherwise, they may be different, because larger clusters contribute more members than smaller clusters do to the first population, but each cluster contributes only one member to the second expectation regardless of its size.

A third type of inference is of interest, if the variation in cluster size is considered to result from missing data, that is, all clusters are actually of equal size and the reason why some observed clusters are smaller is that not all of their members have been observed. From this perspective, there are incomplete clusters and complete clusters, and we want to make inference for the population of all members of all complete clusters (population A) based on the data from our sample of incomplete clusters. That is, we want to know the expectation of Y given $\boldsymbol{X} = \boldsymbol{x}$ for a member drawn at random from the population of all members of all complete clusters. Note that where the variation in cluster size has not arisen from missing data then this hypothetical population will generally not be of interest; for example, when each member is a sexual partnership in the last year, and the cluster is the individual involved in these (Copas et al., 2009).

To see how the populations may differ, continuing with our example, consider that patients in poor health diagnosed with a certain condition are encouraged to attend for monthly health check-ups over the subsequent year. Whilst all patients attend the first check-up, over time some patients drop out of care and are not seen again, so that the number of check-ups varies between patients. We assume that dropout occurs because the patients perceive themselves to be cured of the initial condition because their health has returned to a good level before dropout and remains good after dropout. So, inference for the three populations will be generally different.

Inference for population A could be selected to assess how patient characteristics

are linked to health whether under care or dropped out. Inference for population M could be selected to assess how patient characteristics are linked to the cost of check-ups, aggregating over patients. Inference for population C1 could be selected to assess how patient characteristics are linked to their typical health level whilst under care. Inference for the three populations will differ because the average measure of health is higher in population A than in population C1 as health is good after dropout. If health were assessed in population M, then the average level would be lower even than in population C1 because patients with slow recovery from diagnosis have more check-ups and have poorer health. Unless inference for the three populations is this same, the marginal model above in (5.1) may not apply to all three populations or, if it does, the true parameter values may be different.

Further populations may be selected for inference where the expectation of the outcome given the covariates for the member is related to the covariates values of other members in the same cluster. This has been termed informative covariate structure and discussed in Chapter 3, Section 3.6. In Chapter 3, Section 3.12 we have also discussed how the methods of Huang and Leroux (2011) can be seen as providing inference for the complete clusters. We shall not discuss informative covariate structure further in the current chapter. The approaches discussed in Sections 5.3.2-5.3.5 and also in Section 5.4 are not useful if the covariate structure is informative.

Missing data methods are well known by statisticians; methods for informative cluster size are less well known. A number of authors (Hoffman et al., 2001; Williamson et al., 2003) have referred to the relation between informative cluster size and missing data mechanisms, but have not made clear what this relation is. The target of this chapter is to clarify this relation by considering methods for inference for the complete and observed clusters. We aim to provide intuition about why different methods are generally needed for the two types of inferences. We also describe missing data mechanisms under which methods for observed-cluster inference can be seen as special cases of the methods proposed for complete-cluster inference.

For settings of clustered-data where clusters are considered to be incomplete the notation was introduced in Section 2.9. We define additional notation and describe a special missing data mechanism in Section 5.2. These additional definitions will aid comparisons between complete- and observed-cluster inference and also provide

insight into how the two types of inference can be equivalent. We start by considering marginal inference in Section 5.3. We briefly describe the most frequently used methods for complete- and observed-cluster marginal inference. We investigate which Missing Data Mechanisms (MDMs) may or may not lead to informative cluster size and importantly identify special MDMs where the complete-cluster inference can be equivalent to one for the observed clusters. We discuss the implications of the choice of method for analysis. In Section 5.4 we make analogous comparisons for cluster-specific inference. Importantly, we clarify that the method of Dunson et al. (2003) where the main outcome and the cluster size are modelled jointly (with shared random effects for the two models) should be generally considered as a method for complete-cluster inference.

5.2 Notation and definitions

Depending on the application it may be possible to index each member within every cluster, in which case we say the members are ordered, and otherwise we say the members are unordered. For example if members are check-ups for patients at planned times then these are indexed by time, but for pups in litters there is no 'natural' indexing and these are unordered. If members are unordered then they must be exchangeable within clusters, if they are ordered then they may or may not be exchangeable.

For a given cluster let N denote the number of observed members in the cluster. Depending on the application it may be possible to index members but if not, we imagine an arbitrary indexing is applied. Let $\boldsymbol{D} = (D_1, D_2, ..., D_N)^T$ denote a vector of indices for those N members. In many applications $\boldsymbol{D}$ will equal $(1, \ldots, N)^T$. When members are unordered within clusters an arbitrary indexing is applied in which case $\boldsymbol{D} = (1, 2, \ldots, N)^T$. However, as explained later, if members are ordered and the observed cluster represents the observed part of a larger complete cluster, then $\boldsymbol{D}$ may take other values. Let Y_j and $\boldsymbol{X}_j$ denote the outcome and covariate vector for the jth member of the cluster. Y_j and $\boldsymbol{X}_j$ are observed if and only if $\boldsymbol{D}$ includes j. Some or all elements of $\boldsymbol{X}_j$ may be known even if $\boldsymbol{D}$ does not include j, for example if these are cluster-constant or if they are determined by the indices of the members. Let $\widetilde{\boldsymbol{Y}}^*_{(\boldsymbol{R})i} = (Y_{D_1}, ..., Y_{D_N})^T$, and $\widetilde{\boldsymbol{X}}^*_{(\boldsymbol{R})i} = (\boldsymbol{X}_{D_1}, ..., \boldsymbol{X}_{D_N})^T$. We use i to index the cluster and K to denote the number of clusters. We assume that $(N_i, \boldsymbol{D}_i, \widetilde{\boldsymbol{X}}^*_{(\boldsymbol{R})i}, \widetilde{\boldsymbol{Y}}^*_{(\boldsymbol{R})i})$,

$(i = 1, ...K)$ are i.i.d. observations from the joint distribution of $(N, \boldsymbol{D}, \widetilde{\boldsymbol{X}}^*_{(\boldsymbol{R})}, \widetilde{\boldsymbol{Y}}^*_{(\boldsymbol{R})})$.

In Chapter 3 we defined the *population of typical members of all clusters (C1)* to be the subpopulation of the *population of all members of all clusters (M)* in which each cluster contributes a single member at random. Whilst this provides an intuitive way to consider inference for observed clusters, we here present an alternative definition. Let H_i $(i = 1, ..., K)$ be independent discrete random variables with $P(H = j \mid N, \boldsymbol{D}, \widetilde{\boldsymbol{X}}^*_{(\boldsymbol{R})}, \widetilde{\boldsymbol{Y}}^*_{(\boldsymbol{R})}) = 1/N$ if $\boldsymbol{D}$ includes j and zero otherwise. So, H_i represents the index of a randomly selected member of the ith observed cluster. Inference for the population of typical observed members means estimating the parameters of a model for either $f(Y_H \mid \boldsymbol{X}_H)$ or $E(Y_H \mid \boldsymbol{X}_H)$. Population C1 provides inference for a typical member of a typical cluster which is the focus of the paper by Williamson et al. (2003), though those authors did not provide a formal definition as just described. Inference about population M means estimating the parameters of a model for the following distribution

$$\sum_{\boldsymbol{d}} P(\boldsymbol{D} = \boldsymbol{d}) \sum_{j=1}^{N} f(Y_{d_j} | \boldsymbol{X}_{d_j}) / E(N)$$

or its expectation.

It may be the case that the observed clusters each arose from a corresponding complete cluster containing N_{comp} members. Let $\widetilde{\boldsymbol{Y}}^* = (Y_1, ..., Y_{N_{\text{comp}}})^T$, and $\widetilde{\boldsymbol{X}}^* = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{N_{\text{comp}}})^T$ denote the vector of outcomes and matrix of covariates for a complete cluster. For each cluster let $\boldsymbol{R} = (R_1, \ldots, R_{N_{\text{comp}}})^T$, where $R_j = 1$ if the jth member of the complete cluster is observed and $R_j = 0$ otherwise. So, $N = \sum_{j=1}^{N_{\text{comp}}} R_j$ and $\boldsymbol{D} = \{j : R_j = 1\}$. Let $\overline{D} = (\overline{D}_1, ..., \overline{D}_{N_{\text{comp}}-N})^T$ denote the subvector of $(1, \ldots, N_{\text{comp}})^T$ composed of the indices of the missing members, and let $\widetilde{\boldsymbol{Y}}^*_{(\bar{\boldsymbol{R}})} = (Y_{\overline{D}_1}, ..., Y_{\overline{D}_{N_{\text{comp}}-N}})$ and $\widetilde{\boldsymbol{X}}^*_{(\bar{\boldsymbol{R}})} = (\boldsymbol{X}_{\overline{D}_1}, \ldots, \boldsymbol{X}_{\overline{D}_{N_{\text{comp}}-N}})$. So, $\widetilde{\boldsymbol{Y}}^*$ can be partitioned in $\widetilde{\boldsymbol{Y}}^*_{(\boldsymbol{R})}$ and $\widetilde{\boldsymbol{Y}}^*_{(\bar{\boldsymbol{R}})}$, and similarly for $\widetilde{\boldsymbol{X}}^*$. We assume that $N > 0$ for all clusters, so that there are not additional complete clusters in which no members are observed.

Making inference for the complete clusters means either assuming that $f(Y_j \mid \boldsymbol{X}_j) = f(Y_1 \mid \boldsymbol{X}_1)\ \forall\ j$ and estimating parameters of a model for $f(Y_j \mid \boldsymbol{X}_j)$ or assuming that $E(Y_j \mid \boldsymbol{X}_j) = E(Y_1 \mid \boldsymbol{X}_1)\ \forall\ j$ and estimating parameters of a model for $E(Y_j \mid \boldsymbol{X}_j)$. Whether this model is estimable from the observed data depends on the assumed missingness mechanism (see Chapter 2, Section 2.9).

Recall that the missingness process is said to be monotone, if there exists a permutation $(k_1, \ldots, k_{N_{\text{comp}}})$ of $(1, \ldots, N_{\text{comp}})$ for which $P(R_{k_{j+1}} = 0 \mid R_{k_j} = 0) = 1$ $\forall\, j$. Without loss of generality, we shall assume that if the missing process is monotone then the members have been ordered so that $(k_1, \ldots, k_{N_{\text{comp}}}) = (1, \ldots, N_{\text{comp}})$, and hence for every cluster $\boldsymbol{D} = (1, \ldots, N)^T$. When the process is monotone, N is a 1-1 function of $\boldsymbol{R}$ and so $P(\boldsymbol{R} \mid \widetilde{\boldsymbol{X}}^*, \widetilde{\boldsymbol{Y}}^*)$ defines $P(N \mid \widetilde{\boldsymbol{X}}^*, \widetilde{\boldsymbol{Y}}^*)$ and vice versa. When members are unordered within clusters then the missingness process is considered monotone. Where members are ordered, missing data may be monotone such as dropout in a longitudinal scenario or non-monotone as in data for teeth.

Even if, in reality, the observed clusters are not generated from complete clusters, it is still possible to pretend that they were, to assume a form for the missingness process (e.g. MAR), and then make inference for these hypothetical complete clusters. In this situation, $\boldsymbol{D} = (1, ..., N)^T$ and missingness is monotone.

For the purposes of relating methods of observed-cluster inference to complete-cluster inference we identify a special MDM. We define the *equal-probability MDM* to arise when *within* a complete cluster all members have the same probability of being missing. More formally, this arises when $P(R_j = 1 \mid \widetilde{\boldsymbol{X}}^*, \widetilde{\boldsymbol{Y}}^*) = N/N_{\text{comp}}\ \forall\, j$.

The equal-probability MDM can be seen to operate in two stages. Firstly, the number of observed members, N, is assigned to each of the complete clusters. Secondly, in each of the complete clusters, N members are sampled using simple random sampling to create the observed clusters. In Section 5.3.1 (under the subheading 'Inverse probability weighting') we clarify that if the information that determines N is part of $\boldsymbol{X}$, i.e. $P(N \mid \widetilde{\boldsymbol{X}}^*, \widetilde{\boldsymbol{Y}}^*) = P(N \mid \widetilde{\boldsymbol{X}}^*)$, then the MDM is covariate-dependent MCAR.

5.3 Marginal inference: complete versus observed clusters

5.3.1 Methods for complete cluster inference

Research into the marginal analysis of incomplete data has addressed both the conditions under which standard methods are consistent despite the missing data and also the development of methods which model the missingness process by assuming a certain missing data mechanism. These were discussed in Chapter 2, Section 2.9.

For purposes of comparison with methods for observed-cluster inference, we first briefly review methods which can be used to obtain complete-cluster marginal inference, focusing on scenarios where members are unordered within clusters. In doing this we gain insight of whether these methods may or may not provide an inference which also applies to the observed clusters. This largely depends on the assumed MDM and whether the members within clusters are exchangeable or not.

We define $\mu^A(\boldsymbol{x}) = E(Y_j \mid \boldsymbol{X}_j = \boldsymbol{x}) \; \forall \; j$, to be the marginal expectation of the outcome given covariates in the complete clusters. The assumed model is

$$\mu^A(\boldsymbol{x}) = h^{-1}(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}_1) \tag{5.2}$$

where function $h(.)$ is the link function.

Inverse probability weighting

Inverse probability weighting (IPW) is commonly applied when MAR is assumed. Typically, it requires a model for $\psi_j = P(R_j = 1 \mid \widetilde{\boldsymbol{X}}^*, \widetilde{\boldsymbol{Y}}^*)$, $j = 1, \ldots, N_{\text{comp}}$. Possible forms for this model include logistic regression and probit regression (see Section 2.9.3).

For monotone MAR data, ψ_j can be written as $\psi_j = \prod_{k=1}^{j-1}(1 - \lambda_k)$, where $\lambda_k = P(N = k \mid N \geq k-1, \boldsymbol{X}_1, \boldsymbol{Y}_1, \ldots, \boldsymbol{X}_k, \boldsymbol{Y}_k)$ and a model can be specified for λ_k. If members are ordered, one may assume, for example,

$$h(\lambda_k) = \gamma_k + \boldsymbol{X}_k^T\boldsymbol{\theta}_{k1} + Y_k\theta_{k2}. \tag{5.3}$$

On the other hand, under an equal-probability MDM it will not generally be natural to include Y_k or cluster-varying elements of $\boldsymbol{X}_k$ in the model for λ_k. However, missingness may depend on observed or unobserved *cluster-constant* information, $\boldsymbol{S}$, i.e. $P(N \mid \widetilde{\boldsymbol{Y}}^*, \widetilde{\boldsymbol{X}}^*, \boldsymbol{S}) = P(N \mid \boldsymbol{S})$. We distinguish three cases of interest:

1. If $\boldsymbol{S}$ is observed and is part of $\boldsymbol{X}$, the MDM specified is covariate-dependent MCAR, in which case the standard GEE would provide consistent estimation.

2. If $\boldsymbol{S}$ is observed but for scientific reasons is not part of $\boldsymbol{X}$, then the MDM is not covariate-dependent MCAR. In this case one may model missingness by assuming a model for N in terms of $\boldsymbol{S}$:

$$h(\lambda_k) = \gamma_k + \boldsymbol{S}^T\boldsymbol{\theta}_k. \tag{5.4}$$

The probability of each member being observed is $E(N \mid \boldsymbol{S})/N_{\text{comp}}$. The inverse probability weights for each observed member will be $N_{\text{comp}}/E(N \mid \boldsymbol{S})$ or equivalently $1/E(N \mid \boldsymbol{S})$. Estimates of $E(N \mid \boldsymbol{S})$ can be obtained using consistent estimates of $\boldsymbol{\theta}_k$ and γ_k from fitting (5.4).

3. If $\boldsymbol{S}$ is unobserved, then a model such as (5.4) cannot be specified. In this case an alternative IPW approach which avoids specifying a model for missingness can be applied. The probability of observation for each member, conditional on N, is N/N_{comp}. Consequently the observed members are weighted by N_{comp}/N, or equivalently $1/N$.

Multiple imputation

Multiple imputation (MI) is another approach commonly applied when MAR is assumed. The missing members are imputed from an assumed model for the complete data, whose parameters are estimated from the observed data (using Multivariate Normal Imputation or Chained equations). Auxiliary information (i.e. variables which are not part $\boldsymbol{X}$) could be used in the imputation model. Hot-deck imputation may also be used, where data for a missing member are imputed from another observed member which is judged similar to the member to be imputed.

For monotone ordered MAR data, resulting for example from dropouts in a longitudinal study, MI would impute the missing members using information on observed outcomes and covariates from members which have been observed.

Under an equal-probability MDM and informative cluster size, it is not reasonable to use information from large clusters to impute members in small clusters, because these are considered to be inherently different. Assuming that the members within a complete cluster are exchangeable, a hot-deck-type imputation which imputes missing members from a given cluster using information from other members of the same cluster (or members from other equally sized clusters) could be appropriate, although this is not an approach we would recommend in practice.

For marginal inference, the imputed datasets can be analysed using GEE and the estimates are combined using Rubin's rules.

Selection models

A selection modelling approach for MNAR monotone data would specify, for example, a model for λ_k analogous to the one in equation (5.3) by adding a term $Y_{k+1}\theta_{3k}$. Under an equal-probability MDM, dependence should only be specified in terms of cluster-constant covariates and not in terms of outcomes. In the selection modelling framework this can be achieved by specifying dependence on random effects shared by the model for missingness and main outcome, but such an approach would provide cluster-specific inference. We further discuss the shared random effects approach in Section 5.4.

Pattern-mixture models

Pattern-mixture models (PMMs) are less commonly applied. A model for Y in terms of $\boldsymbol{X}$ in the complete cluster for each missingness pattern is specified and fitted. Typically, untestable restrictions are used to ensure identifiability.

For monotone dropouts, PMMs can be applied as a suitable approach to sensitivity analysis (Molenberghs and Verbeke, 2006). This is because they use identifiability restrictions which make explicit the links between the distribution of $(\boldsymbol{X}, Y)$ for members who have dropped out and members who have not dropped out and belong to different missingness pattern.

Under an equal-probability MDM, the special hot-deck MI considered earlier in this section could be viewed as a PMM with the sole identifiability restriction that the distribution of $(\boldsymbol{X}, Y)$ for missing members in a given cluster is the same as the distribution of $(\boldsymbol{X}, Y)$ for the observed members within the same cluster (or the distribution of $(\boldsymbol{X}, Y)$ for observed members in other, equally sized clusters).

5.3.2 Methods for observed-cluster inference

Research into methods for observed-cluster inference has primarily focused on conditions under which standard methods for clustered data are consistent, or only minimally biased asymptotically, and adaptations of these methods to provide consistent estimation if these conditions are not met. These methods were seen in detail in Chapter 3.

If our interest is in marginal regression, then the model of interest is either for the expectation of Y given $\boldsymbol{X}$ in the population of typical cluster members 1,

$$\mu^{C1}(\boldsymbol{x}) = E(Y_H \mid \boldsymbol{X}_H = \boldsymbol{x}) = h^{-1}(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}_1) \; \forall \; \boldsymbol{x} \tag{5.5}$$

or in the population of all cluster members,

$$\mu^M(\boldsymbol{x}) = E(NY_H \mid \boldsymbol{X}_H = x)/E(N \mid \boldsymbol{X}_H = x) = h^{-1}(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}_1) \; \forall \; \boldsymbol{x}. \quad (5.6)$$

Note that these models allow order effects, i.e. we do not need to assume that $E(Y_j \mid \boldsymbol{X}_j) = E(Y_1 \mid \boldsymbol{X}_1) \; \forall \; j$ as we did in (5.2).

Commonly used methods for marginal observed-cluster inference under informative cluster size for population C1 and M are CWGEE (Williamson et al., 2003) and standard GEE with independence working correlation, respectively (see Section 3.5.1).

5.3.3 Missing data mechanisms and informative cluster size

If each of the observed clusters arose from a complete cluster of common size, then informative cluster size (ICS) can arise from a MDM which may be MAR or MNAR but not MCAR. Consider our introductory example, a longitudinal setting with varying cluster size due to dropout. Dropout may be MAR, occurring with high probability after the first check-up with a high value of Y (health) given $\boldsymbol{X}$ (a vector of patient characteristics such as sex and time). In this case, smaller cluster sizes will be linked to higher values of Y given $\boldsymbol{X}$ and hence ICS. Analogously, if dropout occurs with high probability once the patient perceives that health has reached a high level, between check-ups, then this will also lead to ICS but is an example of an MNAR mechanism.

Note that it is possible to have non-informative cluster size, i.e. $\mu^M(\boldsymbol{x}) = \mu^{C1}(\boldsymbol{x}) \; \forall \; \boldsymbol{x}$, but $\mu^A(\boldsymbol{x})$ not equal to $\mu^M(\boldsymbol{x}) = \mu^{C1}(\boldsymbol{x})$. For example, assume that Y is binary taking values 0 (poor health) and 1 (good health). Also assume the MDM is MNAR, and for every patient dropout occurs between check-ups and just before the first check-up at which $Y = 1$ would have been observed. Then, $\mu^{C1}(\boldsymbol{x}) = \mu^M(\boldsymbol{x}) = 0$ and the cluster size is non-informative, but $\mu^A(\boldsymbol{x}) > 0$.

If the marginal model in equation (5.2) is correctly specified for the complete clusters and either

1. $P(R_j \mid \widetilde{\boldsymbol{Y}}^*, \widetilde{\boldsymbol{X}}^*) = P(R_j \mid \boldsymbol{X}_j)$ (note that MCAR is a special case of this) or
2. a) $P(R_j \mid \widetilde{\boldsymbol{Y}}^*, \widetilde{\boldsymbol{X}}^*) = P(R_j \mid \widetilde{\boldsymbol{X}}^*)$

   b) Pepe and Anderson's condition (see equation 2.15)

are satisfied, then the marginal models in equations (5.5) and (5.6) are correctly specified for populations C1 and M, respectively and $\mu^M(\boldsymbol{x}) = \mu^{C1}(\boldsymbol{x}) = \mu^A(\boldsymbol{x}) \; \forall \; \boldsymbol{x}$.

Under Condition 1, since the information that determines missingness is part of $\boldsymbol{X}$, then *conditional on* $\boldsymbol{X}$ the distribution of Y is independent of R (see also Little, 1995 and Hedeker and Gibbons, 2006, pg. 281-285). So, the subpopulation of members with $\boldsymbol{X} = \boldsymbol{x}$ from observed clusters is a simple random sample of the subpopulation of members with $\boldsymbol{X} = \boldsymbol{x}$ from the complete clusters. Therefore, $\mu^A(\boldsymbol{x}) = \mu^M(\boldsymbol{x}) \; \forall \; \boldsymbol{x}$. Under Condition 2(a), $Y \perp\!\!\!\perp R \mid \widetilde{\boldsymbol{X}}^*$, so $f(Y \mid R, \widetilde{\boldsymbol{X}}^*) = f(Y \mid \widetilde{\boldsymbol{X}}^*) = f(Y \mid \widetilde{\boldsymbol{X}}^*_{(\boldsymbol{R})})$ and therefore $E(Y \mid \widetilde{\boldsymbol{X}}^*) = E(Y \mid \widetilde{\boldsymbol{X}}^*_{(\boldsymbol{R})})$. Using this result and Condition 2(b) it follows that $E(Y \mid \widetilde{\boldsymbol{X}}^*_{(\boldsymbol{R})}) = E(Y \mid \widetilde{\boldsymbol{X}}^*) = E(Y \mid \boldsymbol{X})$, therefore $\mu^A(\boldsymbol{x}) = \mu^M(\boldsymbol{x}) \; \forall \; \boldsymbol{x}$.

As a consequence of Condition 1, the number of members in the observed clusters, N, which is determined by $\boldsymbol{R}$, is also independent of Y given $\boldsymbol{X}$. So, $f(Y \mid \boldsymbol{X}, N) = f(Y \mid \boldsymbol{X})$, the cluster size is not informative and $\mu^{C1}(\boldsymbol{x}) = \mu^M(\boldsymbol{x}) \; \forall \; \boldsymbol{x}$. Similarly, under Condition 2(a), $Y \perp\!\!\!\perp N \mid \widetilde{\boldsymbol{X}}^*$ and consequently $E(Y \mid N, \widetilde{\boldsymbol{X}}^*) = E(Y \mid \widetilde{\boldsymbol{X}}^*)$. Using Condition 2(b) and the result from Condition 2(a), $E(Y \mid \boldsymbol{X}, N) = E(Y \mid \widetilde{\boldsymbol{X}}^*, N) = E(Y \mid \widetilde{\boldsymbol{X}}^*) = E(Y \mid \boldsymbol{X})$. Hence, $\mu^{C1}(\boldsymbol{x}) = \mu^M(\boldsymbol{x}) \; \forall \; \boldsymbol{x}$, also. Note that Conditions 1 and 2 are also sufficient conditions for the consistency of the standard GEE which, under any of these conditions, can be used to provide inference for any of the three populations.

Hoffman et al. (2001), Williamson et al. (2003) vaguely stated that MCAR is equivalent to non-informative cluster size. We have seen that under related conditions, of which MCAR is special case, the cluster size is non-informative and also $\mu^M(\boldsymbol{x}) = \mu^{C1}(\boldsymbol{x}) = \mu^A(\boldsymbol{x}) \; \forall \; \boldsymbol{x}$. We have also seen that the cluster size might not be informative under a wider range of MDMs including MNAR mechanisms.

5.3.4 Failure, in general, of methods for complete cluster inference to provide observed-cluster inference

In general, it cannot be expected that under a MAR mechanism $\mu^A(\boldsymbol{x}) = \mu^{C1}(\boldsymbol{x})$ or $\mu^A(\boldsymbol{x}) = \mu^M(\boldsymbol{x}) \; \forall \; \boldsymbol{x}$. Hence, standard IPW and MI which are known to provide consistent estimation for complete-cluster inference, i.e. for $\mu^A(\boldsymbol{x})$, under a MAR mechanism, will fail to provide inference for the observed clusters, i.e. for either $\mu^{C1}(\boldsymbol{x})$ or $\mu^M(\boldsymbol{x})$.

Consider again the longitudinal example, and the MAR mechanism whereby drop-

out occurs with high probability after the first check-up with a high value of Y (health) given $\boldsymbol{X}$ (e.g. sex and time). Standard MI would impute the values of $(\boldsymbol{X}, Y)$ for members after dropout for each patient, using the values seen after the time of dropout in other patients who did not dropout and had similar measures of health before the time of dropout. Since good health is maintained once it is achieved, these imputed values will be high, reflecting the values that would have been observed without dropout. This imputation approach will not provide inference for population M because on average across all patients such imputed values will be higher than those observed (i.e. before dropout). It will also not provide inference for population C1 because for each patient the values imputed after dropout will be higher than those observed, i.e. before dropout. Using similar arguments it can be seen that standard IPW methods will also not provide inference for neither population C1 nor population M.

5.3.5 Inference for population C1 using methods for complete-cluster inference

In Section 5.3.4 we have seen that standard methods for complete-cluster inference under MAR missing data when the members are ordered and not exchangeable, are not suitable methods to obtain inference for population C1. However, in Section 5.3 we described how the approach of Williamson et al. (2003) of weighting by $1/N$ (recall that we assume $N > 0$) can be seen as an IPW method for complete-cluster inference assuming the probabilities of observation for each member within a given cluster are equal. Under this equal-probability MDM, $\mu^A(\boldsymbol{x}) = \mu^{C1}(\boldsymbol{x}) \ \forall \ \boldsymbol{x}$. This can be seen because in order to form population C1, each cluster contributes one of its observed members at random. Under such MDM, within each cluster the observed members have the same distribution of $(\boldsymbol{X}, Y)$ as the unobserved ones. Hence, the distribution of $(\boldsymbol{X}, Y)$ in populations C1 and A will be equal. Therefore, the approach of Williamson et al. (2003) provides inference which applies both to populations C1 and A.

Similarly, under an equal-probability MDM the special hot-deck imputation approach seen in the end of Section 5.3.1 provides complete-cluster inference which also applies to observed-cluster inference for population C1. In particular, this special imputation relies on similar principles as WCR (Hoffman et al., 2001) and can be seen as equivalent to CWGEE (Williamson et al., 2003) as the number of imputations tends to

infinity.

Often, it will not be reasonable to assume an equal-probability MDM. Nevertheless, an IPW method which makes this assumption, correctly specifies a model for N, and weights each member by $1/E(N)$ (see Section 5.3.1), will provide observed-cluster inference for population C1. If insufficient information is available to correctly specify a model for N under an equal-probability MDM, then the CWGEE method of Williamson et al. (2003) can be used for observed-cluster inference since it does not require a model for N. For example, if members are monotone MAR and are ordered, we may consider using IPW based on a model for $\boldsymbol{R}$ such as the one in equation (5.3) to obtain inference for the complete clusters. If, contrary to reality, we assume an equal-probability MDM, and correctly specify a model for N such as in equation (5.4) or apply CWGEE we obtain inference for the observed clusters.

5.4 Cluster-specific inference: complete versus observed clusters

5.4.1 Methods for complete-cluster inference

If the data are covariate-dependent MAR and a model is correctly specified for $\widetilde{\boldsymbol{Y}}^*$ in terms of $\widetilde{\boldsymbol{X}}^*$ and fitted by maximum likelihood, it will give consistent estimation. A random effects model is commonly used:

$$E(Y_{ij} \mid \widetilde{\boldsymbol{X}}_i^*, \boldsymbol{b}_i) = h^{-1}(\beta_0 + b_{0i} + \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1 + \boldsymbol{X}_{ij}^T\boldsymbol{b}_{1i}), \tag{5.7}$$

where the random terms $\boldsymbol{b}_i$ are assumed to arise from a zero-mean distribution (commonly the multivariate Normal distribution) and $\boldsymbol{b}_i \perp \widetilde{\boldsymbol{X}}_i^*$.

If the data are not covariate-dependent MAR, then joint random effect models based on models for $\widetilde{\boldsymbol{Y}}^*$ and the missingness pattern $\boldsymbol{R}$ can be fitted. Estimation by maximum likelihood will be consistent if the models are correctly specified. As these models share random effects the assumed MDM is MNAR. The model for $\widetilde{\boldsymbol{Y}}^*$ is specified as in (5.7) with the same distributional assumptions. The model for $\boldsymbol{R}$ specifies that $P(\boldsymbol{R} = \boldsymbol{r}|\widetilde{\boldsymbol{X}}^*, \widetilde{\boldsymbol{Y}}^*, \boldsymbol{b}) = f(\boldsymbol{r}, \widetilde{\boldsymbol{X}}^*_{(\boldsymbol{r})}, \boldsymbol{b})\ \forall\ \boldsymbol{r}$ for some function $f(.)$.

In Section 5.3.1 we considered models for $\boldsymbol{R}$ under MAR mechanisms to apply the IPW approach. Similarly here, the model for $\boldsymbol{R}$ can be specified in terms of

ψ_j and λ_k if these are redefined to be $P(R_j = 1|\widetilde{\boldsymbol{X}}^*, \widetilde{\boldsymbol{Y}}^*, \boldsymbol{b})$ and $P(N = k|N \geq k, \boldsymbol{X}_1, Y_1, \ldots, \boldsymbol{X}_k, Y_k, \boldsymbol{b})$, respectively.

So, for monotone missing data and ordered members, the term involving Y_k in equation (5.3) is removed, so the model is

$$h(\lambda_k) = \gamma_k + \boldsymbol{\theta}_{k1}^T \boldsymbol{X}_k + \boldsymbol{\theta}_{k2}^T \boldsymbol{b}.$$

For monotone missing data but unordered members we specify

$$h(\lambda_k) = \gamma_k + \boldsymbol{\theta}_{k1}^T \boldsymbol{S} + \boldsymbol{\theta}_{k2}^T \boldsymbol{b}, \tag{5.8}$$

where $\boldsymbol{S}$ is a subvector of $\boldsymbol{X}$ and it consists of cluster-constant elements of $\boldsymbol{X}$.

5.4.2 Methods for observed-cluster inference

The joint modelling approach of Dunson et al. (2003) was presented in Chapter 3, Section 3.5.4. The Dunson's method specifies a model for the joint conditional distribution of $\widetilde{\boldsymbol{Y}}^*$ and N given $\widetilde{\boldsymbol{X}}^*$ and shared random effects $\boldsymbol{b}$. It corresponds to the shared random effects approach for complete-cluster inference under missing data (see, for example, Diggle et al., 2002, pg. 301-303).

Dunson's method assumes that $\widetilde{\boldsymbol{X}}^* \perp \boldsymbol{b}$ in the complete clusters and also that the missingness depends on $\boldsymbol{X}$ and $\boldsymbol{b}$ and not on Y. It provides inference for the complete clusters, i.e. it estimates the conditional distribution of Y given $\boldsymbol{X}$ and $\boldsymbol{b}$ for members of the complete clusters and also the distribution of $\boldsymbol{b}$ in the population of complete clusters.

Nevertheless, Dunson et al. (2003) presented their method as one for observed-cluster inference. We identify two scenarios under which the inference from the Dunson's approach also applies to the observed clusters.

Case A: If the conditions:

A.1. N depends on $\widetilde{\boldsymbol{X}}^*$ and $\boldsymbol{b}$ but not on Y and

A.2. $\boldsymbol{X}$ is cluster constant,

are satisfied, then Dunson's method provides complete-cluster inference which coincides with observed-cluster inference.

Under these conditions, the conditional distribution of Y_H given $\boldsymbol{X}_H$ and $\boldsymbol{b}$ (see Section 5.2 for the definition of Y_H and $\boldsymbol{X}_H$) is the same as the conditional distribution

of Y given $\boldsymbol{X}$ and $\boldsymbol{b}$. Also, $\boldsymbol{X}_H \perp \boldsymbol{b}$ (recall that $\boldsymbol{X} \perp \boldsymbol{b}$). Furthermore, it is obvious that the distribution of $\boldsymbol{b}$ in the population of observed clusters is the same as the distribution if $\boldsymbol{b}$ in the population of complete clusters since $N > 0$ for all clusters. The implication of these is that the Dunson's model which is correctly specified for the complete clusters, can be considered to also apply to the observed clusters.

Although Dunson et al. (2003) and Chen et al. (2011) applied the method in a dataset from toxicology studies where $\boldsymbol{X}$ was cluster-constant, they seem to suggest that $\boldsymbol{X}$ may also be cluster-varying. In our view, in the presence of cluster-varying covariates, the Dunson's joint modelling approach does not provide observed-cluster inference, at least not without additional conditions. In particular, for cluster varying $\boldsymbol{X}$ we argue that:

Case B: If the conditions:

B.1. N depends on $\boldsymbol{b}$ but not on $\widetilde{\boldsymbol{X}}^*$ and Y and

B.2. $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{N_{\text{comp}}}$ are independent and identically distributed within each cluster,

are satisfied, then Dunson's method provides complete-cluster inference which coincides with the observed-cluster inference.

Under these considerations, again the distribution of Y_H given $\boldsymbol{X}_H$ and $\boldsymbol{b}$ in the observed clusters is the same as the distribution of Y given $\boldsymbol{X}$ and $\boldsymbol{b}$ in the complete clusters. Also, $\boldsymbol{X}_H \perp \boldsymbol{b}$. So the Dunson's model also applies to the observed clusters. Note that the condition $\boldsymbol{X}_H \perp \boldsymbol{b}$ can be false if Conditions (B.1) and (B.2) do not hold. For example, if $\boldsymbol{X}_H$ is not size-balanced (i.e. if the distribution of $\boldsymbol{X}_H$ is not independent of N) and the model for N correctly specifies a dependence of N on cluster-varying elements of $\boldsymbol{X}$, this will result in violation of the condition $\boldsymbol{X}_H \perp \boldsymbol{b}$.

In scenarios other than the ones considered in Cases A and B, Dunson's approach will provide inference for the complete clusters and this will generally not apply to the observed clusters.

5.4.3 Equivalence of cluster-specific inference for populations M and C1

For cluster-specific observed-cluster inference a GLMM such as in (5.7) may be specified, and if correct, then for a randomly selected member of a randomly selected ob-

served cluster, Y_H is independent of N conditional on $\boldsymbol{X}_H$ and $\boldsymbol{b}$.

Analogously to (5.5) and (5.6) we define

$$\mu^{C1}(\boldsymbol{x}, \boldsymbol{b}) = E(Y_H \mid \boldsymbol{X}_H = \boldsymbol{x}, \boldsymbol{b}) \; \forall \; \boldsymbol{x} \text{ and}$$
$$\mu^{M}(\boldsymbol{x}, \boldsymbol{b}) = \frac{E(NY_H \mid \boldsymbol{X}_H = \boldsymbol{x}, \boldsymbol{b})}{E(N|\boldsymbol{X}_H = \boldsymbol{x}, \boldsymbol{b})} = h^{-1}(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}_1 + \boldsymbol{x}^T\boldsymbol{b}) \; \forall \; \boldsymbol{x}$$

for populations C1 and M, respectively. Using the condition $Y_H \perp\!\!\!\perp N \mid \boldsymbol{X}_H, \boldsymbol{b}$ (i.e. cluster size is not informative conditional on $\boldsymbol{b}$ and $\boldsymbol{X}_H$) it can be easily seen that a correctly specified GLMM applies to both populations M and C1 in the sense that $\mu^{C1}(\boldsymbol{x}, \boldsymbol{b}) = \mu^{M}(\boldsymbol{x}, \boldsymbol{b}) = h^{-1}(\beta_0 + b_0 + \boldsymbol{x}^T\boldsymbol{\beta}_1 + \boldsymbol{x}^T\boldsymbol{b}_1) \; \forall \; \boldsymbol{x}$.

5.4.4 Marginal inference from a cluster-specific model

When the link function in model (5.7) is the identity one, $E(Y_H \mid \boldsymbol{X}_H = \boldsymbol{x}, \boldsymbol{b}) = \beta_0 + b_0 + \boldsymbol{x}^T\boldsymbol{\beta}_1 + \boldsymbol{x}^T\boldsymbol{b}_1$. Since $\boldsymbol{X}_H \perp \boldsymbol{b}$, then $E(\boldsymbol{b} \mid \boldsymbol{X}_H = \boldsymbol{x}) = \boldsymbol{0}$. So, $\mu^{C1}(\boldsymbol{x}) = E(Y_H \mid \boldsymbol{X}_H = \boldsymbol{x}) = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}_1$ and the vector of fixed regression coefficients in cluster-specific inference coincides with the ones from marginal regression for population C1. Also, for a random intercepts model with the log-link function, the regression parameters $\boldsymbol{\beta}_1$ in a marginal (for population C1) and a cluster-specific model will again be the equal (see also Section 2.7, last paragraph). For link functions other than the identity and the log-ones, the regression parameters from a marginal and a random effects model are not equal, in general. Nevertheless, $E(Y_H \mid \boldsymbol{X}_H)$ may be obtained by numerical integration over the distribution of $\boldsymbol{b}$ in $E(Y_H \mid \boldsymbol{X}_H, \boldsymbol{b})$.

5.5 Discussion

We have clarified that informative cluster size does not simply correspond to deviations from MCAR as suggested by previous authors, though covariate-dependent MCAR does lead to non-informative cluster size. We have also seen that even under MNAR MDMs, the cluster size may not be informative.

From the viewpoint of different MDMs we have offered insight into why different methods are used according to whether inference is sought for the observed or complete clusters. Methods designed for complete-cluster inference that assume an equal-probability MDM can be considered for inference for observed clusters, and those that have not already been considered for this purpose could be considered further.

This insight has also provided an important note of caution regarding the use of joint random effects models (Dunson et al., 2003) if the observed clusters are of interest. We have seen that the method of Dunson et al. (2003) generally provides inference for the complete clusters. We identified cases where this inference also applies to the observed clusters. The necessary conditions are that (i) the model for N only includes cluster-constant covariates, (ii) the covariates are independent of the random effects and (iii) the model for the expected outcome only includes cluster-constant covariates or i.i.d cluster-varying covariates.

The literature on joint random effects models for informative cluster size problems has not discussed populations for inference nor the link to marginal models. We have established that a correctly specified random effects model provides cluster-specific inference which applies to both population C1 and population M. We have also discussed how marginal inference for population C1 can be obtained from a cluster-specific model.

In this chapter we have considered scenarios where the covariate structure is non-informative. In practice informative covariate structure may arise alongside informative cluster size. In this case, marginal inference for population C1 may be considered uninformative and other populations could be preferred (Huang and Leroux, 2011). In cluster-specific inference, informative covariate distribution would correspond to violation of the assumption of independence between covariates and random effects in which case other methods could be considered (see Sections 2.8 and 3.12).

Chapter 6

Conclusions and further work

In the analysis of clustered data often the cluster size varies. If the variation in cluster size has arisen because some data are missing, then we may seek inference for the population of all members of all complete clusters. However, if the variation in cluster size is an inherent feature of the data, then the observed data are considered to be complete and we seek inference for the observed clusters.

In this work we primarily focused on methods for observed-cluster marginal inference when the cluster size or covariate structure are informative. In Chapter 3 we discussed informative cluster size, introduced the concept of informative covariate structure and additional populations for inference, and proposed estimation methods for marginal inference using weighted independence estimating equations. In Chapter 4 we investigated efficient methods for marginal inference under informative cluster size and informative covariate structure. We discussed an existing efficient method and explained how bias can arise from the use of this method. We also proposed an alternative efficient method. We clarified conditions for the consistency of both methods. In Chapter 5 we have examined the relation between missing data and informative cluster size and attempted to bridge the gap between the two. We have indicated scenarios where observed- and complete-cluster inference might coincide.

In the next three sections we briefly summarise the most important findings in Chapters 3, 4 and 5. We note the limitations of the proposed methodology. In section 6.4 we make recommendations for further work, stemming from the considerations in this thesis.

6.1 Inference under informative cluster size and covariate structure

Informative cluster size has been defined to arise when the expected outcome, $E(Y)$, conditional on covariates, $\boldsymbol{X}$, is not equal to the expected outcome conditional on covariates and the cluster size, N. Early work on marginal inference under informative cluster size (Hoffman et al., 2001; Williamson et al., 2003) primarily focused on scenarios where the covariates are cluster-constant and the effect of $\boldsymbol{X}$ is the same in all clusters. Hoffman et al. (2001) and Williamson et al. (2003) proposed the WCR and CWGEE methods, respectively, to provide inference for the population of typical members 1. Inference for the population of all members can be obtained using the standard GEE with independence working correlation. We clarified that if $\boldsymbol{X}$ is non-size-balanced or the effect of $\boldsymbol{X}$ differs between clusters of different sizes, inference for the population of all members and typical members 1 is different. When $\boldsymbol{X}$ is size-balanced and the effect of $\boldsymbol{X}$ is the same in all clusters, only the intercept terms differ between populations M and C1; the effect of $\boldsymbol{X}$ is the same.

Importantly, we have identified another type of informativeness which might arise when the covariates are cluster-varying. We have defined informative covariate structure to arise when the conditional expectation of the outcome for a member given covariates for that member and the cluster size depends on the covariate values of other members in the cluster where the member in question belongs. Informative covariate structure may arise concurrently with informative cluster size, or can solely arise even when the cluster sizes do not vary. As informative cluster size, informative covariate structure also causes problems for analysis and standard methods are deemed inappropriate.

When the covariate structure is informative and the covariates are categorical we introduced additional populations for inference and proposed estimation methods using weighted independence estimating equations. In particular, when the covariate structure is informative we proposed WIEE for the populations of typical members 2 and 3 (populations C2 and C3). In using WIEE-C2, each cluster member is inversely weighted by the number of members in that cluster with the same value of $\boldsymbol{X}$ as the member in question. WIEE-C3 is analogous to WIEE-C2 but is restricted to clusters

which contain all values of $\boldsymbol{X}$. WIEE-C3 provides a 'matched' analysis and can be seen as a method which estimates the within-cluster effect of $\boldsymbol{X}$ by removing the effect of any measured or unmeasured cluster-level confounders. When a great proportion of clusters do not contain all values of $\boldsymbol{X}$, WIEE-C2 may be used instead, if cluster-confounding due to unmeasured factors is not of great concern.

The WIEE-C3 method consistently estimates the within-cluster effect of $\boldsymbol{X}$ if the effect of $\boldsymbol{X}$ is homogeneous. If the assumption of homogeneous effects is not true, it also requires that all clusters contain all values of $\boldsymbol{X}$. In this case it estimates what can be seen as the 'average' within-cluster effect of $\boldsymbol{X}$. When the effect of $\boldsymbol{X}$ is not homogeneous and not all clusters contain all values of $\boldsymbol{X}$, WIEE-C3 does not consistently estimate the average within-cluster effect of $\boldsymbol{X}$. For such scenarios Huang and Leroux (2011) proposed modelling the frequency distribution of $\boldsymbol{X}$ to obtain suitable weights which up-weight cluster members to represent both themselves and also missing members. In terms of estimation, DWGEE2 are analogous to WIEE-C2 but each cluster member is inversely weighted by the expected (rather than the observed) number of members in that cluster with same value of $\boldsymbol{X}$ as the member in question. So, a pseudo-population of complete clusters is created, where all clusters contain all values of $\boldsymbol{X}$. If the model for the frequency distribution of $\boldsymbol{X}$ is correctly specified, then DWGEE2 method consistently estimates the within-cluster effect of $\boldsymbol{X}$, even if the effect of $\boldsymbol{X}$ is not homogeneous.

Importantly, the DWGEE2 method can only be applied when auxiliary cluster-level covariates are observed. Huang and Leroux (2011) performed simulation studies with a single binary exposure. When the effect of the exposure is not homogeneous and about 10% of the clusters do not experience both exposure levels, WIEE-C3 has been seen to provide little bias. As auxiliary cluster-constant covariates might not always be available or suitable, application of DWGEE2 may not be feasible in practise. Therefore it is important to investigate the amount of bias from the use of WIEE-C3 under non-homogeneous exposure effects for a wider range of scenarios (e.g. larger proportion of clusters with missing exposure levels, categorical exposure with more than 2 categories etc.).

Table 6.2 is a summary of the weighting methods considered or developed in Chapter 3.

Method	Issue - Nature of exposure
	Informative Cluster Size - cluster-constant or cluster-varying size-balanced exposure
IEE	Recommended when modelling costs at the 'aggregate level' (health economics).
WIEE-C1	The experience of the typical patient is of direct interest.
	Informative Covariate Structure - cluster-varying categorical exposure
IEE	Recommended when modelling costs at the 'aggregate level' (health economics).
WIEE-C1	Not intuitive/useful. Does not deal with cluster confounding. Not recommended.
WIEE-C2	More intuitive inference than population C1. Recommended when all exposure levels are present in all clusters; otherwise it does not deal fully with cluster confounding and is not recommended.
WIEE-C3	Estimates the within-cluster effect of the exposure assuming homogeneous exposure effects. Deals with cluster-confounding by unobserved cluster-level confounders. Requires all levels of exposure to be present in all clusters if the effect of the exposure varies across clusters.
DWGEE2	It consistently estimates the within-cluster effect of the exposure whether this is homogeneous or not. It does not require that all clusters contain all levels of the exposure but it requires a correctly specified model for the frequency distribution of the exposure in terms of cluster-constant auxiliary variables.
DWGEE3	It provides causal inference for the potential treatment effect when the exposure is a treatment applied to the cluster. It requires a correctly specified model for treatment allocation in terms of auxiliary variables.

Table 6.1: Weighting methods for informative cluster size and informative covariate structure

6.2 Efficient marginal inference under informative cluster size and structure

The WIEE proposed in Chapter 3 for inference for populations M, C1, C2 and C3, but also the estimators of Huang and Leroux (2011) use independence working correlation. So, the dependence between repeated outcomes is not fully acknowledged.

In Chapter 4 we have discussed efficient methods for informative cluster size and covariate structure. MWCR was proposed by Chiang and Lee (2008) as an extension of WCR to provide consistent estimation with increased efficiency compared to WCR for the population of typical members 1. We proposed an alternative efficient method, WRGEE, which uses a non-diagonal working correlation structure and may offer efficiency gains compared to WIEE. The weights in WRGEE can be selected to obtain inference for the population of all members or the population of typical members 1. We

Method	Issue - Nature of covariates
	Informative Cluster Size - cluster-constant or cluster-varying size-balanced covariates
MWCR	Inference for population C1. Small efficiency gains for cluster-constant covariates. Significant efficiency gains for cluster-varying covariates. Exchangeable is the only 'safe' choice for the working correlation. Caution: additional conditions might be needed under longitudinal-data settings and autoregressive working correlation (see Section 4.2.3, Theorem 1, Conditions 2 and 3)
WRGEE	Inference for either population C1 or population M, using suitable weights. Small efficiency gains for cluster-constant covariates. Significant efficiency gains for cluster-varying covariates. Exchangeable is the only 'safe' choice for the working correlation. Caution: additional conditions needed for longitudinal-data settings and autoregressive working correlation (see Section 4.3, Theorems 2 and 3)
	Informative Covariate Structure - cluster-varying non-size balanced covariates
MWCR	Biased inference for population C1, due to violation of the 'size-balanced' condition. Not recommended.
WRGEE	Biased inference for population C1 and population M, due to violation of the 'size-balanced' condition. Not recommended.
WBGEE	Can provide unbiased inference for populations M, C1, C2 or C3. Moderate efficiency gains. Requires *categorical* exposure. Uses block diagonal correlation matrices. Under non-size-balanced covariates and hence informative covariate structure, the most useful and intuitive inference is for population C2 or C3.

Table 6.2: Use of efficient methods under informative cluster size/structure

also proposed an extension of WRGEE, WBGEE, to obtain efficient inference for population of typical members 2 and 3. WBGEE can also be extended to the populations considered by Huang and Leroux (2011).

A summary in terms of recommendations for practical use of the MWCR and WRGEE methods structure is provided in Table 6.2.

Application of MWCR and WRGEE may result in biased inference in certain scenarios; this bias was not clearly described when MWCR was initially proposed. We clarify conditions, necessary for consistent estimation when using MWCR and WRGEE. These conditions relate to the structure of covariates and the choice of the working correlation structure.

As it is evident from the illustration in Chapter 4, the application of MWCR and WRGEE might be problematic when dealing with longitudinal data, and these methods should be used with caution. In scenarios of clustered data where there is no time or order component among the members of each cluster (such as the ones that arise

in toxicology experiments) the conditions required for consistent estimation are more likely to be satisfied.

When the covariate structure is informative and inference is required for populations M, C2 or C3, WBGEE have been seen to provide small efficiency gains over WIEE. Also, the conditions for the consistency of WRGEE and MWCR limit their applicability in certain scenarios. It is of interest to investigate possible extensions of WRGEE and MWCR or alternative efficient methods which cover a wider range of scenarios and offer more substantial efficiency gains. One possible direction would be to examine whether the locally efficient estimator of Vansteelandt et al. (2007) remains valid under informative cluster size.

6.3 Informative cluster size and missing data

Informative cluster size and missing data have similarities, but the connection between the two has not been made clear in the literature. Methods for missing data are well known by statisticians; methods for informative cluster size are less known. In Chapter 5 we attempted clarifying the relation between the two. This may offer insight to researchers as to which method should be used for a given problem.

When the cluster size varies, we may seek inference for the complete or the observed clusters. We investigated which missing data mechanisms may lead to informative cluster size. Importantly, we identified a special missing data mechanism under which complete- and observed-cluster marginal inferences coincide. We have also described how IPW methods for complete-cluster inference may provide observed-cluster inference.

For cluster-specific inference, we have identified that the method of Dunson et al. (2003) provides inference for the complete clusters, in general. We clarified conditions under which the inference provided also applies to the observed clusters. Importantly we have explained that, when the cluster size is informative and a random effects model is correctly specified, inferences for the population of typical members 1 and the population of all members are equivalent.

6.4 Further work

The work presented in this thesis gives rise to potential further work. The most important topics for future work are presented below:

- *Extension of methods for informative covariate structure to the case of continuous exposures*.

  The weighting methods considered/developed in this thesis for informative covariate structure are only applicable when the exposure is categorical. An important field for further work is their extension to the case of continuous exposure. The two following extensions of the proposed methods are worthy of further investigation:

  - The continuous exposure can be categorised using meaningful cut-off points and the weights for populations C2/C3 can be obtained. The WIEE for population C2 and C3 can be applied using the categorised or the initial continuous version of the exposure. The last choice assumes that within each category the covariate structure is not informative.

  - An alternative strategy which it is worth further investigation would be to derive a version of DWGEE2/3 methods for continuous exposure, X. It can be seen that when $\boldsymbol{L}$ is cluster-constant and the exposure categorical, then DWGEE2 is equivalent to DWGEE3 where the denominator $E(Z_x|\boldsymbol{L}, N)$ can be modelled as $E(Z_x|\boldsymbol{L}, N) = NP(X = x|\boldsymbol{L})$. For the extension of this idea to the case of continuous exposure, the denominator $NP(X = x|L)$ can be substituted by $N\ f(X = x|\boldsymbol{L})$, where $f(X = x|\boldsymbol{L})$ denotes a density. The consistency of this proposed method relies on the correct specification of the model for the distribution of $X|\boldsymbol{L}$. It is expected that this approach can be sensitive to the misspecification of the model for $X|\boldsymbol{L}$ and also that the weights might have high variability resulting in high standard errors of the parameter estimates. These two issues can be tested using suitable simulation studies.

- *Applicability of methods for cluster-confounding in scenarios of informative covariate structure*.

  Informative covariate structure has similarities with what has been termed 'con-

founding by cluster'. It is of interest to investigate whether methods proposed for cluster-confounding can be used when the covariate structure and the cluster size are informative. There is a wide selection of methods for cluster-confounding which involves the Conditional Likelihood method (Neuhaus and MCculloch, 2006), the poor man's method (Neuhaus and Kalbfleisch, 1998) and the Conditional GEE (Goetgeluk and Vansteelandt, 2008). As these methods deal with cluster-confounding and estimate the within-cluster effect of the exposure, they are worthy of consideration alongside the weighting methods developed for dealing with the issue of informative covariate structure. These methods do not require the exposure to be categorical but generally assume non-informative cluster size and homogeneous exposure effects.

- *An alternative DWGEE2: imputing missing exposure levels and outcomes.*
  In the DWGEE2 method, it is assumed that there are complete clusters which have all exposure levels but in some of the observed clusters certain exposure levels have not been observed. The missingness of exposure levels is modelled in terms of cluster-constant auxiliary covariates, $\boldsymbol{L}$. An alternative approach would be to impute the exposure levels and outcomes where the exposure is modelled in terms of the auxiliary variables, $\boldsymbol{L}$. The imputation can be carried out using a fully conditional specification imputation method for clustered data (see, for example, Nevalainen et al. (2009)). Such an approach may offer increased efficiency compared to the DWGEE2 and may also be of use in the case of continuous exposure.

- *A more efficient WBGEE for population C3.*
  When the covariate structure is informative, WIEE-C3 may be the preferred method of analysis in many scenarios. Nevertheless, its efficiency is reduced when clusters which do not include all levels of exposure are discarded. The WBGEE method (proposed in Chapter 4) for population C3 can be used to provide modest efficiency gains compared to WIEE-C2. The WBGEE method assumes that the subclusters defined by the different levels of the exposure in the cluster are independent. One way to further increase its efficiency is to allow the subclusters within a cluster to be correlated. The correlation between the sub-

clusters of the same cluster can be modelled assuming exchangeability between the subclusters. Initial simulation results show efficiency gains of up to 60%; this is a considerable improvement compared to the 15% efficiency gain of the standard WBGEE. Although the newly proposed method shows potential for significant efficiency gains, the conditions required for its consistency need to be investigated.

Bibliography

Aber, V., Aboulker, J. P., Babiker, A. G., and et al (1996). Delta: A Randomized Double-Blind Controlled Trial Comparing Combinations of Zidovudine Plus Didanosine or Zalcitabine With Zidovudine Alone in HIV-Infected Individuals. *Lancet* **347,** 283–291.

Bahadur, R. R. (1961). A representation of the joint distribution of the responses to n dichotomous outcomes. *Studies on item analyis and prediction* pages 155–168.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61,** 962–972.

Begg, M. D. and Parides, M. K. (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine* **22,** 2591–2602.

Bellio, R. and Brazzale, A. R. (2011). Restricted likelihood inference for generalized linear mixed models. *Statistics and Computing* **21,** 173–183.

Benhin, E., Rao, J. N. K., and Scott, A. J. (2005). Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika* **92,** 435–450.

Bishop, Y. M. M., Fiensberg, S. E., and Holland, P. W. (2000). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press, first edition.

Brumback, B. A., Lumley, T., Dailey, A. B., He, Z., Brumback, L. C., and Livingston, M. D. (2010). Efforts to adjust for confounding by neighborhood using complex survey data. *Statistics in Medicine* **29,** 18901899.

Carey, V., Zeger, S. L., and Diggle, P. J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80,** 517–526.

Carriere, I. and Bouyer, J. (2002). Choosing marginal or random-effects models for longitudinal binary responses: application to self-reported disability among older persons. *BMC Medical Research Methodology* **2,** 15.

Chen, Z., Zhang, B., and Albert, P. S. (2011). A joint modeling approach to data with informative cluster size: Robustness to the cluster size model. *Statistics in Medicine* **30,** 1825–1836.

Chiang, T. C. and Lee, K. Y. (2008). Efficient estimation methods for informative cluster size data. *Statistical Sinica* **80,** 121–123.

Copas, A. J., Mercer, C. H., Farewell, V. T., Nanchahal, K., and Johnson, A. M. (2009). Recent heterosexual partnerships and patterns of condom use: a weighted analysis. *Epidemiology (Cambridge, Mass.)* **20,** 44–51.

Copas, A. J. and Seaman, S. R. (2010). Bias from the use of generalized estimating equations to analyze incomplete longitudinal binary data. *Journal of Applied Statistics* **37,** 911–922.

Crowder, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* **82,** 407–410.

De Gruttola, V. and Tu, X. M. (1994). Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics* **50,** 1003–1014.

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)* **39,** 1–38.

Diggle, P. J., Farewell, D., and Henderson, R. (2007). Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. *Journal of The Royal Statistical Society (Series C)* **56,** 499–529.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. (2002). *Analysis of longitudinal data*. Oxford Statistical Scinece Series.

Dunson, D. B., Chen, Z., and Harry, J. (2003). A bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* **59,** 521–530.

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician* **45,** 302–304.

Farewell, D. M. (2010). Marginal analyses of longitudinal data with an informative pattern of observations. *Biometrika* **97,** 65–78.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2009). *Longitudinal Data Analysis*. Chapman & Hall/CRC.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80,** 141–151.

Fitzmaurice, G. M., Lipsitz, S. R., Molenberghs, G., and Ibrahim, J. G. (2001). Bias in estimating association parameters for longitudinal binary responses with drop-outs. *Biometrics* **57,** 15–21.

Ghilagaber, G. (1998). Analysis of survival data with multiple causes of failure - A comparison of hazard- and logistic-regression models with application in demography. *Quality & Quantity* **32,** 297–324.

Godambe, V. P. (1960). An optimum proprty of regular maximum likelihood estimation. *Annals of mathematical Statistics* **31,** 1208–1212.

Goetgeluk, S. and Vansteelandt, S. (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics* **64,** 772–780.

Goldstein, H., Carpenter, J., Kenward, M. G., and Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling* **9,** 173–197.

Gueorguieva, R. V. (2005). Comments about joint modeling of cluster size and binary and continuous subunit-specific outcomes. *Biometrics* **61,** 862–866.

Guo, W., Ratcliffe, S. J., and Ten Have, T. T. (2004). A random pattern-mixture model for longitudinal data with dropouts. *Journal of the American Statistical Association* **99,** 929–937.

Hachen, D. S. (1988). The competing risks model - A method for analysing processes with multiple types of events. *Sociological methods & Research* **17,** 21–54.

Hall, D. B. and Severini, T. A. (1998). Extended generalized estimating equations for clustered data. *Journal of the American Statistical Association* **93,** 1365–1375.

Hedeker, D. and Gibbons, D. R. (2006). *Longitudinal Data Analysis*. Wiley Series in Probability and Statistics.

Hoffman, E. B., Sen, P. K., and Weinberg, C. R. (2001). Within-cluster resampling. *Biometrika* **88,** 1121–1134.

Hogan, J. W. and Laird, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* **16,** 239–257.

Huang, Y. and Leroux, B. (2011). Informative Cluster Sizes for Subcluster-Level Covariates and Weighted Generalised Estimating equations. *Biometrics* **67,** 843–851.

Joffe, M. M., Ten Have, T. R., Feldman, H. I., and Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models: Review and new applications. *The American Statistician* **58,** 272–279.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Statistics in Medicine* **17,** 2723–2732.

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

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,** 624–632.

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

Liang, K.-Y., Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)* **54,** 3–40.

Liao, J. G. and Lipsitz, S. R. (2002). A type of restricted maximum likelihood estimator of variance components in generalised linear mixed models. *Biometrika* **89,** 401–409.

Lipsitz, S. R., Molenberghs, G., Fitzmaurice, G. M., and Ibrahim, J. (2000). GEE with Gaussian estimation of the correlations when data are incomplete. *Biometrics* **56,** 528–536.

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

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

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

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley Series in Probability and Statistics, New York.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley Series in Probability and Statistics, New York, second edition.

Mancl, L. A., Leroux, B. G., and DeRouen, T. A. (2000). Between-subject and within-subject statistical information in dental research. *Journal of Dental Research* **79,** 1778–1781.

McCullogh, P. and Nelder, J. A. (1989). *Generalized linear Models*. Chapman and Hall, second edition.

Molenberghs, G. and Verbeke, G. (2006). *Models for Discrete Longitudinal Data*. Springer.

Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* **54,** 638–645.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1991). A comparison of cluster-specific and population average approaches for analysing binary correlated data. *International Statistical Review* **59,** 25–36.

Neuhaus, J. M. and MCculloch, C. E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **68,** 859–872.

Neuhaus, J. M. and MCculloch, C. E. (2011). Estimation of covariate effects in generalized linear models with informative cluster sizes. *Biometrika* **98,** 147–162.

Nevalainen, J., Kenward, M. G., and Virtanen, S. M. (2009). Missing values in longitudinal dietary data: A multiple imputation approach based on a fully conditional specification. *Statistics in Medicine* **28,** 3657–3669.

Palta, M. and Yao, T. J. (1991). Analysis of longitudinal data with unmeasured covariates. *Biometrics* **47,** 1355–1369.

Panageas, K. S., Schrag, D., Localio, A. R., Venkatraman, E. S., and Begg, C. B. (2007). Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Statistics in Medicine* **26,** 2017–2035.

Park, T. and Shin, D.-Y. (1999). On the use of working correlation matrices in the gee approach for longitudinal data. *Communications in Statistics - Simulation and Computation* **28,** 1011–1029.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58,** 545–554.

Pavlou, M., Copas, A. J., and Seaman, R. S. (2009). Efficient weighted generalised estimating equations when the cluster size or covariate structure are informative. ISCB, Prague 2009. Conference presentation available at: `http://www.iscb2009.info/Text/presentations`.

Pepe, M. S. and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation* **23,** 1939–951.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44,** 1033–1048.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). Analysis of Failure Times in Presence of Competing Risks. *Biometrics* **34,** 541–554.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47,** 825–839.

Robins, J., 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.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11,** 550–560.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63,** 581–592.

Rubin, D. B. (1987). *Multiple imputation for non-response in surveys*. Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall.

Scott, A. J. and Hold, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of American Statistical Association* **77,** 848–854.

Seaman, S. R. and Copas, A. J. (2009). Doubly robust generalized estimating equations for longitudinal data. *Statistics in Medicine* **28,** 937–955.

Su, L., Tom, B. D. M., and T, F. V. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics* **10,** 374–389.

Ten Have, T. R., Landis, J. R., and Weaver, L. S. (1995). Subject-specific and population-averaged continuation ratio logit models for multiple discrete time survival profiles. *Statistics in Medicine* **14,** 413–429.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. Springer.

van Buuren, S. (2000). Multivariate imputation by chained equations: MICE V1.0 User's manual. Technical report.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* **16,** 219–242.

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

Vansteelandt, S., Rotnitzky, A., and Robins, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika* **94,** 841–860.

Verbeke, G. and Molenberghs, G. (2000). *Liner Mixed Models for Longitudinal Data*. Springer.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss–newton method. *Biometrika* **61,** 439–447.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* **30,** 377–399.

Williamson, J. M., Datta, S., and Satten, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59,** 36–42.

Williamson, J. M., Kim, H., and Warner, L. (2007). Weighting condom use data to account for nonignorable cluster size. *Annals of Epidemiology* **17,** 603–607.

Wu, M. C. and Carrol, R. J. (1988). Estimation And Comparison Of Changes In The Presence Of Informative Right Censoring By Modeling The Censoring Process. *Biometrics* **44,** 175–188.

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,** 243–258.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American Statistical Association* **86,** 79–86.

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

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44,** 1049–1060.