ANALYSIS OF CONTINUOUS LONGITUDINAL DATA WITH NON IGNORABLE DROPOUTS

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS
FOR THE MSC IN STATISTICS
IN THE FACULTY OF SCIENCE

By

SIZIBA LILLIAN

SUPERVISOR: MR C. CHIMEDZA

UNIVERSITY OF ZIMBABWE
FACULTY OF SCIENCE
DEPARTMENT OF STATISTICS
DECEMBER 2009

Table of Contents

List of Tables

List of Figures

Acknowledgments

My special thanks go to my supervisor Mr C. Chimedza for his unwavering and ungrudging encouragement, support and help through out all my work on this research. I also wish to express my heartfelt appreciation to Dr F. Matarise, Mr J. Mafodya and Mrs I.S. Nduna for the support and enouragement through out the whole period of study. Special mention goes to Mrs N.T. Chari and Ms M. Tekwa for the moral support. Without forgetting my colleagues Manjoro, Matangi, Maposa, Ruvharo and Mahachi, thank you guys for the support. Last but not least I would like to thank my son Kudzi for being so understanding during the whole period of study even if at times it meant deprivation of my motherly care to him.

Above all I give glory and honour to God Almighty for enabling me to reach this far.

Abstract

Missing responses are very common in longitudinal data. Much research has been going on, on ways to go around this complication in analysing such a data set. The approaches range from simple remedies like: analysing complete cases only, imputing the missing data, available case analysis and many others, to joint modelling of the measurement process and the missing mechanism. The work of Rubin(1976) on classifications of missing mechanisms contributed greatly to the development of researches on joint models, as missingness could now be classified as ignorable or non-ignorable. There are at least three joint modelling approaches, three common forms which are differentiated by the factorisation of the full data density are: selection models, pattern mixture models and shared parameter models.

Missing data in longitudinal studies can be classified into two main categories, which are missing intermittently: which is when a subject has a missing value for one occasion or more but will at a later stage during the study period have observed values, and dropouts: this is when we have a monotone missing pattern that is if we have a missing value at a particular point in time, there after the subject continues to have missing values until the completion of the study. The focus of this research is on the latter. Simulation of the missing pattern was done to produce informative dropouts. The major aim of this research was to compare estimates from different modelling approaches, with the main focus being comparing joint models to complete case basing on how their estimates compared to the complete data model estimates. The first part of this research focuses on linear mixed modelling of a complete longitudinal data set. The extensive modelling process starts from exploration of the data to estimation of parameters forms the baseline of the second part of the research which is the joint modelling process.

The E-M algorithm (Dempster *at al* (1977) formed the backbone of the likelihood estimation under these approaches and at times convergence would not be reached or would be slow, and the in such cases the modified forms like the stochastic E-M algorithm would be used.

The complete case estimates were very close to the complete data estimates. However it is difficult for this researcher to conclude that complete case analysis performs better than the other models. This researcher feels one could reach to a solid conclusion after considering different proportions of dropouts and also different patterns in which the dropouts are distributed throughout the study period.

Chapter 1

Introduction

Longitudinal data is observed over time as well as over space, that is, response is measured repeatedly on a set of units or subjects. Longitudinal designs have become the most convincing and most popular way of collecting change data. Cross sectional studies have one historic context while time series allows for multiple historical context but for only one spatial location (Plewis, 1985), it is to this effect that longitudinal data have an advantage over the earlier two. Longitudinal studies have the capacity to separate changes from baseline, over time within individuals from differences among individuals.

Longitudinal data can be classified into categories, with the two major categories being: panel data and time series cross-section data. The major differences between panel data and time series cross section data are that panel data have a large number of cross-sections with each unit observed only a few times where as time series cross sectional data has a reasonably sized number of observations and not very large number of cross sections. There are other types of longitudinal data like historical data, that is, data collected in retrospect and pseudo panel data (repeated cross sections). The problem of missing responses is common in longitudinal researches. Missing values arise whenever one or more of the sequence of measurements from units within the study are incomplete in the sense that the intended measurements are not taken, are lost or are otherwise unavailable. In the absence of missing values many multivariate statistical analysis would start by reducing the data matrix $\mathbf{Y}$, to the mean vector, $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{S} = [s_{jk}]$. But now, missing values result in unbalanced data and our challenge would be the estimation of $\boldsymbol{\mu}$ and $\mathbf{S}$ since some of the y_{ij} values are now missing. We also have to address the questions:

- Why the values are missing?
- Whether or not their being missing has any bearing on the practical questions being posed by the data?

Whenever, not all planned measurements are observed, a level of complexity is added to the analysis of longitudinal data. Several remedies to the analysis of incomplete data have been suggested. The most common approaches are (1) to discard subjects with incomplete sequences and (2) imputing missing values. The first approach has an advantage of simplicity but it is an inefficient use of information, since information on completers only is used, in a trivial sense it describes response conditional on a subject having all measurements taken. The principle of imputation is that observed values are used to impute values for missing observations. There are several ways of imputing values ranging from simple to complex. These methods use more data than the complete case analysis but they suffer from some drawbacks.

There are several approaches to estimation of parameters, which can be classified as likelihood based and non likelihood based approaches. Likelihood estimates are known to be asymptotically unbiased and sufficient (Diggle, 2002).

1.1 Motivation

A lot of research has been done on statistical methods and models for the analysis of longitudinal data, with much emphasis on models for balanced or complete data and in recent years there is a lot of focus on the area of joint modelling of the measurement process and the missingness mechanism (Molenberghs and Verbeke, 2005). Dropouts occur frequently in longitudinal data, and in most fields where the subjects are living organisms, dropouts in the form of deaths are usually of interest. It becomes of paramount importance to closely consider and also model the missing value mechanism when modeling the measurement process. Quick and easy remedies have commonly been used to deal with data which has dropouts, using such methods as, the complete case analysis, the available case analysis, imputing conditional or unconditional means. These methods have their advantages and disadvantages. But one may really want to know if the cause of dropout is related to or is saying something about the experiment and hence it becomes important to consider the dropping out as a stochastic process, which also needs to be modelled.

A subject's pattern of response in a study is likely to depend on many characteristics of that subject including some which are unobservable. Debates about relative merits of certain models over others should be mainly contextually based. An important consideration in formulating a model for a longitudinal data set with dropouts would be to think of what casual relationship might plausibly exist amongst the three stochastic processes namely the measurements, dropout times and the unobserved characteristics(random effects). It is upon this background that this research is made, and is geared to explore and compare different procedures for modeling longitudinal data with dropouts. Focus will be put on the likelihood approach.

1.2 Aims and Objectives

1.2.1 Aims

The major aim of this research is to produce a comparison of models for longitudinal data with non-ignorable dropouts obtained using likelihood approaches.

1.2.2 Objectives

The objectives of this research are to:

1. Produce models for longitudinal data with dropouts using Complete case analysis

2. Produce an integrated analysis of the measurement process and the dropout mechanism for a:
    - (i) Selection model
    - (ii) Pattern mixture model
    - (iii) Shared parameter model
3. Compare models obtained from integrated analysis and those obtained from complete case analysis.

1.3 Significance of the study

In practical situations it is difficult to justify a particular missing data mechanism, and it maybe hard to distinguish whether it is random or not. Unless missing data are a deliberate feature of the study design it is important to try to limit them during data collection, since any method for compensating for the missing data requires some unverifiable assumptions that maybe or may not be justified. However, since data are likely to be missing despite all these efforts, it is important to try and collect covariates that are predictive of the missing value so that an adequate adjustment can be made. In real situations assumptions of random missingness are in-testable and non-ignorable models offer a conservative approach. This research will contribute to on going researches on longitudinal data with dropouts. No one approach may be said to cover all forms in which the practical problems pose to researchers. It is therefore upon this background that this research is geared to produce a comparison of the possible modelling approaches to longitudinal data with non ignorable dropouts.

1.4 Organisation

Chapter one consists of the general introduction, motivation, aims and objectives, which is in actual fact the definition of work to be covered in this research. Chapter two covers a detailed literature review i.e, longitudinal data, missingness mechanisms, the quick solutions to modelling longitudinal data and likelihood based modelling. Chapter three gives an outline of the methods to be used in data analysis. Chapter four consists of analysis and the results. The last chapter, chapter five gives recommendations and conclusions. At the end of the document is the Appendix and Bibliography. The appendix covers the Data and Programs used in this research

Chapter 2

LITERATURE REVIEW

2.1 Longitudinal Data

Longitudinal data is observed over time as well as over space, and therefore a longitudinal design would produce a rectangular data matrix, say $\mathbf{Y} = [y_{ij}]$, where y_{ij} is the value of the variable y_j for unit i, $j = 1, 2, 3...n_i$ and $i = 1, 2, 3...m$, so y_{ij} denotes the j^{th} measurement of the i^{th} of the m units. Often, the primary objective of longitudinal data analysis is to describe the mean response as a function of time, treatment effects and possibly covariates attached to the units or individual measurement.

Likelihood based approaches lean strongly on the assumption of independent observations, but repeated observations on the same unit are seldom independent. Therefore, the assumption of independence should not be assumed but tested. Naturally there is heterogeneity due to unmeasured factors across subjects.

2.1.1 Notation

This section presents the basic notation to be followed in this study (adopted from Diggle (2002)). We let Y_{ij} represent a response variable and $\mathbf{X}_{ij}$ be a $p \times 1$ vector of explanatory variables observed at times t_{ij} for the observation $j = 1, 2, \ldots, n_i$ on subject $i = 1, 2, \ldots, m$. The mean of Y_{ij} is represented by $E(\mathbf{Y}_i) = \mu_{ij}$ and the variance of Y_{ij} is represented by $Var(Y_{ij}) = \upsilon_{ij}$. The set of repeated outcomes for subject i are collected into vector $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ with mean $E(\mathbf{Y}_i) = \boldsymbol{\mu}_i$ and the $n_i \times n_i$ covariance matrix $Var(\mathbf{Y}_i) = \boldsymbol{\nu}_i$, where the $(jk)^{th}$ element of $\boldsymbol{\nu}_i$ is the covariance between Y_{ij} and Y_{ik} denoted by $Cov(Y_{ij}, Y_{ik}) = \upsilon_{ijk}$. Responses for all units constitute a matrix $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m)'$.

The basic linear regression model becomes

$$Y_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \ldots + \beta_p x_{ijp} + \epsilon_{ij} \tag{2.1.1}$$

where x_{ijk} is k^{th} explanatory variable, with $k = 1, 2, 3, \ldots, p$. This can be expressed in matrix notation as

$$Y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \epsilon_{ij}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a p vector of unknown regression coefficients, and ϵ_{ij} is a zero mean random variable representing the deviation of responses from $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$.

2.1.2 Modelling Issues

Repeated observations on same unit are seldom independent, so the independence assumption should be tested and not assumed. Another important issue to consider when modelling longitudinal data is the fact that naturally there is heterogeneity due to unmeasured factors across subjects, (Diggle,2002).

Model Fitting Process

Model fitting aims at answering questions about the process which generated the data. A model contains a relatively small number of parameters whose values can be interpreted as answers to the scientific questions being posed by the data. The model fitting process can be divided into four stage, which are:

(i) *Formulation*: This stage involves choosing the general form of the model. The focus at this stage is the mean and the covariance structure. The time series plots of observed averages within treatment groups are simple and effective tools to model formulation in instances where data is well replicated and non parametric smoothing is helpful where there are few measurements at any one time. Time series plots, scatter plot matrices and variogram plots of residuals can be used to give an idea of the underlying covariance structure.

(ii) *Estimation*: The aim at this stage is to attach numerical values to the parameters in the model whose general form is

$$\mathbf{Y}_i \sim \mathbf{N}_p(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2\mathbf{V}_i(\boldsymbol{\phi})) \tag{2.1.2}$$

where $\mathbf{V}(\boldsymbol{\phi})$ is the covariance structure

(iii) *Inference*: The aim of this stage is to make inference about the parameter space $\boldsymbol{\beta}$.

(iv) *Diagnostic checking*: This stage aims at comparing the observed data with the estimated data to highlight any discrepancies.

Types of Models

With repeated measurements we often strategise to reduce the repeated values into one or two summaries and then analyse each summary variable as a function of covariates $\mathbf{X}_i$. The vector of repeated measurements for each subject is summarised by a vector of a relatively small number of estimated subject specific regression coefficients, then multivariate regression techniques can be used to relate these estimates to known covariates and compare the estimates across groups. This is the so called two-stage or derived variable analysis. The three main approaches in applying the two stage analysis are:

(a) *Marginal Models*: The main feature of the marginal models is that the regression of the response on explanatory variables is done separately from the within subject correlation. The marginal expectation $E(Y_{ij})$ is modelled as a function of explanatory variables, where the marginal expectation means the average response over a subpopulation that share a common value of $\mathbf{X}$, i.e. $E(Y_{ij}) = \mu_{ij}$ depends on explanatory variables X_{ij} through the function $\eta(\mu_{ij}) = \mathbf{X}'_{ij}\boldsymbol{\beta}$ where η is a known link function, such as the identity link for Gaussian or the logit link for binary responses and the log link for counts. The marginal variance depends on the marginal mean according to $Var(Y_{ij}) = v(\mu_{ij})\phi$ where $v(.)$ is a known variance function which needs to be estimated.

Repeated values are not likely to be independent, therefore the assumptions about the form of correlations are to be included. The correlation of Y_{ij} and Y_{ik} is usually a function of marginal means and perhaps of additional parameters $\boldsymbol{\alpha}$, that is, $Corr(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}, \boldsymbol{\alpha})$. Marginal models are appropriate if the population average is the focus.

(b) *Random effects models*: There is natural heterogeneity across subjects due to unmeasured factors. This heterogeneity is reflected in their regression coefficients which vary from one individual to the next. This heterogeneity can be represented by a probability distribution. Another factor is that observations from the same subject are seldom independent, they share unobservable variables say $\mathbf{U}_i$. For Gaussian data the linear random effects model does the job and the basic ideas extend to regression models for discrete and non-Gaussian continuous responses. The random effects Generalised linear model assumes that data for a subject are independent observations following a Generalised linear model (GLM) but that the regression coefficients vary from subject to subject according to a distribution $\mathbf{F}$. The general specification of a random effects GLM is as follows: Given $\mathbf{U}_i$, responses $Y_{i1}, \ldots, Y_{in_i}$ are mutually independent and follow a GLM (also known as the exponential family of distributions) with density

$$f(y_{ij}|\mathbf{u}_i) = exp[\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi)} + c(y_{ij}, \phi)], \qquad (2.1.3)$$

the conditional moments for the canonical exponential distribution in 2.1.3 are;

$$\mu_{ij} = E(y_{ij}|\mathbf{u}_i) = b'(\theta_{ij})$$

and

$$v_{ij} = Var(y_{ij}|\mathbf{u}_i) = b''(\theta_{ij})/a(\phi)$$

satisfying $h(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}^* + \mathbf{d}'_{ij}\mathbf{u}_i$, and $v_{ij} = \upsilon(\mu_{ij})\phi$ where h and υ are link and variance functions respectively, $\mathbf{d}_{ij}$ is a subset of $\mathbf{x}_{ij}$. β^* represents the effects of the explanatory variables on individuals' response. Random effects models are appropriate when the aim is to make inference about subjects rather than the population.

(c) *Transition models*: Since repeated observations on a subject are seldom independent, correlation among $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ exists because the past values,

$Y_{i1}, Y_{i2}, \ldots, Y_{ij-1}$, influence the present observation Y_{ij}. The past observations are treated as additional predictor variables. In the general transition model, the conditional distribution of Y_{ij} given the past is modelled as an explicit function of the preceding responses. The general transition model can be specified as follows: let $H_{ij} = \{Y_{i1}, Y_{i2}, \ldots, Y_{ij-1}\}$ represent past responses for the i^{th} subject, also let $\mu_{ij}^c{=}E(Y_{ij}|H_{ij})$ and $v_{ij}^c{=}Var(Y_{ij}|H_{ij})$ be the conditional mean and variance of Y_{ij} given past responses and the explanatory variables assuming $h(\mu_{ij}^c) = \mathbf{x}_{ij}'\boldsymbol{\beta}^{**} + \sum_{r=1}^{s} f_r(H_{ij}, \boldsymbol{\alpha})$, and that $v_{ij}^c = v(\mu_{ij}^c)\phi$ where h and v are link and variance functions respectively while $\boldsymbol{\beta}^{**}$ represents change per unit change in $\mathbf{x}$.

2.2 The General Linear Mixed Effect Model

The general linear mixed effect model can viewed as a combination of models from a two stage analysis where: The first stage assumes that $\mathbf{Y}_i$ satisfies a linear regression model

$$\boldsymbol{Y}_i = \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \qquad (2.2.4)$$

where $\mathbf{Z}_i$ is an appropriate design matrix. This model shows how the response evolves over time for the i^{th} subject where $\boldsymbol{\beta}_i$ is a $q - dimensional$ vector of unknown subject specific regression coefficients and $\boldsymbol{\epsilon}_i$ is a vector of the residual components $\epsilon_{ij}, j = 1, 2, 3, \cdots, n_i$, usually assumed to be normally distributed with mean zero and covariance matrix $\mathbf{V}_i$. The model is completed by specifying the covariance structure, which can be homogeneous or heterogeneous.

Commonly used homogeneous covariance structures are;

- When $\mathbf{V}_i = \sigma^2 \mathbf{I}_{n_i}$ for $\mathbf{I}_{n_i}$ denoting the identity matrix of dimension n_i. This is so under the strong assumption that all repeated measurements are independent though repeated measurements within the same subject are seldom independent

- The first order autoregressive model which assumes that the covariance between two measurements Y_{ij} and Y_{ik} from the same subject i is of the form $\sigma^2 \rho^{|t_{ij} - t_{ik}|}$ for unknown parameters σ^2 and ρ.

- Compound symmetry which assumes that the covariance is of the form $\sigma^2 + \gamma \delta_{ij}^2$ for unknown parameters σ^2 and $\gamma > -\sigma^2$, and where δ_{jk} equals 1 for $j = k$ and zero otherwise.

The second stage is a multivariate regression model of the form

$$\boldsymbol{\beta}_i = \mathbf{K}_i \boldsymbol{\beta} + \mathbf{b}_i \tag{2.2.5}$$

which models variability between the subjects with respect to their subject specific regression coefficients, $\boldsymbol{\beta}_i$, $\mathbf{K}_i$ is a $(q \times p)$ matrix of covariates, $\mathbf{b}_i' s$ are assumed to be independent following a $q - \mathit{dimensional}$ normal distribution with mean zero and general covariance structure $\mathbf{D}$.

Substituting for equation 2.2.5 in 2.2.4 we get the general linear mixed model as

$$\mathbf{Y}_i = \mathbf{Z}_i\mathbf{K}_i\boldsymbol{\beta}_i + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \tag{2.2.6}$$

which simplifies to

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \tag{2.2.7}$$

where $\mathbf{X}_i = \mathbf{Z}_i\mathbf{K}_i$ is a $n_i \times p$ matrix of known covariates. Model 2.2.7 has linear mixed effects: with fixed effects $\boldsymbol{\beta}$ which are population specific, that is the same for all subjects and random effects b_i which are subject specific. The $\mathbf{b}_i's$ are assumed to be random because subjects are selected randomly from a population.

It follows that $\mathbf{Y}_i$ conditional on random effects $\mathbf{b}_i$ is normally distributed with mean $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i$ with covariance matrix $\boldsymbol{\Sigma}_i$, and $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$. The marginal density of $\mathbf{Y}_i$ is therefore given by

$$f(\mathbf{y}_i) = \int f(\mathbf{y}_i|\mathbf{b}_i)\, f(\mathbf{b}_i)\, d\mathbf{b}_i \tag{2.2.8}$$

where $f(\mathbf{y}_i|\mathbf{b}_i)$ is the conditional density of $\mathbf{Y}_i$ given $\mathbf{b}_i$ and $f(\mathbf{b}_i)$ is the density of $\mathbf{b}_i$. We have $f(\mathbf{y}_i)$ as a density of an n_i dimensional normal distribution with mean vector $\mathbf{X}_i\boldsymbol{\beta}$ and with covariance matrix $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \boldsymbol{\Sigma}_i$. The mean structure depends on covariates $\mathbf{X}_i$ and the covariance structure depends on $\mathbf{Z}_i$.

2.3 Incomplete Data

Whenever there is a situation that not all planned measurements are observed, a level of complexity is added to the analysis of longitudinal data. Questions like, why the values are missing and whether missingness has any bearing on the practical question being posed by the data, have to be addressed.

According to Diggle (2002) missing values can be classified as whether missing intermittently or as dropouts. Missing values occur as dropouts when observations on a subject are obtained until a certain point in time, after which all measurements are missing, that is, if we intend to take a sequence of measurements on a particular unit, and whenever y_j is missing so are y_k for all $k > j$, otherwise missing values are said to be intermittent.

Intermittent missing values can arise through a known censoring mechanism or the reason for their being missing is often known since the subject in question remains in the study. In some cases it will be reasonable to assume that the missingness is unrelated to the measurement process. When such is the case analysis of the data can be done by any method which accommodates unbalanced data.

Dropouts are often lost to any form of follow up and there is a possibility that dropouts arise for reasons directly or indirectly related to the measurement process. When there is any kind of relationship between the measurement process and the dropout process, the interpretation trends in mean response over time can be problematic.

Also, resulting from dropouts, subjects are lost from the study, which results in a decreasing sample size, this increases variability which in turn decreases precision.

2.4 Simple Missing Data Methods

With missing values in longitudinal data, inference will often be invalidated when the observed measurements do not constitute a simple random subset of the complete set of measurements. Also often standard software work with complete arrays of data. Often quick and simple ways are found round the problem of modelling data with missing values. Two most common fixes will be reviewed below.

2.4.1 Complete Case Analysis

A complete case analysis includes only those cases into the analysis for which all n_i measurements were recorded, that is discard all incomplete sequences. The advantages of this methods are that it is very simple to describe and since the data structure would be complete arrays, standard statistical software can be used. The method suffers drawbacks, among others it is obviously wasteful of data especially if the dropout process is unrelated to the measurement process. The method performs differently under different missing mechanism and therefore a partial check on missing mechanism assumptions can be made.

2.4.2 Imputation

Imputation is an alternative way to obtain a complete data set instead of discarding subjects with incomplete sequences. The principle is that observed values are used to impute values for missing observations. There are several ways of imputing ranging from simple to some complex. Some commonly used methods of imputing values include:

1. *Last observation carried forward*: This procedure uses information on the subject to impute the values for missing observations of that particular subject. The method consists of extrapolating the last observed measurement for the subject in question to the remainder of their intended time sequence. Very strong and unrealistic assumptions have to be made to ensure validity of this method, like that the subject's measurement stays at the same level from the moment of dropout onward.

   The method overestimates precision by treating imputed and actually observed values on equal footing.

2. *Unconditional mean imputations*: This is termed unconditional because information is borrowed from other subjects to impute a value for a missing observation. One does not use information on the subject for which an imputation is generated. The missing value is replaced by the average of the observed values on the same variable over the other subjects. Other methods of imputation

which are a bit complex like Buck's method and multiple imputation can be used to impute values.

3. *Available case method*: The method uses all available values to the j^{th} variable disregarding their response status at the other measurement occasions. This method uses more information than the complete case.

2.5 Modelling the Dropout Process

In order to incorporate incompleteness into the modelling process we need to reflect on the nature of the missing value mechanism and its implications for statistical inference.

Little and Rubin (1987) made important distinctions between different missing value processes. Let $\mathbf{Y}_i$ denote the complete set of measurements for the i^{th} subject which would have been obtained were there no missing values that is, we assume that for each subject i in the study a sequence of measurements Y_{ij} is designed to be measured at occasions $j = 1, \ldots, n_i$ and the outcomes are grouped into a vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$.

For each occasion j define

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed} \\ 0 & \text{Otherwise} \end{cases}$$

The missing data indicators R_{ij} are grouped into a vector $\mathbf{R}_i$ which is of the same length as $\mathbf{Y}_i$. Now $(\mathbf{Y}_i, \mathbf{R}_i)$ is the full data that is, the complete data together with the missingness indicators. $\mathbf{Y}_i$ can be partitioned into $\mathbf{Y}_i = (\mathbf{Y}_i^o, \mathbf{Y}_i^m)$ where $\mathbf{Y}_i^o$ denotes measurements actually observed,that is the vector containing those Y_{ij} for which $R_{ij} = 1$ and $\mathbf{Y}_i^m$ denotes those measurements which would have been available had they not been missing. $\mathbf{Y}_i^o$ and $\mathbf{Y}_i^m$ can be referred to as the observed and missing components respectively. The process generating $\mathbf{R}_i$ is referred to as the missing data process. To note is the fact that unless all components of $\mathbf{R}_i$ equal 1, the data components are never jointly observed. One observes the measurements of $\mathbf{Y}_i^o$ together with the missingness indicators $\mathbf{R}_i$.

We can partition $\boldsymbol{\mu}_i, \mathbf{V}_i$ and $\mathbf{x}_i$ in a similar manner to obtain $(\boldsymbol{\mu}_i^o, \boldsymbol{\mu}_i^m)$, $(\mathbf{V}_i^o, \mathbf{V}_i^m)$ and $(\mathbf{x}_i^o, \mathbf{x}_i^m)$ respectively.

Lets consider the full data density $f(\mathbf{y}_i, \mathbf{r}_i | \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}, \boldsymbol{\varphi})$ where $\mathbf{X}_i$ is the design matrix for fixed effects and $\mathbf{Z}_i$ is the design matrix for random effects. $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are vectors that parameterise the joint distribution where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\phi}')$ (fixed and covariance parameters) describes the measurement process and $\boldsymbol{\varphi}$ describes the missingness process. In the case of dropouts: Adopting the notation that for any subject the complete set of observed measurements is $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{in_i})$. Let the scalar D_i be the dropout indicator obeying the relationship $D_i = 1 + \sum_{j=1}^{n_i} R_{ij}$, where R_{ij} is as defined earlier on. D_i denotes the dropout time and $2 \leq D_i \leq n_i + 1$ with $D = n + 1$ indicating that the subject in question has not dropped out. Further more, we introduce another indicator T_i, indicating the number of measurements taken for subject i, where

$$T_i = \sum_{j=1}^{n_i} R_{ij} = D_i - 1.$$

Rubin and Little (1987) classify missing value mechanisms as

1. *Completely Random Dropout (CRD)* if D_i is independent of both $\mathbf{Y}_i^o$ and $\mathbf{Y}_i^m$. The dropout and the measurement process are independent. Consider for example, the joint density of d_i given $\mathbf{Y}_i$ and $\boldsymbol{\varphi}$, $f(d_i|\mathbf{y}_i, \boldsymbol{\varphi})$, with φ being the vector of the non response model then if CRD

$$f(d_i|\mathbf{y}_i, \boldsymbol{\varphi}) = f(\mathbf{d}_i|\boldsymbol{\varphi}).$$

2. *Random Dropout (RD)* if d_i is independent of $\mathbf{Y}_i^m$. The probability of non response depends on the observed response $\mathbf{Y}_i^o$ but not on the missing values $\mathbf{Y}_i^m$ and

$$f(d_i|\mathbf{y}_i, \varphi) = f(d_i|\mathbf{y}_i^o, \boldsymbol{\varphi})$$

3. *Informative Dropout (ID)* if d_i is dependent on $\mathbf{Y}_i^m$. The dropout depends on unobserved measurements, i.e. those that would have been observed if the the unit had not dropped out and

$$f(d_i|\mathbf{y}_i, \varphi) = f(d_i|\mathbf{y}_i^m, \mathbf{y}_i^o, \boldsymbol{\varphi})$$

We can afford to ignore the first two dropout patterns since the dropout process either depends on the observed values only or is completely independent of the measurement process. Hence CRD and RD constitute the ignorable class of dropouts. However for ID, the dropout process depends on the unobserved values of the measurement process, meaning that what ever the measurement value would have been, had the subject not dropped out, the value has a causal effect to the subject's dropout, hence the the dropout process is non-ignorable.

2.6 Joint Modelling of Measurements and Missingness

Likelihood-based and non-likelihood-based approaches can be used to model data with non responses. In the case of maximum likelihood, there exist three approaches based on the factorisation of the full data density or equivalently the likelihood function. These are:

- Selection models, which are a result of outcome dependent factorisation of the joint or full data density. Missingness indicators are conditioned on the values of the measurement process.

- Pattern mixture models, which are a result of pattern dependent factorisation. The distribution of the measurement process is a mixture of distributions for subjects within distinct subgroups determined by patterns of missingness.

- Shared parameter models (random effects models), which are a result of parameter dependent factorisation. The measurement process and the missingness indicators are conditional independent given a group of parameters shared by the two parties.

The following graphical illustration from Diggle (2002) gives a kind of thought experiment that the data analyst must conduct in deciding how to deal with the dropouts. The three stochastic processes in question are **Y**, the measurement process, **D**, the drop out process and **U**, the unobserved characteristics or random effects.

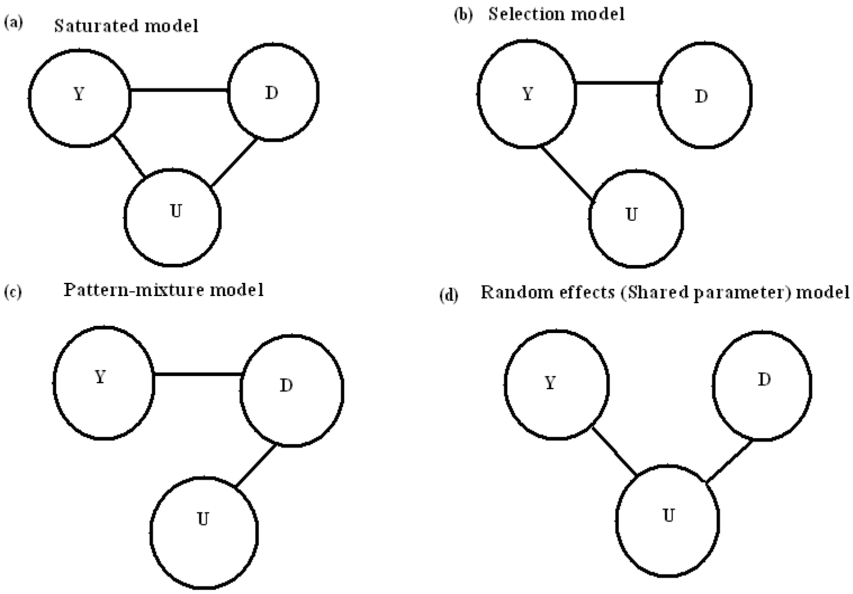
(a) Saturated model
Y
D
U
(b) Selection model
Y
D
U
(c) Pattern-mixture model
Y
D
U
(d) Random effects (Shared parameter) model
Y
D
U

Figure 2.1: Graphical representation of the models

In the diagram, (a) represents a denial that any simplifying assumptions are possible. Under this circumstance one would be more or less compelled to express a model for the data as a collection of joint distributions of **Y** and **U** conditional on each of the possible values of **D**. Diagram (b) can be interpreted as a situation in which the random effects or subject specific characteristics, **U**, influence the properties of the measurement process, **Y**, for the subject in question, with the propensity to drop out subsequently determined by the realisation of the measurement process. Diagram (c) can be interpreted as a situation where the subject-specific characteristics initially determine the propensity to drop out, with a consequential variation in the measurement process between different, predestined dropout cohorts (patterns). Diagram (d) suggests that the measurement and the dropout processes are a joint response to subject-specific characteristics, which could be thought of as under identified explanatory variables. The natural parameters of the three models have different meanings.

2.7 Likelihood Based Estimation

Let the data denoted by Y, where Y may be a scalar, vector or matrix. For longitudinal data Y would be a matrix. The data are assumed to be generated by a model described by a probability function $f(Y|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the vector of parameters. The following definitions and terminology shall be used:

(i) $f(Y|\boldsymbol{\theta})$ is a function that gives probabilities or densities of various Y value sets for fixed $\boldsymbol{\theta}$ values.

(ii) The likelihood function $L(\boldsymbol{\theta}|Y)$ is any function proportional to $f(Y|\boldsymbol{\theta})$. $L(\boldsymbol{\theta}|Y)$ is a function of the parameter $\boldsymbol{\theta}$ for fixed Y.

(iii) The log likelihood function $\ell(\boldsymbol{\theta}|Y)$ is the natural logarithm of the log likelihood function $L(\boldsymbol{\theta}|Y)$.

2.7.1 Maximum Likelihood Estimation

The idea behind maximum likelihood parameter estimation is to determine the parameters that maximise the probability (likelihood) of the sample data. Consider the full data density $f(\mathbf{y}, \mathbf{r}; \boldsymbol{\theta}, \boldsymbol{\varphi})$ where $\boldsymbol{\theta}$ and φ are unknown parameter spaces which need to be estimated. The likelihood function,

$$L(\theta, \varphi|\mathbf{y}, \mathbf{r}) = \prod f(\mathbf{y}_i, \mathbf{r}_i|\boldsymbol{\theta}, \boldsymbol{\varphi}). \tag{2.7.9}$$

Now since we are dealing with, incomplete data $\mathbf{y} = (\mathbf{y}^o, \mathbf{y}^m)$ where $\mathbf{y}^o$ denotes the observed data and $\mathbf{y}^m$ denotes the observed data. Let

$$f(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y}^o, \mathbf{y}^m|\boldsymbol{\theta}) \tag{2.7.10}$$

denote the joint density of $\mathbf{y}^o$ and $\mathbf{y}^m$. The marginal pdf of $\mathbf{y}^o$ can be found by integrating out $\mathbf{y}^m$ from the joint density, that is,

$$f(\mathbf{y}^o|\boldsymbol{\theta}) = \int f(\mathbf{y}^o, \mathbf{y}^m|\boldsymbol{\theta})\, d\mathbf{y}^m. \tag{2.7.11}$$

The logarithm likelihood function is given by

$$\ell = lnL = \sum lnf(\mathbf{y}_i, \mathbf{r}_i|\boldsymbol{\theta}, \boldsymbol{\varphi}). \tag{2.7.12}$$

If the likelihood is differentiable, the Maximum Likelihood estimators of $\boldsymbol{\theta}$ are obtained by differentiating the likelihood or the log likelihood with respect to $\boldsymbol{\theta}$, setting the result equal to zero and solving for $\boldsymbol{\theta}$.

$$S(\boldsymbol{\theta}|\mathbf{y}) = \frac{\partial \ell}{\partial \boldsymbol{\theta}_j} = 0 \tag{2.7.13}$$

is the likelihood equation. $S(\boldsymbol{\theta}|\mathbf{y})$ is called the score function. $I(\boldsymbol{\theta}|\mathbf{y})$ is the observed information where $I(\boldsymbol{\theta}|\mathbf{y}) = -\frac{\partial^2 \ell}{(\partial \boldsymbol{\theta})^2}$. The covariance matrix is the inverse of the observed information evaluated at $\hat{\theta}$: $C = I^{-1}(\hat{\boldsymbol{\theta}}|Y)$.

The likelihood of $\boldsymbol{\theta}$ based on $\mathbf{Y}^o$, ignoring the missing data mechanism is a function of $\boldsymbol{\theta}$ proportional to $f(\mathbf{Y}^o|\boldsymbol{\theta})$ implying that inference about $\boldsymbol{\theta}$ can be based on this likelihood, $L(\theta|\mathbf{Y}^o)$.

Consider $\mathbf{Y}$ and the missing data indicator R, one way of factorisation of their joint distribution is expressing it as a product of the densities of $\mathbf{Y}$ and the conditional distribution of $\mathbf{R}$ given $\mathbf{Y}$, that is,

$$f(Y, R|\boldsymbol{\theta}, \boldsymbol{\varphi}) = f(Y|\boldsymbol{\theta})\, f(R|Y, \boldsymbol{\varphi}) \tag{2.7.14}$$

where $\boldsymbol{\varphi}$ is unknown parameter space.

The actual observed data consists of the values of the variables$(\mathbf{Y}^o, R)$. Their joint distribution is obtained by integrating Y^m out of the joint density of $Y = (Y^o, Y^m)$ and R i.e.,

$$f(Y^o, R|\boldsymbol{\theta}, \boldsymbol{\varphi}) = \int f(Y^o, Y^m|\boldsymbol{\theta})\, f(R|Y^o, Y^m, \boldsymbol{\varphi})\, dY^m. \tag{2.7.15}$$

The likelihood function of $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$, $L(\boldsymbol{\theta}, \boldsymbol{\varphi}|Y^o, R) \propto f(Y^o, R|\boldsymbol{\theta}, \boldsymbol{\varphi})$.

From Rubin's work, if the distribution of the missing data mechanism does not depend on the missing values Y^m then the factor $f(R|Y^o, Y^m, \boldsymbol{\varphi})$ in equation 2.7.14 simplifies to $f(R|Y^o, \boldsymbol{\varphi})$ then

$$f(Y^o, R, \boldsymbol{\theta}, \boldsymbol{\varphi}) = f(R|Y^o, \boldsymbol{\varphi}) \int f(Y^o, Y^m, \boldsymbol{\theta})\, dY^m = f(R|Y^o, \boldsymbol{\varphi}) f(Y^o|\boldsymbol{\theta}) \tag{2.7.16}$$

If $f(R|Y^o, Y^m, \boldsymbol{\varphi}) = f(R|Y^o, \boldsymbol{\varphi})$ then the probability that a particular component of Y is missing cannot depend on the value of the component when it is missing.

When a closed form solution of equation 2.7.15 above can not be found, iterative methods can be applied. Such methods like the New-Raphson algorithm and the Expectation-Maximisation (EM) algorithm can be used as an alternative to direct maximisation. Let $\theta^{(0)}$ be the initial estimate of $\boldsymbol{\theta}$, and let $\boldsymbol{\theta}^{(t)}$ be the t^{th} iteration. The Newton-Raphson algorithm is given by the equation

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + I^{-1}(\boldsymbol{\theta}^{(t)}|Y^o)\, S(\boldsymbol{\theta}^{(t)}|Y^o) \tag{2.7.17}$$

where $I(\boldsymbol{\theta}|Y^o)$ is the observed information and $I(\boldsymbol{\theta}|Y^o) = -\frac{\partial^2 \ell(\boldsymbol{\theta}|Y^o)}{(\partial \boldsymbol{\theta})^2}$. Alternatively the method of Scoring can be used where the observed information in equation 2.7.17

is replaced by the expected information.

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + J^{-1}(\boldsymbol{\theta}^{(t)}|Y^o)\, S(\boldsymbol{\theta}^{(t)}|Y^o) \tag{2.7.18}$$

where $J(\boldsymbol{\theta}) = -E\left\{I(\boldsymbol{\theta}|Y^o)|\boldsymbol{\theta}\right\} = \int \frac{\partial^2 \ell(\boldsymbol{\theta}|Y^o)}{(\partial \boldsymbol{\theta})^2} f(Y^o|\boldsymbol{\theta})\, dY^o$. Both these methods involve calculating the matrix of second derivatives of the log likelihood. The EM algorithm does not require second derivatives.

2.8 Ignorability of Missingness

Since inference has to be based on what was obtained we focus on outcome dependent missingness and according Rubin (1976) missing values are ignorable when r_i is independent of $\mathbf{y}_i^m$, given $\mathbf{y}_i^o$ and $\mathbf{X}_i$ and also when $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are distinct. Under this ignorability the log-likelihood function for $\boldsymbol{\theta}$ can be separated from the log-likelihood for $\boldsymbol{\varphi}$.

$$\ell(\boldsymbol{\theta}, \boldsymbol{\varphi}|\mathbf{y}_i^o, \mathbf{r}_i) = \ell(\boldsymbol{\theta}|\mathbf{y}_i^o) + \ell(\boldsymbol{\varphi}|\mathbf{y}_i^o, \mathbf{r}_i)$$

where $\ell = lnL$ and

$$L = L(\boldsymbol{\theta}, \boldsymbol{\varphi}|\mathbf{y}_i^o, \mathbf{r}_i) \propto \Pi \int f(\mathbf{y}_i, \mathbf{r}_i|\boldsymbol{\theta}, \boldsymbol{\varphi}) d\mathbf{y}_i^o.$$

Under this condition ignorability is equivalent to the union of RD and CRD hence non ignorability becomes a synonym for ID in this context. So under ignorability the CRD and RD provide the same fitted measurement but however, as discussed by Verbeke and Molenberghs (1997) this does not imply that inferences under CRD and RD are equivalent. Yang (2006) defines ignorability as a condition under which observed data can be used to estimate $\boldsymbol{\theta}$ without bias.

2.8.1 Expectation-Maximisation (EM) Algorithm

When dealing with likelihood based estimation the process of maximisation becomes a challenge when the data is incomplete. Often patterns of incomplete data do not have particular forms that allow explicit maximum likelihood estimates to be calculated by exploiting factorisations of the likelihood.

Special forms of the algorithm have been proposed about half a century ago but the first unifying and formal account was given by Dempster, Laird and Rubin(1977). The EM algorithm consists of an Expectation (E step) and a Maximisation (M step).

The E Step

Given the current value $\theta^{(t)}$ of the parameter vector, the E step computes the expected value of the complete data log likelihood, given the observed data and the current parameters to give what is called the objective function. If we consider the complete data set $Y = (Y^o, Y^m)$, the joint likelihood density function is :

$$f(\mathbf{y};\boldsymbol{\theta}) = f(y^o;\boldsymbol{\theta})\, f(y^m|y^o;\boldsymbol{\theta})).$$

So the likelihood function is

$$logL(\boldsymbol{\theta};y) = logL(\boldsymbol{\theta};y^o) + logL(\boldsymbol{\theta};y^m).$$

First by using initial values $\boldsymbol{\theta}^{(0)}$ of the parameter vector we can compute the objective function for ignorable data at the t^{th} iteration:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int \ell(\boldsymbol{\theta}|Y) f(Y^m|Y^o, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)})\, dY^m \tag{2.8.19}$$

which simplifies to

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E\left\{\ell(\boldsymbol{\theta}|Y)|Y^o, \boldsymbol{\theta}^{(t)}\right\}. \tag{2.8.20}$$

For non ignorable models we find the initial estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$, $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\varphi}^{(0)}$ respectively. At the iteration t, given the current estimates ($\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\varphi}^{(t)}$) of $(\boldsymbol{\theta}, \boldsymbol{\varphi})$ the E step calculates

$$Q(\boldsymbol{\theta}, \boldsymbol{\varphi}|\boldsymbol{\theta}^{(t)}, \boldsymbol{\varphi}^{(t)}) = \int \ell(\boldsymbol{\theta}, \boldsymbol{\varphi}|Y^o, Y^m, R)\, f(Y^m|Y^o, R, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}, \boldsymbol{\varphi} = \boldsymbol{\varphi}^{(t)})\, dY^m \tag{2.8.21}$$

The M Step

The M step determines the parameter vector that maximises the respective objective function, $\boldsymbol{\theta}^{(t+1)}$ and $(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\varphi}^{(t+1)})$ for ignorable and non-ignorable missing data respectively. Formally $\boldsymbol{\theta}^{(t+1)}$ satisfies

$$Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \tag{2.8.22}$$

for all $\boldsymbol{\theta}$, and $(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\varphi}^{(t+1)})$ satisfies

$$Q(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\varphi}^{(t+1)}|\boldsymbol{\theta}^{(t)}, \boldsymbol{\varphi}^{(t)}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\varphi}|\boldsymbol{\theta}^{(t)}, \boldsymbol{\varphi}^{(t)}), \tag{2.8.23}$$

for all $(\boldsymbol{\theta}, \boldsymbol{\varphi})$. One then makes iterations of the E and M steps until convergence. The E step entails calculating the conditional expectation, this maybe infeasible in many situations. Celuex and Diebolt (1985) provided a possible alternative in which the expectations are calculated via simulations. In their stochastic EM algorithm, at each iteration the missing data is imputed with a single draw from the conditional distribution of the missing data given the observed and the current parameter estimates. This imputation of missing values is based on all our current information about θ and hence provides us with a plausible pseudo-complete data. Once we have a pseudo-complete data we can directly maximise its log likelihood to obtain updated estimates. The process is iterated for a sufficient number of iterations.

2.8.2 Theory of the EM Algorithm

Assuming ignorability of missingness the distribution of the complete data Y can be factored as $f(Y|\boldsymbol{\theta}) = f(Y^o, Y^m|\boldsymbol{\theta}) = f(Y^o|\boldsymbol{\theta})f(Y^m|Y^o, \boldsymbol{\theta})$ where $f(Y^o|\boldsymbol{\theta})$ is the density of of the observed data Y^o. $f(Y^m|Y^o, \boldsymbol{\theta})$ is the density of the missing data given the observed data. The corresponding decomposition of the likelihood becomes

$$\ell(\boldsymbol{\theta}|Y) = \ell(\boldsymbol{\theta}|Y^o, Y^m) = \ell(\boldsymbol{\theta}|Y^o) + lnf(Y^m|Y^o, \boldsymbol{\theta}). \tag{2.8.24}$$

We wish to estimate $\boldsymbol{\theta}$ by maximising the incomplete data likelihood $\ell(\boldsymbol{\theta}|Y^o)$ with respect to $\boldsymbol{\theta}$ for fixed Y^o

$$\ell(\theta|Y^o) = \ell(\boldsymbol{\theta}|Y) - lnf(Y^m|Y^o, \boldsymbol{\theta}) \tag{2.8.25}$$

where $lnf(Y^m|Y^o,\boldsymbol{\theta})$ is the missing part of the complete data log-likelihood. The expectation of equation 2.7.25 above over the distribution of the missing data Y^m, given the observed data Y^o and a current estimate of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(t)}$ is

$$\ell(\boldsymbol{\theta}|Y^o) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \tag{2.8.26}$$

where $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int \ell(\boldsymbol{\theta}|Y) f(Y^m|Y^o, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)})\, dY^m$ and

$$H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int lnf(Y^m|Y^o,\boldsymbol{\theta})\, f(Y^m|Y^o,\boldsymbol{\theta}^{(t)})\, dY^m \tag{2.8.27}$$

It can be shown by Jensen's inequality that $H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)} \leq H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$. The difference in values of $\ell(\boldsymbol{\theta}|Y^o)$ at successive iterations is given by

$$\ell(\boldsymbol{\theta}^{(t+1)}|Y^o) - \ell(\boldsymbol{\theta}^{(t)}|Y^o) = \left[Q(\boldsymbol{\theta}^{(t+1)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})\right] - \left[H(\boldsymbol{\theta}^{(t+1)}) - H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})\right]. \tag{2.8.28}$$

The EM algorithm chooses $\boldsymbol{\theta}^{(t+1)}$ to maximise $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$with respect to $\boldsymbol{\theta}$.

2.8.3 The Stochastic E-M Algorithm

The E-M algorithm is a common approach for parameter estimates in complete data sets, however calculating the conditional expectation required in the E-step maybe infeasible in many situations. Celuex and Diebolt (1985) brought the general idea of a stochastic E-M algorithm where the expectations are performed via simulations. The idea behind the stochastic E-M algorithm is that at each iteration the missing data is imputed with a single draw from the conditional distribution of the missing data given the observed and the current parameter estimates. So the imputation is based on all current information about $\boldsymbol{\theta}$, and hence provides a pseudo-complete data set which can be therefore maximised to obtain current estimates. The main advantage of the stochastic E-M over the E-M is that the earlier avoids the evaluations of the integrations, in the E step of the E-M algorithm. The stochastic E-M algorithm has the S-step and the M-step which are developed as follows:

- **S-step**: The first missing value is simulated from the conditional distribution $f(Y_i^m|Y_i^o, D_i, \boldsymbol{\theta}^{(t)}, \varphi(t))$. This distribution does not have a plausible form and hence it is not possible to use direct simulation. to overcome this problem the following procedure is used
  - A candidate value y^* is generated from the conditional distribution function $f(Y_i^m|Y_i^o, \boldsymbol{\theta}^{(t)})$ which is normal distribution.
  - Calculate the probability of dropout for Y^* according to the dropout model, where the parameters φ are fixed at current values. Lets denote this probability P_i^*
  - A random variate U is simulated from the uniform distribution on the interval $[0, 1]$, then $Y_i^m = Y^*$, if $U \leq P_i^*$: otherwise repeat first step.
- **M-step**: The M-step consist of two sub steps, which are,
  - M1-step which caters for estimation of the dropout parameters, e.g. the

logistic step in the case where we assume the dropout process can be modelled by a logistic model. These parameters can be obtained using an iterative method for likelihood estimation of binary data models.

- M2-step, maximum likelihood estimates of the measurement process are obtained using an appropriate optimisation approach.

2.9 Models for Data with Dropouts

Let us denote $Y_i^* = (Y_{i1}^*, Y_{i2}^*, \ldots, Y_{in}^*)^T$ be the $n-$element complete vector of measurements on the i^{th} subject and $t_i = (t_{i1}, \ldots, t_{in})^T$ be the corresponding times at which the measurements are made. Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})^T$ denote the vector of observed measurements with missing values coded as 0.

2.9.1 Selection Models

Under the Selection model one uses the functional form of $f(d_i|y_i)$ to discriminate between different types of dropout processes. The full data density can be factorised as below

$$f(\mathbf{y}_i, d_i|\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}, \boldsymbol{\varphi}) = f(\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}) f(d_i|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\varphi}) \tag{2.9.29}$$

where $f(\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta})$ is the marginal density of the measurement process, $f(d_i|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\varphi})$ is the density for the missingness process conditional on the outcomes. The latter factor corresponds to the self selection of individuals into 'observed' and 'missing' groups which is the basis of selection models. This factor can be expressed in the form

$$f(d_i|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\varphi}) = f(d_i|\mathbf{y}_i^o, \mathbf{y}_i^m, \mathbf{X}_i, \boldsymbol{\varphi}) \qquad (2.9.30)$$

If $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are disjoint in the sense that the parameter space $(\boldsymbol{\theta'}, \boldsymbol{\varphi'})'$ is the product of the parameter spaces $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ then inference can be based on the marginal observed data density only, which is the separability condition. The observed data likelihood can be expressed as

$$f(\mathbf{y}_i^o, d_i|\boldsymbol{\theta}, \boldsymbol{\varphi}) = \int f(\mathbf{y}_i, d_i|\boldsymbol{\theta}, \boldsymbol{\varphi})\, d\mathbf{y}_i^m = \int f(\mathbf{y}_i^o, \mathbf{y}_i^m|\boldsymbol{\theta}) f(d_i|\mathbf{y}_i^o, \mathbf{y}_i^m, \boldsymbol{\varphi}) d\mathbf{y}_i^m \qquad (2.9.31)$$

If the density of missingness $f(d_i|\mathbf{y}_i^o, \mathbf{y}_i^m, \boldsymbol{\varphi})$ is independent of the measurements (both $\mathbf{y}_i^o$ and $\mathbf{y}_i^m$), when it assumes the form $f(d_i|\boldsymbol{\varphi})$ then the process is termed *Completely Random Dropout (CRD).*

If $f(d_i|\mathbf{y}_i^o, \mathbf{y}_i^m, \boldsymbol{\varphi})$ is independent of the unobserved measurements $\mathbf{y}_i^m$, but depends on the observed measurements $\mathbf{y}_i^o$, there by assuming the form $f(d_i|\mathbf{y}_i^o, \boldsymbol{\varphi})$ then the process is referred to as *random drop out(RD).*

If $f(d_i|\mathbf{y}_i^o, \mathbf{y}_i^m, \boldsymbol{\varphi})$ depends on values of $\mathbf{y}_i^m$, the process is referred to as *informative drop out (ID).* ID process is allowed to depend on $\mathbf{y}_i^o$. It is important to note that for selection models to hold the separability condition has to be satisfied such that the measurement model $f(y_i^o)$ and the dropout model $f(d_i)$ or $f(d_i|y_i)$ can be fitted separately.

2.9.2 A Selection Model for Dropouts

Let us denote t_{d_i} as the dropout time for the i^{th} subject, where $2 \leq d_i \leq n+1$, where $n+1$ identifies no drop out. Then, r_i is a vector of $d_i - 1$ consecutive ones followed by $n+1-d_i$ consecutive zeros. The crucial assumption we make for the dropout process is the one made by Diggle and Kenward (1994) which says that if a subject is still in the study at time t_k, its associated sequence of measurements $Y_j : j = 1, \ldots, k$ follows the same joint distribution as that of the corresponding $Y_j^* : j = 1, \ldots, k$. The working relationship between Y and Y^* is as follows:

$$Y_j = \begin{cases} Y_j^* & \text{if } j = 1, \ldots, D-1 \\ 0 & j \geq D \end{cases}$$

where $2 \leq D \leq n$. Now let $f^*(y, \boldsymbol{\beta}, \phi)$ denote the joint probability density function of Y^*, which follows the multivariate Gaussian model. Let $H_k = (y_1, y_2, \ldots, y_{k-1}$ denote an observed sequence of measurements up to time t_{k-1} and y_k^* the value that would be observed at time t_k if the unit did not drop out. Diggle and Kenward(1994)'s model allows the conditional probability of drop out at time d to depend on the history of the measurement process up to and including time t_d, so that for $d \leq n$

$$P(D = d|history) = p_d(H_d, y_d^*; \boldsymbol{\varphi}), \tag{2.9.32}$$

where $\boldsymbol{\varphi}$ is a vector of unknown parameters. Now the joint distribution of the observed sequence $\mathbf{Y}$ via the sequence of conditional distributions for Y_k given $(Y_1, Y_2, \ldots, Y_{k-1}) = H_k$. Let $f_k^*(y|H_k^*; \boldsymbol{\beta}, \boldsymbol{\phi})$ denote the conditional univariate probability density function of Y_k^* given $(Y_1^*, Y_2^*, \ldots, Y_{k-1}^*) = H_k^*$. Also let $f_k(y|H_k; \boldsymbol{\beta}, \boldsymbol{\phi})$ denote the the conditional probability density function of Y_k given $(Y_1, Y_2, \ldots, Y_{k-1}) = H_k$, then it follows that

$$P(Y_k = 0|H_k, Y_{k-1} = 0) = 1 \tag{2.9.33}$$

and

$$P(Y_k = 0|H_k, Y_{k-1} \neq 0) = \int p_k(H_k, y, \boldsymbol{\varphi}) f_k^*(y|H_k; \boldsymbol{\beta}, \boldsymbol{\phi})\, dy \tag{2.9.34}$$

and for $Y_k = y \neq 0$,

$$f_k(y|H_k; \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\varphi}) = \{1 - p_k(H_k, y; \boldsymbol{\varphi})\}\ f_k^*(y|H_k; \boldsymbol{\beta}, \boldsymbol{\phi}) \tag{2.9.35}$$

Then for a complete sequence $\mathbf{Y} = (Y_1, \ldots, Y_n)$, and suppressing the dependency on the parameters $\boldsymbol{\beta}$, $\boldsymbol{\varphi}$ and $\boldsymbol{\phi}$,

$$\begin{aligned} f(\mathbf{y}) &= f_1^*(y_1) \prod_{k=2}^{n} f_k(y_k|H_k) && (2.9.36) \\ &= f^*(\mathbf{y}) \prod_{k=2}^{n} \{1 - p_k(H_k, y_k)\} && (2.9.37) \end{aligned}$$

For an incomplete sequence $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_{d-1}, 0, \ldots, 0)$ with dropout at time t_d

$$\begin{aligned} f(\mathbf{y}) &= \left\{ f_1^*(y_1) \prod_{k=2}^{d-1} f_k(y_k|H_k) \right\} P(Y_d = 0|H_d) && (2.9.38) \\ &= f_{d-1}^*(\mathbf{y}) \left[\prod_{k=2}^{d-1} \{1 - p_k(H_k, y_k)\} \right] P(Y_d = 0|H_d) && (2.9.39) \end{aligned}$$

, where $f_{d-1}^*(y)$ denotes the joint pdf of the first $d-1$ non zero elements of Y^*

If we let $\mu(\boldsymbol{\beta})$ and $\mathbf{V}(\boldsymbol{\phi})$ be the mean response vector and the covariance matrix for a complete sequence of measurements $\mathbf{Y}^* = (Y_1, Y_2, \ldots, Y_n)^t$ on a single subject, respectively and $\mathbf{y}$ be the realisation of $\mathbf{Y}^*$, then we let $\mu^{(k)}$ denote the first k elements of μ. Also let $\mathbf{V}^{(k)}$ denote the leading $k \times k$ sub matrix of $\mathbf{V}$, $\mathbf{c}^{(k)} = (c_1^{(k)}, \ldots, c_k^{(k)})$ for k-element c-vector of covariances, where

$$c_j^{(k)} = cov(Y_j^*, Y_{k+1}^*), j = 1, \ldots, k$$

and $v_{kk} = var(\mathbf{Y}_k^*$. Then for each $k = 1, \ldots, n-1$, the joint pdf of $\mathbf{y}$ is k variate Gaussian with mean vector $\mu^{(k)}$ and variance $\mathbf{V}^{(k)}$. As cited earlier in section 2.1, $\mathbf{V}$ can take various forms.

Considering the dropout process, there are several ways of modelling the dropout process and specification of conditional probability, $p_k(H_k, y; \boldsymbol{\varphi})$ in equation 2.9.32. A logistic linear model is proposed as an empirical model,

$$logit\left\{p_k(H_k, y; \boldsymbol{\varphi})\right\} = \varphi_0 + \varphi_1 y + \sum_{j=2}^{k} \varphi_j y_{k+1-j} \tag{2.9.40}$$

where φ can be allowed to depend on covariates or on time in which case model above can be extended by making φ_0 a function of covariates say, w_{qk} time t_k. The relationship can be for example linear :

$$\varphi_0 = \sum_{q=1}^{r} \varphi_{q0} w_{qk} \tag{2.9.41}$$

Now for the likelihood function, let $\mathbf{y}_i = \{y_{ij} : j = 1, \ldots, d_i - 1\}$ denote the observed measurements on the ith subject, with d_i indicating the dropout time. Let $f^*(\mathbf{y}_i)$ denote the joint pdf of the $d_i - 1$ available measurements from the ith subject, we have

$$f_i^*(\mathbf{y}_i) = (\sqrt{2\pi})^{-(d_i-1)}(\mathbf{V}^{(d_i-1)}(\boldsymbol{\phi})^{-\frac{1}{2}})exp\left\{-\frac{1}{2}(\mathbf{y}_i - \mu_i)^T\mathbf{V}^{(d_i-1)}(\boldsymbol{\phi})^{-1}(\mathbf{y}_i - \mu_i)\right\} \tag{2.9.42}$$

and applying logarithms we get

$$logf_i^*(\mathbf{y}_i) = -\{(d_i-1)/2\}\,log(2\pi) - \frac{1}{2}log\left|\mathbf{V}^{(d_i-1)}(\boldsymbol{\phi})\right| - \frac{1}{2}(\mathbf{y}_i - \mu^{(i)})^T\mathbf{V}^{(d_i-1)}(\boldsymbol{\phi})^{-1}(\mathbf{y}_i - \mu^{(i)}) \tag{2.9.43}$$

where $\mu^{(i)}$ represent the relevant $d_i - 1$ elements of the mean response vector. Model 2.9.32 can be expressed in the form

$$log\left\{\frac{p_k(H_k, y; \boldsymbol{\varphi})}{1 - p_k(H_k, y; \boldsymbol{\varphi})}\right\} = \varphi_0 + \varphi_1 y + \sum_{j=2}^{k}\varphi_j y_{k+1-j} \tag{2.9.44}$$

implying that

$$log\{(1 - p_k(H_k, y; \boldsymbol{\varphi}))\} = -log\left\{1 + exp\left(\varphi_0 + \varphi_1 y + \sum_{j=2}^{k}\varphi_j y_{k+1-j})\right)\right\} \tag{2.9.45}$$

Then the log-likelihood for $\boldsymbol{\beta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\varphi}$ based on $y_i : i = 1, \ldots, m$ is

$$L(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\varphi}) = L_1(\boldsymbol{\beta}, \boldsymbol{\phi}) + L_2(\boldsymbol{\varphi}) + L_3(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\varphi}) \tag{2.9.46}$$

where

$$L_1(\boldsymbol{\beta}, \boldsymbol{\phi}) = \sum_{i=1}^{m} logf^*(\mathbf{y}_i) \tag{2.9.47}$$

$$L_2(\boldsymbol{\beta}) = \sum_{i=1}^{m}\sum_{k=2}^{d-1} log\{1 - p_k(H_{ik}, y_{ik})\} \tag{2.9.48}$$

and

$$L_3(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\beta}) = \sum_{i:d_i \leq n} logP(D = d_i|\mathbf{y}_i) \tag{2.9.49}$$

2.9.3 Pattern-Mixture Models

Pattern mixture models offer an alternative way of factorising the joint density. For longitudinal data the classification of subjects according to dropout time divides subjects into subgroups after the event. One would want to know whether the response characteristics which are of primary interest do or do not vary between these subgroups. The usual assumption is that subjects with the same drop out time are more alike than those with different dropout times, thus those with the same dropout time share a common response distribution. In other words, the response is a mixture over patterns. However, this assumption may be too strong in many circumstances. Pattern-mixture models work with the factorisation of the joint distribution of $\mathbf{Y}_i$ and D_i, the full data density into the marginal density of D_i and the conditional distribution of $\mathbf{Y}_i$ given D_i. Thus

$$f(\mathbf{y}_i, d_i|\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}, \boldsymbol{\varphi}) = f(\mathbf{y}_i|d_i, \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta})f(d_i|\mathbf{X}_i, \boldsymbol{\varphi}). \tag{2.9.50}$$

The rationale of the pattern mixture models is that each subject's dropout time is somehow predestined and that the measurement process varies between dropout cohorts. The typical feature of pattern mixture models is that the distribution of the missingness mechanism only depends on the covariates only and not on the outcome variable.

Suppressing the dependence on covariates in equation 2.9.50 above, we obtain

$$f(\mathbf{y}_i, d_i|\boldsymbol{\theta}, \boldsymbol{\varphi}) = f(\mathbf{y}_i|d_i, \boldsymbol{\theta})f(d_i|\boldsymbol{\varphi}). \tag{2.9.51}$$

Equivalently using T_i introduced earlier on,

$$f(\mathbf{y}_i, t_i|\boldsymbol{\theta}, \boldsymbol{\varphi}) = f(\mathbf{y}_i|t_i, \boldsymbol{\theta})f(t_i|\boldsymbol{\varphi}). \tag{2.9.52}$$

where $T_i = D_i - 1$ indicates the pattern in which t measurements are obtained. The models above imply a different distribution for each dropout time. For a continuous response with a Gaussian distribution $\mathbf{y}_i|t_i \sim N(\boldsymbol{\mu}(t_i), \boldsymbol{\Sigma}(t_i))$ where

$$\boldsymbol{\mu}(t) = \begin{bmatrix} \mu_1(t) \\ \mu_2(t) \\ \cdots \\ \mu_n(t) \end{bmatrix}$$

and

$$\boldsymbol{\Sigma}(t_i) = \begin{bmatrix} \sigma_{11} & \sigma_{21} & \cdots & \sigma_{n1} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_{nn} \end{bmatrix}$$

for $t = 1, 2, \ldots, m$ (where t indicates the length of the sequence). Let $P(t) = \pi_t = f(t_i|\varphi)$, this implies that the marginal distribution of response is a mixture of normal means, i.e

$$\boldsymbol{\mu} = \sum_{t=1}^{m} \pi_t \mu(t) \tag{2.9.53}$$

This produces a saturated pattern mixture model. Pattern-mixture model factorisation stresses those aspects of the model which are assumption driven rather than data driven. This means that pattern mixture model parameters can not be identified without placing restrictions on the conditional distributions $f(\mathbf{y}|\mathbf{t})$.

Little (1993) discusses the use of a less stringent restriction called *complete case missing value (CCMV).* CCMV corresponds to assuming that for each $t < n+1$ and $T = t < j$

$$f(y_j|y_1, \ldots, y_{j-1}, T = t) = f(y_j|y_1, \ldots, y_{j-1}, T = n) \qquad (2.9.54)$$

He shows how these constraints can be used to identify all the parameters in the model and also to obtain estimates for these and the marginal probabilities. The CCMV restrictions equate the conditional distributions beyond time t i.e those unidentifiable from this dropout group, with the same conditional distributions from the completers. Molenberghs *et al* (1998) show that RD dropout corresponds to a stronger restriction in which the conditional distributions beyond time t is equated with all conditional distributions from all those who dropout after time t, which they call the *available case missing value (ACMV)* restrictions.

$$f(y_j|y_1, \ldots, y_{j-1}, T = t) = f(y_j|y_1, \ldots, y_{j-1}, T > j) \qquad (2.9.55)$$

Usually the choice of restriction will need to be guided by the context of the scientific question being posed by the data. Also the form of the data requires a more structured model for the response which incorporates covariates. Models for $f(t_i|\boldsymbol{\varphi})$

can be constructed in many different ways but most authors assume that the dropout process is fully observed and that T_i satisfies a parametric model. Some authors even extend to cases where the dropout is allowed to be right censored with no parametric restriction put on the dropout times. Hogan and Laird (1997)'s conditional model for $\mathbf{Y}_i^o$ given T_i is a linear mixed model with dropout time as one of the covariates in the mean structure. To handle the incomplete covariates due to right censoring Hogan and Laird (1997) use the Expected Maximisation (EM) algorithm for Maximum Likelihood estimation. Under CRD, $f(\mathbf{y}, \mathbf{d}) = f(\mathbf{y})f(\mathbf{d})$, the pattern-mixture models and the selection models coincide.

RD can be expressed in pattern mixture model framework through restrictions CCMV and ACMV. CCMV can be defined as the condition that for $t \geq 2$ and $j < t$

$$f(y_t|y_1, \ldots, y_{t-1}, d = j+1) = f(y_t|y_1, \ldots, y_{t-1}, d = n+1) \tag{2.9.56}$$

that is, the conditional density of unobserved components given a particular set of observed components is equal to the corresponding density in the subgroup of completers. ACMV can be defined as the condition that for $t \geq 2$ and $j < t$

$$f(y_t|y_1, \ldots, y_{t-1}, d = j+1) = f(y_t|y_1, \ldots, y_{t-1}, d > t) \tag{2.9.57}$$

that is, the conditional density of unobserved components given a particular set of observed components is equal to the conditional density calculated from the subgroup of all patterns for which all required components have been observed.

2.9.4 A Pattern Mixture Model for Dropouts

As seen in the previous section, a pattern mixture model factorises the joint distribution $f(\mathbf{y}_i, \mathbf{d}_i|\theta, \varphi)$ into the product of the conditional density of the measurements given the dropout pattern, $f(\mathbf{y}_i|\mathbf{X}_i, \theta^{(t_i)})$, and the marginal density describing the dropout mechanism, $f(\mathbf{t}_i|X_i, \varphi)$, where $t_i = 1, \ldots, n$ indicate the dropout time. The dropout process $f(\mathbf{t}_i|X_i, \varphi)$ is the probability to belong to a particular dropout pattern. The measurement models depends on dropout and take the general form:

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{X}_i\boldsymbol{\beta}(t_i) + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i & (2.9.58) \\ \mathbf{b}_i &\sim N(\mathbf{0}, \mathbf{D}(t_i)) & (2.9.59) \\ \epsilon_i &\sim N(\mathbf{0}, \Sigma(t_i)). & (2.9.60) \end{aligned}$$

In this model the fixed effects as well as the covariance parameters are allowed to change with dropout pattern. The likelihood contribution of the i^{th} subject based on observed data (y_i^o, t_i) is proportional to

$$f(y_i^o, t_i) = f(t_i)\, f(y_i^o|t_i).$$

This now requires specifying the marginal model for the dropout process and the conditional model for observed outcomes given the drop out pattern as in 2.9.58. For any subject with $t_i < n$, the sub vector of $\theta^{(t_i)}$ describing y_i^m is unidentified. Identifying restrictions such as available-case missing value (ACMV) proposed by Molenberghs *et al.*(1998), the complete case missing value (CCMV) proposed by Little (2003) and the neighbouring case missing value (NCMV) can be applied. The distribution of $\mathbf{d}_i$ depend on $\mathbf{X}_i$ and the distribution of $\mathbf{y}_i$ conditional on $\mathbf{d}_i$ and $\mathbf{X}_i$ is normally distributed.

2.9.5 Shared Parameter Models

For the shared parameter models the full data density is factorised as follows:

$$f(\mathbf{y}_i, d_i, \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}, \boldsymbol{\varphi}) = f(\mathbf{y}_i|\boldsymbol{\xi}_i, \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta})\, f(d_i|\mathbf{X}_i, \boldsymbol{\xi}_i, \boldsymbol{\varphi}). \tag{2.9.61}$$

where ξ_i are the shared parameters which play the role of the confounder for the relationship between $\mathbf{y}_i$ and d_i. They can be either observable variables or latent variables (e.g random effects). For the case when you observable variables the shared parameter model becomes a pattern mixture model. Jason Roy (2003) introduced a special shared parameter model he called the latent dropout class model. His model combines features of pattern mixture models and latent-class models where he assumes that there exists a small number of dropout classes and class membership which is unobserved but the probability of being in any particular latent dropout class is determined by the drop out times themselves.

2.9.6 A Shared Parameter Model for Dropouts - The Latent Dropout Class Model

Adopting the notation introduced in section 2.5 for T_i, and indicating the number of measurements taken for subject i, that is the length of time the i^{th} unit has been in the study, where

$$T_i = \sum_{j=1}^{n_i} R_{ij} = D_i - 1$$

the latent dropout class model factorises the joint distribution as

$$f(\mathbf{Y}, \mathbf{T}) = f(\xi|\mathbf{T}) f(\mathbf{Y}|\xi)$$

We have to specify the models for the latent dropout class conditional on the drop out and for the response conditional on the latent class.

If we assume that T_i determines the probability that a subject in one of the $M < n$ latent dropout classes, where n is the number of intended observations on a unit and denote $\xi_i = (\xi_{i1}, \dots, \xi_{iM})^T$ where $\xi_{ik}, k = 1, 2, \dots, M$ are indicator variables for each latent class such that if a subject i is in class j then $\xi_{ij} = 1$ and $\xi_{ij'} = 0$ for $j \neq j'$. The latent variable ξ is considered ordinal in the sense that increases in the dropout time are assumed to monotonically increase or decrease the chances of being in one of the first k dropout classes. Under this assumption the probability of being in a given class is assumed to be determined by T_i through the regression model

$$P(\sum_{j=1}^{k} \xi_{ij} = 1|T_i) = h(T_i, \lambda_k), k = 1, 2, \ldots, M-1$$

where h is a monotone function and λ_k is a vector of parameters. The probability of being in dropout class less than or equal to k is some monotone function of the dropout time. Therefore, $\xi_i|T_i \sim \mathit{Multinomial}$ with probabilities p_{ij}. For $j = 2, 3, \ldots, M-1$, $p_{ij} = h(T_i; \lambda_j) - h(T_i; \lambda_{j-1}$,

$$p_{i1} = h(T_i; \lambda_1) \tag{2.9.62}$$

and

$$p_{iM} = 1 - h(T_i; \lambda_{M-1}). \tag{2.9.63}$$

Common examples of $h(.)$ are the probit and the inverse of the logit links.

Now specifying the model for the response conditional on the latent class, we let $\mathbf{X}_{ij}$ be a vector of covariates whose effects on Y do not depend on the latent class. Let $\mathbf{Z}_{ij}$ be a vector of covariates whose effects on Y vary with the latent class ξ_i. The complete response conditional on the latent class is assumed to be normally distributed with mean

$$E(Y_{ij}|\xi_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, T_i) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + (\mathbf{z}_{ij} \otimes \xi_i)^T \alpha_i$$

for $j = 1, 2, \ldots, n$, where α and $\boldsymbol{\beta}$ are vectors of parameters. The term $(\mathbf{z}_{ij} \otimes \xi_i)$ corresponds to interactions between levels of the latent class and covariates. No dependence between the variance and covariance of $(Y_{i1}, Y_{i2}, \ldots, Y_{in})$ and the latent class ξ_i is assumed. The conditional distribution of Y_i given ξ_i is

$$(\mathbf{Y}_i|\xi_i) \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i(\xi)\alpha,\, V_i(\boldsymbol{\phi}))$$

where $V_i(\boldsymbol{\phi})$ is the covariance matrix parameterised by the vector $\boldsymbol{\phi}$.

Now deriving the likelihood function, $\mathbf{Y} = (\mathbf{Y}^o, \mathbf{Y}^m)$, where $\mathbf{Y}$ is the complete vector, $\mathbf{Y}^o$ is the observed vector and $\mathbf{Y}^m$ is the missing vector. Let $\eta = (\boldsymbol{\beta}^T, \alpha^T, \boldsymbol{\phi}^T, \lambda^T)^T$ represent the set of parameters of interest. Then the likelihood function for the i^{th} subject is

$$\begin{aligned} L(\eta, \pi, Y_i^o, T_i) &= \sum_{\xi} \int L(\boldsymbol{\beta}, \alpha, \boldsymbol{\phi}; Y_i|\xi_i, T_i)\, L(\lambda_i; \xi_i|T_i)\, L(\eta; T_i)\, dY_i^m \quad (2.9.64) \\ &= L(\pi; T_i) \sum_{\xi} L(\boldsymbol{\beta}, \alpha, \boldsymbol{\phi}; Y_i^o|\xi)\, L(\lambda; \xi|T_i) \quad (2.9.65) \end{aligned}$$

Where π are parameters characterising the distribution of T_i. Maximising the logarithm of $L(\eta, \pi, Y_i^o, T_i)$ using the normal proceedure is not feasible so the EM algorithm or other numerical methods like the Newton-Raphson algorithm are applied to obtain the parameters.

2.10 Testing for completely random dropouts

The main objective here is to test the hypothesis that the dropouts are completely random, that is, the probability that a unit drops out at time t_j is independent of the observed sequence of measurements of that unit at times $t_1, t_2, \ldots, t_{j-1}$ where we assume that a complete set of measurements on a unit would be taken at times $t_j,\ j = 1, 2, \ldots, n$ but dropout occur. The available data on the i^{th} of the m units are $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})$ with $n_i < n$ and y_{ij} taken at times t_j. Let p_{ij} denote the probability that the i^{th} unit drops out time t_j. Under the CRD assumption $\mathbf{p}_{ij}$ may depend on time, treatment or other explanatory variables but cannot depend on the observed measurement $\mathbf{y}_i$. Diggle (1989) develops a method to test this assumption which consists of applying separate tests at each time with each treatment group and analysing the sample of p-values for departure from the uniform distribution on $(0, 1)$. Formal tests like Kolmogorov-Smirnov test or Benard's Monte Carlo test can be used to investigate whether the complete set of p-values behaves like a random sample from the uniform distribution.

Chapter 3

Methodology

3.1 Data used in the research

This research will be based on real life data. It will start with analysis on the complete data set, followed by non-ignorable dropout pattern simulated on the data set to produce an incomplete data set. The data set, ***Diabetes***, is obtained from observing 191 diabetes patients over a period of nine years, with measurements taken annually. The main research objective is to model the body-mass index (BMI) of the patients in relation to mainly time, with the main grouping factor observed being gender. Other factors and covariates observed are hypertensive status, family history of diabetics status, sugar level (hba1c) and age in years. The body-mass index was invented by Aldophe Quetelet (Wikipedia). It compares a person's weight and height. The mathematical formulae is

$$BMI = \frac{W}{H^2} \tag{3.1.1}$$

Table 3.1: **BMI classifications**

BMI values(x)	$x \leq 18.5$	$18.5 < x \leq 25$	$25 < x \leq 30$	$x > 30$
Classification	Underweight	Normal	Overweight	Obese

where W is the person's weight in *kilograms* and H is the person's height in *metres* hence the BMI unit is kg/m^2. Based on the BMI value a person can then be classified either as underweight, or normal weight,or overweight or obese. Table 3.1 show the categories and it is mostly used as tool to estimate a healthy body weight based on how tall a person is:

3.2 Model building

The focus of this research is fitting general linear mixed models. Under the linear mixed model the data vector $\mathbf{Y}_i$ for the i^{th} subject is assumed to be normally distributed with mean $X_i\beta$ and covariance matrix of the form $V_i = Z_iDZ_i' + \Sigma_i$. When fitting a general linear mixed model, appropriate mean and covariance structures have to be specified. Exploratory Data analysis provides the crucial initial detective work for model specification. Two approaches can be adopted, exploring the marginal distribution or exploring the subject specific profiles. This research will focus on graphical methods to explore the mean structure, variance function and the correlation structure. This will form a basis for selection of the preliminary mean structure, fixed effects structure, random effects structure and the residual covariance structure.

This research is going to focus on likelihood estimation methods, the main attracting property being that likelihood estimates known to be asymptotically unbiased and asymptotically efficient. Since the samples to be used in this research are large, estimates obtained can be safely considered to be good. Model fitting and parameter estimation will be considered as follows:

1. *Estimation on complete data*: Linear mixed models will be fitted also with the possibility of serial correlation under consideration.

2. *Estimation of the joint models*: The three main different approaches to joint modelling of the measurement process and the dropout process, owing to different factorisations that can be done to the full data density, will be done. This research's main focus is on the measurement process given a non-ignorable missing pattern, which is more conservative since the assumptions of ignorable missing patterns are in most cases in-testable, for comparison purposes it would be good to also consider models under ignorability assumptions, so time permitting models under ignorability assumptions are going to be considered also. Under ignorability assumptions the dropout process is assumed to be ignorable and if focus is on the marginal of the observed vector, $\mathbf{Y}_i^o$ only. Under non-ignorability assumptions, dropouts are no longer ignorable, the dropout could be related to unobserved responses implying that the treatment effects can no longer be estimated without taking the drop out model into account. A

marginal model is now required for the complete vector, $\mathbf{Y}_i$.

3.3 Summary of methods to be used

1. Working with the Complete longitudinal data set. The initial step would be Exploratory Data Analysis of the data set. This mainly involves the graphical approach, looking at plots for individual profiles, average mean over time and variance plots. These will help come up with plausible tentative structures for the random effect, fixed effects, covariance and correlation.

2. Model building and estimation on the complete data set would follow with a linear mixed effect model fitted to the complete data set.

3. The next step would be to simulate dropout patterns on the data set.

4. Jointly model the measurement process and the dropout process is then done using the three approaches to factorisation of the joint density, which are, selection model, pattern mixture model and shared parameter model. For the pattern mixture model and the shared parameter model, the Latent dropout class model will be fitted as a representative of both models. The E-M algorithm will be used to obtain likelihood based estimates .

5. Diagnostic checks are done on the fitted models using exploratory or graphical approach, aiming at detecting serious departures from model assumptions and

identifying potentially influential cases

6. A comparison of the results from the four models based on measurement process estimates will be done. In this research only the bias of estimates from the three models on the incomplete data from the estimates of the complete data set will be considered.

Chapter 4

Analysis and Results

4.1 Introduction

In this research, analysis is mainly centred on the incomplete data set, however the first part of this chapter focuses on linear mixed effects models on the complete data set and the second part focuses on joint modeling for the incomplete data. Analysis starts with exploring the data, mainly aiming at coming up with tentative mean and covariance structures, followed by model estimation and lastly model adequacy checking.

4.2 Complete data set analysis

The main research question is modeling the BMI values by gender over time.

4.2.1 Exploration of the data

Profile plots for BMI values were plotted against time for (a) all patients, (b) male patients and (c) female patients:

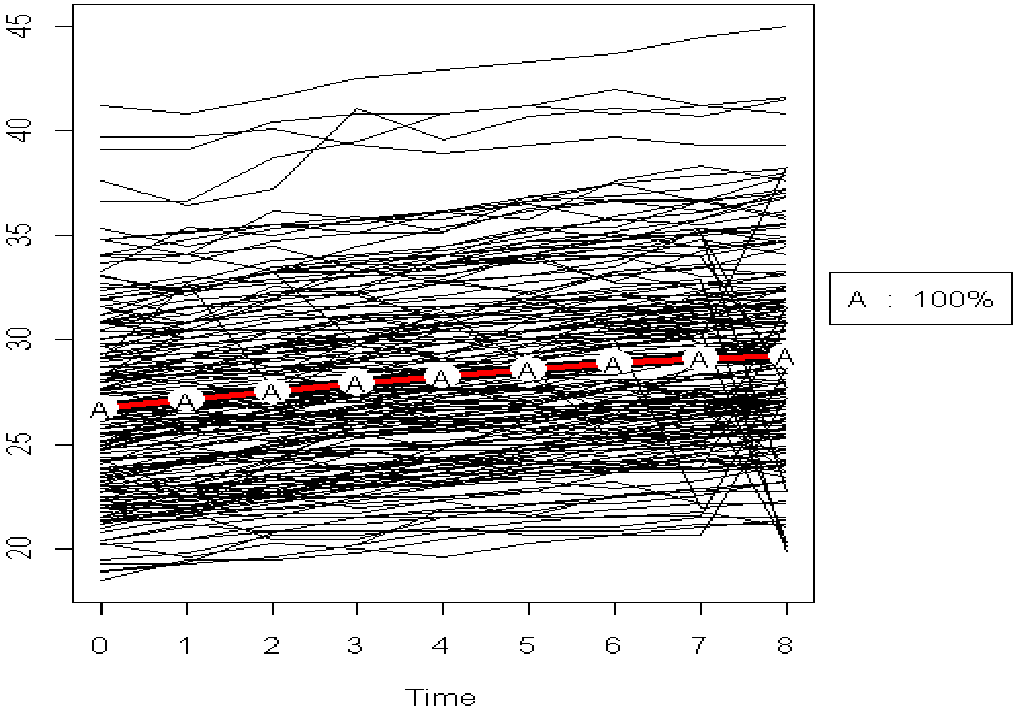

Figure 4.1: Bmi profiles for all the patients

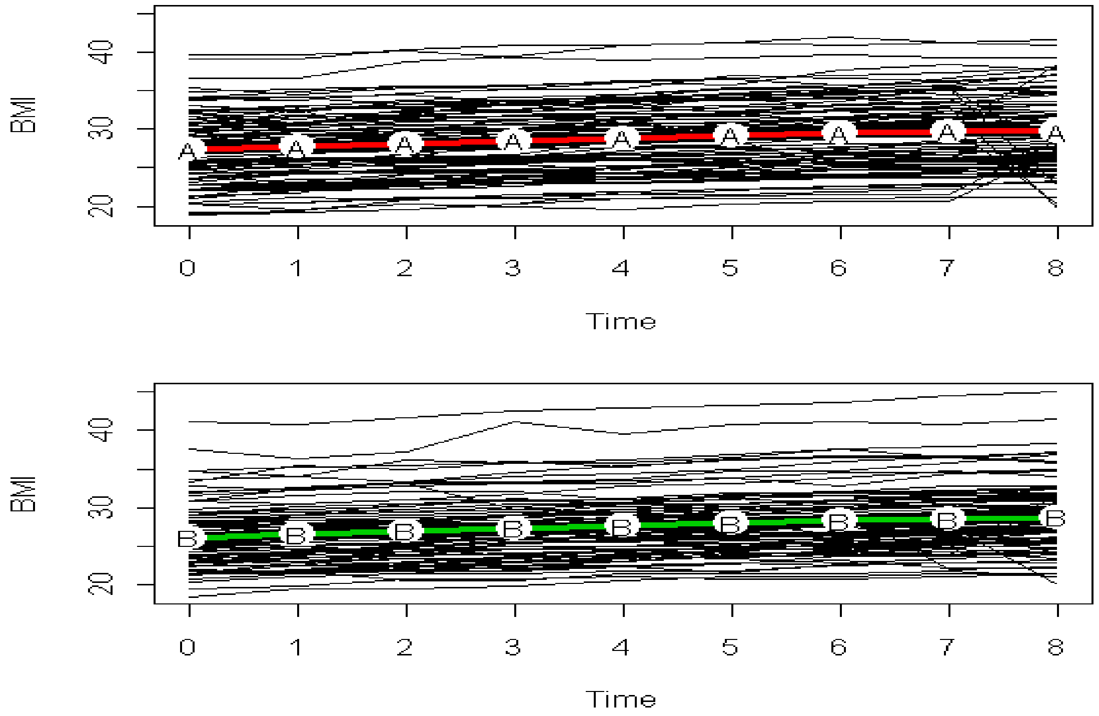

Figure 4.2: BMI profiles for patients by Gender

There seems to be a difference in mean profiles for males and females, with mean profiles for females depicting slightly higher values than mean profiles for males. Though the number of observations for each subject are not so many, there is an indication of some within subject-pattern, the BMI values increase with time with a flattish slope though. This would warrant for some autoregressive serial correlation in the models.

Regressions of BMI values on time are fitted for each subject to pursue these impressions. The box plots of regression coefficients are shown in Figure 4.3. The intercept represents the mean value at the start of the study. The median intercept for female patients is higher than that for male patients and there is more variation among females than there is among males. The median slopes are almost equal, with females slopes showing slightly more variation. For both groups slopes are skewed to the positive values. The T-tests for comparing the slopes and intercepts for males and females produced the following results(full results in Appendix) showing that males have a slightly higher average BMI growth rate(0.3298) than that of females(0.3182) but it is not significantly different from the females BMI growth rate. Comparing the average BMI values at the the start of the study period, results show that we have some evidence($\alpha = 0.1$ level of significance) to conclude that females have a higher

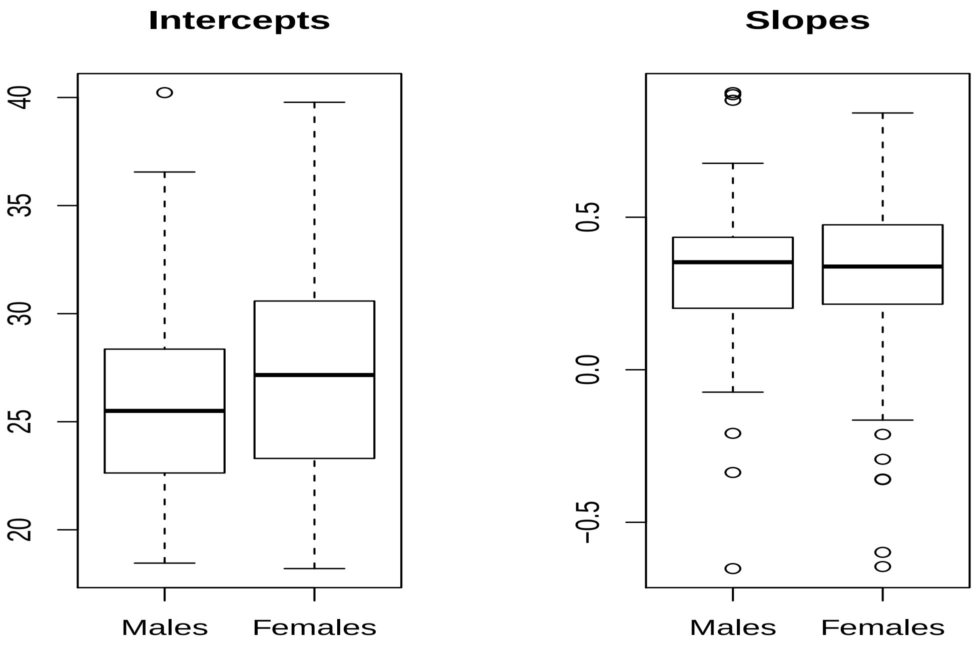

Figure 4.3: Coefficients for within subject regression of BMI values on time

average BMI value.

4.2.2 Fitting the linear mixed model

Our starting point is the general liner mixed effect model, suppose each subject has response $\mathbf{BMI}_i$, a vector of length n_i, which is modelled as:

$$\mathbf{BMI}_i \sim \mathbf{N}_p(\mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i, \sigma^2\mathbf{V}_i(\phi)) \tag{4.2.1}$$

where $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$ and $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{\Sigma}_i$. With m individuals and assuming the errors and the random effects between individuals are uncorrelated we model using

$$\mathbf{BMI} \sim \mathbf{N}_p(\mathbf{X}\beta, \sigma^2\mathbf{V}(\phi)) \tag{4.2.2}$$

where $\mathbf{V} = \mathbf{ZDZ}' + \mathbf{\Sigma}$ with $\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \cdots \\ \mathbf{y}_{191} \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \cdots \\ \mathbf{x}_{191} \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \cdots \\ \mathbf{b}_{191} \end{bmatrix}$,
$\mathbf{D} = diag(D, D, \cdots, D)$, $\mathbf{Z} = diag(Z_1, Z_2, \cdots, Z_{191})$ and $\Sigma = diag(\Sigma_1, \Sigma_2, \cdots, \Sigma_{191})$
Fitting a mixed effects model to the data, that included fixed effects of time and sex and the interaction between time and sex, and the random intercepts and slope gave the following output:

```
Linear mixed-effects model fit by REML
 Data: d3
       AIC      BIC    logLik
  5589.245 5632.822 -2786.622
Random effects:
 Formula: ~time | serialno
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
```

```
(Intercept) 4.4402609 (Intr)
time        0.2307473 -0.248
Residual    0.8149142
Fixed effects: BMI ~ time + sex + time:sex
                    Value Std.Error   DF  t-value p-value
(Intercept)     25.921709 0.4670258 1526 55.50380  0.0000
time             0.329764 0.0264395 1526 12.47241  0.0000
sexFemales       1.200317 0.6486941  189  1.85036  0.0658
time:sexFemales -0.011566 0.0367242 1526 -0.31494  0.7529
 Correlation:
                (Intr) time   sxFmls
time            -0.272
sexFemales      -0.720  0.196
time:sexFemales  0.196 -0.720 -0.272

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-10.81015463  -0.29931969   0.01131231   0.32404452   6.33807709

Number of Observations: 1719
Number of Groups: 191
```

There is a statistically significant upward trend for groups signified by the coefficient of *time*. The average for the female group is slightly higher than that of females as shown by the coefficient of $sexFemales$ which is not significant at 5% but is significant at 10%. The interaction coefficient is not statistically significant, indicating that there is not much difference between the gender groupings in changes in BMI values with time. The model is updated leaving out the interaction, there are slight changes in fixed effects parameters which all remain significant. The AIC value decreased slightly showing that the latter is a better model.

To test whether the random effects are necessary (the intercept and slope), models are refitted omitting one each time in turn from the model. The refitted models are contrasted with the original model by calculating a likelihood-ratio statistic. The following results were obtained: (i)For testing the necessity of the random slope, $bm.lme.1$ is the original model and $bm.lme.2$ is the model without the random slope:

```
          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
bm.lme.1      1  7 5582.571 5620.705 -2784.286
bm.lme.2      2  5 5958.211 5985.449 -2974.105 1 vs 2 379.6395  <.0001
```

(ii)For testing the necessity of the random intercept, $bm.lme.1$ is the original model and $bm.lme.3$ is the model without the random intercept:

```
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
bm.lme.1      1  7 5582.571 5620.705 -2784.286
bm.lme.3      2  5 8371.795 8399.033 -4180.897 1 vs 2 2793.223  <.0001
```

The tests are highly significant, suggesting that both the random intercept and slope are necessary. A plot of the residuals of the model in Fig 4.4 indicates that there is some serial correlation, as supported by the plot of residuals against fitted values.

Since observations are taken longitudinally on the same subject, there is a high likelihood that the within subject errors are correlated. A continuous first order autoregressive (CAR1) process in the errors is assumed. According to CAR1, suppose that $\epsilon_{i,t}$ and $\epsilon_{i,t+s}$ are errors for subject i separated by s units of time, then the correlation between these two errors is $\rho(s) = \phi^{|s|}$ where $0 \leq \phi \leq 1$ (Fox, 2002). A CAR1 model is fitted to the data, by updating the original model. The results (see

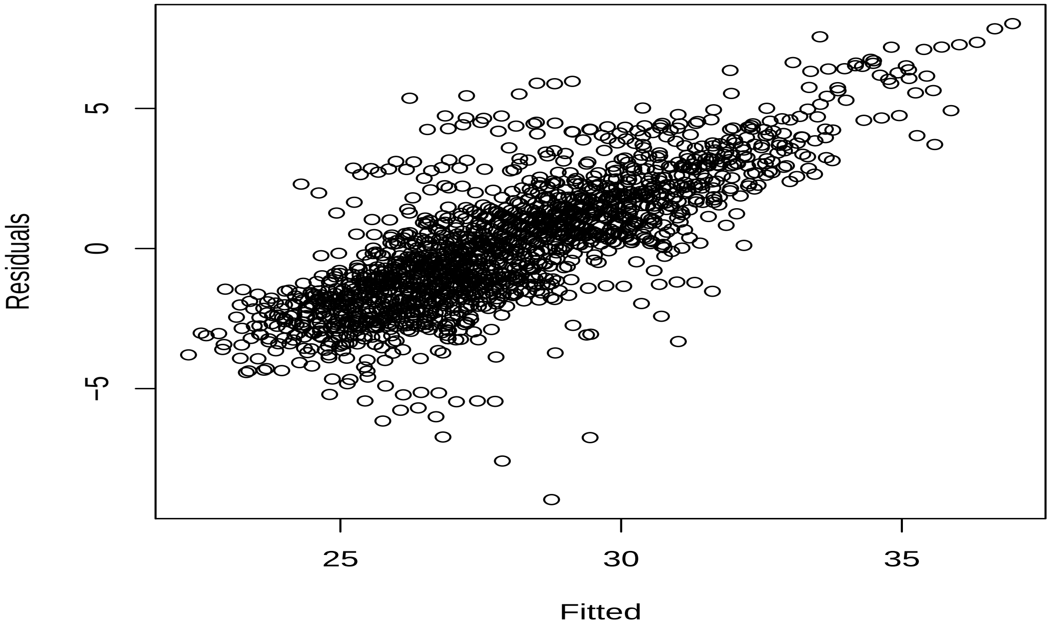

Figure 4.4: Residuals vs fitted values for all subjects

Appendix) show that the estimated autocorrelation parameter, $\hat{\phi} = 0.78$ is quite large and the estimated fixed effects coefficients have changed slightly, with the coefficient for gender decreasing by 25% from 1.45. A test for the statistical significance of the error autocorrelation was performed with *bm.lme*.1 being the original model and *bm.lme*.4 being the model with autocorrelation. The following output was obtained:

```
          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
bm.lme.1      1  7 5582.571 5620.705 -2784.286
bm.lme.4      2  8 5404.411 5447.993 -2694.205 1 vs 2 180.1603  <.0001
```

This test shows that the error autocorrelation is of significance to the model. With the new development we reconsider the necessity of the random slope and intercept. Note: *bm.lme*.5 is the autocorrelated model without random slope and *bm.lme*.6 is the autocorrelated model without the random intercept.

```
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
bm.lme.4     1  8 5404.411 5447.993 -2694.205
bm.lme.5     2  6 5404.519 5437.206 -2696.260 1 vs 2 4.108473  0.1282
```

```
        Model df      AIC      BIC    logLik   Test L.Ratio p-value
bm.lme.4     1  8 5404.411 5447.993 -2694.205
bm.lme.6     2  6 5404.943 5437.629 -2696.471 1 vs 2  4.5319  0.0037
```

The results show that the random time term can now be removed from the model giving the following output:

```
Linear mixed-effects model fit by REML
 Data: d3
       AIC      BIC    logLik
  5404.519 5437.206 -2696.260

Random effects:
 Formula: ~1 | serialno
        (Intercept) Residual
StdDev:    3.111951 3.211055

Correlation Structure: Continuous AR(1)
 Formula: ~time | serialno
 Parameter estimate(s):
      Phi
0.9527864
Fixed effects: BMI ~ time + sex
                Value Std.Error   DF  t-value p-value
(Intercept) 25.899078 0.4612001 1527 56.15584   0.000
time         0.315824 0.0232650 1527 13.57506   0.000
sexFemales   1.046601 0.6198911  189  1.68836   0.093
 Correlation:
           (Intr) time
time       -0.252
sexFemales -0.697  0.000

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.79065293 -0.47224039  0.00957101  0.47758716  2.50013692

Number of Observations: 1719
Number of Groups: 191
```

The fitted model is

$$BMI_{ij} = 25.899 + 0.3158t_j + 1.0466sexFemales + b_{0i} + 0.9528\epsilon_{i,j-1} + a_{ij} \tag{4.2.3}$$

The average BMI value for females at baseline is 26.945 compared to an average of 25.899 for males. On average both male and female diabetic patients are overweight.

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 25.899 \\ 0.3158 \\ 1.0466 \end{bmatrix}$$

$$\mathbf{Z}_i = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

a 9 × 1 vector of ones, the random effects variance $d_{11} = 9.684$ and

$$\mathbf{X}_i = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 4 & 1 \\ 1 & 5 & 1 \\ 1 & 6 & 1 \\ 1 & 7 & 1 \\ 1 & 8 & 1 \\ 1 & 9 & 1 \end{bmatrix}$$

for female subjects and

$$\mathbf{X}_i = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 4 & 0 \\ 1 & 5 & 0 \\ 1 & 6 & 0 \\ 1 & 7 & 0 \\ 1 & 8 & 0 \\ 1 & 9 & 0 \end{bmatrix}$$

for male subjects. The estimated autocorrelation parameter, $\hat{\phi} = 0.95278$, which is quite large. Fitted values for the males and females are produced using the CAR1 model without a random slope to explore the fixed effects across the two groups. Figure 4.5 shows a plot of the fitted values for the two groups:

The average BMI values of females is higher than that of males from the start of the study and this gap is constantly maintained through out the observation period. Figure 4.6 shows that the residuals are normally distributed.

4.3 Incomplete data analysis

As an initial step, Trellis plots of the two groups of patients (Figures 4.7 and 4.8), were obtained, with missing values were set at zero so as to visualise the dropouts on the plots. Generally there seems to be no distinct difference in the way dropouts occurred for between the two gender groupings.

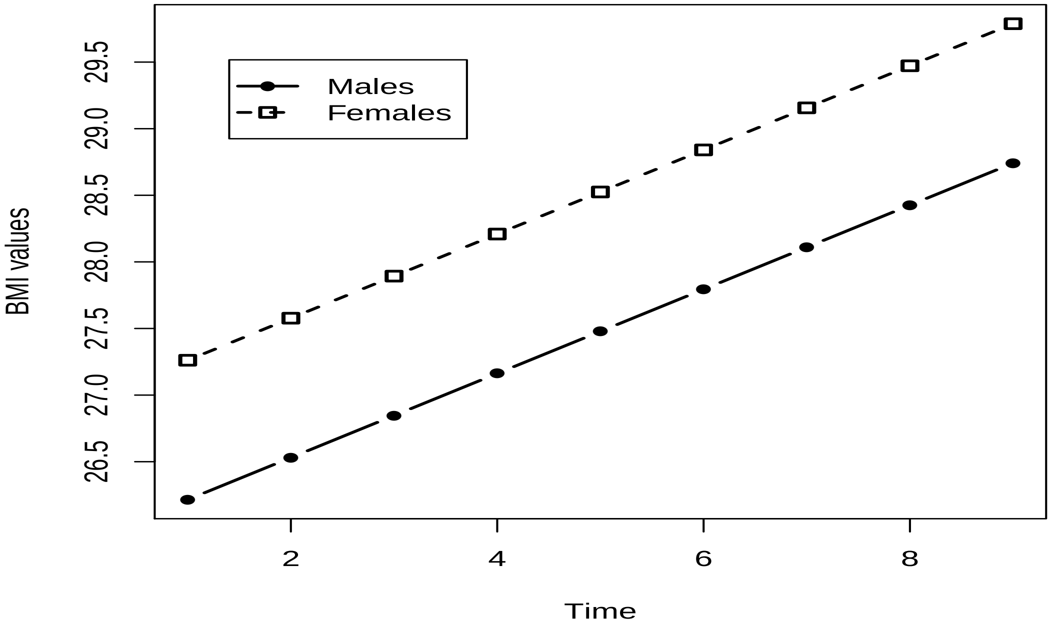

Figure 4.5: Fitted values representing the effects of time

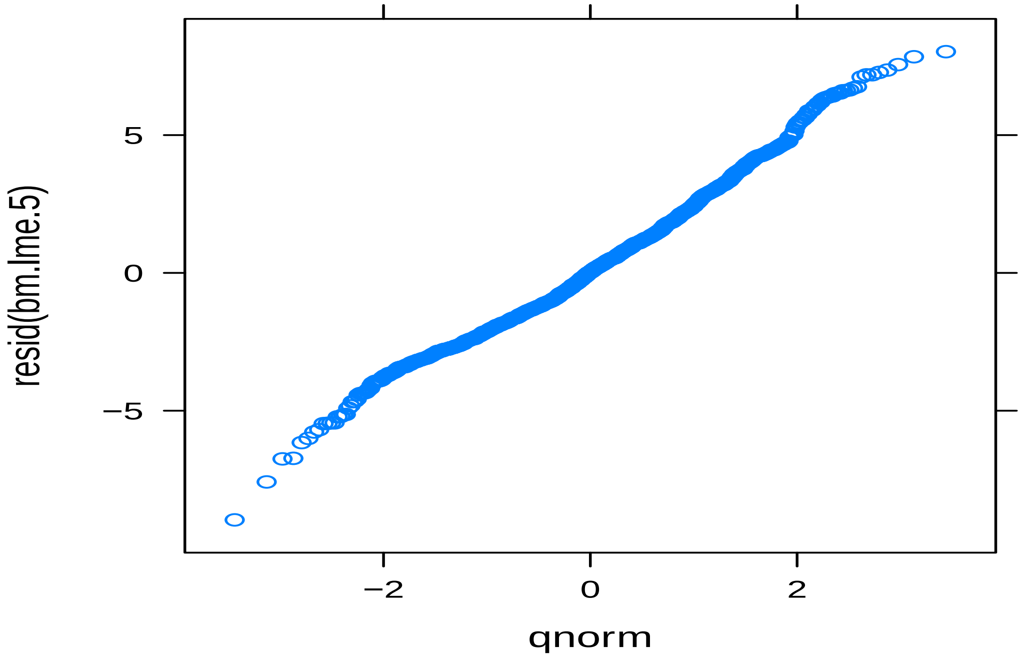

Figure 4.6: Q-Q plot for residuals

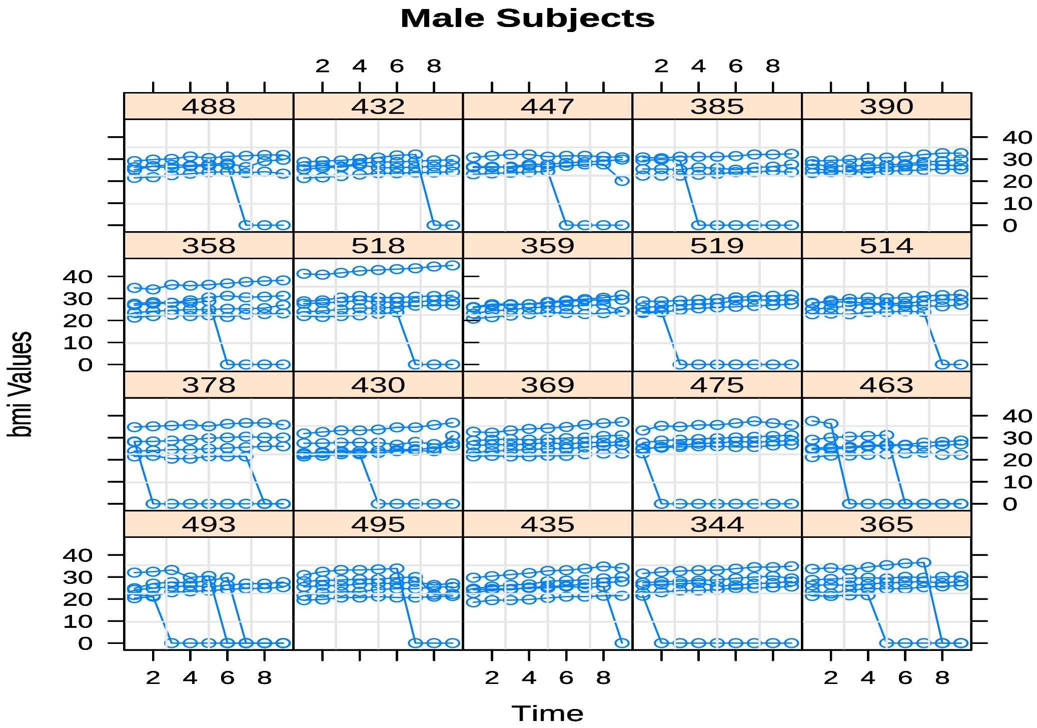

Figure 4.7: Trellis plots for male patients

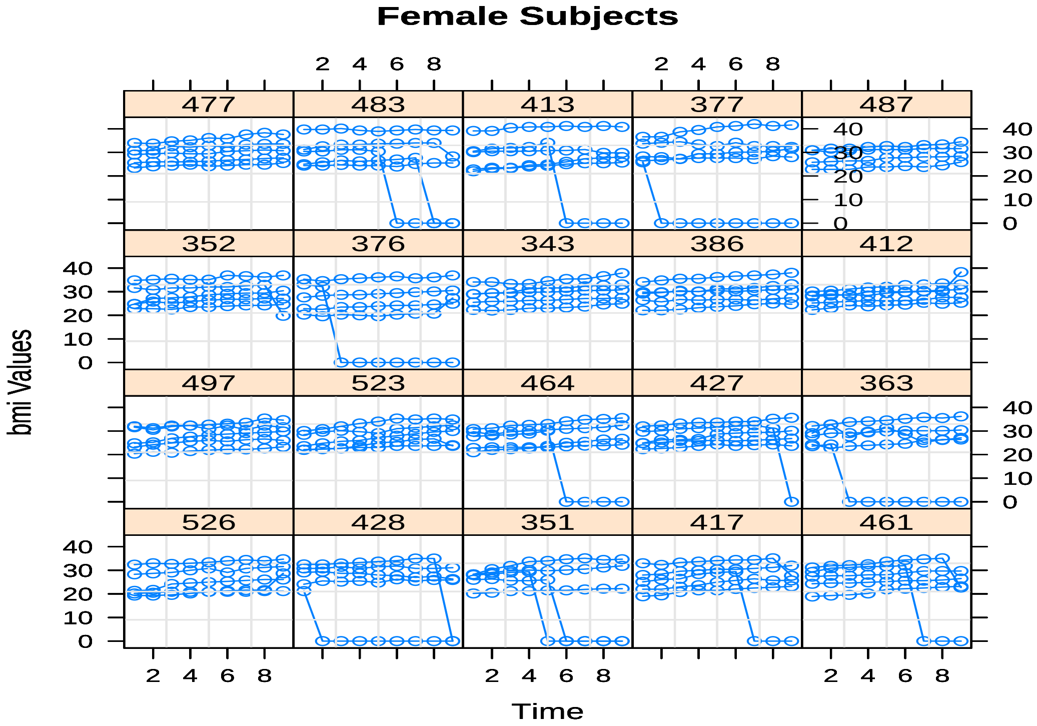

Figure 4.8: Trellis plots for female patients

4.3.1 Complete case analysis

Discarding all incomplete sequences and fitting a linear mixed model with first order auto regression CAR(1), the following output was obtained:

```
Linear mixed-effects model fit by REML
 Data: d5
      AIC      BIC   logLik
  4888.68 4920.784 -2438.34
Random effects:
 Formula: ~1 | serialno
        (Intercept) Residual
StdDev:    3.202934 3.080305

Correlation Structure: Continuous AR(1)
 Formula: ~time | serialno
 Parameter estimate(s):
      Phi
0.9512138
Fixed effects: bmNew2 ~ time + sex
                Value Std.Error   DF  t-value p-value
(Intercept) 25.846433 0.4594018 1368 56.26106  0.0000
time         0.313317 0.0239154 1368 13.10104  0.0000
sexFemales   1.148478 0.6197868  189  1.85302  0.0654
 Correlation:
           (Intr) time
time       -0.234
sexFemales -0.699 -0.008

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.8927681 -0.4456124  0.0201981  0.4600157  2.4357658

Number of Observations: 1560
Number of Groups: 191
```

The following estimates were obtained:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 25.846 \\ 0.3133 \\ 1.1484 \end{bmatrix}$$

the random effects variance $d_{11} = 10.25856$, autocorrelation parameter $\phi = 0,9512$.

4.3.2 Selection model for the diabetes data

Assuming separability condition to be met, modelling is done in two stages, modelling the measurement process and modelling the dropout process.

Modelling the BMI profiles

Using the EM algorithm built in the Linear mixed models ("**lmm**" library by Schafer J.L (2009)), the following results were obtained for the estimates of $(\boldsymbol{\beta}', \phi)$

```
$beta
        int         time       gender
25.844756       0.3112734        1.0901032

$sigma2
[1] 3.0717747

$psi
         int
int 10.75789
$phi
[1] 0,92356
$converged
[1] TRUE
$iter
[1] 16
$loglik
 [1] -2688.206 -2033.486 -1805.148 -1662.309 -1633.426 -1627.091 -1626.209
 [8] -1626.196 -1626.070 -1626.068 -1626.068 -1626.068 -1626.068 -1626.068
```

```
[15] -1626.068 -1626.068
```

The log likelihood is seen decrease from first iteration to the fourth iteration there after an exponential decay is seen in the log likelihood figures which stabilises at a value of -1626.068. The EM algorithm converged in 16 cycles giving

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 25.645 \\ 0.278 \\ 0.901 \end{bmatrix}$$

the random effects variance $d_{11} = 10.75789$, autocorrelation parameter $\phi = 0,92356$.

Modelling the dropout process

In this research, dropout is restricted to relate to the factor hypertensive and a covariate age in years and current and previous observations only. In line with Diggle and Kenward (1994), we assume that the probability of a dropout at occasion j, $(j = 2, \ldots, n_i)$, given that the subject was still in the study the previous occasion follows a logistic model

$$logit[P(D_i = j|D_i \geq d)] = \varphi_0 + \varphi_1 BMI_{ij} + \varphi_2 BMI_{i,j-1} + \varphi_3 hypertens_{ij} + \epsilon_{ij}$$

4.3.3 Pattern-mixture model /Shared parameter model for the diabetes data

We will consider Roy (2003)'s Latent Dropout Class Model. According to Roy (2003) the Latent Dropout class model shares properties of both Pattern-mixture model and Shared parameter model and will therefore consider it as a representative of both types of models. Assuming the dropout pattern to be sampled from multinomial distribution with support $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, where class $T_i = 9$ contains all completers. The associated multinomial probability vector is denoted by $\pi = (\pi_1, \, \pi_2, \, \cdots, \pi_9)'$. The model for Y_i^o conditional on T_i is of the form 4.2.1.

Dropout occasion:t	1	2	3	4	5	6	7	8	9
Fitted prob:$\hat{\pi} = P(T_i = t)$	$\frac{5}{191}$	$\frac{5}{191}$	$\frac{1}{191}$	$\frac{3}{191}$	$\frac{8}{191}$	$\frac{6}{191}$	$\frac{5}{191}$	$\frac{3}{191}$	$\frac{155}{191}$

Table 4.1: Fitted probabilities under the Multinomial dropout model

Let the latent dropout classes be, **early** for those subjects which drop out on the second or third occasion, **mid** for those which dropout on the fourth, fifth or sixth occasion, **late** for those which drop out on the seventh or eighth occasion and **completers** for those which did not dropout. Consider p_{ij} the probability of being in dropout class less or equal to k where $k = 1, 2, 3, 4$. Then $\hat{p}_{ij}$ is distributed as follows:

k	1	2	3	4
$\hat{p}_{ij}$	$\frac{10}{191}$	$\frac{22}{191}$	$\frac{36}{191}$	$\frac{191}{191}$

Table 4.2: Cumulative probabilities for dropout classes

These probabilities can then be modelled using probit link, with the drop out time as a covariate. The conditional distribution of $\mathbf{Y}_i$ on latent class is given by the conditional distribution of $\mathbf{BMI}_i$ given ξ_i is

$$(\mathbf{BMI}_i | \xi_i) \sim N(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i(\xi)\alpha, \, V_i(\boldsymbol{\phi}))$$

Fitting the model using Schafer (2009)'s **lmm** package, in which a dummy variable is included for each dropout class gave the following output:

```
> ecmeml.result
beta
int       time     gender   beta.hat[1]  beta.hat[2]  beta.hat[3]   beta.hat[4]
25.594776 0.300751  0.910328    0.889131   0.751231    0.541465    0.321489
sigma2
[1] 9.559070
phi
[1] 0,955634
psi

[1] 10.516733
converged
[1] TRUE

iter
[1] 8

loglik
[1] -2458.209 -2247.236 -1789.412 -1627.283 -1621.271 -1592.063 -1592.063
[8] -1592.063
```

4.4 Comparison of the estimates

The main focus of this research is on the measurement process, that BMI values estimates under various models. Since the dropout pattern was a simulated one, it will only serve as a place holder for modelling purposes but it has no practical meaning. Estimates from the model on the complete data set will be the comparing factor and therefore are compared to the estimates obtained from complete case analysis, selection model and latent dropout class model. The table below shows the estimates obtained from the different models and their percentage deviation from the complete data model estimates. The complete data average BMI value for females at baseline is 26.945 compared to an average of 25.899 for males. The change in BMI value from one point in time to the next was found not to be significantly different for the gender groupings, with a universal increment of 0.3458 for both male and female subjects. The variance for the random effects was found to be 9.684 and the $AR(1)$ parameter, $\phi = 0.9528$. There is a mixture of results in terms of how the estimates from the different models on incomplete data compare to the estimates from the model on the complete dataset. There is variation in performance from one estimate comparison to another and there seems to be no clear trend on performance of the models in terms

$Model\ (estimates)$	**Complete data** LMM	**Complete case** LMM	**Selection Model**	**Latent dropout class model**
Parameter				
$\beta_0(constant)$	25.899	25.846(-0.2%)	24.847(-4.1%)	24.595(-3.2%)
$\beta_1(time)$	0.346	0.313(-9.4%)	0.311(-10.0%)	0.301(-13.0%)
$\beta_2(sexF)$	1.047	1.148(9.7%)	1.090(4.2%)	0.910(-13.0%)
$\beta_{11}(early)$	-	-	-	0.889
$\beta_{12}(mid)$	-	-	-	0.751
$\beta_{13}(late)$	-	-	-	0.541
$\beta_{14}(completer)$	-	-	-	0.321
$d_{11}(r.e\ var)$	9.684	10.259(5.92%)	10.758(11.1%)	10.577(9.2%)
$\sigma(res.\ var)$	3.211	3.080(-4.1%)	3.072(-4.3%)	3.092(-3.7%)
$\phi(autoregr)$	0.953	0.952(-0.2%)	0.924(-3.1%)	0.956(0.3%)
$loglik$	-2696.20	-2438.34	-1626.068	-1592.063

Table 4.3: Estimates from the models

of the ”bias” looked at.

Chapter 5

Conclusions and Recommendations

5.1 Conclusions

The major aim of this research was to compare estimates from different modelling approaches, with the main focus being comparing joint models to complete case basing on how their estimates compared with the complete data model estimates. The data nature and the experimental question to be answered gave the direction and main focus of the research. Considering that initially the data produced a complete data array for the subjects and the dropouts were simulated, it was reasonable to focus on the measure process estimation, which is the *bmi* profiles for the subjects by gender over time. Since dropouts were simulated, it would be difficult to attach a practical meaning to them, but had they been actual observed it would have been of importance to also focus on drop out models.

The linear mixed modelling was implementation was not much of a challenge because of the the availability of several R libraries that can fit these models, which include ***Lmm*** by J. L. Schafer (May 2009), ***Lme4*** by D. Bates (July 2008), ***nlme*** by J. Pinheiro *at al* (2009), Linear mixed models were fitted well on the complete data set and the complete case analysis. The complete case estimates were very close to the complete data estimates. However it is difficult for this researcher to conclude that complete case analysis performs better than the other models. The EM algorithm for the Latent Dropout Class model converged faster than the algorithm for the selection model, actually the Latent Class Dropout model algorithm was twice as faster as the algorithm for the selection model. For all models the average profiles at baseline for both male and female subjects fell in the overweight class and there is a steady increment as we move in time.

The main reason for comparing the models on the incomplete data set to the model on the complete data set was come a conclusion on which model produced estimates that are more closer to the full data estimates. This conclusion can not be explicitly reached since there was no clear cut trend in performance of the models and looking at fact that this comparison was just in terms of percent deviation of incomplete data estimate from the from the full data estimate. This researcher feels one could reach to a solid conclusion after considering different proportions of dropouts and also different patterns in which the dropouts are distributed throughout the study period. The log likelihood variations of the models indicates that joint models are better and

I think with a finer refinement on assumptions of the models and also consideration of the dropout model estimates a more satisfying conclusion could be reached on the comparison of the complete case analysis to the joint modelling for the measurement process and the dropout process under non-ignorable dropouts. In most practical situations, the assumptions on randomness or non randomness of the dropout process are sometimes in-testable and therefore the assumption of non-ignorable dropouts is a conservative approach. Though in this research the complete case analysis might have produced estimates more closer to the joint models I still feel with refinement of the model assumptions and considerations the joint models might perform better. And of course the basis of comparison in this research might not be of much statistically significance, and the findings of this research can serve as a foundation to deeper comparisons. However it is not in all cases that the joint models are necessary, the remedies can fare better especially considering that the joint models can make some rigid and in-testable assumptions.

5.2 Recommendations

The joint modelling of the measurement process and the dropout process is of great importance to most fields of study like in biology, medicine and other fields where the subjects are living organism, and in such cases if not due to censoring, a dropout might be due to loss of life. It becomes of paramount importance to establish if the dropout is or is not related to the measurement process. under these circumstances modelling of the measurement process and of the dropout processes are both of great importance.

Incomplete longitudinal data poses challenges related to sensitivity of modelling assumptions and therefore there is need for a sensitivity analysis. Molenberghs *et al* (2000) defines sensitivity analysis as one in which several statistical models are simultaneously fitted and/or where a statistical model is further scrutunised using specialised tools. This entails fitting a selected number of (non random) models which are deemed plausible or in which a preferred (primary) analysis is supplemented with a number of variations. Though different analysis methods are likely to have distinct impacts on conclusions, I think there can be found a common ground for comparing these models. No one approach may be said to cover all forms in which the practical problem pose to researchers.

I recommend that modelling should be done under all the three missing mechanisms and them comparisons of estimates cane be made linking to the practical question of the experiment unless in situations where randomness can be verified.

Bibliography

1. Bates D.M. and Watts D.G.(1988), *Nonlinear regression analysis and its Applications*, Wiley, New York.

2. Cox D.R. and Oakes D.(1984), *Analysis of survival data*, Chapman and Hall, London

3. Dalgaard Peter(2002), *Introductory Statistics with R*, Springer-Verlag, New York

4. Diggle P.J.(1988),An approach to analysis of repeated measures. *Biometrics*, **44**, 959-971.

5. Diggle P.J., Heagerty Patrick J.,Zeger and Liang Kung-Yee(2002), *Analysis of Longitudinal Data*, Second Edition, Oxford University Press.

6. Diggle P.J. and Kenward M.G.(1994), Informative Drop-out in Longitudinal Data Analysis, *Applied Statistics*, **43**, 49-73.

7. Fitzmaurice G.M. and Birmingham Jolene(2002), A Pattern-Mixture Model for Longitudinal Binary Responses with Nonignorable Nonresponse, *Biometrics* **58**, 989-996.

8. Hogan J.W. and Laird N.M.(1997), Mixture models for joint distribution of repeated measures and event times, *Statistics in Medicine*, **16,** 239-258.

9. Little R.J.A.(1993), Pattern -mixture models for multivariate incomplete data, *Journal of American Statistical Association*, **88**, 125-134.

10. Little R.J.A. and Rubin D.B (1987), *Statistical analysis with missing Data*, John Wiley, New York.

11. Little R.J.A. (2003), *Statistical analysis with missing Data* (2*nded.*), John Wiley, New York.

12. Molenberghs G., Kenward M.G., and Lesaffre E., (1997) The analysis of ordinal longitudinal data with informative dropouts. *Biometrika*, **84**, 33-44.

13. Molenberghs G., Michiels B., Kenward M.G., and Diggle P.J. (1998) Missing data mechanism and pattern-mixture models. *Statistica Neerlandica*, **52**, 153-161.

14. Molenberghs G. and Verbeke G.(2005), Models for Discrete Longitudinal Data, Springer-Verlag 2005.

15. Plewis I.(1985), *Analysing change*, John Wiley, New York.

16. Robins J.M.(1987), Non response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine*, **16**, 21-38.

17. Roy J.(2003), Modelling Longitudinal Data with Nonignorable Dropouts , *Biometrics* **59**, 829-836.

18. Rubin D.B.(1976) Inference and missing data, *Biometrika*, **63**, 581-592.

19. Rubin D.B.(1977) Bayesian inference for casual effects: the role of randomisation., *Annals of Statistics*, **54**, 221-226.

20. Stiratelli R. Laird N. and Ware J.(1984), Random effects models for serial observation with dichotomous response. *Biometrics*, **40**, 961-972.

21. Verbeke G and Molenberghs G.(2000), *Linear Mixed Models for Longitudinal Data*, Springer-Verlag 2000.

22. Verby A.P. and Cullis B.R.(1990) Modelling in repeated measures experiments. *Applied Statistics*, **39**, 341-356.

23. White H.(1982) Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1-25.

24. Yang Xiaowei, Li Jinhui and Shoptaw Steven (2006), Multiple Partial Imputation for Longitudinal Data with Missing values in Clinical Trials, *Statistical Methods in Medical Research*, **59**, 33-45.

25. Yi Grace Y. and Thompson Mary E.(2005), Marginal and association regression models for longitudinal binary data with drop-outs:a likelihood-based approach, *The Canadian Journal of Statistics* **Vol 33**, No. 1 , 3-20.

26. Zeger S.C, Liang K.Y. and Albert P.S.(1988), Models for Longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049-1060.

5.3 R codes (Syntax)

Importing the data

```
> library(foreign) # Loading the required package to import data.
> library(nlme) # Loading the non-linear mixed effects models library.
> d1=read.spss("C:/diabetes.sav", use.value.labels=TRUE) # Importing
an spss data file
> attach(d1) # Putting the file in the search path.
> d2<-data.frame(d1) #Converting to data frame format.
> d2
```

Complete data set analysis **Profile plots for the complete data set**

```
> library(longitudinalData) #Required package
> bmi1= array(bmi, c(9,191))# Restructure bmi values : rows
showing time series data for a subject.
> bmit = t(bmi1) # Transpose:columns represent individual profiles.
> bmilong = as.longData(bmit,id =1:191,timeCol = c(1:9))# Converting
 to longData class (data format that can be handled by package)
> part = partition(clusters=c(rep("M",92), rep("F",99),nbClusters=2))
#Partion data by gender, M-males, F-females (Note: Data already sorted
by gender).
> plot(bmilong, type= "l" ,xlab = "Time", ylab = "Bmi",
+ main= "Individual profile plots") # Plot individual profiles for
all subjects on one plot.
> plotSubGroups(bmilong, part, type= "l" ,xlab = "Time", ylab = "Bmi",
+    main= "Individual profile plots ) # plot individual profiles by
gender

Profile plots
> library(lattice)
> xyplot(bmi ~ time | serialno, d3,type="l", strip=FALSE)

Trellis plots
> Msample<-sample(unique(d2$serialno[d2$sex=="Males"]),92)

> Malegroup<-groupedData(bmi ~ time | serialno,
+     data=d2[is.element(serialno, Msample),])

> Fsample<-sample(unique(d2$serialno[d2$sex=="Females"]),99)
> Femalegroup<-groupedData(bmi ~ time | serialno,
+     data=d2[is.element(serialno, Fsample),])


> print(plot(Malegroup, main= "Male Subjects",
+      xlab="Time", ylab="bmi Values",
+      layout= c(5,4), aspect=1.0),
+      position= c(0, 0, 1 , 1), more=T)
```

Modelling

```
> d3=data.frame(d1)
> Msample1<-sample(unique(d3$serialno[d3$sex=="Males"]),92)
> Malegroup1<-groupedData(bmi ~ time | serialno,
+     data=d3[is.element(d3$serialno, Msample1),]) # All Male
 patients
> Fsample1<-sample(unique(d3$serialno[d3$sex=="Females"]),99)
> Femalegroup1<-groupedData(bmi ~ time | serialno,
+     data=d3[is.element(d3$serialno, Fsample1),]) # All
Female patients
> mlist<-lmList(bmi ~ time | serialno,
+     subset= sex=="Males", data=d3)   # Fitting a
regression(within individual) of bmi values on time for males
> flist<-lmList(bmi ~ time | serialno,
+     subset= sex=="Females", data=d3) # Fitting a regression
 of bmi  values on time for females
> mlist.coef<-coef(mlist) # Array of regression coefficients
for the male group.
> flist.coef<-coef(flist) # Array of regression coefficients for
the male group.

> new<-par(mfrow= c(1,2)) # Arrangement format on a page

> boxplot(mlist.coef[,1],flist.coef[,1] ,main="Intercepts",
+     names=c("Males", "Females")) # Boxplot of Intercepts
> boxplot(mlist.coef[,2],flist.coef[,2] ,main="Slopes",
+      names=c("Males", "Females")) # Boxplots of Slopes
> par(new)

> t.test(mlist.coef[,1],flist.coef[,1]) T-test for intercepts.
> t.test(mlist.coef[,2],flist.coef[,2]) T-test for  slopes.
```

```
Fitting a linear mixed model
> bm.lme.1<-lme(bmi ~ time,  # Time as the major covariate.
+           random = ~ time |serialno,# Time as random effects
+           data = d3)

summary(bm.lme.1) # Detailed model output.

> bm.lme.2<-update(bm.lme.1, random = ~ 1|serialno)#
Refitting a lmm without the random slope.
> bm.lme.3<-update(bm.lme.1, random = ~ time -1|serialno)#
Refitting a lmm without the random intercept.
> anova(bm.lme.1, bm.lme.2)# Testing  for significance of the
random slope.
> anova(bm.lme.1, bm.lme.3) # Testing for significance of
the random intercept.

> bm.lme.4<-update(bm.lme.1, correlation = corCAR1(form=~ time |serialno))
# Factoring in autocorrelation (AR(1)).

> anova(bm.lme.1, bm.lme.4) # Testing for the necessity of autocorrelation.

> bm.lme.5<-update(bm.lme.4, random = ~ 1|serialno) # Autocorrelated mixed
 model without random slope.
> bm.lme.6<-update(bm.lme.4, random = ~time- 1|serialno) # Autocorrelated
mixed model without random intercept.

> anova(bm.lme.4, bm.lme.5)  # Testing for the necessity of random slope
in the autocorrelated mixed model.
> anova(bm.lme.4, bm.lme.6)  # Testing for the neccesity of the random
intercept in the autocorrelated mixed model.
> summary(bm.lme.5) # Summary of the fitted model(adopted).

> bdata<-expand.grid(time=seq(1 ,9, by=1),sex=c("Males", "Females"))
# Fitted values.
> bdata$bmi<-predict(bm.lme.5, bdata, level=0) # level=0 produces
estimates for fixed effects.


> plot(bdata$time, bdata$bmi, type="n",
+ xlab="Time", ylab="BMI values")
> points(bdata$time[1:9], bdata$bmi[1:9], type="b", pch=16, lwd=2)
> points(bdata$time[10:18], bdata$bmi[10:18], type="b",pch=0,lty=2,lwd=2)
> legend(locator(1), c("Males", "Females"),pch=c(16,0), lty=c(1,2),lwd=2)
```

Incomplete data analysis

```
> for(i in 2:1719){       # Begin loop
+ if(d2$hypertens[(i-1)]=="Non-hypertensive"&d2$hypertens[i]=="Hypertensive"
 & d2$serialno[i]==d2$serialno[(i-1)])  d2$bmi[i]<- 0# Inducing drop out
 times.
+              else d2$bmi[i]= d2$bmi[i]
+ }                                 # end loop
> d2$bmi

> for(i in 2:1719){
+ if (d2$serialno[i] ==d2$serialno[(i-1)] & d2$bmi[(i-1)]==0)
d2$bmi[i]<-0 else d2$bmi[i]=d2$bmi[i]
+ } # Inducing missing values on all observation of the same subject
after a missing value.
> d2$bmi

> bmNew<-d2$bmi
> ind2=(bmNew <1)
> bmNew[ind2]=NA  # replacing zero's with NA
```

Constructing the missing value indicator vector r

```
> r<-d2$bmi
> ind3=(r>0)
> r[ind3]=1
```

Constructing a vector with length of series for each subject $\mathbf{T}_i$

```
> d3<-data.frame(d2[,-2],bmNew,r)
> T_i<-c()
> r1= array(r, c(9,191))
> rt = t(r1)
> T_i<-rowSums(rt)
Obtag profile plots for the incomplete data set
```

```
> bmi2= array(bmNew, c(9,191))
> bmit2 = t(bmi2)
> bmilong2 = as.longData(bmit2, id =1:191, timeCol = c(1:9))
> part = partition(clusters=c(rep("M",92), rep("F",99),nbClusters=2))
> plotSubGroups(bmilong2, part, type= "l" ,xlab = "Time", ylab = "Bmi",
+main= "Individual profile plots ) # plot individual profiles by gender
```

Modelling

Complete case analysis

```
> bmNew2<-bmNew[!is.na(bmNew)]
> bmNew1=!is.na(bmNew)
> serialno<-d2$serialno[bmNew1]
> bmNew2<-bmNew[!is.na(bmNew)]
> bmNew1=!is.na(bmNew)
> serialno<-d2$serialno[bmNew1]
> sex<-d2$sex[bmNew1]
> T_iNew<-T_i[bmNew1]
> R_New<-r[bmNew1]
> time<-d2$time[bmNew1]

> d5<-data.frame(d2$serialno, d2$sex, bmNew, r, T_i, d2$time)
> d5<-data.frame(serialno, sex, bmNew2, R_New,  time)
> bm.lme.5mis<-lme(bmNew2 ~ time + sex,
+                  random=~1|serialno,
+                  correlation=corCAR1(form=~time|serialno),
+                  data=d5)
>
```

Selection modelling

```
> pred<-cbind(int=rep(1,1719),time=d2$time,gender=1*(d2$sex=="Females"))
> xcol=1:3
> zcol=1
> ecmeml.result <- ecmeml.lmm(bmNew2,serialno,pred,xcol,zcol)
```

The dropout process model

```
> glm.drop<-glm(drop.tbl ~ hypertens_j+bmi_j+bmi_{j-1}, binomial}
Latent dropout class
pred <- cbind(int=rep(1,1719),time=d2$time,gender=1*(d2$sex=="Females"),
+ dummy1=1*(ldc==1),dummy2=1*(ldc==2),dummy3=1*(ldcc==3),dummy4=1*(ldcc==4))
xcol <- 1:7
zcol <- 1

>ecmeml.result <- ecmetml.lmm(bmNew2,serialno,pred,xcol,zcol)
```

5.4 Results

T-tests for comparing slopes and intercepts for males and females

```
 Welch Two Sample t-test
data:  mlist.coef[, 2] and flist.coef[, 2]
t = 0.3169, df = 187.175, p-value = 0.7516
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.06042439  0.08355607
sample estimates:
mean of x mean of y
0.3297645 0.3181987
```

T-tests for comparing average BMI value at the start of the research period

```
 Welch Two Sample t-test
data:  mlist.coef[, 1] and flist.coef[, 1]
t = -1.8605, df = 187.919, p-value = 0.06437
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.47297561  0.07234185
sample estimates:
mean of x mean of y
 25.92171  27.12203
```

Linear mixed model without the interaction term:

```
 Linear mixed-effects model fit by REML
 Data: d3
       AIC      BIC    logLik
  5582.571 5620.705 -2784.286
Random effects:
 Formula: ~time | serialno
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 4.4394672 (Intr)
time        0.2300857 -0.247
Residual    0.8149142

Fixed effects: BMI ~ time + sex
                Value Std.Error   DF  t-value p-value
(Intercept) 25.950531 0.4579313 1527 56.66905  0.0000
time         0.323770 0.0183062 1527 17.68632  0.0000
sexFemales   1.144711 0.6242038  189  1.83387  0.0682
 Correlation:
           (Intr) time
time       -0.192
sexFemales -0.707  0.000

Standardized Within-Group Residuals:
          Min            Q1           Med            Q3           Max
-10.818759145  -0.298818811   0.008918353   0.322969586   6.335455115

Number of Observations: 1719
Number of Groups: 191
```

CAR1 model

```
Linear mixed-effects model fit by REML
 Data: d3
       AIC      BIC    logLik
  5404.411 5447.993 -2694.205
Random effects:
 Formula: ~time | serialno
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 4.2444606 (Intr)
```

```
time        0.2014551 -0.207
Residual    1.4459284

Correlation Structure: Continuous AR(1)
 Formula: ~time | serialno
 Parameter estimate(s):
      Phi
0.7764626
Fixed effects: BMI ~ time + sex
                Value Std.Error   DF  t-value p-value
(Intercept) 25.930210 0.4585112 1527 56.55305  0.0000
time         0.317344 0.0225144 1527 14.09516  0.0000
sexFemales   1.099906 0.6194670  189  1.77557  0.0774
 Correlation:
           (Intr) time
time       -0.232
sexFemales -0.700  0.000

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-4.95249419 -0.23640978  0.01752392  0.30762015  4.10779415
Number of Observations: 1719
Number of Groups: 191
```