Heidi Elisabeth Sando

Statistical methods for the analysis of randomized controlled trials with repeated measurements and missing data

Master's thesis in Applied Physics and Mathematics
Supervisor: Turid Follestad
January 2019

Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Heidi Elisabeth Sando

Statistical methods for the analysis of randomized controlled trials with repeated measurements and missing data

Master's thesis in Applied Physics and Mathematics
Supervisor: Turid Follestad
January 2019

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Summary

Randomized controlled trials (RCTs) is a widely used method in medical research. The method aims to analyze the effect of a treatment over time. This type of research is exposed to missing data. The aim of this report is to compare different methods to analyze longitudinal RCTs with different assumptions about the missing data. The methods used to analyze RCTs are: Comparisons of follow-up scores, change score analysis, analysis of covariance (ANCOVA) and constrained longitudinal analysis (cLDA). The reason for why data are missing is one of the main challenges associated with missing data. This is called the missing data mechanism. The different missing data mechanisms are: Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The percentage of missing data is also of importance when analyzing missing data. Different scenarios of missing data in longitudinal trials are simulated to compare the methods used to analyze RCTs together with different methods to handle missing data. The methods are compared with respect to bias, power and confidence interval coverage. The methods of cLDA and ANCOVA are also applied on a real clinical trial. The methods of comparisons of follow-up score and change score analysis are most commonly used on RCTs with only two time points. The results of the method of change score depends highly on the correlation between the time points, while the method of follow-up does not take the correlation or the baseline values into account. With low correlation the method of follow-up is just as good as the method of ANCOVA and cLDA, and superior to the method of change score. With high correlation the method of change score analysis is equally good as the method of ANCOVA and cLDA, and superior to the method of follow-up. In general the methods of ANCOVA and cLDA are superior to the other methods. With no missing data produces the methods equally good results. When there are missing data more information is used in the analysis using the method of cLDA compared to the method of ANCOVA. Thus, the method of cLDA may be regarded as the method of choice compared to the method of ANCOVA. However, when multiple imputation is used together with the method of ANCOVA are the results comparable to the method of cLDA. The missing data mechanisms MAR and MCAR can produce unbiased estimates when using the methods of ANCOVA and cLDA, but the power decreases when there are missing data. The power of the methods also decreases when the percentage of missing data increases. In addition, when the percentage of missing data reaches 20%, the results may be biased, regardless the choice of method.

Preface

This project constitutes the master thesis in Industrial Mathematics at the Norwegian University of Science and Technology (NTNU).
The topic of this thesis is to analyze longitudinal randomized controlled trials with missing data. My motivation for this project is to combine my interest for medical research with my interest for statistics.
I would like to give great thanks to my supervisor, Turid Follestad, for all the advice she has given me and all hours spent guiding me through this project. I would also like to thank my co-supervisor Mette Langaas, which have given me good feedback during the process. Thanks to Charlotte B. Ingul for letting me use her material to analyze a real clinical trial. At last, special thanks to Audun Tufte Larsen and Bergitte Viste for using many hours for proofread my thesis.

Contents

List of Tables

List of Figures

Abbreviations

ANCOVA	=	Analysis of covariance
BMI	=	Body mas index
CI	=	Confidence interval
cLDA	=	Constrained longitudinal data analysis
HIIT	=	High intensity interval training
ICC	=	Intraclass correlation
ITT	=	Intention-to-treat
LDA	=	Longitudinal data analysis
LMM	=	Linear mixed models
LOCF	=	Last observation carried forward
MAR	=	Missing at random
MCAR	=	Missing completely at random
MI	=	Multiple imputation
MICT	=	Moderate intensity continuous training
MLE	=	Maximum likelihood estimator
MNAR	=	Missing not at random
RCT	=	Randomized controlled trials
REML	=	Restricted maximum likelihood estimation
GLS	=	Resting left ventricular global longitudinal strain
LVS	=	Resting left ventricular peak systolic tissue velocity

Chapter 1

Introduction

This project aims to describe and compare methods to analyze longitudinal randomized controlled trials (RCTs) with missing data. RCT is a method to conduct clinical trials with the aim to identify the effect of a treatment. This is done by measuring an outcome variable of interest before intervention and by doing several follow-up measurements after intervention. The outcome variable is assumed to be continuous. The methods for analyzing RCTs that are presented and compared are: Comparison of follow-up score, change score analysis, analysis of covariance (ANCOVA) and constrained longitudinal data analysis (cLDA). Missing data in clinical trials are unavoidable. Because of their frequent occurrence they must be accounted for such that the results will not be affected. The reason behind missing data is called the missing data mechanism, which can highly affect the results. The missing data mechanism is divided into three types: Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). To overcome the bias potentially caused by the missing data mechanism valid methods to handle this should be used, depending on the type of missing data mechanism. Invalid methods for handling missing data are commonly used in medical research. Methods to handle the missing data mechanisms MAR and MCAR will be the main focus in this project. Especially the methods of multiple imputation (MI) and mixed-effects models. The different methods to analyze RCTs are compared by conducting simulation studies with different missing data mechanisms and where different methods to handle missing data are used. The methods are in addition used on a real clinical trial with missing data.
The structure of this project is as follows: In Chapter 2 RCTs and methods to analyze RCTs are introduced. In Chapter 3 the mathematical background of longitudinal data is presented in addition to the statistical background used in the simulation study. In Chapter 4 the concept of missing data is introduced, including the description of missing data mechanisms and methods to deal with missing data. In Chapter 5 the methods used to analyze RCTs are described mathematically. In Chapter 6 the simulation studies that are conducted are described and the results are presented. In Chapter 7 the real clinical example and the results of the analysis are presented. At last, in Chapter 8 the simulations are discussed.

Chapter 2

Randomized controlled trials

A randomized controlled trial (RCT) is a scientific experiment with the aim to analyze the effect of a certain variable introduced in the trial (Matthews, 2006, p. 1). The variable introduced may be a new type of treatment to improve a certain medical condition for a patient. Further on, this variable will be referred to as ”the treatment”. Analyzing trials using the method of RCT is regarded as ”the gold standard” when conducting clinical trials and is therefore widely used (Schulz et al., 2010). In an RCT, a sample of participants is collected and randomly divided into different groups, where the type of treatment differs between the groups. When proper randomization is achieved, the effect of the variable of interest is possible to compare between the groups. This randomization will minimize the selection bias (Matthews, 2006, p. 17). Selection bias is defined as the bias introduced if the selection of individuals in a trial is done in a way such that proper randomization is not achieved. Thus, the sample used in the trial is not representative for the population to be studied. The minimization of selection bias is one of the strengths of RCTs and is also an assumption when analyzing RCTs (Matthews, 2006, p. 17). In many medical trials confounders could be problematic. A confounder is a known or unknown underlying variable that is associated with both the exposure and the outcome. This is known as a factor that can have major impact on the results of a trial (Attia, 2005). Since the confounders are equally distributed between the groups it is not necessary to take it into account when analyzing RCTs.

Although RCTs often are the method of choice when conducting clinical trials are there some disadvantages. An RCT can be expensive both in time and cost. The effect of the treatment can usually only be able to assess the average effect of the whole sample, not on an individual level. RCTs are also limited by ethical questions. A variable that is known to be potentially damaging for the patients can never be investigated in an RCT. In addition it is not allowed to give an inferior treatment to a patient when better options are known (Matthews, 2006, p. 4-5).

There are different types of RCTs: A crossover RCT is a trial where more than one treatment is given to each participant (Matthews, 2006, p. 193-194). Another example is cluster randomized trial where individuals are not randomly divided into groups but whole clus-

ters of individuals are (Matthews, 2006, p. 202-203). The most common is RCT with parallel-group design. This is the type of RCT we are going to look further into.
The structure of the preceding sections of this chapter is as follows: In Section 2.1 is a presentation on RCTs are carried out. A description of a parallel-group RCT is presented in Section 2.2. At last, different methods to analyze RCTs are presented in Section 2.3.

2.1 How an RCT is carried out

In an RCT, a sample of participants is collected from a given population of interest depending on the type of the trial. When the sample of participants is found the participants are randomly divided into different groups. Because of the randomization of the groups it is not expected to be any differences between the groups at the beginning of the trial (**?**). Thus, the variable of interest should be equally distributed between the groups at the first measurement. After randomization and the first measurement the researcher introduces the treatment. The introduction of this variable is called the intervention. Since there are differences between the groups, the effect of the treatment can be analyzed. Examples of a treatment difference between the groups are getting a treatment or not getting a treatment (or getting placebo), giving different doses of the treatment or giving different types of treatments. Each participants is measured several times, where each measurement should be done in the same way, and should contain the outcome variable of interest. The outcome variable of interest should represent whether or not the treatment has an effect. The first measurement is called the baseline measurement, and is the measurement before intervention. At this point, due to randomization, the variable of interest and all other variables should be equally distributed in all groups. After baseline, the treatment is introduced and there should be one variable that differ between the groups which can cause an effect on the parameter of interest. The following measurements are conducted and called the follow-up measurements. When more time-points than two, i.e. there are more than one follow-up measurement in addition to the baseline measurement, the trial is called longitudinal RCT. Even though it is possible to have several follow-up measurements, a small amount of follow-up measurements is normal, typically $2 - 4$ follow-ups (Coffman et al., 2016). Trials with baseline measurements are commonly used to analyze the effects of treatments (Vickers and Altman, 2001; Liu et al., 2009). The differences of the variables of interest between the groups are analyzed in the end of the trial.

2.1.1 Intention-to-treat and per protocol analysis

When conducting an RCT, an intention-to-treat (ITT) analysis should be conducted. In an ITT analysis, all the participants that once are assigned to a group in the RCT should be a part of the group analysis, no matter how much of the trial that is completed (Hollis and Campbell, 1999). This is important if the proper randomization should be archived (Schulz and Grimes, 2002). If the participants that did not complete the trial were excluded in the analysis, or other arrangements for the drop-outs were made, the distributions of the groups could differ and bias could be introduced (Schulz and Grimes, 2002). Thus, an ITT analysis should be the primary analysis for the results, and is regarded as the gold standard (Veieroed et al., 2012). Without ITT the effect of the treatment can be overestimated. As a

secondary analysis, a per protocol analysis could be conducted. This analysis only includes the participants that completed the trial given their originally treatment (Sedgwick, 2013). This type of analysis can tell us something about the maximum potential benefits of the treatment given. But the results can be biased (Sedgwick, 2013). Therefore, per protocol analysis should never be presented alone, but only used as a secondary analysis (Schulz and Grimes, 2002).

2.2 Parallel-group RCT

An illustration of how a parallel-group RCT can be carried out is shown in Figure 2.1. In a parallel-group RCT, each participant is randomly allocated into one group. During

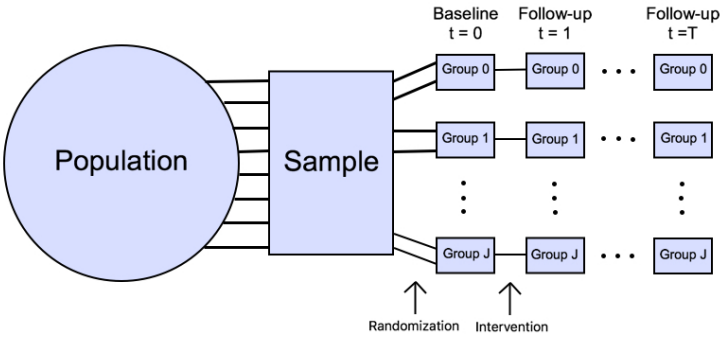


Figure 2.1: Model design for a parallel-group RCT.

the trial, the participants cannot change to another group, thus each participants is only introduced to one type of treatment. This is the main difference between parallel-group RCT and other types of RCTs. The aim of comparing different groups could be to detect the difference between a new and presumably better treatment compared to a traditional and more commonly used treatment. This is done by detecting the difference between the types of treatments and not getting a treatment at all, or by comparing the effect of different doses of the same treatment. One of the groups is usually a control group. This is the group not getting any treatment at all, getting placebo or the group getting the traditional treatment. The other groups are getting different types of treatments, and are called the treatment groups. In a parallel-group RCT, the following steps are as described in a general RCT: The participants in the groups are measured with respect to the outcome variable of interest at baseline, i.e. before intervention. The expected baseline outcome variables are assumed to be equal between the groups because of randomization. Then, the different treatments are given to the participants in the different groups. After a certain time, the first follow-up measurement is conducted in the same way as the baseline measurement with the same outcome variables of interest measured. This is repeated at several follow-up times if wanted. By comparing the differences between the groups before and after

intervention the effect of the treatment can be analyzed. The simplest form of a parallel-group RCT is an RCT with two groups and two time points. This situation is illustrated in Figure 2.2.

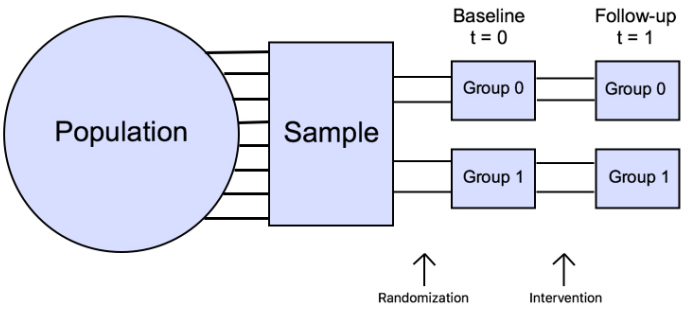


Figure 2.2: Model design for a parallel-group RCT with two groups and two time points.

The steps using a parallel-group RCT can be summarized as:

1. A population of patients that are of interest for the specific RCT is found.
2. A sample from the population is recruited to be a part of the experiment.
3. The sample is now formally entered in the program and is randomly divided into the control group and the treatment groups. From now on it is possible to get dropouts from the trial.
4. When randomized into groups, the first measurement of the variables of interest are often done at baseline before treatment is given.
5. Treatment is given to the different groups after baseline and before the first follow-up time point. This is called the intervention.
6. The first measurement after intervention is done at the (first) follow-up time point. If the RCT consists of several follow-up time points, the measurements are repeated until the end of the trial.
7. The results are analyzed at the end of the trial.

2.3 Methods to analyze parallel-group RCTs

Parallel-group RCTs can be analyzed by different methods. When analyzing RCTs with two time-points and two groups, there are more methods that are commonly used compared to analysis of longitudinal RCTs. We are going to look further into four different methods when analyzing RCTs with two time-points: Comparison of means at follow-up

score, change score analysis, ANCOVA and cLDA. When analyzing longitudinal RCTs, only longitudinal ANCOVA and cLDA can be used, unless only baseline and one follow-up time point is taken into account in the analysis. In this section, the descriptions of the methods are given, while the mathematical models and the statistical analysis of the methods will be presented in Section 5. When analyzing RCTs with two time points, the difference in mean value between the groups at follow-up is the parameter of interest. When analyzing longitudinal RCTs, the main parameter of interest in our case is the mean value difference between the groups at the last follow-up time point. There can be several parameters of interest, at different follow-up time points. The analyses are equal independent of follow-up time point.

2.3.1 Comparison of mean at follow-up

A simple method to analyze the differences between the groups is to simply look at the mean differences at the last follow-up time point (Vickers and Altman, 2001). This method is called the comparison of mean at follow-up, or simply just the method of follow-up. Using this method, the differences between the groups are measured by comparing the means at follow-up. In a parallel-group RCT with two time points, only half of the data are used in the analysis when the method of follow-up is used. One might argue that in a well conducted RCT the mean values at baseline are assumed to be equal, thus the baseline values does not make any difference. Still, there is a disadvantage not to control for the baseline values when the information is available. This method is not recommended to use in a longitudinal RCT, because too much information is lost in the analysis when only one time point is taken into account in the analysis. Since the method of follow-up only takes one measurement into account it is better to use the method when the correlation between the measurements are low (Vickers and Altman, 2001).

2.3.2 Change score analysis

To extend the method of follow-up by taking the baseline values into account, the differences between the follow-up score and the baseline scores between the groups can be compared (Vickers and Altman, 2001). This method is called the change score analysis, or just the method of change score. Using this method, all of the data available are taken into account in a parallel-group RCT with two time points. This method is usually not used to analyze longitudinal RCTs, because much information is lost when not taking the other measurements into account. If the mean values at baseline for both groups are equal, the estimated treatment effect will give the same result as using the method of follow-up (Vickers and Altman, 2001). This will be shown in the mathematical presentation of the method of change score in Section 5.2.2. In an RCT, the baseline values are assumed to be equal in both groups. Thus, both the method of follow-up and the method of change score are assumed to give the same treatment effect. But the bias and power will differ depending on the correlation between baseline and follow-up (Vickers and Altman, 2001). If the correlation is high and the method of follow-up score is used, important information is lost. In this case, the analysis of the method of change score will result in higher power. On the other hand, if the correlation is low will the method of change score add variation to the analysis. In this case will the method of follow-up score result in higher power. Even

though the method of change score takes the baseline values into account, the method does not correct for the differences at baseline for the two groups. This is because of the regression towards the mean (Vickers and Altman, 2001). Regression towards the mean is the phenomena that if a given baseline score is extreme, the follow-up score is expected to be closer to the mean of all individuals than the baseline value (Bland and Altman, 1994). If the correlation between baseline and follow-up is low, the effect will be high and the other way around. To take this factor into account, other methods than the method of change score and the method of follow-up should be used.

2.3.3 ANCOVA

Analysis of covariance (ANCOVA) has traditionally been one of the most commonly used statistical methods for analyzing RCTs (Liu et al., 2009). It is a conditional regression model which adjust the follow-up measurements to the baseline measurement, thus it is unaffected by the baseline differences between the groups (Coffman et al., 2016; Vickers and Altman, 2001). This means that the method of ANCOVA creates unbiased estimates even if there are differences at baseline in the case of correlation between baseline and follow-up time points (Liu et al., 2009). The baseline values are treated as covariates in the model. Thus, missing baseline data are not accepted in the model. In the case of missing data at baseline, an imputation method can be used to create a complete baseline data set. When there are RCTs with two time points, the method is called ANCOVA. When there are more than one follow-up time point, the method is known as longitudinal ANCOVA. Compared to the methods of follow-up score and change score is the method of ANCOVA is superior (Frison and Pocock, 1992). For example, the method of ANCOVA has smaller variance and greater statistical power when analyzing the treatment effect (Vickers and Altman, 2001; Frison and Pocock, 1992). Thus, to detect the effect of a treatment, a smaller sample size is needed to get the same effect as for the method of follow-up or the method of change score. However, the difference between the method of change score and ANCOVA is low when the correlation between the time points is high (Vickers, 2001). The correlation is said to be high when $\rho > 0.8$. In the situation with stable chronic conditions is the correlation often high (Vickers and Altman, 2001). In practice the bias of the treatment effect obtained by using the method of ANCOVA is small, and may be even more reduced if there are more than one follow-up time point (Frison and Pocock, 1992). If there are missing data at baseline or at all follow-up time points, the method of ANCOVA may give biased results (Vickers, 2001).

2.3.4 cLDA

Constrained longitudinal data analysis (cLDA) is in contrast to ANCOVA an unconditional analysis method. Both the baseline values and the follow-up values are assumed to be dependent variables, thus modelled as the response variable in a regression model (Coffman et al., 2016). This is a special case of longitudinal data analysis (LDA) where the baseline values are constrained to be equal in both groups. This assumption is reasonable when analyzing RCTs since proper randomization before intervention is assumed. The effect of including the baseline variable dependent of the group is discussed by Liang and L. Zeger (2000). The difference between adjusting for baseline or not is described as an example of

a trade-off between bias and precision in statistical inference.
The advantage of cLDA compared to ANCOVA is that the method of cLDA can be used even if there are missing values. Thus, there is no need to use imputation methods. This is because there should be enough information to estimate the model since both the baseline values and follow-up values are dependent variables even in the case of (a reasonable amount of) missing data. In both the methods of cLDA and ANCOVA an individual covariance matrix can be modelled to take into account the correlation within a subject. The advantage with the method of cLDA compared to ANCOVA is that this matrix can be different for each treatment group, which makes the method more flexible (Liu et al., 2009).
By comparing ANCOVA and cLDA, we touch into a topic which have been discussed widely. Should the outcome variable for the baseline values be included as a covariate or as an outcome variable? The question of how baseline should be modelled was first introduced by Lord (1967) who presented the paradox where the same data with different modelling strategies resulted in different statistical inferences. Dinh and Yang (2011) have presented different articles with different assumptions about this topic. It is plausible to assume that the baseline values cannot be part of the outcome vector, because the measurement is done before intervention. Thus, it cannot define the treatment effect when the treatment is not introduced. In addition, the baseline values are often used as an inclusion criteria. Therefore, the baseline measurements can be truncated compared to the follow-up scores (Dinh and Yang, 2011). Liu et al. (2009), on the other hand, recommend using the method of cLDA with baseline values as part of the outcome vector, because implementing it as a covariate would lead to loss of efficiency. Even if one would assume that the baseline values should be part of the response vector or modelled as a covariate, Dinh and Yang (2011) showed that both of the methods of ANCOVA and cLDA were preferable compared to similar methods. The methods are compared by efficient treatment effect estimates and the robustness of the statistic inferences. Liu et al. (2009) showed that the mean difference between the groups in an RCT conditional on baseline using the method of ANCOVA is equal to the unconditional mean difference between the groups using the method of cLDA. In addition the maximum likelihood estimators for both methods are equal. This means that in the case of no missing data the methods of cLDA and ANCOVA should result in approximately the same results (Coffman et al., 2016; Liu et al., 2009). The variance of the point estimates of the mean difference between the groups by using the method of cLDA will always be equal or smaller than in the case of using the method of ANCOVA (equal if the baseline means are the same for both groups) (Liu et al., 2009). This makes the method of cLDA more powerful than the method of ANCOVA, but in practice this difference is small. In the case of the missing data mechanisms MAR and MCAR concludes Coffman et al. (2016) that the method of cLDA is the method of choice compared with ANCOVA. The method of cLDA results in unbiased estimates when the missing data mechanism is MCAR or MAR (Liu et al., 2009).

Chapter 3

Longitudinal data

The theory presented in this chapter is based on the books *Analysis of Longitudinal Data* by Diggle et al. (2001) and *Regression* by Fahrmeir et al. (2013).

A longitudinal trial is defined as a trial with repeated measurements of individuals on the same parameter of interest over time (Diggle et al., 2001, p. 1). In Figure 3.1, an example of longitudinal data is illustrated.

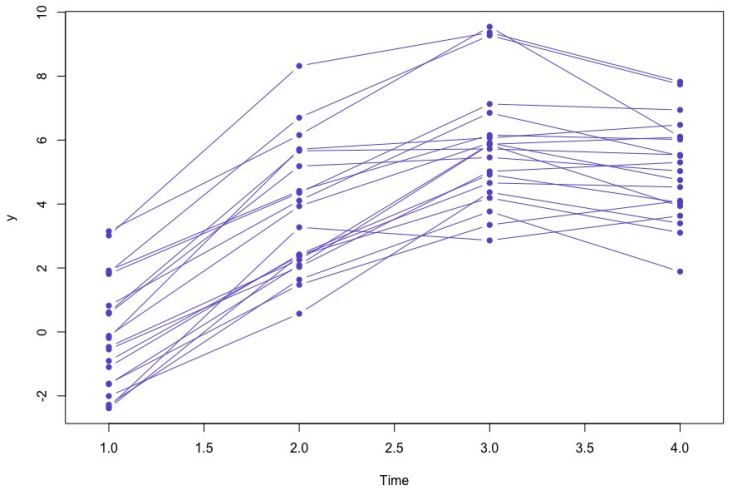


Figure 3.1: An illustration of longitudinal data with 20 individuals, each represented with a line.

Longitudinal trials are commonly used in medical research. It is a key feature in RCTs. The longitudinal data can reveal how the health status changes over time when a treatment

is introduced. A longitudinal trial is a more powerful method compared to a cross-sectional study, where the outcome variable is measured only once on each individual since the treatment changes over time. The advantage is that the analysis can separate the analysis of parameter of interest and how the outcome varies with time (Diggle et al., 2001, p. 1). Thus, parallel-group RCTs are more powerful than cross-sectional RCTs. When analyzing longitudinal data it is necessary to apply special statistical methods because the data tends to be intracorrelated. This means that there are correlations between the time points within each individual but not necessarily between the individuals. If intracorrelation is not taken into account the statistical inference of the data can be invalid (Diggle et al., 2001, p. 2). Longitudinal data are distinguished from time series. A time series is defined as a sequence of observations in order where the ordering is often in time (Wei, 2006, p. 1). This is also true for longitudinal data but the time series often have much higher number of observations. In addition longitudinal data consists of observations from several individuals which can be assumed to be independent (Diggle et al., 2001, p. 2). This makes the analysis of longitudinal data simpler than time series analysis and more robust. This is because it is possible to detect patterns across the individuals (Diggle et al., 2001, p. 2).
This chapter is structured up as follows: The notation for longitudinal data is presented in Section 3.1. The different regression methods for modelling longitudinal data are introduced in Section 3.2 and Section 3.3. Parameter estimators for the different models are given in Section 3.4 and Section 3.5. At last, the hypothesis tests and confidence interval for the parameter estimates are presented in Section 3.6.

3.1 Notation

Now follows an introduction of the notation used. Random variables are denoted with big letters and specific observations are denoted with small letters. In a longitudinal trial there are several individuals, given as $i = 1, 2, ..., n$, where n is the total number of individuals. All individuals are measured at several time points given as $t = 0, 1, ..., T$, where $t = 0$ is the time of the first measurement and $t = T$ of the last measurement. The observations of the outcome variable of interest are given as y_{it} for individual i at time t. In addition a set of covariates is related to each individual at each time point, given as x_{ijt}, where $j = 1, ..., p$ represents the set of all covariates. The set of covariates is assumed to be equal for all individuals at all time points, thus $x_{ijt} = x_{ij} \; \forall \, t$. An individual in a longitudinal trial is thus associated with a set of observed outcome variables and a set of covariates. This is given as

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i0} \\ y_{i1} \\ \vdots \\ y_{1T} \end{pmatrix} \quad \text{and} \quad \boldsymbol{x}_{it} = \begin{pmatrix} x_{i10} & x_{i11} & \dots & x_{i1T} \\ x_{i20} & x_{i21} & \dots & x_{i2T} \\ \vdots & \vdots & & \vdots \\ x_{ip0} & x_{ip1} & \dots & x_{ipT} \end{pmatrix}, \tag{3.1}$$

where $\boldsymbol{y}_i$ is a vector of length $(T + 1)$ and $\boldsymbol{x}_i$ is a matrix of dimension $(p \times (T + 1))$, where the elements are often equal for all t. All the individuals can be represented into

one vector representation. The observed outcome variables for all individuals are given as

$$\boldsymbol{y} = \begin{pmatrix} y_{10} \\ \vdots \\ y_{1T} \\ \vdots \\ \vdots \\ y_{n0} \\ \vdots \\ y_{nT} \end{pmatrix}, \tag{3.2}$$

and the covariates are given as

$$\boldsymbol{x} = \begin{pmatrix} x_{110} & x_{111} & \dots & x_{11T} & \dots & \dots & x_{n10} & x_{n11} & \dots & x_{n1T} \\ x_{120} & x_{121} & \dots & x_{12T} & \dots & \dots & x_{n20} & x_{n21} & \dots & x_{n2T} \\ \vdots & \vdots & & \vdots & & & & & & \\ x_{1p0} & x_{1p1} & \dots & x_{1pT} & \dots & \dots & x_{np0} & x_{np1} & \dots & x_{npT} \end{pmatrix}. \tag{3.3}$$

3.2 Linear regression

A longitudinal trial is usually modelled by a regression model (Diggle et al., 2001, p. 15). The linear regression model is an example. Using a regression model it is possible to analyze the effect of the covariates on the outcome variable (Fahrmeir et al., 2013, p. 12). An illustration of a simple linear regression model is given in Figure 3.2

Representation for one observation

A linear regression model for an individual at a given time point can be given as

$$Y_{it} = \beta_0 + \beta_1 x_{i1t} + ... + \beta_p x_{ipt} + \varepsilon_{it} = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_{it}, \tag{3.4}$$

where

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{and} \quad \boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{i1t} \\ \vdots \\ x_{ipt} \end{pmatrix}. \tag{3.5}$$

Here $\boldsymbol{\beta}$ is a vector of length $(p+1)$ representing the unknown regression coefficients with β_0 as the intercept. The design vector, $\boldsymbol{x}_i$, is of length $(p+1)$. It consists of all covariates and in addition an element associated to the intercept, β_0. The error term, ε_{it}, is a random variable which represents the deviation of the response from the model. It is assumed to be distributed as a zero-mean normal variable with constant variance. This is given as

$$\varepsilon_{it} \sim \mathcal{N}(0, \sigma^2). \tag{3.6}$$

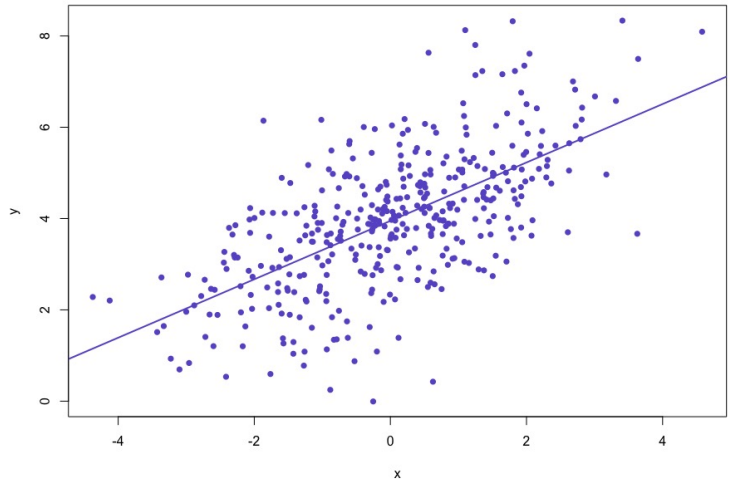


Figure 3.2: An illustration of a simple linear regression model with one covariate. The dots represents the observations for each individual, and the line represents the linear regression line.

These error terms are independent between the individuals, i, but dependent within each individual when assumed to be correlation within each individual. The expected value and variance for each observation in the linear regression model in Equation (3.4) are given as

$$\mathrm{E}(Y_{it}) = \mathrm{E}(\boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_{it}) = \boldsymbol{x}_i\boldsymbol{\beta},$$

and

$$\mathrm{Var}(Y_{it}) = \mathrm{Var}(\boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_{it}) = \mathrm{Var}(\varepsilon_{it}) = \sigma^2.$$

Using this regression model, the observed variables, y_{it}, given in Equation (3.1) are assume to be are realizations of the random variable Y_{it} (Diggle et al., 2001), which is assumed to be multivariate normally distributed as

$$Y_{it} \sim \mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta}, \sigma^2).$$

Using a linear regression model, it is possible to extend the model to include interaction terms. This means that one covariate is dependent of at least on other covariate. As an example, a linear regression model with two covariates with interaction is given as

$$Y_{it} = \beta_0 + \beta_1 x_{i1t} + \beta_2 x_{i2t} + \beta_3 x_{i1t} x_{i2t} + \varepsilon_{it}.$$

Here, $\beta_1 x_{i1t}$ and $\beta_2 x_{i2t}$ depend only on one covariate while $\beta_3 x_{i1t} x_{i2t}$ depends on two covariates (Fahrmeir et al., 2013, p. 98).

Representation for one individual

The vector representation for all time points for one individual of the linear regression model given in Equation (3.4) is given as

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \tag{3.7}$$

where $\boldsymbol{Y}_i$ is a vector of length $(T+1)$, given as

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i0} \\ Y_{i1} \\ \vdots \\ Y_{iT} \end{pmatrix}. \tag{3.8}$$

The design matrix, X_i, and the vector consisting of the error terms, $\boldsymbol{\varepsilon}_i$ are given as

$$X_i = \begin{pmatrix} 1 & x_{i10} & \dots & x_{ip0} \\ \vdots & \vdots & & \vdots \\ 1 & x_{i1T} & \dots & x_{ipT} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i0} \\ \vdots \\ \varepsilon_{iT} \end{pmatrix}. \tag{3.9}$$

The design matrix, X_i, is of dimension $((T+1)\times(p+1))$, while the vector consisting of the error terms, $\boldsymbol{\varepsilon}_i$, is of length $(T+1)$. The error terms are assumed to have a distribution given by

$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 V_*),$$

where V_* represents the correlation matrix within each individual. This correlation matrix describes the dependence between each time point for a given individual. The dimension of V_* is $((T+1)\times(T+1))$. The correlation matrix can have different structures. In general, this is given as

$$\text{Corr}(Y_{it}, Y_{i't'}) = \begin{cases} 1 & \text{if } i = i' \text{ and } t = t', \\ \rho(t,t') & \text{if } i = i' \text{ and } t \neq t', \\ 0 & \text{else.} \end{cases}$$

We are going to describe two different correlation structures that are usual to assume in the case of longitudinal data. The first correlation structure is the compound symmetry correlation (or uniform correlation). This is modelled as equal correlation between each pair of observations independent of time, given as

$$\text{Corr}(Y_{it}, Y_{i't'}) = \begin{cases} 1 & \text{if } i = i' \text{ and } t = t', \\ \rho & \text{if } i = i' \text{ and } t \neq t', \\ 0 & \text{else.} \end{cases}$$

Thus, the compound symmetry matrix, V_0, is given as

$$V_0 = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix}. \tag{3.10}$$

Here, the time difference between each observation does not affect the correlation. It is assumed that the correlation is constant within each individual.
The second correlation structure presented is the exponential correlation structure. Using this structure the correlation is assumed to decay towards zero as time separation increases. This is given as

$$\text{Corr}(Y_{it}, Y_{i't'}) = \begin{cases} 1 & \text{if } i = i' \text{ and } t = t', \\ e^{-\phi|t-t'|} & \text{if } i = i' \text{ and } t \neq t', \\ 0 & \text{else,} \end{cases}$$

where ϕ represents the speed of the correlation decay. This is given as a number between 0 and 1, where higher values of ϕ gives faster decay. The correlation matrix for the exponential correlation model is thus given as

$$V_e = \begin{pmatrix} 1 & e^{-\phi|t_0-t_1|} & \dots & e^{-\phi|t_0-t_T|} \\ e^{-\phi|t_1-t_0|} & 1 & \dots & e^{-\phi|t_1-t_T|} \\ \vdots & & \ddots & \vdots \\ e^{-\phi|t_T-t_0|} & e^{-\phi|t_T-t_1|} & \dots & 1 \end{pmatrix}. \tag{3.11}$$

Going back to the linear regression representation given in Equation (3.7), the expected values and the covariance matrix are given as

$$\text{E}(\boldsymbol{Y}_i) = \text{E}(X_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i) = X_i\boldsymbol{\beta}$$

and

$$\text{Cov}(\boldsymbol{Y}_i) = \text{Cov}(X_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i) = \text{Cov}(\boldsymbol{\varepsilon}_i) = \sigma^2 V_*.$$

The outcome variable representing one individual is assumed to be normally distributed, given as

$$\boldsymbol{Y}_i \sim \mathcal{N}(X_i\boldsymbol{\beta}, \sigma^2 V_*).$$

Representation for all individuals

The overall matrix representation of all individuals at all time points is given as

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3.12}$$

The vector $\boldsymbol{Y}$ is of length $n(T+1)$ and is given by

$$\boldsymbol{Y} = \begin{pmatrix} Y_{10} \\ \vdots \\ Y_{1T} \\ \vdots \\ \vdots \\ Y_{n0} \\ \vdots \\ Y_{nT} \end{pmatrix}. \tag{3.13}$$

The design matrix, X, of dimension $((T+1)n \times (p+1))$ and the vector, $\boldsymbol{\varepsilon}$, of length $n(T+1)$ are given as

$$X = \begin{pmatrix} 1 & x_{110} & \dots & x_{1p0} \\ \vdots & \vdots & & \vdots \\ 1 & x_{11T} & \dots & x_{1pT} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & x_{n10} & \dots & x_{np0} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1T} & \dots & x_{npT} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{10} \\ \vdots \\ \varepsilon_{1T} \\ \vdots \\ \vdots \\ \varepsilon_{n0} \\ \vdots \\ \varepsilon_{nT} \end{pmatrix}. \tag{3.14}$$

The vector consisting of the error terms, $\boldsymbol{\varepsilon}$, is distributed as

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 V),$$

where V is a $(n(T+1) \times n(T+1))$ block-diagonal matrix with V_* on the main diagonal. The matrix V_* is the correlation matrix within each individual, as given in either Equation (3.10) or as in Equation (3.11) depending on what type of correlation structure that is assumed. Thus, the matrix V is given as

$$V = \begin{pmatrix} [V_*] & & 0 \\ & \ddots & \\ 0 & & [V_*] \end{pmatrix}. \tag{3.15}$$

This correlation structure represents the situation we assume, where there are correlation within each individual but not between the individuals. The expected value and covariance matrix for the vector $\boldsymbol{Y}$, given in Equation (3.12), are given as

$$\mathrm{E}(\boldsymbol{Y}) = \mathrm{E}(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = X\boldsymbol{\beta},$$

and

$$\mathrm{Cov}(\boldsymbol{Y}) = \mathrm{Cov}(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 V.$$

The outcome variable $\boldsymbol{Y}$ given in (3.4) is assumed to be normally distributed, given as

$$\boldsymbol{Y} \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2 V). \tag{3.16}$$

3.3 Linear mixed models

Linear mixed models (LMM) are extensions of the linear regression model, given in Equation (3.4). The difference between linear regression model and LMM is that there is added a random effect in the model in addition to the fixed effects in the LMM model. This is also called the random effects models. Using this random effects the observations can be modelled in clusters. This way some observations may be modelled more similar than others, depending on which cluster the individual belongs to. LMM is a commonly used and popular method when analyzing longitudinal data (Fahrmeir et al., 2013, p. 349).

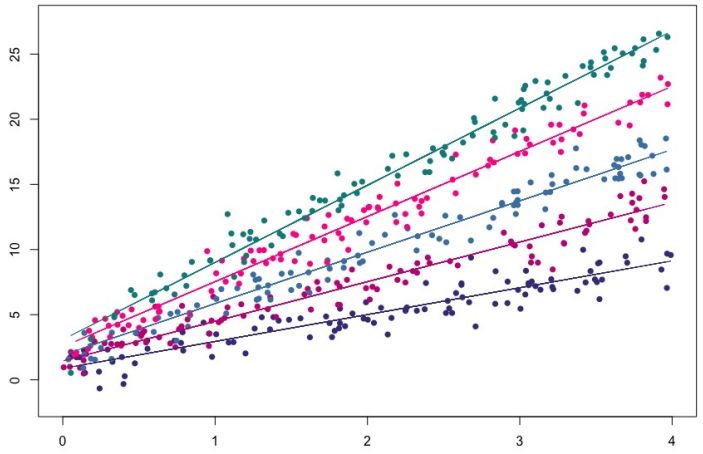


Figure 3.3: An illustration of a linear mixed model with 100 individual in each cluster. Each cluster is represented with its own color and has its own regression line.

Representation for one observation

A linear mixed model can be given as

$$Y_{it} = \beta_0 + \beta_1 x_{i1t} + ... + \beta_p x_{ipt} + \gamma_{i0} + \gamma_{i1} u_{i1t} + ... + \gamma_{iq} u_{iqt} + \varepsilon_{it}, \tag{3.17}$$

or in a more compact notation given as

$$Y_{it} = \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{u}_{it}^T \boldsymbol{\gamma}_i + \varepsilon_{it}. \tag{3.18}$$

Here, $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ are the fixed effects given in Equation (3.5) and ε_{it} is given in Equation (3.6). These are the same terms as in the linear regression model given by Equation (3.4). The random effects are γ_{ik} for $k = 0, 1, ..., q$, which is the cluster specific terms. These terms are equal for each observation belonging to the same cluster. The cluster specific term are zero-mean random variables with equal variance within the group. This is given as

$$\gamma_{ik} \sim \mathcal{N}(0, \tau_k^2)$$

The random effect vector, $\boldsymbol{\gamma}_i$, is of length $(q+1)$ and is given as

$$\boldsymbol{\gamma}_i = \begin{pmatrix} \gamma_{i0} \\ \gamma_{i1} \\ \vdots \\ \gamma_{iq} \end{pmatrix}. \tag{3.19}$$

The distribution of $\boldsymbol{\gamma_i}$ is given as

$$\boldsymbol{\gamma}_i \sim \mathcal{N}(0, Q),$$

where Q is the covariance matrix of dimension $((q+1)\times(q+1))$ for the vector of random effects. If we assume that the random effects are independent of each other, Q is a diagonal matrix, with $\tau_0^2, \tau_1^2, ..., \tau_q^2$ on the main diagonal and zero elsewhere, given as

$$Q = \begin{pmatrix} \tau_0^2 & & 0 \\ & \ddots & \\ 0 & & \tau_q^2 \end{pmatrix}. \tag{3.20}$$

The vector $\boldsymbol{u}_{it}$ is of length $(q+1)$ and is given by

$$\boldsymbol{u}_{it} = \begin{pmatrix} 1 \\ u_{i1t} \\ \vdots \\ u_{iqt} \end{pmatrix}.$$

The elements of the design vector, u_{ikt}, is given as

$$u_{ikt} = \begin{cases} 1 & \text{if observation } Y_{it} \text{ is in cluster } k \text{ at time } t, \\ 0 & \text{else.} \end{cases}$$

Often, the observations are in the same cluster at all time points, thus u_{ikt} is independent of time, t. The expected values and variance for the linear mixed model given in Equation (3.18), are given as

$$\mathrm{E}(Y_{it}) = \mathrm{E}(\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{u}_{it}^T\boldsymbol{\gamma}_i + \varepsilon_{it}) = \boldsymbol{x}_i^T\boldsymbol{\beta},$$

and

$$\mathrm{Var}(Y_{it}) = \mathrm{Var}(\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{u}_{it}^T\boldsymbol{\gamma}_i + \varepsilon_{it}) = u_{it}^T Q u_{it} + \sigma^2.$$

The correlation between two time points within one individual is called the intraclass correlation (ICC). Given that individual Y_{it} is in cluster k, and that Equation (3.20) holds, the ICC is given as

$$\mathrm{ICC} = \mathrm{Corr}(Y_{it}, Y_{it'}) = \begin{cases} 1 & \text{if } t = t', \\ \frac{\sigma^2}{\tau_k^2+\sigma^2} & \text{if } t \neq t'. \end{cases}$$

The linear mixed model, given in Equation (3.18) is assumed to be normally distributed, given by

$$y_{it} \sim \mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta}, u_{it}^T Q u_{it} + \sigma^2).$$

Representation for one individual

Writing the linear mixed model as the total vector representation for one individual yields

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + U_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i. \tag{3.21}$$

Here, $\boldsymbol{Y}_i$ is given as in Equation (3.8), while X_i and $\boldsymbol{\varepsilon}_i$ are given in Equation (3.9). The design matrix, U_i, for the cluster specific terms is a $((T+1)\times(q+1))$ matrix, given by

$$U_i = \begin{pmatrix} 1 & u_{i10} & \dots & u_{iq0} \\ \vdots & \vdots & & \vdots \\ 1 & u_{i1T} & \dots & u_{iqT} \end{pmatrix}. \tag{3.22}$$

The expected value and covariance matrix of $\boldsymbol{y}_i$ are given as

$$\mathrm{E}(\boldsymbol{Y}_i) = \mathrm{E}(X_i\boldsymbol{\beta} + U_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon_i}) = X_i\boldsymbol{\beta},$$

and

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{Y}_i) &= \mathrm{Cov}(X_i\boldsymbol{\beta} + U_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i)\\ &= \mathrm{Cov}(U_i\boldsymbol{\gamma}_i) + \mathrm{Cov}(\boldsymbol{\varepsilon}_i)\\ &= U_iQU_i^T + \sigma^2V_*,\end{aligned}$$

where V_* is given by Equation (3.10) or Equation (3.11). If we assume that there are independence between the random effects, Q is given as in Equation (3.20) and the covariance matrix is given by

$$\mathrm{Cov}(\boldsymbol{Y}_i) = U_iQU_i^T + \sigma^2V_*.$$

Here, the matrix J_{T+1} is the matrix of dimension $((T+1)\times(T+1))$ consisting of only ones. The linear mixed model, given in Equation (3.21) is assumed to be normally distributed, given as

$$\boldsymbol{Y}_i \sim \mathcal{N}(X_i\boldsymbol{\beta}, U_iQU_i^T + \sigma^2V_*).$$

Representation for all individuals

The total vector representation of all individuals at all time points for a linear mixed model given in Equation (3.17) is given by

$$\boldsymbol{Y} = X\boldsymbol{\beta} + U\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{3.23}$$

$\boldsymbol{Y}$ is given in Equation (3.13) while X and $\boldsymbol{\varepsilon}$ are given in Equation (3.14). The design matrix U is a $(n(T+1)\times n(J+1))$ block-diagonal matrix with the design matrix U_i given in Equation (3.22) on the main diagonal and zero elsewhere, given as

$$U = \begin{pmatrix}[U_1] & & 0\\ & \ddots & \\ 0 & & [U_n]\end{pmatrix}.$$

The vector of random effects, $\boldsymbol{\gamma}$, is of length $n(T+1)$ and is given as

$$\boldsymbol{\gamma} = \begin{pmatrix}\gamma_{10}\\ \vdots\\ \gamma_{1q}\\ \vdots\\ \gamma_{n0}\\ \vdots\\ \gamma_{nq}\end{pmatrix}. \tag{3.24}$$

This vector consisting of all the random effects is multivariate normally distributed, given as

$$\boldsymbol{\gamma} \sim \mathcal{N}(0, G),$$

where G is a block-diagonal matrix of dimensions $(n(q+1) \times n(q+1))$ with Q on the main diagonal and zero elsewhere, given as

$$G = \begin{pmatrix} [Q] & & 0 \\ & \ddots & \\ 0 & & [Q] \end{pmatrix}.$$

The expected value of the vector and the covariance matrix of $\boldsymbol{Y}$ given in Equation (3.23) are given as

$$\mathrm{E}(\boldsymbol{Y}) = \mathrm{E}(X\boldsymbol{\beta} + U\boldsymbol{\gamma} + \boldsymbol{\varepsilon}) = X\boldsymbol{\beta},$$

and

$$\mathrm{Cov}(\boldsymbol{Y}) = \mathrm{Cov}(X\boldsymbol{\beta} + U\boldsymbol{\gamma} + \boldsymbol{\varepsilon}) = \mathrm{Cov}(U\boldsymbol{\gamma}) + \mathrm{Cov}(\boldsymbol{\varepsilon}) = UGU^T + \sigma^2 V.$$

The outcome variable $\boldsymbol{Y}$, given in Equation (3.23) is assumed to be multivariate normally distributed, given as

$$\boldsymbol{Y} \sim \mathcal{N}(X\boldsymbol{\beta}, UGU^T + \sigma^2 V). \tag{3.25}$$

3.4 Parameter estimators for the linear regression model

When modelling longitudinal data using a linear regression model or a linear mixed effects model the regression coefficients, $\boldsymbol{\beta}$, given in Equation (3.5), are unknown and must be estimated. For the linear regression model given in Equation (3.4) the predictor for Y_{it} is given as

$$\hat{Y}_{it} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1t} + ... + \hat{\beta}_p x_{ipt} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}.$$

Here, $\hat{\boldsymbol{\beta}}$ is the vector of estimators of the unknown regression coefficients. These estimators may be found by using the method of least squares or the method of maximum likelihood. The estimators for the error terms, ε_{it}, is called the residuals and are given by

$$\hat{\varepsilon}_{it} = y_{it} - \hat{Y}_{it} = y_{it} - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}},$$

where y_{it} are the observed variables. To find the estimators for $\boldsymbol{\beta}$, and thus also $\boldsymbol{\varepsilon}$, the vector representation of all observations, given in Equation (3.4) is used. This is given as

$$\hat{\boldsymbol{Y}} = X\hat{\boldsymbol{\beta}}$$

and

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{Y}} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}.$$

Here, $\boldsymbol{y}$ is the set of all observed variables, given in Equation (3.2) and X is the design matrix given in Equation (3.14).

Method of least squares for $\boldsymbol{\beta}$

A method that is widely used when finding estimators for the regression coefficients $\boldsymbol{\beta}$ is the method of least squares (Fahrmeir et al., 2013). This method aims to minimize the difference between the observed values and the estimated values with respect to $\boldsymbol{\beta}$. This is given as

$$\text{LS}(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^T V \boldsymbol{\varepsilon} = (\boldsymbol{y} - X\boldsymbol{\beta})^T V (\boldsymbol{y} - X\boldsymbol{\beta}),$$

where V is the matrix given in Equation (3.15) and is assumed to be known. This equation can be rewritten as

$$\text{LS}(\boldsymbol{\beta}) = \boldsymbol{y} V \boldsymbol{y}^T - 2\boldsymbol{y}^T V X \boldsymbol{\beta} + \boldsymbol{\beta}^T X^T V X \boldsymbol{\beta}.$$

This equation is minimized by setting the derivatives of $\boldsymbol{\beta}$ equal to zero. By doing this and rewriting the equation in terms of the estimated regression coefficients, the result is given by

$$\hat{\boldsymbol{\beta}} = (X^T V X)^{-1} X^T V \boldsymbol{y}. \tag{3.26}$$

This is the estimator for the regression coefficients. The expected value of the estimator is given as

$$\begin{aligned} \text{E}(\hat{\boldsymbol{\beta}}) &= \text{E}((X^T V X)^{-1} X^T V \boldsymbol{y}) \\ &= ((X^T V X)^{-1} X^T V)\text{E}(\boldsymbol{y}) \\ &= (X^T V X)^{-1} X^T V X \boldsymbol{\beta} = \boldsymbol{\beta}. \end{aligned}$$

Since $\text{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ the estimator is said to be unbiased. The bias of an estimator is the deviation of the estimator from the true value. This is given as

$$\text{Bias}(\hat{\theta}) = \theta - \hat{\theta}.$$

An estimator is unbiased if the bias of the estimator is zero. The covariance matrix of the estimator is given as

$$\begin{aligned} \text{Cov}(\hat{\boldsymbol{\beta}}) &= \text{Cov}((X^T V X)^{-1} X^T V \boldsymbol{y}) \\ &= \sigma^2((X^T V X)^{-1} X^T V) V (V X (X^T V X)^{-1}) \\ &= \sigma^2 (X^T V^{-1} X)^{-1}. \end{aligned}$$

The true covariance cannot be calculated if σ^2 is unknown. Thus, an estimator of the covariance matrix for $\hat{\boldsymbol{\beta}}$ is given as

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}_\varepsilon^2 (X^T V X)^{-1} = \frac{\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} (X^T V X)^{-1}}{n(T+1) - (p+1)}, \tag{3.27}$$

where the diagonal elements are the estimated variances of $\hat{\boldsymbol{\beta}}$.

Maximum likelihood estimator for $\boldsymbol{\beta}$ and σ^2

The maximum likelihood estimator (MLE) for Equation (3.12) use the assumption that the equation is assumed to be normally distributed. The likelihood for Equation (3.12) is given as

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{(n(T+1))/2}} e^{-\frac{1}{2\sigma^2}(\boldsymbol{y}-X\boldsymbol{\beta})^T V(\boldsymbol{y}-X\boldsymbol{\beta})}.$$

The log-likelihood is thus given as

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n(T+1)}{2}\big(\log(2\pi) - \log(\sigma^2)\big) - \frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{y} - X\boldsymbol{\beta})^T V(\boldsymbol{y} - X\boldsymbol{\beta}). \quad (3.28)$$

By maximizing the log-likelihood with respect to $\boldsymbol{\beta}$ and σ^2 we find the most likely estimators for $\boldsymbol{\beta}$ and σ^2. This is done by differentiating $\boldsymbol{\beta}$ and σ^2 and solving the set of equations. This results for $\boldsymbol{\beta}$ is the same estimator as the estimator found by using the method of least square, given in Equation (3.26) (Fahrmeir et al., 2013, p. 107). The estimator for the variance of the error terms, σ^2, is given by

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n(T+1)}{2\sigma^2} + \frac{1}{\sigma^4}(\boldsymbol{y} - X\boldsymbol{\beta})^T V(\boldsymbol{y} - X\boldsymbol{\beta}) = 0.$$

By inserting $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, this results in

$$\begin{aligned}\frac{\partial l(\hat{\boldsymbol{\beta}}, \sigma^2)}{\partial \sigma^2} &= -\frac{n(T+1)}{2\sigma^2} + \frac{1}{\sigma^4}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T V(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) \\ &= -\frac{n(T+1)}{2\sigma^2} + \frac{1}{\sigma^4}\hat{\boldsymbol{\varepsilon}}^T V \hat{\boldsymbol{\varepsilon}}.\end{aligned}$$

By setting this equation equal to zero the resulting expression for the MLE of σ_ε^2 is given as

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^T V \hat{\boldsymbol{\varepsilon}}}{n(T+1)}.$$

However, Fahrmeir et al. (2013) showed that this estimator is biased because

$$\mathrm{E}(\hat{\sigma}^2) = \frac{n(T+1) - (p+1)\sigma^2}{n(T+1)} \neq \sigma^2. \quad (3.29)$$

An improved estimator for σ^2 is the restricted maximum likelihood estimator (Fahrmeir et al., 2013, p. 109).

Restricted maximum likelihood estimation

Restricted maximum likelihood estimation (REML) is a method which produces unbiased estimation of the variance for the error terms. Given the expected value given in Equation (3.29) the REML estimator for σ^2, which is unbiased, is given as

$$\hat{\sigma}_\varepsilon^2 = \frac{\hat{\boldsymbol{\varepsilon}}^T V \hat{\boldsymbol{\varepsilon}}}{n(T+1) - (p+1)}.$$

3.5 Model parameter estimators for LMM

Marginal and conditional model

When analyzing RCT by using a linear regression model only the fixed effects, $\boldsymbol{\beta}$, is of interest to estimate. In this case the marginal model can be used, given in Equation (3.16). Estimating parameters by the marginal model were used in Section 3.4. In the case of a LMM, the random effects, $\boldsymbol{\gamma}$, are also of interest to estimate. For this to be done, the conditional model is needed (Fahrmeir et al., 2013, p. 371). The marginal model for a LMM is given in Equation (3.25). The conditional model of $\boldsymbol{Y}$ given the random effects $\boldsymbol{\gamma}$ is given by

$$\boldsymbol{Y}|\boldsymbol{\gamma} \sim \mathcal{N}(X\boldsymbol{\beta} + U\boldsymbol{\gamma}, \sigma^2 V).$$

Likelihood estimation

In the case of known covariance matrices V and G we rewrite equation (3.25) to

$$\boldsymbol{Y} \sim \mathcal{N}(X\boldsymbol{\beta}, R),$$

where $R = UGU^T + \sigma^2 V$ is known. According to Fahrmeir et al. (2013) the unknown parameters can be estimated by maximizing the joint log-likelihood of $\boldsymbol{Y}$ and $\boldsymbol{\gamma}$ with respect to both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ at the same time (Fahrmeir et al., 2013, p. 371). The log-likelihood is given as

$$\log(p(\boldsymbol{Y}, \boldsymbol{\gamma})) = \log(p(\boldsymbol{Y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})) = \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta} - U\boldsymbol{\gamma})^T V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta} - U\boldsymbol{\gamma}) + \boldsymbol{\gamma} G^{-1}\boldsymbol{\gamma},$$

where $p(\cdot)$ is the distribution of the given variable. This is the maximum likelihood estimator. The maximum likelihood estimator is equivalent to the least square estimator (Fahrmeir et al., 2013, p. 371). This is given as

$$\frac{\partial \log(p(\boldsymbol{Y}, \boldsymbol{\gamma}))}{\partial(\boldsymbol{\beta}, \boldsymbol{\gamma})} = (\boldsymbol{y} - X\boldsymbol{\beta} - U\boldsymbol{\gamma})^T V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta} - U\boldsymbol{\gamma}) + \boldsymbol{\gamma}^T G^{-1}\boldsymbol{\gamma} = 0.$$

The derivative of the least square estimator is taken on both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. By setting the derivative of all parameters equal to zero this results in the mixed models equations, given by

$$\begin{pmatrix} X^T V^{-1} X & X^T V^{-1} U \\ U^T V^{-1} X & U^T V^{-1} U + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} X^T V^{-1} \boldsymbol{y} \\ U^T V^{-1} \boldsymbol{y} \end{pmatrix}.$$

Which results in the estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, given as

$$\hat{\boldsymbol{\beta}} = (X^T R^{-1} X)^{-1} X^T R^{-1} \boldsymbol{y},$$

and

$$\hat{\boldsymbol{\gamma}} = G U^T R^{-1}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}).$$

In addition Fahrmeir et al. (2013) has shown that the covariance for the parameter estimates are given as

$$\widehat{\text{Cov}} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = (\boldsymbol{C}^{\boldsymbol{T}} V^{-1} \boldsymbol{C} + \boldsymbol{B})^{-1},$$

where

$$\boldsymbol{B} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & G \end{pmatrix} \quad \text{and} \quad \boldsymbol{C} = (X, U).$$

3.6 Hypothesis testing

When estimators for the regression parameters are found, the significance of the estimates has to be found. This can be done by performing a hypothesis test. A general hypothesis for a regression model is given as

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{d} \quad \text{vs} \quad H_1 : \boldsymbol{C\beta} \neq \boldsymbol{d},$$

where $\boldsymbol{C}$ is a $(p+1)$ row vector that represents which elements of $\boldsymbol{\beta}$ that is tested by the hypothesis. The hypothesis of significance for only one regression parameter is given as

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0, \tag{3.30}$$

where the hypothesis is if β_j is equal to zero or not. Here, $\boldsymbol{d} = 0$ and $\boldsymbol{C}$ is a vector with 1 at the $(j+1)$-th position and zero elsewhere, given as

$$\boldsymbol{C} = (0, 0, \dots 0, \underbrace{1,}_{(j+1)-\text{th position}} 0, \dots 0)^T.$$

This hypothesis can be evaluated by using the t-statistic. The test statistic for evaluating the hypothesis is thus given by

$$t_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}}. \tag{3.31}$$

The null hypothesis is rejected if

$$|t_j| > t_{(1-\alpha/2),(n(T+1)-(p+1))}. \tag{3.32}$$

Here, α is the significance level of the test. Further on, for simplicity, we define

$$t_t = t_{(1-\alpha/2),(n(T+1)-(p+1))}. \tag{3.33}$$

If the hypothesis test of interest is to compare two regression coefficients, the hypothesis is given as

$$H_0 : \beta_j - \beta_k = 0 \quad \text{vs} \quad H_1 : \beta_j - \beta_k \neq 0. \tag{3.34}$$

Here, $\boldsymbol{d} = 0$ as before, while $\boldsymbol{C}$ is now the vector with 1 at the $(j+1)$-th position and -1 at the $(k+1)$-th position, given as

$$\boldsymbol{C} = (0, \dots 0, \underbrace{1,}_{(j+1)-\text{th position}} 0, 0, \dots 0, \underbrace{-1,}_{(k+1)-\text{th position}} 0, \dots 0)^T.$$

The hypothesis is evaluated by the t-statistic, where the test statistic now is given by

$$t_{jk} = \frac{\boldsymbol{C}\hat{\boldsymbol{\beta}}}{\sqrt{\widehat{\text{Cov}}(\boldsymbol{C}\hat{\boldsymbol{\beta}})}}. \tag{3.35}$$

The null hypothesis is rejected if

$$|t_{jk}| > t_t, \tag{3.36}$$

where t_t is given in Equation (3.33).

3.7 Confidence intervals

A $(1-\alpha)100\%$-confidence interval for a parameter represents the region where we are $(1-\alpha)100\%$ confident to find the true value of the parameter. This is related to the hypothesis test given in Equation (3.30) and is found by combining the Equations (3.31) and (3.32). The confidence interval for the regression coefficients is given by

$$[\hat{\beta}_j - t_t\sqrt{\widehat{\text{Var}}(\hat{\beta}_j)},\ \hat{\beta}_j t_t\sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}].$$

The confidence interval for a difference between two regression coefficients, given in the Equation (3.34), is given by the Equations (3.35) and (3.36). This is given as

$$[\boldsymbol{C}\hat{\boldsymbol{\beta}} - t_t\sqrt{\widehat{\text{Cov}}(\boldsymbol{C}\hat{\boldsymbol{\beta}})},\ \boldsymbol{C}\hat{\boldsymbol{\beta}} + t_t\sqrt{\widehat{\text{Cov}}(\boldsymbol{C}\hat{\boldsymbol{\beta}})}].$$

Chapter 4

Missing data

This chapter is inspired by the book ”Medical Statistics” by Veieroed et al. (2012).

Missing data are defined as the data that were intended to be collected in a trial but for some reason were not possible to collect or are lost (Carpender and Kenward, 2007; Diggle et al., 2001). This means that there actually exist meaningful values for the missing data but there is no available information about the missing data. Thus, if a patient dies during a trial the data are no longer available and are therefore not regarded as missing (Veieroed et al., 2012, p. 431). In this case the data are known as censored data. In almost every RCT there are missing data. This is a potential source of bias in the results and will lead to reduced power of the analysis (Altman and Bland, 2007; Sterne et al., 2009). Many statistical analyses assume complete data and do not take the missing data into account (Altman and Bland, 2007; Veieroed et al., 2012). Recent studies have shown that most researchers which conduct an RCT fail to use one of the best approaches to deal with missing data or do not deal with missing data at all (Zhang et al., 2017; Rombach et al., 2016). Most of the researchers either take the easy way out and only deal with the complete cases (i.e. remove all units with missing data) or use methods that are not valid. A method is said to be valid if both the estimates of the parameters of interest and the estimated variance of the estimates are approximately unbiased. (Veieroed et al., 2012, p. 429) In addition, Zhang et al. (2017) showed that only a small amount of the researchers discussed the risk of bias which may have been caused by missing data in their articles. Missing data is a bigger problem in studies which are retrospective (Altman and Bland, 2007). If the assumption of missing data is taken into account before the trial is conducted, a larger sample size could reduce the potential power loss of the study. However, larger sample size does not control for the potential bias missing data can introduce (Altman and Bland, 2007). RCTs are prospective studies and should therefore take the assumption of missing data into account before conducting the study.
The chapter is structured as follows: The different missing data patterns and the consequences of the amount of missing data are presented in Section 4.1. The different types of missing data mechanisms are introduced and discussed in Section 4.2. Sensitivity analy-

ses when handling missing data are discussed in Section 4.3. At last, different methods to handle missing data are presented and discussed in Section 4.4.

4.1 Missing data patterns

4.1.1 Interim and withdrawal missing data

When analyzing missing data we distinguish between interim missing data and withdrawal missing data (Carpender and Kenward, 2007, p. 4). Interim missing data is defined as the missing data when a patient has missed one follow-up measurement but has at least one follow-up measurement available later on. Withdrawal missing data is defined as the missing data when there are no measurements available after a certain time point. Thus, if y_{it} is missing, then $y_{it'}$ is missing for all $t' > t$. Withdrawal missing is sometimes referred to as monotone missingness or just drop-outs. Withdrawal missing data affects the results to the highest degree (Carpender and Kenward, 2007, p. 5). However, interim missing data are more difficult to deal with because there is more variation in the missing data patterns (Diggle et al., 2001, p. 284).

4.1.2 Item and unit nonresponse

Missing data can also be divided into item and unit nonreseponse. Item nonresponse means that some of the data are available for the items of a subject but not all. Unit nonresponse is the situation where no information are available for a unit. A unit can for example be an individual. Unit nonrespons can typically be the case where no questions are filled out in a questionnaire. Unit nonresponse is typically dealt with by using weighting methods as direct standardization, while imputation is commonly used when the case is item nonresponse. (Veieroed et al., 2012, p. 431). Item nonresponse is typically the type of missing data medical studies are exposed for. Unit nonresponse is a bigger problem in observational studies than in RCTs. Further on, the focus will be on item nonresponse, since this is most common in RCTs.

4.1.3 How much missing data is too much missing data?

The amount of missing data is of interest when analyzing a trial with missing data. Schulz and Grimes (2002) and Fielding et al. (2012) present the following assumptions about the amount of missing data: A trial with less than 5% missing data is regarded as a small amount and thus the bias will be minimal. Anything between 5% and 20% is an intermediate amount of missing data and could cause problems. In this case, missing data should be discussed and analyzed. More than 20% missing data is problematic and reduces the validity of the trial (Altman and Bland, 2007). Trials with more than 20% missing data are refused by some journals (Schulz and Grimes, 2002).
In 2014 Zhang et al. (2017) conducted a study to investigate how researchers handled missing data in 200 RCTs with continuous outcome. The study showed that 10% was the average amount of missing data in the trials. The amount of missing data in RCTs is usually so high that it should always be taken into account when planning the trial.

4.2 Missing data mechanisms

When there are missing data in a trial, it is important to investigate why the data are missing. This is called the missing data mechanisms and explains if the probability of missing data is dependent on some variables or not. When analyzing trials with missing data, the missing data mechanism is important to include in the analysis to get valid results. However, according to Zhang et al. (2017) and Rombach et al. (2016) less than 10% of the RCTs investigated in their trial report what type of the missing data mechanism they have assumed. The results of the analysis in a trial with missing data can be biased and the bias may be dependent on the missing data mechanism. A broadly used terminology is presented by Little and Rubin (2002) where the missing data mechanism is divided into three types: Missing completely at random, missing at random and missing not at random. Before presenting the missing data mechanisms we will describe missing data in mathematical terms.

Consider a data set with n units. For each unit p variables are observed. The response for unit $i = 1, 2, ..., n$ and variable $j = 1, 2, ..., p$ is given as y_{ij}. If the value of y_{ij} is missing, the value of y_{ij} is exchanged with a question mark, "?". This can be presented as a matrix, $\boldsymbol{y}$, of dimensions $(n \times p)$, given in Equation (4.1), where the values y_{12} and y_{n2} are missing as an illustration.

$$\boldsymbol{y} = \begin{bmatrix} y_{11} & ? & \dots & y_{1p} \\ y_{21} & y_{22} & \dots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & ? & \dots & y_{np} \end{bmatrix} \tag{4.1}$$

We divide this data set into two parts, one with missing data and one with observed data, given as

$$\boldsymbol{y} = (\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}). \tag{4.2}$$

We now define $\boldsymbol{R}$ as the $(n \times p)$ matrix, i.e. the same size as $\boldsymbol{y}$. The elements in $\boldsymbol{R}$ is given as R_{ij}, where each element indicates whether there are missing observations or not in the matrix $\boldsymbol{y}$. This is given as

$$R_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is missing,} \\ 0 & \text{else.} \end{cases} \tag{4.3}$$

Thus, the matrix $\boldsymbol{R}$, representing the missing values given the matrix in Equation (4.1) is given as

$$\boldsymbol{R} = \begin{bmatrix} 0 & 1 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \dots & 0 \end{bmatrix}.$$

The relation of how $\boldsymbol{R}$ is dependent of $\boldsymbol{y}$ describes the different missing data mechanisms.

4.2.1 Missing completely at random (MCAR)

Missing completely at random is defined as the situation where the probability that an observation is missing is not dependent of neither the observed nor unobserved data (Sterne et al., 2009). By the model given in Equations (4.1) and (4.3) a MCAR data is given as

$$P(\boldsymbol{R}|\boldsymbol{y}) = P(\boldsymbol{R}).$$

Whether the measured data are observed or not observed does not yield any systematic differences. This means that the unobserved data are not dependent of any variables observed in the trial, thus not dependent on the covariates or the parameter of interest. MCAR data will not affect the results of the analysis in any other way than less precise estimates. In this case, it is possible to use only the complete cases in the analysis without introducing bias in the results (Veieroed et al., 2012, p. 437). However, even if the assumption about MCAR data is fulfilled, it is not sufficient enough to show that the data truly are MCAR. This is because there may always be some unmeasured variables that are related to the missing values which we cannot take into account when checking if the data are MCAR. Thus, we can never be certain that the data are MCAR, and in practical trials data are seldom MCAR (Zhang et al., 2017). Assuming a MCAR situation when it is not the case can lead to severe errors on the results. Figure 4.1 illustrates the missing data in a MCAR situation where there is no systematic difference in the missing data.

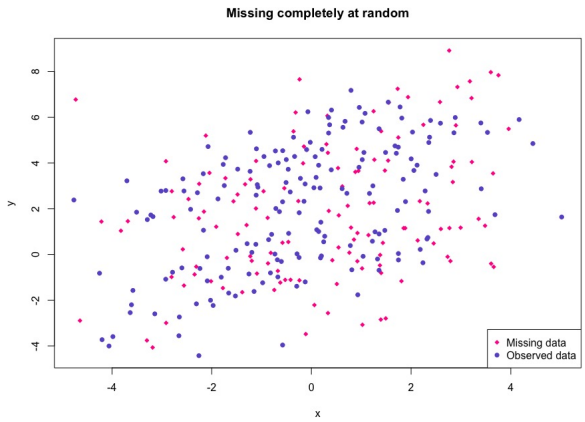


Figure 4.1: Illustration of data missing completely at random.

4.2.2 Missing at random (MAR)

Seldom there is no relation between the missing data and the observed or unobserved data as in the MCAR situation. When there is a systematic difference between the probability of unobserved data and observed values, the missing data mechanism is called missing at random (Sterne et al., 2009). This means that it is possible to find a systematic relation between the unobserved data and the observed data. By the equations given in the Equations

(4.1), (4.2) and (4.3), a MAR situation is given as

$$P(\boldsymbol{R}|\boldsymbol{y}) = P(\boldsymbol{R}|\boldsymbol{y}_{obs}).$$

Unbiased analysis of the data can still be carried out in the case of MAR. However, this is only true if we know that there are no relation between the unobserved data and the probability of missing data. If there is a MAR situation, there can for example be a relation between the missing data and the covariates. If there is a relation between a covariate and the missing data a true MAR situation is possible to detect. The probability of missing data, given this particular covariate should have the missing data mechanism MCAR. Thus, the missing data should be dependent only on the observed covariate and not some unobserved data. In the case of the missing data mechanism MAR, the analysis of the data should not be done only on the complete cases (remove all the units with unobserved values). This is statistically invalid and may produce biased results. The analysis should be done by conditioning on the observed variables which is related to the probability of missing data. By not taking the missing data mechanism MAR into account the results may be biased (Zhang et al., 2017). Assuming MAR when the data are truly not MAR, will often only have a minor impact on the results. MAR and MCAR data are said to be ignorable missing data mechanisms (Veieroed et al., 2012, p. 433). Figure 4.2 illustrates the situation with the missing data assumption MAR.

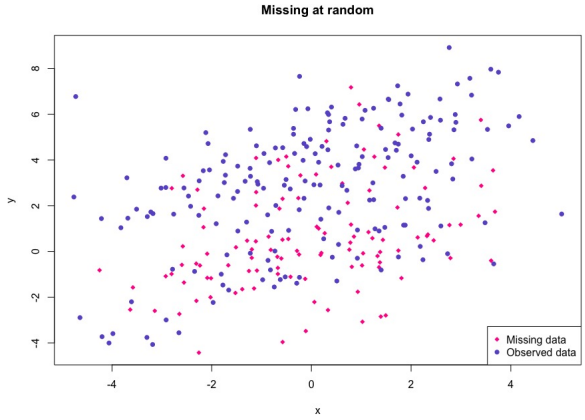


Figure 4.2: Visualization of data missing at random.

4.2.3 Missing not at random (MNAR)

Missing not at random is the missing data mechanism that is neither MAR nor MCAR. This means that even after taking the observed data into account, there is still differences between the probabilities of the missing data. Thus, the probability of missing data is dependent of unobserved (and observed) data. This is given as

$$P(\boldsymbol{R}|\boldsymbol{y}) = P(\boldsymbol{R}|(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis})).$$

Thus, we cannot get all the information needed to find enough information to do the best analysis. If the patients are lost to follow-up before the end of the study, i.e. withdrawal missing data, the situation is probably a MNAR situation (Altman and Bland, 2007). MNAR data are more difficult to analyze than MCAR and MAR data because the distribution differences between participants with missing data and without must be described to get valid analysis. This information is often not available to us because the information exists only in the missing data. Thus, there are no best way to deal with MNAR data, so sensitivity analyses should be conducted to compare the results of the estimates under different assumptions (Zhang et al., 2017; Sterne et al., 2009). Figure 4.3 illustrates the situation with MNAR missing data mechanism.

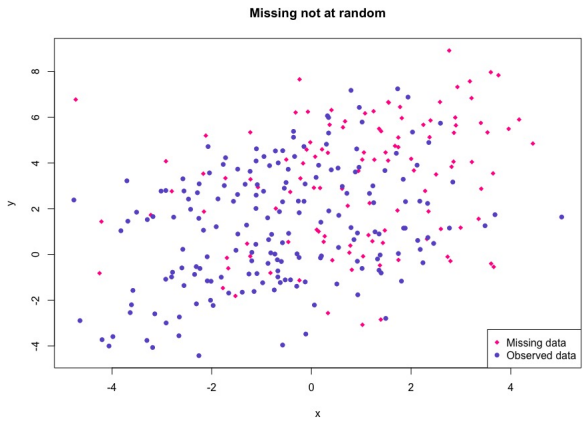


Figure 4.3: Visualization of data missing not at random.

4.3 Sensitivity analysis

A sensitivity analysis is defined by Thabane et al. (2013, p. 2) as ”a method to determine the robustness of an assessment by examining the extent to which results are affected by the changes in methods, models, values of unmeasured variables, or assumptions with the aim of identifying results that are most dependent on questionable or unsupported assumptions”. In other words, a sensitivity analysis describes how the results differ when different assumptions are made in a trial. If the assumptions do not affect the results in a trial, the results are said to be robust (Thabane et al., 2013). A sensitivity analysis is recommended when missing data occur in RCTs (Fielding et al., 2012; Rombach et al., 2016; Zhang et al., 2017). This can be done by conducting different strategies to handle missing data and will be presented in Section 4.4 (Fielding et al., 2012). Another way to conduct a sensitivity analysis is to assume that the missing data are extreme in one way or another and to see how the results varies (Zhang et al., 2017; Sterne et al., 2009). The choice of methods when conducting a sensitivity analysis also depends on the missing data mechanism (Thabane et al., 2013). When dealing with data that are MCAR and MAR,

there are good methods to analyze the data to get valid results. In the case of MNAR data, there is no good way to analyze the data to get valid results. In this case a sensitivity analysis is a good option to investigate how the results vary according to what methods are used to deal with missing data (Sterne et al., 2009). According to Zhang et al. (2017), less that 10% of the researchers investigated in their review had conducted a sensitivity analysis. They also reported that sensitivity analysis should always be conducted in the case of MNAR (and MAR) data.

4.4 Methods to handle missing data

There are several methods to deal with missing data, but as Zhang et al. (2017) have shown, not all methods that are commonly used are valid. Valid methods can be divided into three groups: Weighting procedures, imputation procedures and direct model based analysis. Weighting procedures adjust each observation based on the distribution of the sample. Imputation methods replace missing data with a substitute to make a complete data set. Direct model based analysis try to model the vairables for the data set with missing data to find the overall distribution of the data. If missing data depend on the variable of interest, then imputation procedures and direct model analysis are valid methods. If the missing data depend on auxiliary variables, then weighting methods are also valid methods that can be used. Auxiliary variables are variables that are included in the analysis only to improve the performance of the missing data method (Collins et al., 2002). This information does not need to be of interest of the trial. The method of choice when analyzing missing data depends also on whether there is unit or item nonresponse, and of what type of missing data mechanism there is. For example, weighting procedures are commonly used when there is unit nonresponse, while item nonresponse is often handled with imputation. Since item nonresponse is most common in RCTs, the focus will not be on weighting methods, but rather on imputation methods and direct model based procedures. Even though there are several methods to handle missing data, there is no solution that substitutes a complete data set. Thus, Altman and Bland (2007) highlights the importance of maximizing the data collection. Rombach et al. (2016) presents four commonly used categories of methods for handling missing data: Complete case analysis, single imputation, multiple imputation and model based approaches such as mixed effects models. These methods are widely used to handle missing data (Zhang et al., 2017; Fielding et al., 2012; Veieroed et al., 2012). Not all of these methods are recommended, but since also the methods that are not valid are also commonly used, all methods will be presented and discussed in the following subsections. In addition selection models and pattern mixture models will be presented.

4.4.1 Complete case analysis and available case analysis

The simplest method to deal with missing data is to exclude all units with missing data from the analysis. This is called complete case analysis. The advantage is that the method is simple, and most statistical methods can be used. The disadvantage is that the sample size is reduced, and the point estimates may be less precise. This means that confidence intervals can be wider and the p-values can be higher, due to higher variance. In addition the power of the analysis will decrease. Complete case analysis is only valid when the

missing data mechanism is MCAR, else bias due to missing data can be introduced. The percentage of missing data should be small if complete case analysis should be used. If not, the statistical power may be reduced (Altman and Bland, 2007). Sterne et al. (2009) makes one exception: If the missing data occur only in the outcome variable and the outcome variable is only measured once for each individual, then complete case analysis can be applied to MAR data. Even though complete case analysis is a method which is only valid under restricted limitations, many researchers use the method to conduct analysis (Sterne et al., 2009). According to Zhang et al. (2017), 67% of the analyzed RCTs used available data only in the analysis. An available case analysis takes all the available information for exactly that analysis and conducts complete case analysis on the subset of data needed for that specific analysis (Veieroed et al., 2012, p. 437). This means that for two different analyses, two different subsets can be used to conduct the analysis. Available case analysis can therefore get higher power and precision on the estimates compared to complete case analysis, but the analysis is based on different samples and sample sizes. As in the complete case analysis, the results will be biased if the missing data mechanism is not MCAR. If not data should be excluded from the analyses, some alternative method should be used to substitute the elements of the missing data. This can be done by using imputation methods.

4.4.2 Single imputation

Imputation methods are methods to create complete data sets with the purpose to use statistical methods where complete data is assumed. This is done by imputing data where there are missing values. The imputed values should be predicted using the observed data (Veieroed et al., 2012, p. 442). The values that are imputed should come from the predictive distribution of the data set, conditioning on the observed values. When an imputation method is used, all variables in the analysis should also be in the the imputation model. This includes interactions and covariates and in particular the variables that are known to be related to the missing data (Sterne et al., 2009). Other variables available, but not used in the analysis could also be included in the imputation model. When single imputation is done right, unbiased results can be carried out in the case of the missing data mechanism MAR (Zhang et al., 2017). However, the variance is often estimated to be too small (Sterne et al., 2009). Fielding et al. (2012) do not recommend the usage of simple imputation. Single imputation is illustrated in Equation (4.4), where the missing values are indicated with "?" and the imputed value for y_{ij} are indicated with $\hat{y}_{ij}$.

$$\begin{bmatrix} y_{11} & ? & y_{13} & y_{14} \\ y_{21} & y_{22} & y_{23} & y_{24} \\ y_{31} & ? & ? & y_{34} \\ ? & ? & y_{43} & y_{44} \end{bmatrix} \xrightarrow{\text{Imputation}} \begin{bmatrix} y_{11} & \hat{y}_{12} & y_{13} & y_{14} \\ y_{21} & y_{22} & y_{23} & y_{24} \\ y_{31} & \hat{y}_{32} & \hat{y}_{33} & y_{34} \\ \hat{y}_{41} & \hat{y}_{42} & y_{43} & y_{44} \end{bmatrix} \tag{4.4}$$

We are going to describe a few commonly used imputation methods to illustrate how imputation can be done. However, there are several other imputation methods available.

Last observation carried forward

The method of last observation carried forward (LOCF) is a type of single imputation that is done by replacing the missing values with the last observed value for the given participant. Given the situation illustrated in Equation (4.4) where y_{12} is missing, when using the method of LOCF then $\hat{y}_{12} = y_{11}$. This method is commonly used in longitudinal data analysis and for dropouts. This is a popular method to use because of its simplicity, but it is only valid under very special circumstances, and will never be valid for the multivariate normal distribution or any of the standard distributions (Carpender and Kenward, 2007, p. 33). Even though this method is rarely valid, it is still a commonly used method when analyzing RCTs in medical research (Zhang et al., 2017). It have been showed that the method leads to biased results and that it is also potentially worse than other methods (Zhang et al., 2017). Thus, this method should not be used under regular circumstances.

Hot-deck imputation

Using the method of Hot-deck imputation, the observed values constitutes a so called "donor pool", from which the imputed values are drawn (Veieroed et al., 2012, p. 443). The donor size and content of the donor pool can vary. This variation is dependent of which part of the observed data that are used to construct the donor pool. Each imputed value is randomly drawn from the donor pool. In the case of the missing data mechanism MCAR with no auxiliary information available, a missing observation is just imputed by a randomly drawn value from all of the observed values. If the missing data mechanism is MAR, the missing data is dependent of some information given in the observed values (this can for example be a group variable). In this case the donor pool should depend on the variable related to the probability of missing data. For example, given the situation where there are more missing data in one group, the donor pool should depend on the group variable. Hot-deck imputation is only recommended when there are no auxiliary information available. Auxiliary information are information that is available but not used in the analysis.

Regression imputation

When auxiliary variables are available, regression imputation is a commonly used method for item nonresponse (Veieroed et al., 2012). By using regression imputation, each missing observation is imputed by a predicted variable based on a regression model. In the case of a one-dimensional Y with one auxiliary variable x, the linear regression model is given by

$$Y = \alpha + \beta x + \varepsilon,$$

with the expected value and variance given by

$$\mathrm{E}(Y|x) = \alpha + \beta x \quad \text{and} \quad \mathrm{Var}(Y|x) = \mathrm{Var}(\epsilon) = \sigma^2.$$

Assuming the model is predicted by the response sample r and the least squares estimates given by $\hat{\alpha}_r$ and $\hat{\beta}_r$, the imputed value for the missing observation Y_i is given by

$$\hat{Y}_i = \hat{\alpha}_r + \hat{\beta}_r x_i.$$

This regression imputation is called the conditional mean imputation, and is one example of regression imputations. A drawback with this method is that there is not enough variation in the model to take into account the variance of the missing data. This problem can be taken into account by using residual regression imputation. For all the imputed values given in Equation (4.4.2), the residuals are computed by the equation

$$e_j = Y_j - \hat{Y}_j = Y_j - (\hat{\alpha}_r + \hat{\beta}_r x_j), \quad j \in r.$$

Thus, each imputed value is given by

$$\hat{Y}_i = \hat{\alpha}_r + \hat{\beta}_r x_i + e_i^*,$$

where e_i^* is drawn randomly from the complete set of all the residuals. Regression imputation is also a good imputation method when there are withdrawal missing data (Veieroed et al., 2012, p. 449). Say that the value y_t, $t > 1$ and the succeeding values $t+1, t+2, ...$ are missing. The value y_t may be imputed based on the value y_{t-1}, and so on.

4.4.3 Multiple imputation

Single imputation can give unbiased results in the case of the missing data assumptions MCAR or MAR, but the estimation of the variance is often too low (Rubin, 1987, p. 11). Thus, the method can turn out to be invalid. Multiple imputation (MI) is an improvement of single imputation. MI produces unbiased results, and the variance is reasonable high. Thus, it is known as a valid method in the case of MAR data. The interest and usage of the method have increased over the last years, due to the potential to increase the validity of the results of the analysis in a trial (Sterne et al., 2009). MI is a very flexible method and it can mimic a full model based analysis, which is regarded as the gold standard method (Veieroed et al., 2012, p. 430). By using MI, different imputation procedures are done several times to find an overall estimate for each imputed value. Thus, the uncertainty for the imputed values is also taken into account. Since also the imputed values are randomly drawn from a distribution, the method of MI increases the efficiency of the estimation compared to single imputation (Rubin, 1987, p. 16). An additional advantage of MI is that when using the method, m different imputations are generated for the same model. This means that in itself, the method of MI conducts a sensitivity analysis of the estimates (Rubin, 1987, p. 16). It is recommended to use MI in the case of MAR and MCAR data, but in the case of MNAR data MI may lead to biased results that can even be bigger than for the complete case analysis (Sterne et al., 2009). When conducting MI, each imputed data set is analyzed by a statistical method assuming complete data, which results in m estimates of the parameter of interest and m estimates for the variance of the estimates. This way, the analysis ignores the difference between the observed units with missing data and without missing data (Rubin, 1987). The number of repeated imputations, m, is usually set to $m = 20$ or more (Veieroed et al., 2012; Marshall et al., 2009; Sterne et al., 2009). Continuous data are assumed to be normally distributed when conducting MI, if not, transformation can be done before analyzing the data (Sterne et al., 2009).
When conducting MI in a trial, the parameter of interest, θ, is estimated m times. These estimates are given as

$$\hat{\theta}_j, \quad j = 1, ..., m. \tag{4.5}$$

The associated estimated variances are given by

$$\hat{\sigma}_j^2 = \text{Var}(\hat{\theta}_j), \quad j = 1, ..., m.$$

The calculation of the overall estimates of the parameter of interest by using MI was first presented by Rubin (1987), and is widely used today under the name Rubin's rule. The calculation arises from a Bayesian approach. Thus, it is assumed that the imputations are drawn to simulate a Bayesian posterior distribution of the missing data. By combining the analysis of each imputed data set, the result will be approximately valid. For complete data, it is assumed that

$$(\theta - \hat{\theta}) \sim \mathcal{N}(0, \sigma^2).$$

The m sets of repeated imputations are drawn to make m complete data sets, which results in the estimates for the parameter of interest and the variance, given in the Equation (4.5). The overall average estimate of the parameter of interest is given by

$$\bar{\theta} = \frac{1}{m} \sum_{j=1}^{m} \hat{\theta}_j.$$

The overall average of the variances is given by

$$W = \frac{1}{m} \sum_{j=1}^{m} \hat{\sigma}_j^2.$$

This is the variance within each imputation combined. The variance between the m estimates is given by

$$B = \frac{1}{m-1} \sum_{j=1}^{m} (\hat{\theta}_j - \bar{\theta})^2.$$

The total variance is given by the within imputation variance and between imputation variance. This is the total variance estimate of $(\theta - \hat{\theta})$, and is given by

$$T = W + (1 + \frac{1}{m})B.$$

The statistic $(\bar{\theta} - \theta)/\sqrt{T}$ is approximately distributed as a Student's t-distribution with ν degrees of freedom, given as

$$\nu = (m-1)(1 + r^{-1})^2. \tag{4.6}$$

Here, r represents the relative increase in variance due to missing data and is given as

$$r = (1 + m^{-1})\frac{B}{W}.$$

The $100(1-\alpha)\%$ confidence interval for the estimate $\bar{\theta}$ is given as

$$\left[\bar{\theta} - t_{(1-\alpha/2),\nu}\sqrt{T}, \quad \bar{\theta} + t_{(1-\alpha/2),\nu}\sqrt{T}\right].$$

The hypotheses for testing the significance of the parameter of interest is given as

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0. \tag{4.7}$$

The test statistic for the hypothesis test is given as

$$(\theta - \theta_0)^2/T,$$

and will be evaluated by comparing this to the F-statistic with 1 and ν degrees of freedom. Thus, the hypothesis test given in Equation (4.7) is rejected if

$$F_{1,\nu} < (\theta - \theta_0)^2/T.$$

Both the confidence interval and the p-values cannot be combined using Rubin's rule, because they change systematically with the sample size (Veieroed et al., 2012, p. 448). Statistics that can be estimated by Rubin's rule are the mean and standard deviation, and in addition proportions or regression coefficient. When applying Rubin's rule on the odds ratio, the parameters should be log-transformed before combining the different analysis. When calculating the correlation a Fisher's z-transformation should be used before combining the result. When applying Rubin's rule, the estimates of the parameters of interest must be different for each imputed data set. If the interest is to compare different estimates which comes from the same distribution and may be dependent, the same procedure is used. Rubin's rule can also be used in the case of k parameters of interest. Then the estimates from the m imputations are given as

$$\hat{\boldsymbol{\theta}}_j, \quad j = 1, ..., m,$$

where $\hat{\boldsymbol{\theta}}_j$ is a vector of length k. The associated covariance matrices are given as

$$\text{Var}(\hat{\boldsymbol{\theta}}_j) = \hat{\Sigma}_j,$$

which is of dimension $(k \times k)$. It is assumed that

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \sim \mathcal{N}(0, \Sigma),$$

The estimates found by Rubin's rule are given by

$$\bar{\boldsymbol{\theta}} = \frac{1}{m} \sum_{j=1}^{m} \hat{\boldsymbol{\theta}}_j,$$

with the associated variance of the estimates given by

$$\text{Var}(\hat{\boldsymbol{\theta}}_j) = \Sigma_j.$$

The average within covariance matrix is given by

$$\boldsymbol{W} = \frac{1}{m} \sum_{j=1}^{m} \Sigma_j,$$

The between variance is given as

$$\boldsymbol{B} = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\boldsymbol{\theta}}_j - \bar{\boldsymbol{\theta}})^T(\hat{\boldsymbol{\theta}}_j - \bar{\boldsymbol{\theta}}),$$

and thus the total variance is given by

$$\boldsymbol{T} = \boldsymbol{W} + (1 + m^{-1})\boldsymbol{B}.$$

The degrees of freedom is given as in Equation (4.6), but with r replaced with r_k given as

$$\boldsymbol{r_k} = \frac{1}{k}(1 + m^{-1})Tr(\boldsymbol{B}\bar{\boldsymbol{\theta}}^{-\mathbf{1}}),$$

where $Tr(\cdot)$ is the trace of the matrix, which is the sum of the diagonal elements of a matrix. The hypothesis for testing the significance of the parameter of interest is given by

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{vs} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

When m is large compared with k, say $m \geq 5k$, the test statistic for the hypothesis is given by

$$D = \frac{1}{k}(\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}})\boldsymbol{T}^{-1}(\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}})^T.$$

The hypothesis given in Equation (4.4.3) is rejected if

$$F_{k,\nu} < D.$$

Using this model, it is assumed that the covariance matrix, Σ_j, of the estimates $\hat{\boldsymbol{\theta}}$ is available. In practice the Σ_j are not always accessible, especially when k is large. Other method, which are presented by Marshall et al. (2009), is to use the method to combine χ^2 statistics or to use the method for combining the likelihood ratio χ^2 statistics. These methods are not as good as the one presented by Rubin (1987) and should only be used when necessary (Marshall et al., 2009).

4.4.4 Mixed-effects models

Mixed-effects models, which were presented in Section 3.3, can be used to analyze the data in the case of missing data in longitudinal trials (Rombach et al., 2016). This is a direct model based analysis. Using a mixed-effects model includes the unobserved variables that characterize the subject which are included in a random effect term different for the subjects (Diggle et al., 2001, p. 301). Using this method, imputation is not needed because complete data set is not necessary. Each outcome variable is modelled after several variables. Thus, if one observation is missing, the model can still be fitted (although not as precise) (Rombach et al., 2016). On the other hand, if there are missing data in the covariates, problems in the analysis could appear. This problem could be handled by using imputation methods only in the covariates. Using a mixed-effects model, the results can be unbiased if the missing data mechanism is MCAR or MAR (Sterne et al., 2009). Both multiple imputation and mixed-effects models have been proved to have good results

when handling missing data. However, Zhang et al. (2017) showed that these methods where not used any more frequent than poorer methods like LOCF. In addition Zhang et al. (2017) showed that the method of complete case analysis was used much more frequently than mixed-effects models. When missing data are MAR, mixed-effects models have been proved to have good statistical properties (Dinh and Yang, 2011).

4.4.5 Selection models and pattern mixture models

When the missing data mechanism is MNAR, none of the preceding models yield valid results in the case of missing data. A method that can be used is a model based approach (Veieroed et al., 2012, p. 425). We are only going to describe it briefly.
When a model based approach is used, assumptions for a model of joint distribution must be made. The joint distribution should be made for the variables and the response (Veieroed et al., 2012, p. 425). This can be done either by selection models or pattern mixture models. Selection models use the information from the missing data mechanism. The selection model takes into account the information about the missing data mechanism in the observed data and analyzes how it depends on the missing data. This changes the probability distribution of the missing data and thus the results from the analysis. Pattern mixture model is the method that looks at the distribution of the missing data depending on the observed data. The method of pattern mixture models analyze if the 'pattern' of the data differs from those with elements with missing data and those without.

Chapter 5

Statistical models for analyzing longitudinal RCTs

In Chapter 3 the theory for longitudinal data was introduced. In this chapter, this theory is used to make statistical models for the methods used to analyze RCTs presented in Section 2.3: Comparison of follow-up score, Change score analysis, ANCOVA and cLDA. These methods aim to analyze both longitudinal parallel-group RCTs and parallel-group RCTs with two time points. First, the mathematical model for the parallel-group longitudinal RCT is presented. Then, the methods to analyze the RCTs will be described mathematically. The parameter of interest is the expected difference between the two groups at a certain follow-up time point. This time point should be equal for both groups. For each method a hypothesis test will be conducted. The null hypothesis is given as "there are no difference between the groups", while the alternative hypothesis is that "there are differences between the groups". When describing the different methods to analyze RCTs, only the situation with two groups (one treatment group and one control group) is presented. In the situation with more than two groups, there are several parameter of interest; the pairwise differences between all the different groups. The analysis are similar to the situation with only two groups and is therefore easy to extend to more than two groups.

5.1 Mathematical Model

Notation

In a parallel-group RCT a total of n individuals are included in the trial. These are indexed as $i = 1, 2, ..., n$. In a longitudinal RCT, each individual is measured at $T + 1$ different time points, given as $t = 0, 1, ..., T$. Here, the baseline measurement, the measurement before intervention, is labeled $t = 0$. The time point $t = T$ is the last follow-up time point of the study. We assume that the time point of measurements are equal for all individuals. In addition, the difference between two time points, t and t', is assumed to be equal for

all individuals. The observations of the outcome variable of interest are given as y_{it} for individual i at time t. The n individuals are divided into J different groups, given by $j = 1, 2, ..., J$. The first group, $j = 1$, is given as the control group. The succeeding $J - 1$ groups, $j = 2, ..., J$, are given as the treatment groups. The total number of individuals in group j is given as n_j. Thus,

$$\sum_{j=1}^{J} n_j = n.$$

The total number of individuals measured at time t in group j are given as n_{jt}. If the data are complete, i.e. no data are missing, we have $n_{jt} = n_j$. The total number of observations for all the groups at all time points is given as

$$n_{tot} = \sum_{j=1}^{J} \sum_{t=0}^{T} n_{jt}.$$

If the data are complete, the total number of observations is given as

$$n_{tot} = \sum_{j=1}^{J} \sum_{t=0}^{T} n_{jt} = \sum_{j=1}^{J} \sum_{t=0}^{T} n_j = (T+1) \sum_{j=1}^{J} n_j = (T+1)n.$$

A group variable is assigned to each individual. This is given as

$$x_i = j \; = \; \begin{cases} 1 & \text{if } i \text{ in the control group,} \\ 2 & \text{if } i \text{ in treatment group 1,} \\ \vdots & \\ J & \text{if } i \text{ in treatment group } J-1. \end{cases} \tag{5.1}$$

This group variable is independent of time, since the individuals cannot change which group it is allocated to during a trial. The individual i has a set of observations, given as

$$\boldsymbol{y}_i = (y_{i0}, y_{i1}, ..., y_{iT})^T.$$

Expected values, variance, correlation and covariance

We assume that the observations, y_{it}, are normally distributed and that the expected values of the elements y_{it} are dependent of group, $x_i = j$, and time, t. This is given as

$$\mathrm{E}(y_{it}|x_i = j) = \mu_{jt} = \begin{cases} \mu_{10} & \text{if control group at baseline,} \\ \vdots & \\ \mu_{1T} & \text{if control group at last follow-up,} \\ \vdots & \\ \vdots & \\ \mu_{J0} & \text{if treatment group } J \text{ at baseline,} \\ \vdots & \\ \mu_{JT} & \text{if treatment group } J \text{ at last follow-up.} \end{cases} \tag{5.2}$$

In some trials the expected values for the control group are assumed to be equal for all time point, t, because the observed outcome variables, y_{it}, may not vary over time when no treatment is introduced. In this situation, the expected value is given as

$$\mathrm{E}(y_{it}|x_i = 1) = \mu_{1t} = \mu_1 \quad \forall\, t.$$

In other situations, the expected value for the control group is allowed to change over time, for example due to natural improvement over time. Then, the expected values are given as in Equation (5.2). The variance is assumed to be equal independent of group, $x_i = j$, and time, t, and is given as

$$\mathrm{Var}(y_{it}|x_i = j) = \sigma^2 \quad \forall\, t \text{ and } j. \tag{5.3}$$

Since each individual has several measurements over time, the observations within each individual are assumed to be dependent on each other. Thus, the outcome variables y_{it} at different time points within the same individual are assumed to be correlated. The correlation is given by

$$\mathrm{Corr}(y_{it}, y_{i't'}) = \begin{cases} 1 & \text{if } t = t' \text{ and } i = i', \\ \rho(t,t') & \text{if } t \neq t' \text{ and } i = i', \\ 0 & \text{else.} \end{cases} \tag{5.4}$$

Here, $\rho(t,t')$, is the correlation between different time points within an individual, which can be dependent on time. Thus, each individual can be modelled by a correlation matrix, V_*, of dimension $((T+1)\times(T+1))$. Examples of different correlation matrix structures are given in the Equations (3.10) and (3.11). The covariance is given as

$$\mathrm{Cov}(y_{it}, y_{i't'}) = \begin{cases} 1 & \text{if } t = t' \text{ and } i = i', \\ \sigma^2\rho(t,t') & \text{if } t \neq t' \text{ and } i = i', \\ 0 & \text{else.} \end{cases} \tag{5.5}$$

Full model

The set of all the observations are assumed to be normally distributed, given as

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma), \tag{5.6}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_{10} \\ \vdots \\ y_{1T} \\ \vdots \\ \vdots \\ y_{n0} \\ \vdots \\ y_{nT} \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_{10} \\ \vdots \\ \mu_{1T} \\ \vdots \\ \vdots \\ \mu_{J0} \\ \vdots \\ \mu_{JT} \end{pmatrix} \quad \text{and} \quad \Sigma = \sigma^2 \begin{pmatrix} [V_*] & & 0 \\ & \ddots & \\ 0 & & [V_*] \end{pmatrix}.$$

Here, $\boldsymbol{y}$ and $\boldsymbol{\mu}$ are vectors of length $n(T+1)$ consisting of the observations and the expected values. The covariance matrix Σ, is of dimension $(n(T+1) \times n(T+1))$. It is a block-diagonal matrix with the correlation matrix, V_*, for each individual on the main diagonal and zero elsewhere.

Model

The total set of observations in a longitudinal parallel-group RCT is given as in Table 5.1.

Id	**Group**	**Observations**		
i	$x_i = j$	y_{i0}	$\cdots$	y_{iT}
1	1	y_{10}	$\cdots$	y_{1T}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n_1	1	$y_{n_1 0}$	$\cdots$	$y_{n_1 T}$
n_1+1	2	$y_{(n_1+1)0}$	$\cdots$	$y_{(n_1+1)T}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n_2	2	$y_{(n_2)0}$	$\cdots$	$y_{(n_2)T}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$n_{(J-1)+1}$	J	$y_{(n_{J-1}+1)0}$	$\cdots$	$y_{(n_{J-1}+1)T}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	J	y_{n0}	$\cdots$	y_{nT}

Table 5.1: Observations in a longitudinal trial.

5.2 Methods to analyze RCTs

5.2.1 Comparison of mean at follow-up

The method of comparisons of the means of the follow-up scores can be used at parallel-group RCTs with two time-points: Baseline and follow-up. However, using this method only the follow-up scores are used in the analysis. The mean values at follow-up for the two groups are given by

$$\mathrm{E}(Y_{i1}|x_i = j) = \begin{cases} \mu_{11} & \text{if } j = 1, \\ \mu_{21} & \text{if } j = 2, \end{cases}$$

and the variance is given as

$$\mathrm{Var}(Y_{i1}|x_i = j) = \sigma^2 \quad \forall\, i, j.$$

The parameter of interest is given as

$$\theta_f = E(Y_{i1}|x_i = 2) - E(Y_{i'1}|x_{i'} = 1) = \mu_{21} - \mu_{11}.$$

The hypotheses to be tested are

$$H_0 : \theta_f = 0 \qquad \text{vs.} \qquad H_1 : \theta_f \neq 0. \tag{5.7}$$

The estimators for the expected values are the empirical mean values, given as

$$\hat{\mu}_{11} = \frac{1}{n_1}\sum_{i=1}^{n_1}(Y_{i1}|x_i = 1) \quad \text{and} \quad \hat{\mu}_{21} = \frac{1}{n_2}\sum_{i=1}^{n_2}(Y_{i1}|x_i = 2).$$

These two equations give us an estimator for the parameter of interest, θ_f. This is given as

$$\hat{\theta}_f = \hat{\mu}_{21} - \hat{\mu}_{11}.$$

The estimators of the variances for the two groups are given by

$$S_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(Y_{i1} - \hat{\mu}_{j1})^2 \quad \text{for } j = 1, 2.$$

A two-sample t-test for two independent groups with assumed equal variance is used to test the hypothesis given in Equation (5.7). The test statistic is given as

$$T_f = \frac{\hat{\theta}_f}{S_f\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

and follows a t_{n-2} distribution under H_0. Here is S_f is the pooled standard deviation, given as

$$S_f = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}.$$

A $(1 - \alpha)100\%$ confidence interval for the parameter θ_f is given as

$$\left[\hat{\theta}_f - t_{\alpha/2,\nu}\sqrt{S_f^2/n},\ \hat{\theta}_f + t_{\alpha/2,\nu}\sqrt{S_f^2/n}\right],$$

where $\nu = n - 2$ is the number of degrees of freedom and $n = n_1 + n_2$.

5.2.2 Change score analysis

The method of comparison of change scores is another method that can be used when there are only two time points. Here we introduce a new variable, the difference between follow-up and baseline for each participant. This is given as

$$D_i = Y_{i1} - Y_{i0}.$$

The expected mean values for the change from baseline to follow-up are given as

$$\mu_j = E(D_i|x_i = j) = E((Y_{i1}|x_i = j) - (Y_{i0}|x_i = j)).$$

Thus,

$$\mu_j = \begin{cases} \mu_{11} - \mu_{10} & \text{if} \quad j = 1, \\ \mu_{21} - \mu_{20} & \text{if} \quad j = 2. \end{cases}$$

The variance for the change score is independent of group and is given as

$$\begin{aligned} \text{Var}(D_i) &= \text{Var}(Y_{i1} - Y_{i0}) \\ &= \text{Var}(Y_{i1}) + \text{Var}(Y_{i0}) - 2 \cdot \text{Cov}(Y_{i1}, Y_{i0}) \\ &= \sigma^2 + \sigma^2 - 2\rho\sigma^2 = 2\sigma^2(1-\rho). \end{aligned}$$

The expected difference between the two groups (the parameter of interest) is given as

$$\theta_c = E(D_i | x_i = 2) - E(D_i' | x_i' = 1) = \mu_1 - \mu_0.$$

Since the change score is the difference between two normally distributed variables, the change score is also normally distributed, given as

$$\boldsymbol{D} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma),$$

where

$$\boldsymbol{D} = \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} 2\sigma^2(1-\rho) & & 0 \\ & \ddots & \\ 0 & & 2\sigma^2(1-\rho) \end{pmatrix}.$$

In $\boldsymbol{\mu}$, the first n_1 entries are given as μ_1, and the last n_2 entries are given as μ_2. The hypotheses to be tested for the method of change scores analysis are given as

$$H_0 : \theta_c = 0 \qquad \text{vs.} \qquad H_1 : \theta_c \neq 0. \tag{5.8}$$

The estimates for the expected values are the empirical mean values given as

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} D_i = \begin{cases} \hat{\mu}_1 & \text{if} \quad j = 1, \\ \hat{\mu}_2 & \text{if} \quad j = 2. \end{cases}$$

This results in the estimator for the parameter of interest, given as

$$\hat{\theta}_c = \hat{\mu}_2 - \hat{\mu}_1.$$

The parameter estimators for the variances of the change score is given as

$$S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} ((D_i) - \hat{\mu}_j)^2 \quad \text{for} \quad j = 1, 2.$$

A two-samples t-test for two independent groups with assumed equal variance is used to test the hypothesis given in Equation (5.8). The test statistic is given as

$$T_c = \frac{\hat{\theta}_c}{S_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

which follows a t_{n-2} distribution. Here S_c is the pooled standard deviation given as

$$S_c = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}.$$

The $(1-\alpha)100\%$ confidence interval for the parameter of interest, θ_c, is given as

$$\left[\hat{\theta}_c - t_{\alpha/2,\nu} S_c \sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\ \hat{\theta}_c + t_{\alpha/2,\nu} S_c \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right],$$

with number of degrees of freedom equal to $\nu = n_1 + n_2 - 2$. If the expected value at baseline for the two groups are equal ($\mu_{10} = \mu_{00}$), as assumed for RCTs, the expected difference between the two groups by using the method of change score is equal to the difference found by using the method of follow-up. This is given as

$$\theta_c = \mu_{21} - \mu_{20} - (\mu_{11} - \mu_{10}) = \mu_{21} - \mu_{20} - \mu_{11} + \mu_{20} = \mu_{21} - \mu_{11} = \theta_f. \quad (5.9)$$

5.2.3 ANCOVA

We are now moving on to the methods that are commonly used to analyze longitudinal RCTs. When using the method of longitudinal ANCOVA, a linear mixed effect model is used. Here, the outcome variables are conditioning on the baseline observation. The group variable and time variable are modelled as interaction terms.
The model for the method of ANCOVA is given as

$$Y_{it} = \alpha_t Y_{i0} + \beta_{jt}\mathrm{I}(\text{time} = t)\mathrm{I}(x_i = j) + \varepsilon_{it} + u_i, \quad t = 1, ..., T, \quad (5.10)$$

where the error terms, ε_{it} and u_i, are assumed to be normally distributed with zero-mean and equal variance. This is given as

$$\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2) \quad \text{and} \quad u_i \sim \mathcal{N}(0, \sigma_u^2).$$

Here, ε_{it} represents the within variance for an individual. The error term u_i is the random intercept, which represents the between individual variance. The marginal mean at time t conditional on baseline, Y_{i0}, given as

$$\mathrm{E}(Y_{it}|Y_{i0} = y_{i0}, x_i = j) = \alpha_t y_{i0} + \beta_{jt} \quad t = 1, 2, ..., T.$$

For each time point, $\alpha_t Y_{i0}$ is assumed to be equal (since we assume equal baseline values in both groups), while β_{jt} is different for the groups. This is the effect of treatment j at time t after adjusting for baseline effect. The variance of the regression model is given as

$$\mathrm{Var}(Y_{it}|Y_{i0}, x_i = j) = \mathrm{Var}(\alpha_t Y_{i0} + \beta_{jt}\mathrm{I}(\text{time} = t)\mathrm{I}(x_i = j) + \varepsilon_{it} + u_i) = \sigma_\varepsilon^2 + \sigma_u^2.$$

The covariance between two outcome variables is given as

$$\mathrm{Cov}(Y_{it}, Y_{i't'}) = \begin{cases} \sigma_u^2 + \sigma_\epsilon^2 & \text{if } i = i', t = t', \\ \sigma_u^2 & \text{if } i = i', t \neq t', \\ 0 & \text{else,} \end{cases}$$

thus the correlation given as

$$\text{Corr}(Y_{it}, Y_{i't'}) = \frac{\text{Cov}(Y_{it}, Y_{i't'})}{\sqrt{\text{Var}(Y_{it})}\sqrt{\text{Var}(Y_{i't'})}} = \begin{cases} 1 & \text{if } i = i', t = t', \\ \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2} & \text{if } i = i', t \neq t', \\ 0 & \text{else.} \end{cases}$$

The intraclass correlation is given as

$$\text{ICC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}.$$

The Equation (5.10) represents one individual i at one time point t. This equation can be extended to the set of all individuals at all time points by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1T} \\ \vdots \\ \vdots \\ Y_{n1} \\ \vdots \\ Y_{nT} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1T} \\ \vdots \\ \vdots \\ \varepsilon_{n1} \\ \vdots \\ \varepsilon_{nT} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_T \\ \beta_{11} \\ \vdots \\ \beta_{1T} \\ \beta_{21} \\ \vdots \\ \beta_{2T} \end{pmatrix}.$$

Here, $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$ are nT-vectors and $\boldsymbol{\beta}$ is a $3T$-vector. The matrix $\boldsymbol{X}$ is of dimensions $(3T \times nT)$, given as

$$\boldsymbol{X} = \begin{pmatrix} Y_{10} & \dots & Y_{10} & \text{I}(x_1 = 1)\text{I}(t = 1) & \dots & \text{I}(x_1 = 1)\text{I}(t = T) & \text{I}(x_1 = 2)\text{I}(t = 1) & \dots & \text{I}(x_1 = 2)\text{I}(t = T) \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ Y_{10} & \dots & Y_{10} & \text{I}(x_1 = 1)\text{I}(t = 1) & \dots & \text{I}(x_1 = 1)\text{I}(t = T) & \text{I}(x_1 = 2)\text{I}(t = 1) & \dots & \text{I}(x_1 = 2)\text{I}(t = T) \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ Y_{n0} & \dots & Y_{n0} & \text{I}(x_n = 1)\text{I}(t = 1) & \dots & \text{I}(x_n = 1)\text{I}(t = T) & \text{I}(x_n = 2)\text{I}(t = 1) & \dots & \text{I}(x_n = 2)\text{I}(t = T) \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ Y_{n0} & \dots & Y_{n0} & \text{I}(x_n = 1)\text{I}(t = 1) & \dots & \text{I}(x_n = 1)\text{I}(t = T) & \text{I}(x_n = 2)\text{I}(t = 1) & \dots & \text{I}(x_n = 2)\text{I}(t = T) \end{pmatrix}$$

The parameter of interest, θ_t, is the difference between the groups, at time t, adjusted for baseline. This is given as

$$\begin{aligned} \theta_t &= \text{E}(Y_{it}|Y_{i0}, x_i = 2) - \text{E}(Y_{it}|Y_{i0}, x_i = 1) \\ &= (\alpha_t Y_{i0} + \beta_{2t}) - (\alpha_t Y_{i0} + \beta_{1t}) \\ &= \beta_{2t} - \beta_{1t}. \end{aligned}$$

The hypotheses for the parameters of interest, θ_t, are given as

$$H_0 : \theta_t = 0 \quad \text{vs} \quad H_1 : \theta_t \neq 0, \tag{5.11}$$

or given in terms of the regression coefficients

$$H_0 : \beta_{2t} - \beta_{1t} = 0 \quad \text{vs} \quad H_1 : \beta_{2t} - \beta_{1t} \neq 0.$$

This can be written in matrix notation as

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = 0 \quad \text{vs} \quad H_1 : \boldsymbol{C}\boldsymbol{\beta} \neq 0,$$

where $\boldsymbol{C}$ is a $3T$ row vector given as

$$\boldsymbol{C} = \begin{cases} -1 & \text{if element number } T + t \text{ in } \boldsymbol{\beta}, \\ 1 & \text{if element number } 2T + t \text{ in } \boldsymbol{\beta}, \\ 0 & \text{else.} \end{cases}$$

An estimator for $\boldsymbol{\beta}$ is found by using the method of maximum likelihood as given in Equation (3.26). This is given by

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\alpha}_1 & \dots & \hat{\alpha}_T & \hat{\beta}_{11} & \dots & \hat{\beta}_{1T} & \hat{\beta}_{21} & \dots & \hat{\beta}_{2T} \end{pmatrix}^T = (\boldsymbol{X}^T V \boldsymbol{X})^{-1} \boldsymbol{X}^T V \boldsymbol{y}.$$

Here, V is the covariance matrix given in Equation (3.15). The estimator for the parameter of interest, θ_t, is thus

$$\hat{\theta}_t = \hat{\beta}_{2t} - \hat{\beta}_{1t} = \boldsymbol{C}\hat{\boldsymbol{\beta}} = \boldsymbol{C}(\boldsymbol{X}^T V \boldsymbol{X})^{-1} \boldsymbol{X}^T V \boldsymbol{y}.$$

The estimator of the variance is given as

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \varepsilon_{it}^2.$$

The covariance matrix for $\boldsymbol{C}\hat{\boldsymbol{\beta}}$ is given as

$$\widehat{\text{Cov}}(\boldsymbol{C}\hat{\boldsymbol{\beta}}) = \hat{\sigma}_\varepsilon^2 \boldsymbol{C}(\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1} \boldsymbol{C}^T.$$

When assuming normal errors under H_0, the test statistic is given as

$$T_a = \frac{\boldsymbol{C}\hat{\boldsymbol{\beta}}}{\sqrt{\widehat{\text{Cov}}(\boldsymbol{C}\hat{\boldsymbol{\beta}})}} = \frac{\hat{\beta}_{2t} - \hat{\beta}_{1t}}{\sqrt{\hat{\sigma}_\varepsilon^2 (\boldsymbol{C}(\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1} \boldsymbol{C}^T)}}.$$

The hypothesis in Equation (5.13) is rejected if

$$T_a > t_{(1-\alpha/2),1,n-3T}.$$

A confidence interval for $\boldsymbol{C}\hat{\boldsymbol{\beta}} = (\hat{\beta}_{1t}, \hat{\beta}_{2t})$ is thus given by

$$\left[\boldsymbol{C}\hat{\boldsymbol{\beta}} - t_{(1-\alpha/2),1,n-3T} \sqrt{\widehat{\text{Cov}}(\boldsymbol{C}\hat{\boldsymbol{\beta}})},\ \boldsymbol{C}\hat{\boldsymbol{\beta}} + t_{(1-\alpha/2),1,n-3T} \sqrt{\widehat{\text{Cov}}(\boldsymbol{C}\hat{\boldsymbol{\beta}})} \right].$$

5.2.4 cLDA

The method of cLDA can also be used on longitudinal data. Here, the baseline values and follow-up values are assumed to be jointly multivariate normally distributed, and a linear mixed effects model is used. The parameter estimates when using the method of cLDA are asymptotically unbiased when based on full likelihood functions (Liu et al., 2009). The difference between cLDA and ANCOVA is that the method of ANCOVA condition on the baseline values. Using the method of cLDA, the baseline value is modelled as a part of the outcome vector similar to all other outcome values. The model for the method of cLDA is given by

$$Y_{it} = \gamma_0 + \gamma_{jt}\mathrm{I}(x_i = j)\mathrm{I}(\text{time} = t, t > 0) + \varepsilon_{it} + u_i \quad t = 0, ..., T. \tag{5.12}$$

This is a special case of the linear mixed effects model, where the random effect is a random intercept. There are two error terms in this model: ε_{it} and u_i, both are assumed to be zero-mean normal variables, given as

$$\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2) \quad \text{and} \quad u_i \sim \mathcal{N}(0, \sigma_u^2).$$

The first error term, ε_{it}, represents the within variance in an individual, while u_i, is the random intercept which represents the between individual variance. The expected value of Y_{it} from Equation (5.12) is given as

$$\mathrm{E}(Y_{it}|x_i = j) = \gamma_0 + \gamma_{jt}\mathrm{I}(x_i = j)\mathrm{I}(\text{time} = t, t > 0) \quad t = 0, ..., T.$$

Thus, if $t = 0$, the expected values are assumed to be equal for all i independent of group, x_i. The variance of the model is given as

$$\mathrm{Var}(Y_{it}|x_i = j) = \mathrm{Var}(\gamma_0 + \gamma_{jt}\mathrm{I}(x_i = j)\mathrm{I}(\text{time} = t, t > 0) + \varepsilon_{it} + u_i) = \sigma_\varepsilon^2 + \sigma_u^2.$$

The covariance between two outcome variables is given as

$$\mathrm{Cov}(Y_{it}, Y_{i't'}) = \begin{cases} \sigma_u^2 + \sigma_\epsilon^2 & \text{if } i = i', t = t', \\ \sigma_u^2 & \text{if } i = i', t \neq t', \\ 0 & \text{else,} \end{cases}$$

thus the correlation given as

$$\mathrm{Corr}(Y_{it}, Y_{i't'}) = \frac{\mathrm{Cov}(Y_{it}, Y_{i't'})}{\sqrt{\mathrm{Var}(Y_{it})}\sqrt{\mathrm{Var}(Y_{i't'})}} = \begin{cases} 1 & \text{if } i = i', t = t', \\ \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2} & \text{if } i = i', t \neq t', \\ 0 & \text{else.} \end{cases}$$

The intraclass correlation is given as

$$\mathrm{ICC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}.$$

The Equation (5.12) can be represented in matrix notation for all individuals at all time points, given as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} + \boldsymbol{u},$$

where

$$\boldsymbol{Y} = \begin{pmatrix} Y_{10} \\ \vdots \\ Y_{1T} \\ \vdots \\ \vdots \\ Y_{n0} \\ \vdots \\ Y_{nT} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{10} \\ \vdots \\ \varepsilon_{1T} \\ \vdots \\ \vdots \\ \varepsilon_{n0} \\ \vdots \\ \varepsilon_{nT} \end{pmatrix}, \quad \boldsymbol{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_1 \\ \vdots \\ \vdots \\ u_n \\ \vdots \\ u_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\gamma} = \begin{pmatrix} \gamma_0 \\ \gamma_{11} \\ \vdots \\ \gamma_{1T} \\ \gamma_{21} \\ \vdots \\ \gamma_{2T} \end{pmatrix}.$$

The vectors $\boldsymbol{Y}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\gamma}$ are of length $(n(T+1))$. Each element in $\boldsymbol{u}$, u_i, is repeated $T+1$ times. The vector of regression coefficients, $\boldsymbol{\gamma}$, is of length $2T+1$. The design matrix, $\boldsymbol{X}$, is a $(n(T+1) \times (2T+1))$ matrix, given as

$$\boldsymbol{X} = \begin{pmatrix} 1 & \mathrm{I}(x_1 = 1)\mathrm{I}(t = 1) & \dots & \mathrm{I}(x_1 = 1)\mathrm{I}(t = T) & \mathrm{I}(x_1 = 2)\mathrm{I}(t = 1) & \dots & \mathrm{I}(x_1 = 2)\mathrm{I}(t = T) \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & \mathrm{I}(x_1 = 1)\mathrm{I}(t = 1) & \dots & \mathrm{I}(x_1 = 1)\mathrm{I}(t = T) & \mathrm{I}(x_1 = 2)\mathrm{I}(t = 1) & \dots & \mathrm{I}(x_1 = 2)\mathrm{I}(t = T) \\ \vdots & & \vdots & & & \vdots & \\ \vdots & & \vdots & & & \vdots & \\ \vdots & & \vdots & & & \vdots & \\ 1 & \mathrm{I}(x_n = 1)\mathrm{I}(t = 1) & \dots & \mathrm{I}(x_n = 1)\mathrm{I}(t = T) & \mathrm{I}(x_n = 1)\mathrm{I}(t = 1) & \dots & \mathrm{I}(x_n = 2)\mathrm{I}(t = T) \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & \mathrm{I}(x_n = 1)\mathrm{I}(t = 1) & \dots & \mathrm{I}(x_n = 1)\mathrm{I}(t = T) & \mathrm{I}(x_n = 2)\mathrm{I}(t = 1) & \dots & \mathrm{I}(x_n = 2)\mathrm{I}(t = T) \end{pmatrix}$$

The parameter of interest is the difference between the groups at a certain time point, t, given as

$$\begin{aligned} \theta_t &= \mathrm{E}(Y_{it}|x_i = 2) - \mathrm{E}(Y_{it}|x_i = 1) \\ &= (\gamma_0 + \gamma_{2t}\mathrm{I}(x_i = 2)\mathrm{I}(\text{time} = t, t > 0)) - (\gamma_0 + \gamma_{1t}\mathrm{I}(x_i = 1)\mathrm{I}(\text{time} = t, t > 0)) \\ &= \gamma_{2t}\mathrm{I}(x_i = 2)\mathrm{I}(\text{time} = t, t > 0) - \gamma_{1t}\mathrm{I}(x_i = 1)\mathrm{I}(\text{time} = t, t > 0) \\ &= \gamma_{2t} - \gamma_{1t}. \end{aligned}$$

The hypotheses for the parameters of interest, θ_t, are given as

$$H_0 : \theta_t = 0 \quad \text{vs} \quad H_1 : \theta_t \neq 0, \tag{5.13}$$

or given in terms of the regression coefficients

$$H_0 : \gamma_{2t} - \gamma_{1t} = 0 \quad \text{vs} \quad H_1 : \gamma_{2t} - \gamma_{1t} \neq 0.$$

This can be written in matrix notation as

$$H_0 : \boldsymbol{C}\boldsymbol{\gamma} = 0 \quad \text{vs} \quad H_1 : \boldsymbol{C}\boldsymbol{\gamma} \neq 0,$$

where $\boldsymbol{C}$ is a $2T+1$ row vector given as

$$\boldsymbol{C} = \begin{cases} -1 & \text{if element number } t+1, \\ 1 & \text{if element number } T+t+1, \\ 0 & \text{else.} \end{cases}$$

An estimator for $\boldsymbol{\gamma}$ is found by using the method of maximum likelihood as given in Equation (3.26). This is given by

$$\hat{\boldsymbol{\gamma}} = \begin{pmatrix} \hat{\gamma}_0 & \hat{\gamma}_{11} & \dots & \hat{\gamma}_{1T} & \hat{\gamma}_{21} & \dots & \hat{\gamma}_{2T} \end{pmatrix}^T = (\boldsymbol{X}^T V \boldsymbol{X})^{-1} \boldsymbol{X}^T V \boldsymbol{y},$$

where V is the covariance matrix of $\boldsymbol{Y}$. The estimator for the parameter of interest, θ_t, is thus

$$\hat{\theta}_t = \gamma_{2t} - \gamma_{1t} = \boldsymbol{C}\hat{\boldsymbol{\gamma}} = \boldsymbol{C}(\boldsymbol{X}^T \hat{V} \boldsymbol{X})^{-1} \boldsymbol{X}^T \hat{V} \boldsymbol{y}.$$

Here, $\hat{V}$, is the REML estimator of V when unknown. The estimator of the variance is given as

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n(T+1)} \sum_{i=1}^{n} \sum_{t=0}^{T} \varepsilon_{it}^2.$$

The covariance matrix for $\boldsymbol{C}\hat{\boldsymbol{\gamma}}$ is given as

$$\widehat{\text{Cov}}(\boldsymbol{C}\hat{\boldsymbol{\gamma}}) = \hat{\sigma}_\varepsilon^2 \boldsymbol{C}(\boldsymbol{X}^T \hat{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{C}^T.$$

When assuming normal errors under H_0, the test statistic is given as

$$T_c = \frac{\boldsymbol{C}\hat{\boldsymbol{\gamma}}}{\sqrt{\widehat{\text{Cov}}(\boldsymbol{C}\hat{\boldsymbol{\gamma}})}} = \frac{\hat{\gamma}_{2t} - \hat{\gamma}_{1t}}{\sqrt{\hat{\sigma}_\varepsilon^2 (\boldsymbol{C}(\boldsymbol{X}^T \hat{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{C}^T)}}.$$

The hypothesis in Equation (5.13) is rejected if

$$T_c > t_{(1-\alpha/2),1,n-(2T+1)}.$$

A confidence interval for $\boldsymbol{C}\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_{1t}, \hat{\gamma}_{2t})$ is thus given by

$$\left[\boldsymbol{C}\hat{\boldsymbol{\gamma}} - t_{(1-\alpha/2),1,n-(2T+1)} \sqrt{\widehat{\text{Cov}}(\boldsymbol{C}\hat{\boldsymbol{\gamma}})},\ \boldsymbol{C}\hat{\boldsymbol{\gamma}} + t_{(1-\alpha/2),1,n-(2T+1)} \sqrt{\widehat{\text{Cov}}(\boldsymbol{C}\hat{\boldsymbol{\gamma}})} \right].$$

Chapter 6

Simulation study

A simulation study have been conducted in this report to reveal result differences between the different methods to analyze RCTs that are described in Chapter 5.2. The methods are comparisons of means of follow-up scores, change score analysis, ANCOVA and cLDA. The different simulated scenarios have been conducted with and without missing data and with different missing data mechanisms. In addition, different correlation structures and different percentages of missing data have been simulated. The simulation study is inspired by the simulation study conducted in the article by Liu et al. (2009). This chapter is arranged as follows: The model used in the simulations are described in Section 6.1. The different scenarios for the simulations are presented in Section 6.2. At last, the results are presented in Section 6.3.

6.1 Model for simulations

In each simulation it is assumed that there are two groups, one control group ($x_i = 1$) and one treatment group ($x_i = 2$). The aim of the simulations is to find the parameter of interest, θ_t, which is the difference between the two groups at time t. The parameter of interest, θ_t, is given by

$$\theta_t = \mathrm{E}(Y_{it}|x_i = 2) - \mathrm{E}(Y_{it}|x_i = 1).$$

Of main interest is the difference in means at the last follow-up time point, θ_T, but the differences for $t < T$ are also included in the results of the simulations with longitudinal data. There are $m = 5000$ simulations conducted for each simulated scenario. All simulations are conducted with $n = 100$ individuals, with $n/2$ individuals in each group. These numbers are chosen according to the simulation study conducted by Liu et al. (2009), and should result in approximately $80-90\%$ power in the case of no missing data at the parameter of interest at the last follow-up time point. This is a typical requirement in a simulation study (Vickers, 2001). The power represents the proportion of false negative findings of the hypothesis, and is an important quantity when analyzing a statistical method (Vickers,

2001). The result of the simulations are the average bias of the parameter of interest for each of the m simulations, in addition to the average power of the estimate and the coverage of the confidence interval with significance level $\alpha = 0.05$.
The simulations have been conducted with different types of missing data mechanisms: No missing data, missing completely at random, missing at random and missing not at random. The simulations are conducted in *R* (R Core Team, 2017). The data sets with missing data have been generated by the function *ampute* in the package *mice* (van Buuren and Groothuis-Oudshoorn, 2011). The missing data are modelled as interim missing data. The simulations are conducted with a standard amount of missing data of 10%, since this is reported as the average amount of missing data in RCTs (Zhang et al., 2017). However, simulation studies with different amount of missing data have also been conducted. When analyzing data by using the method of ANCOVA, missing data cannot occur at baseline since this is a part of the design matrix. Thus, both analyses conducted by available case analysis (analysis of only the data where baseline value exists) and analysis with multiple imputation at missing baseline values have been conducted. Using the method of cLDA, all data are used in the analysis. For both the method of change score and the method of follow-up, complete case analysis is used. Multiple imputation is done with $M = 20$ imputations by the function *mice* in the *mice* package (van Buuren and Groothuis-Oudshoorn, 2011). When MI is used, the different outcome variables and the group variable were used in the imputation model.
Both the method of cLDA and ANCOVA are modeled using the function *lme* in the package *nlme* (Pinheiro et al., 2018). The resulting estimates, p-values and confidence intervals of the group differences were found by using the function *estimable* in the package *gmodels* (Warnes et al., 2015). Simulations have been conducted both with longitudinal data ($T = 4$) and with only baseline and one follow-up time point ($T = 2$). The simulations with two time point are conducted to compare the methods of follow-up score and change score analysis. The number of time points at longitudinal data equal to $T = 4$ is also used in the simulation study conducted by Liu et al. (2009). In addition has Coffman et al. (2016) stated that $2 - 4$ follow-up measurements are common in longitudinal RCTs. The simulated values are drawn from a multivariate normal distribution. The expected values at baseline are equal for both groups and are given as $\mu_{j0} = 0 \quad \forall\, j$. The expected value at the last follow-up for the control group is also equal to the baseline value, $\mu_{1T} = 0$. While the expected value for the control group is continuous increasing and given as $\mu_{2T} = 1$. In the case of two time points, this is given as

$$\boldsymbol{\mu}_1 = (0, 0)^T \quad \text{and} \quad \boldsymbol{\mu}_2 = (0, 1)^T. \tag{6.1}$$

In the situation with longitudinal data ($T = 4$) this is given as

$$\boldsymbol{\mu}_1 = (0, 0, 0, 0)^T \quad \text{and} \quad \boldsymbol{\mu}_2 = (0, 0.333, 0.667, 1)^T. \tag{6.2}$$

The variance is given as $\sigma^2 = 4 \quad \forall\, t, j$. The correlation structure used in the simulations between the time points is exponential decreasing, where the structure is given in Equation (3.11). The decay rate given in the simulation is $\phi = 0.8$. This results in the exponential

correlation matrix are

$$V_e = \begin{pmatrix} 1 & 0.45 & 0.20 & 0.09 \\ 0.45 & 1 & 0.45 & 0.20 \\ 0.20 & 0.45 & 1 & 0.45 \\ 0.09 & 0.20 & 0.45 & 1 \end{pmatrix}. \tag{6.3}$$

There have also been conducted simulations with a compound symmetry correlation structure. This matrix structure is presented and described in Equation (3.10). With correlation equal to ρ between each time point, this is given as

$$V_0 = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}. \tag{6.4}$$

The null hypothesis in all simulation studies is that there are no differences between the groups in the average parameter of interest. The alternative hypothesis is that there is a difference.

6.2 Simulated scenarios

6.2.1 Simulations with two time points

Simulations with two time points were conducted to look at the performance of the methods of follow-up score and change score analysis in addition to the methods of ANCOVA and cLDA. The amount of 10% missing data was used. The expected values for the groups are given in Equation (6.1). The simulations were done with different values of the correlation, given as

$$\boldsymbol{\rho} = (0.1, 0.2, ..., 0.9).$$

The method of follow-up score does not take the baseline values into account, thus the analysis are not dependent on the correlation. The simulations were only conducted once.

6.2.2 Simulations of different correlation structures

A simulation study was conducted to compare two different correlation structures: Compound symmetry correlation structure and exponential correlation structure. The simulations was conducted on longitudinal data ($T = 4$) with 10% missing data. The expected values for the groups are given in Equation (6.2) and the different correlation structures are given in the Equations (6.3) and (6.4). The analysis were conducted by using the method of ANCOVA with MI on baseline, the method of ANCOVA without MI on baseline and the method of cLDA.

6.2.3 Simulations with different percentage of missing data

The last simulation study was conducted to compare the methods with different percentages of missing data. The expected values for the groups are given in Equation (6.2) and

the exponential correlation structure was used, given in Equation (6.3). The simulation was conducted with longitudinal data and the analysis was conducted by cLDA and ANCOVA with and without MI at baseline. The percentages of missing data are given as

$$\text{Percentage missing data} = (5, 10, 15, 20).$$

6.3 Results of the simulation studies

6.3.1 Simulations with two time points

The results using the methods of ANCOVA and cLDA are given in Table 6.2, while the results using the methods of change score analysis and follow-up are given in Table 6.1. The results of the power are visualized in the graphs given in Figure 6.1. The Tables 6.2 and 6.1 shows that the results of the confidence interval coverage are in general the expected results (0.95) with $\alpha = 0.05$. The results of the bias are also low for all methods and all correlations.

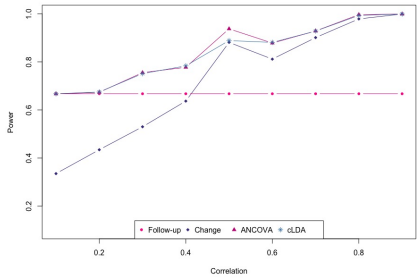
		ρ								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
		Change score								
No missing:	Bias	-0.039	-0.019	-0.031	0.001	0.059	-0.015	0.030	-0.009	0.015
	Power	0.335	0.434	0.530	0.637	0.881	0.811	0.901	0.979	0.999
	CI-coverage	0.952	0.959	0.953	0.949	0.997	0.958	0.939	0.934	0.943
MCAR:	Bias	-0.080	-0.073	-0.027	0.008	0.046	-0.014	0.021	-0.006	0.022
	Power	0.147	0.391	0.483	0.567	0.727	0.741	0.806	0.937	0.999
	CI-coverage	0.952	0.921	0.974	0.959	0.998	0.948	0.946	0.931	0.933
MAR:	Bias	-0.020	-0.031	-0.030	0.004	0.109	-0.013	0.010	-0.007	0.020
	Power	0.288	0.353	0.466	0.538	0.777	0.723	0.822	0.934	0.999
	CI-coverage	0.952	0.939	0.953	0.945	0.997	0.953	0.944	0.930	0.925
MNAR:	Bias	-0.070	-0.001	-0.046	-0.016	0.016	-0.040	0.033	-0.012	0.011
	Power	0.382	0.411	0.445	0.550	0.778	0.712	0.872	0.955	0.999
	CI-coverage	0.906	0.957	0.966	0.955	0.997	0.958	0.939	0.937	0.934
		Follow-up								
No missing:	Bias	0.050								
	Power	0.667								
	CI-coverage	0.952								
MCAR:	Bias	0.035								
	Power	0.667								
	CI-coverage	0.952								
MAR:	Bias	0.054								
	Power	0.714								
	CI-coverage	0.952								
MNAR:	Bias	-0.010								
	Power	0.666								
	CI-coverage	0.953								

Table 6.1: Results of simulations with two time points with different types of missing data mechanisms and different value of correlation for the method of change score and follow-up score. The results are given by bias, power and confidence interval coverage.

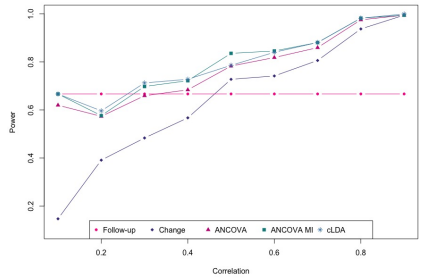
		ρ								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
		ANCOVA without MI								
No missing:	Bias	0.038	-0.025	-0.014	-0.012	0.057	-0.006	0.018	-0.006	0.016
	Power	0.667	0.673	0.755	0.777	0.937	0.878	0.928	0.996	0.999
	CI-coverage	0.952	0.940	0.952	0.952	0.996	0.964	0.938	0.958	0.943
MCAR:	Bias	0.011	-0.049	-0.004	-0.004	0.069	-0.001	0.007	-0.003	0.024
	Power	0.620	0.573	0.659	0.684	0.783	0.818	0.859	0.975	0.999
	CI-coverage	0.952	0.958	0.958	0.953	0.997	0.960	0.950	0.958	0.914
MAR:	Bias	0.013	-0.059	-0.028	-0.009	0.096	-0.005	-0.001	-0.007	0.019
	Power	0.619	0.611	0.630	0.679	0.883	0.812	0.880	0.962	0.999
	CI-coverage	0.952	0.976	0.945	0.956	0.997	0.964	0.930	0.937	0.943
MNAR:	Bias	-0.042	0.001	-0.040	-0.040	0.035	-0.044	0.001	-0.022	0.010
	Power	0.619	0.629	0.617	0.665	0.882	0.789	0.887	0.989	0.999
	CI-coverage	0.859	0.939	0.971	0.951	0.996	0.957	0.934	0.948	0.953
		ANCOVA with MI								
MCAR:	Bias	0.022	-0.049	-0.007	-0.001	0.063	-0.007	0.013	-0.001	0.022
	Power	0.666	0.576	0.698	0.722	0.835	0.845	0.880	0.982	0.999
	CI-coverage	0.952	0.958	0.961	0.965	0.997	0.954	0.951	0.944	0.914
MAR:	Bias	0.046	-0.038	-0.020	-0.005	0.093	-0.005	0.009	-0.009	0.023
	Power	0.761	0.594	0.694	0.726	0.935	0.849	0.889	0.966	0.999
	CI-coverage	0.952	0.920	0.949	0.952	0.997	0.964	0.934	0.937	0.915
MNAR:	Bias	-0.016	-0.018	-0.039	-0.039	0.041	-0.033	0.015	-0.018	0.015
	Power	0.667	0.613	0.696	0.695	0.883	0.817	0.900	0.979	0.999
	CI-coverage	0.953	0.939	0.949	0.955	0.996	0.957	0.947	0.951	0.952
		cLDA								
No missing:	Bias	0.038	-0.024	-0.013	-0.011	0.058	-0.006	0.018	-0.007	0.014
	Power	0.667	0.675	0.751	0.783	0.889	0.881	0.928	0.993	0.999
	CI-coverage	0.952	0.940	0.960	0.951	0.996	0.958	0.938	0.955	0.943
MCAR:	Bias	0.023	-0.047	-0.006	-0.001	0.060	-0.006	0.009	-0.004	0.019
	Power	0.666	0.596	0.712	0.728	0.786	0.840	0.880	0.982	0.999
	CI-coverage	0.952	0.959	0.968	0.961	0.997	0.957	0.960	0.958	0.933
MAR:	Bias	0.045	-0.038	-0.021	-0.004	0.094	-0.007	0.005	-0.008	0.020
	Power	0.760	0.652	0.693	0.746	0.935	0.847	0.889	0.969	0.999
	CI-coverage	0.952	0.920	0.949	0.952	0.998	0.967	0.929	0.937	0.934
MNAR:	Bias	-0.016	-0.020	-0.038	-0.038	0.047	-0.034	0.008	-0.021	0.013
	Power	0.665	0.669	0.704	0.721	0.884	0.838	0.900	0.986	0.999
	CI-coverage	0.906	0.939	0.960	0.950	0.996	0.957	0.943	0.951	0.962

Table 6.2: Results of simulation with two time points with different types of missing data mechanisms and different value for the correlation for the method of ANCOVA and cLDA. The results are given as bias, power and confidence interval coverage.

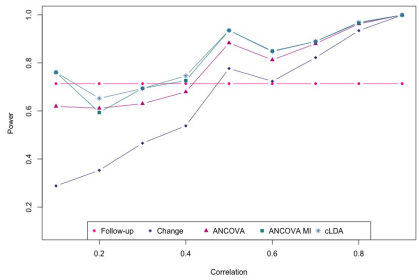
The results show that the power increases when the correlation increases for the method of change score analysis. Using the method of follow-up, the power is equal as using the method of change score when the correlation is between $\rho = 0.4$ and $\rho = 0.5$. Both the method of cLDA and ANCOVA result in higher power than the method of change score and follow-up. However, when the correlation is high using the method of change score the power are almost on the same level as using the methods of ANCOVA and cLDA. The results of the power using method of cLDA are higher than the method of ANCOVA without MI. With MI the method of ANCOVA and cLDA are almost identical with respect to power. Also when there are no missing data the method of ANCOVA and cLDA are almost identical with respect to power. The different missing data mechanisms also affect the results of the power. No missing data result in higher power than missing data, MCAR data result generally in higher power than MAR data, and MNAR results generally in slightly lower power than MAR data. However, the differences between the missing data mechanisms MCAR, MAR and MNAR are small.



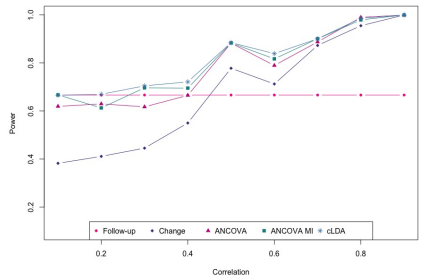
(a) No missing data



(b) MCAR



(c) MAR

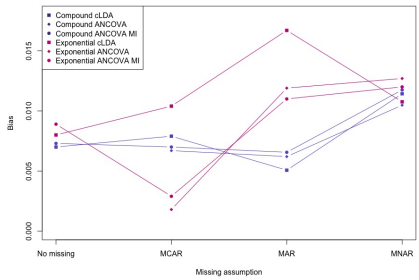


(d) MNAR

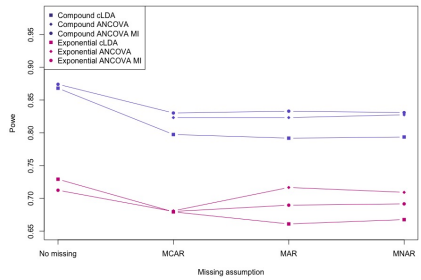
Figure 6.1: Graph of correlation and power with two time points using different methods with different missing data mechanisms.

6.3.2 Simulations of different correlation structures

The results of the simulations are given in Table 6.3. The results of the power and bias are visualized in Figure 6.2.



(a) Bias



(b) Power

Figure 6.2: Graphs of bias and power of two different correlation structures.

		Compound correlation structure								
		ANCOVA MI			ANCOVA			cLDA		
		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
No missing:	Bias				-0.002	0.002	-0.007	-0.002	0.002	-0.007
	Power				0.182	0.550	0.868	0.185	0.553	0.874
	CI-coverage				0.946	0.953	0.946	0.945	0.954	0.943
MCAR:	Bias	-0.001	0.002	-0.007	-0.001	0.003	-0.008	-0.001	0.002	-0.007
	Power	0.171	0.501	0.823	0.153	0.464	0.797	0.172	0.505	0.830
	CI-coverage	0.947	0.951	0.944	0.949	0.952	0.946	0.948	0.951	0.945
MAR:	Bias	-0.002	-0.001	-0.006	-0.002	0.001	-0.005	-0.003	-0.001	-0.007
	Power	0.164	0.488	0.823	0.152	0.458	0.792	0.161	0.497	0.833
	CI-coverage	0.941	0.950	0.946	0.948	0.953	0.948	0.942	0.951	0.947
MNAR:	Bias	-0.003	0.001	-0.010	-0.005	-0.002	-0.011	-0.005	-0.001	-0.012
	Power	0.165	0.496	0.827	0.154	0.465	0.794	0.163	0.510	0.831
	CI-coverage	0.948	0.954	0.946	0.950	0.954	0.947	0.947	0.953	0.945
		Exponential correlation structure								
		ANCOVA MI			ANCOVA			cLDA		
		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
No missing:	Bias				0.020	-0.002	0.008	0.021	-0.001	0.009
	Power				0.124	0.381	0.729	0.152	0.371	0.712
	CI-coverage				0.981	0.939	0.954	0.979	0.945	0.961
MCAR:	Bias	0.020	0.008	-0.002	0.019	0.013	0.010	0.025	0.016	0.003
	Power	0.120	0.355	0.681	0.119	0.331	0.679	0.157	0.348	0.680
	CI-coverage	0.975	0.943	0.932	0.990	0.954	0.938	0.975	0.938	0.938
MAR:	Bias	0.017	-0.003	0.012	0.023	0.007	0.017	0.017	-0.005	0.011
	Power	0.103	0.376	0.717	0.108	0.326	0.661	0.131	0.349	0.690
	CI-coverage	0.985	0.949	0.928	0.976	0.939	0.947	0.984	0.944	0.935
MNAR:	Bias	0.013	0.004	0.013	0.005	0.009	0.011	0.009	-0.001	0.012
	Power	0.115	0.360	0.709	0.103	0.341	0.668	0.151	0.353	0.692
	CI-coverage	0.980	0.959	0.954	0.969	0.950	0.948	0.978	0.954	0.960

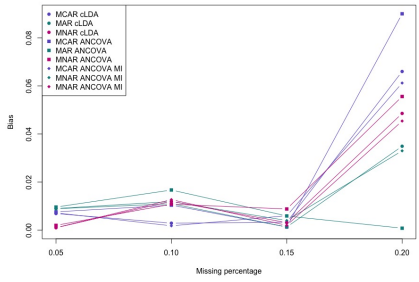
Table 6.3: Results of simulations with different correlation structures. The parameter of interests at different time points are given as θ_t.

Figure 6.2 shows that the bias is in general higher when using exponential correlation structure and the power is lower when using exponential correlation structure. The bias increases when the missing data mechanism MNAR is used for compound correlation symmetry using all methods. When using exponential correlation structure, the bias varies more between the methods and between different missing data mechanisms. The results of confidence interval coverage are all approximately equal to the expected value, 0.95.

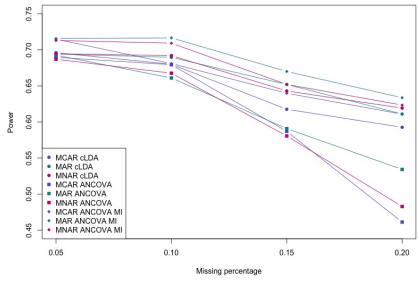
6.3.3 Simulations with different percentages of missing data

The results from the simulation study with different amounts of missing data are presented in Table 6.4. The results are visualized in Figure 6.3. The results from the simulation with percentage of missing data equal 10% are the same as the results when comparing the compound correlation structure and exponential correlation structure given in Table 6.3, since all the variables were the same. The results in Figure 6.3a shows that when the percentage of missing data reaches 20%, the bias is high for all methods and all missing data mechanisms. Figure 6.3b shows that the power decreases when the percentage of missing data increases. Figure 6.3c shows how the bias changes when the missing data mechanism changes for the simulations with missing data percentage equal or lower than 15%. The simulations with missing data percentage equal to 20% are not included in the figure since the differences were too high. This is shown in Figure 6.3a. The bias is low when there are no missing data and varies according to method and percentage missing

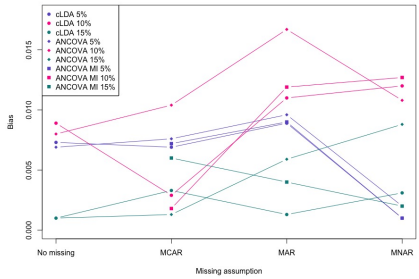
data. Figure 6.3d shows that the power decreases when there are missing data, but does not differ according to what type of missing data mechanism used. The results of confidence interval coverage are approximately as expected (0.95) for all simulations.



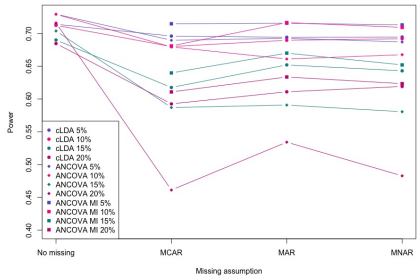
(a) Bias vs. missing percentage.



(b) Power vs. missing percentage



(c) Bias vs. missing data mechanisms.



(d) Power vs. missing data mechanisms.

Figure 6.3: Results of longitudinal simulations with different percentage of missing data.

		Missing percentage: 5%								
		ANCOVA MI			ANCOVA			cLDA		
		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
No missing:	Bias				-0.002	-0.001	0.007	-0.001	-0.001	0.007
	Power				0.118	0.421	0.729	0.152	0.407	0.714
	CI-coverage				0.955	0.946	0.927	0.940	0.949	0.935
MCAR:	Bias	-0.002	-0.001	0.007	0.001	-0.001	0.008	-0.003	-0.001	0.007
	Power	0.119	0.403	0.715	0.119	0.377	0.690	0.147	0.391	0.696
	CI-coverage	0.951	0.940	0.926	0.959	0.946	0.931	0.942	0.949	0.935
MAR:	Bias	-0.002	0.001	0.009	-0.001	0.001	0.010	-0.002	0.001	0.009
	Power	0.121	0.402	0.716	0.115	0.387	0.693	0.149	0.392	0.694
	CI-coverage	0.958	0.940	0.920	0.960	0.944	0.930	0.942	0.946	0.935
MNAR:	Bias	-0.003	-0.010	-0.001	-0.007	-0.012	-0.002	-0.004	-0.011	-0.001
	Power	0.124	0.404	0.713	0.120	0.378	0.687	0.153	0.390	0.695
	CI-coverage	0.953	0.948	0.926	0.956	0.945	0.932	0.941	0.954	0.940

		Missing percentage: 10%								
		ANCOVA MI			ANCOVA			cLDA		
		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
No missing:	Bias				0.020	-0.002	0.008	0.021	-0.001	0.009
	Power				0.124	0.381	0.729	0.152	0.371	0.712
	CI-coverage				0.981	0.939	0.954	0.979	0.945	0.961
MCAR:	Bias	0.020	0.008	-0.002	0.019	0.013	0.010	0.025	0.016	0.003
	Power	0.120	0.355	0.681	0.119	0.331	0.679	0.157	0.348	0.680
	CI-coverage	0.975	0.943	0.932	0.990	0.954	0.938	0.975	0.938	0.938
MAR:	Bias	0.017	-0.003	0.012	0.023	0.007	0.017	0.017	-0.005	0.011
	Power	0.103	0.376	0.717	0.108	0.326	0.661	0.131	0.349	0.690
	CI-coverage	0.985	0.949	0.928	0.976	0.939	0.947	0.984	0.944	0.935
MNAR:	Bias	0.013	0.004	0.013	0.005	0.009	0.011	0.009	-0.001	0.012
	Power	0.115	0.360	0.709	0.103	0.341	0.668	0.151	0.353	0.692
	CI-coverage	0.980	0.959	0.954	0.969	0.950	0.948	0.978	0.954	0.960

		Missing percentage: 15%								
		ANCOVA MI			ANCOVA			cLDA		
		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
No missing:	Bias				0.002	-0.024	-0.001	0.003	-0.024	-0.001
	Power				0.131	0.389	0.704	0.161	0.380	0.690
	CI-coverage				0.962	0.948	0.951	0.942	0.954	0.954
MCAR:	Bias	-0.004	-0.028	-0.006	-0.003	-0.025	0.001	-0.003	-0.029	-0.003
	Power	0.117	0.340	0.640	0.092	0.285	0.587	0.144	0.337	0.618
	CI-coverage	0.960	0.947	0.953	0.953	0.938	0.940	0.951	0.949	0.956
MAR:	Bias	0.003	-0.016	-0.004	-0.004	-0.015	-0.006	0.009	-0.015	-0.001
	Power	0.135	0.342	0.670	0.103	0.304	0.591	0.158	0.347	0.652
	CI-coverage	0.949	0.948	0.929	0.948	0.942	0.914	0.938	0.947	0.940
MNAR:	Bias	0.005	-0.026	-0.002	-0.001	-0.030	-0.009	0.006	-0.028	-0.003
	Power	0.128	0.356	0.652	0.097	0.295	0.581	0.143	0.352	0.643
	CI-coverage	0.957	0.951	0.945	0.959	0.959	0.944	0.948	0.957	0.951

		Missing percentage: 20%								
		ANCOVA MI			ANCOVA			cLDA		
		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
No missing:	Bias				0.016	-0.030	-0.040	0.017	-0.028	-0.040
	Power				0.135	0.361	0.716	0.182	0.357	0.685
	CI-coverage				0.981	0.958	0.934	0.963	0.972	0.940
MCAR:	Bias	0.003	-0.024	-0.061	-0.001	-0.019	-0.090	0.009	-0.021	-0.066
	Power	0.104	0.297	0.611	0.076	0.245	0.461	0.104	0.325	0.593
	CI-coverage	0.972	0.953	0.909	0.981	0.954	0.935	0.963	0.954	0.926
MAR:	Bias	0.027	-0.037	-0.033	0.040	-0.026	0.001	0.023	-0.030	-0.035
	Power	0.121	0.291	0.634	0.102	0.286	0.534	0.131	0.302	0.611
	CI-coverage	0.968	0.971	0.935	0.953	0.944	0.948	0.954	0.971	0.939
MNAR:	Bias	0.029	-0.008	-0.045	0.039	-0.035	-0.056	0.027	-0.015	-0.049
	Power	0.085	0.334	0.623	0.108	0.249	0.483	0.131	0.342	0.619
	CI-coverage	0.958	0.976	0.940	0.967	0.962	0.917	0.953	0.967	0.944

Table 6.4: Results from simulations of longitudinal data with different percentages of missing data. The parameter of interests at different time points is given as θ_t.

Chapter 7

Application

The methods of ANCOVA and cLDA were used to analyze a clinical trial. The trial is presented in the article "Effect of high intensity interval training on cardiac function in children with obesity: A randomised controlled trial" by Ingul et al. (2018). The trial is presented in Section 7.1. In Section 7.2 the analyses conducted in this report is presented. At last, the results are presented in Section 7.3.

7.1 The trial

The aim of this trial was to compare the effect of high intensity interval training (HIIT) to moderate intensity continuous training (MICT) and only getting nutrition advises on cardiovascular health of obese children. It has earlier been showed that HIIT has superior effect on cardiovascular health for adults compared to MICT. The hypothesis is thus that HIIT has better effect than MICT. The trial is an RCT, so the children were randomly divided into a group and stratified by age and sex. They were measured three times over 12 weeks; One time before intervention, one time after 3 weeks and one time after 12 weeks. The HIIT and MICT groups were the different treatment groups, while the group of children getting nutrition advice was the control group. The treatment groups also received nutrition advice. There were 99 children included in the analysis of the trial, which were measured at least one time after intervention started. The children were in the age group $7-16$ years old, with a body mass index (BMI) $\geq$ percentile curves that passed through $30\text{kg}/\text{m}^2$ at age 18. The trial is a multicenter trial, thus the children were collected from different universities: The University of Queensland, Brisbane, Australia and The Norwegian University of Science and Technology (NTNU), Trondheim, Norway. The trial was conducted between March 2012 and February 2017. The intervention was continuous, thus both the nutrition advice lessons and the training were continuously conducted during the trial. The group of HIIT had 4×4 min bouts at $85-90\%$ of maximal heart rate 3 times/week. The group of MICT had 44 min training at $60-70\%$ of maximal heart rate 3 times/week. At least two of the three exercises each week were supervised. The nutrition advice lessons were given $4-6$ times during the trial. The baseline mean values

were assumed to be equal for all groups. For the intention-to-treat analysis, the groups consisted of the following numbers of participants: HIIT: $n = 33$, MISC: $n = 32$ and nutrition advises only: $n = 34$. This is the total amount of participants in each group. For the per protocol analysis, the numbers of participants were: HIIT: $n = 17$, MICT: $n = 24$ and nutrition advice only: $n = 21$. For the per protocol analysis were the children assumed to complete at least 80% of the exercise and nutrition sessions. Since an LMM was conducted, all available data were used in the ITT analysis. The data were assumed to be missing at random, thus the results were assumed to be unbiased using a LMM. The results showed that MICT and HIIT were superior compared to nutrition advice only on the effect of cardiovascular health on obese children, but that there were no significant difference between MICT and HIIT.

7.2 The analysis

In the analysis of the trial conducted here, only the children with obesity collected from NTNU were included in the analysis. Thus, 68 children were included. The baseline values are complete. The analysis are conducted on two different outcome variables: Resting left ventricular global longitudinal strain (GLS) and resting left ventricular peak systolic tissue velocity (LVS). Both of these variables are associated with cardiovascular health. The mean values, variance, size of groups and amount of missing data for the different groups for the outcome variable GLS are given in Table 7.1. The same values for the outcome value LVS are given in Table 7.2.

	Diet group - GLS			
	Number of participants	Number of missing	Mean	Variance
Y_0	23	0	-17.21	10.49
Y_1	23	10	-17.08	6.95
Y_2	23	10	-18.57	7.82
	MICT group - GLS			
	Number of participants	Number of missing	Mean	Variance
Y_0	22	0	-18.17	8.04
Y_1	22	5	-18.98	5.73
Y_2	22	10	-20.28	5.93
	HIIT group - GLS			
	Number of participants	Number of missing	Mean	Variance
Y_0	23	0	-18.16	7.40
Y_1	23	9	-19.67	2.60
Y_2	23	14	-19.42	5.13

Table 7.1: Mean values, variance, group size and amount of missing data for the different groups with the outcome variable GLS.

7.3 Results

Results for GLS

A histogram and a Q-Q plot were carried out of the residuals from the analysis conducted by using the method of ANCOVA and cLDA. This was done to approve the assumption

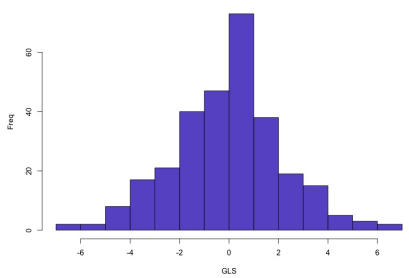
	Diet group - LVS			
	Number of participants	Number of missing	Mean	Variance
Y_0	23	0	8.31	0.86
Y_1	23	8	8.69	1.64
Y_2	23	10	8.94	2.43
	MICT group - LVS			
	Number of participants	Number of missing	Mean	Variance
Y_0	22	0	8.61	1.85
Y_1	22	5	9.90	3.47
Y_2	22	10	10.67	6.69
	HIIT group - LVS			
	Number of participants	Number of missing	Mean	Variance
Y_0	23	0	8.49	1.83
Y_1	23	8	9.95	2.25
Y_2	23	14	9.62	3.45

Table 7.2: Mean values, variance, group size and amount of missing data for the different groups with the outcome variable LVS.

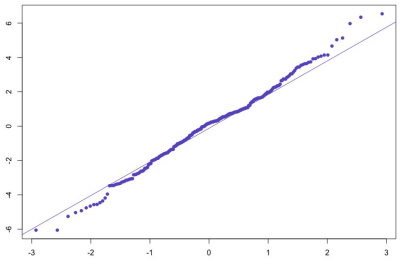
about normal distributed data on the outcome variable GLS. These are given in Figure 7.1. Perfect normally distributed should lie on a straight line on the Q-Q plot, and the histogram should look like a normal distribution. The results in Figure 7.1 are approximately normally distributed, so the assumption is fulfilled. A histogram and a Q-Q plot were also carried out for each group, with equal results. The results after using the methods of ANCOVA and cLDA on the outcome variable GLS are given in Table 7.3. The method of ANCOVA was used without MI, since there were no missing data at baseline. A compound correlation structure was assumed in the model. An exponential correlation structure was also tried and the results were almost identical. The change in the variable of interest for the different groups are visualized in the Figure 7.2. The results shown in Table 7.3 shows that using both the method of ANCOVA and cLDA results in similar results. There is a significant difference between the MICT group and the diet group at both time points. The HIIT group and the diet group have a significant difference at the first follow-up time point, but not at the last follow-up. The MICT group and the HIIT group are not significant different at any of the two follow-up time points.

	ANCOVA		
	Estimate	Confidence interval	p-value
HIIT vs DIET θ_1	-2.329	$[-3.967, -0.692]$	0.0070
HIIT vs DIET θ_2	-0.688	$[-2.537, 1.160]$	0.450
HIIT vs MICT θ_1	-0.851	$[-2.384, 0.681]$	0.263
HIIT vs MICT θ_2	1.066	$[-0.821, 2.953]$	0.255
MICT vs DIET θ_1	-1.4910	$[-3.0.637, 0.094]$	0.064
MICT vs DIET θ_2	-1.755	$[-3.454, -0.055]$	0.0435
	cLDA		
	Estimate	Confidence interval	p-value
HIIT vs DIET θ_1	-2.461	$[-4.307, -0.614]$	0.010
HIIT vs DIET θ_2	-0.641	$[-2.702, 1.420]$	0.537
HIIT vs MICT θ_1	-0.816	$[-2.554, 0.923]$	0.353
HIIT vs MICT θ_2	1.352	$[-0.741, 3.444]$	0.202
MICT vs DIET θ_1	-1.645	$[-3.410, 0.120]$	0.067
MICT vs DIET θ_2	-1.992	$[-3.899, -0.086]$	0.041

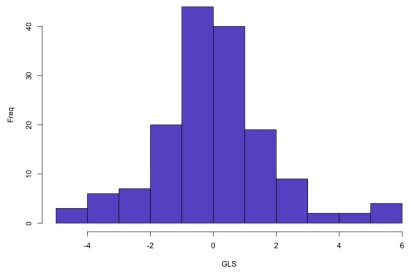
Table 7.3: Results using the methods of ANCOVA and cLDA on the outcome variable GLS.



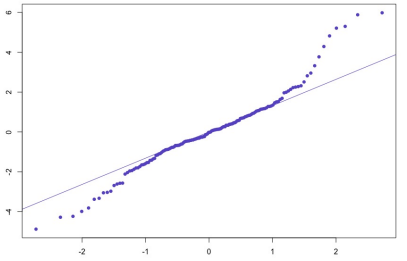
(a) Histogram, cLDA.



(b) Q-Q plot, cLDA.



(c) Histogram, ANCOVA.

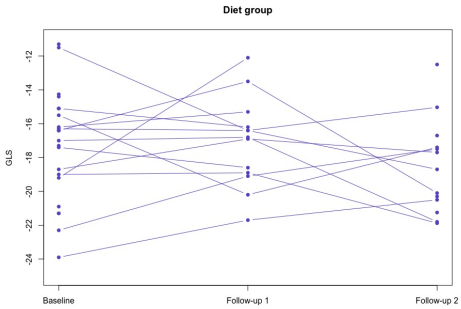


(d) Q-Q plot, ANCOVA.

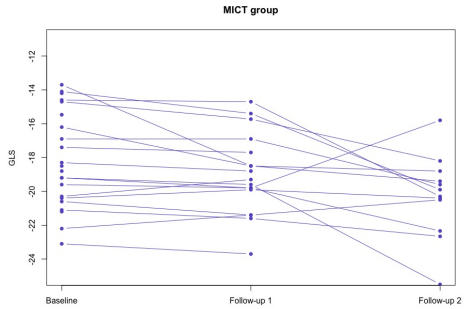
Figure 7.1: Q-Q plots and histograms from the residuals using the methods of ANCOVA and cLDA on the outcome variable GLS.

Results for LVS

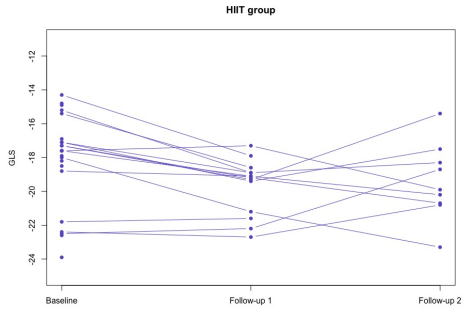
The outcome variable LVS was log-transformed before analyzing the data. A histogram and a Q-Q plot were carried out of the data to approve the assumption about normal distributed data. This was done on the residuals from the methods of ANCOVA and cLDA. Histograms and Q-Q plots for the different groups were also carried out with equal results. These are shown in Figure 7.4 and confirms the assumption of normality. The change in the variable of interest for the different groups are visualized in the Figure 7.5. The results when analyzing the data by the methods of ANCOVA and cLDA are given in Table 7.4. A compound correlation structure was assumed in the analysis. The results of the analysis from the outcome variable GLS given in Table 7.4 shows that the methods of ANCOVA and cLDA generates similar results. However, the method of cLDA results in lower p-values. There are significant differences between the control group and the treatment groups at both time points using the method of cLDA. The methods of ANCOVA results in significant difference between the HIIT group and diet group, and between the MICT group and the diet group at the last follow-up time point.



(a) Diet group.



(b) MICT group.

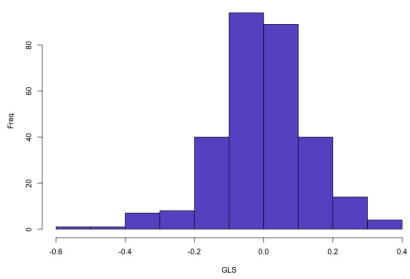


(a) HIIT group.

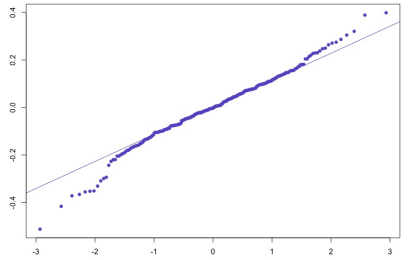
Figure 7.3: Change in the outcome variable GLS.

	ANCOVA		
	Estimate	Confidence interval	p-value
HIIT vs DIET θ_1	0.119	$[0.007, 0.231]$	0.037
HIIT vs DIET θ_2	0.110	$[-0.022, 0.242]$	0.099
HIIT vs MICT θ_1	0.032	$[-0.077, 0.141]$	0.548
HIIT vs MICT θ_2	-0.037	$[-0.172, 0.097]$	0.570
MICT vs DIET θ_1	0.087	$[-0.222, 0.197]$	0.114
MICT vs DIET θ_2	0.148	$[0.026, 0.269]$	0.019
	cLDA		
	Estimate	Confidence interval	p-value
HIIT vs DIET θ_1	0.123	$[0.022, 0.223]$	0.017
HIIT vs DIET θ_2	0.109	$[-0.008, 0.227]$	0.067
HIIT vs MICT θ_1	0.026	$[-0.072, 0.124]$	0.600
HIIT vs MICT θ_2	-0.040	$[-0.159, 0.079]$	0.505
MICT vs DIET θ_1	0.097	$[-0.001, 0.194]$	0.051
MICT vs DIET θ_2	0.150	$[0.041, 0.258]$	0.007

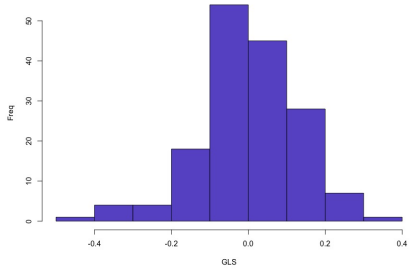
Table 7.4: Results using the methods of ANCOVA and cLDA on the outcome variable LVS.



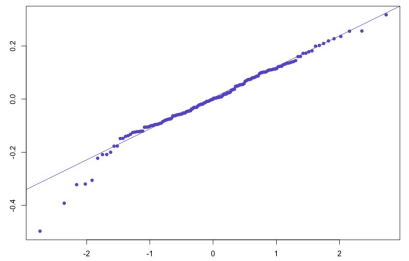
(a) Histogram, cLDA.



(b) Q-Q plot, cLDA.

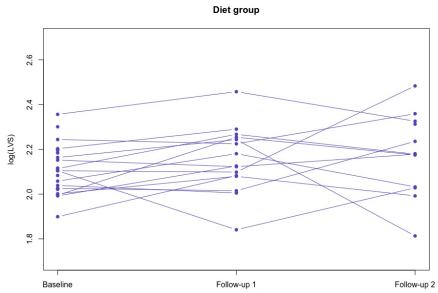


(c) Histogram, ANCOVA.

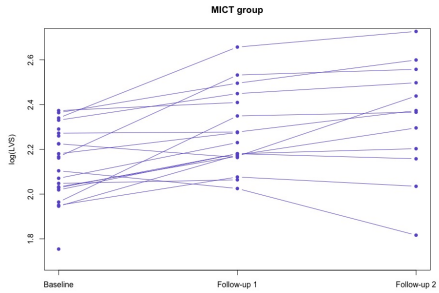


(d) Q-Q plot, ANCOVA.

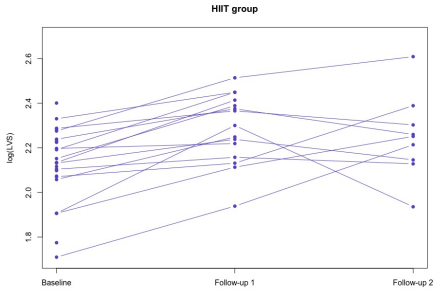
Figure 7.4: Q-Q plots and histograms from the residuals using the methods of ANCOVA and cLDA on the outcome variable LVS.



(a) Diet group.



(b) MICT group.



(c) HIIT group.

Figure 7.5: Change in the outcome variable LVS for the different groups.

Chapter 8

Discussion and conclusion

In this chapter the results from Chapter 6 are discussed. The methods presented in Section 2.3, comparisons of means of follow-up score, change score analysis, ANCOVA and cLDA are with respect to bias, power and confidence interval coverage compared. The missing data mechanisms, correlation, correlation structures and differences in percentage of missing data are the main focus.
This chapter is structured as follows: The results of the simulations with two time points will be discussed in Section 8.1. The results from the different correlation structures presented in Section 6.3.2 are discussed in Section 8.2. The results from the simulated scenario with longitudinal data and different missing percentages presented in Section 6.3.3 are discussed in Section 8.3. At last, in Section 8.4 the conclusion is presented.

8.1 RCTs with two time points

The methods used to analyzed RCTs with two time points presented in Section 2.3 are: Comparisons of means of follow-up score, change scores analysis, ANCOVA and cLDA. The results of the simulations are given in Tables 6.2 and 6.1, and in Figure 6.1.
The least complex method is the method of follow-up score. A disadvantage of this method is that only half of the data are used in the analysis. The baseline values are not used in the analysis. The method of follow-up results in a reasonable amount of power, shown in Figure 6.1. This confirms the results in the simulation study by Vickers (2001). An mentioned in Section 2.3, the method of follow-up will loose variation when the correlation is high, and show less significant results, while the method of change score will add variation when the correlation is high (Vickers and Altman, 2001). This is confirmed in Figure 6.1. The method of follow-up should thus only be considered to be used in situations with low correlation between observations ($\rho < 0.4$). In this case the method of follow-up is the preferred method compared to the method of change score and it can also be comparable to the method of ANCOVA and cLDA. When there are missing data, using the method of follow-up all observations with observed follow-up values will be used (available case analysis). Thus, more data can be used compared to using the method of change score,

where a complete case analysis is used. This is confirmed in Figure 6.1. Using the method of change score when there are missing data may result in reduced power, while the difference between no missing data and missing data for the method of follow-up is smaller. When there are no missing data, the method of follow-up score is only comparable with the methods of ANCOVA and cLDA when the correlation is quite low ($\rho < 0.2$). When there are missing data, the method of follow-up score is comparable to the methods of cLDA and ANCOVA with higher value of correlation ($\rho = 0.4$), as shown in Figure 6.1. However, in medical research the correlation may be assumed to be reasonable high. For example, in the case of research on stable chronic diseases, the correlation is assumed to be high. Thus, the method of change score is the method of choice compared to the method of follow-up.
The method of change score is the method most sensitive to correlation differences. The method is comparable to the method of cLDA and ANCOVA when the correlation is high, but the results when the correlation is low are poor. The method of ANCOVA is the method of choice compared to the methods of change score and follow-up. This is also found by Vickers (2001). Liu et al. (2009) stated that the method of ANCOVA and cLDA would estimate identical point estimates in the case of no missing data. This is confirmed in Figure 6.1a. In the case of missing data, Coffman et al. (2016) argued that the method of cLDA is the method of choice compared to the method of ANCOVA, and Dinh and Yang (2011) argued that using baseline as a covariate would lead to loss of efficiency. The method of ANCOVA cannot use observations where the baseline value is missing. Thus, using the method of ANCOVA some information is lost. The point estimates may be less precise and the variation may increase. This is confirmed in Figure 6.1. The power using the method of cLDA is higher than the power using the method of ANCOVA. However, using MI on the missing baseline values results in power almost identical to using the method of cLDA. Thus, using MI combined with the method of ANCOVA produces equally good results with respect to power compared to the method of cLDA.

8.2 Different correlation structures

In Section 6.3.2 the results of the correlation structures were presented. A compound symmetry correlation structure and an exponential correlation structure were simulated. The results are presented in Figure 6.2 and in Table 6.3. The figure and the table show the bias and power differ between the two correlation structures. Using the exponential correlation structure, the results of the bias are more variable and in general higher. The power is lower compared to the compound symmetry correlation structure. When modelling a LMM with a correlation structure, the compound symmetry correlation structure is more stable than the exponential correlation structure. This can explain the differences in the results. The simulations were not conducted with different values of ρ for the compound symmetry correlation structure, and ϕ for the exponential correlation structure. This is a drawback with the simulations conducted. By varying the variables ρ and ϕ, more information about how the correlation structures differ could be found. When modelling RCTs in medical research, the exponential correlation structure may be more realistic than the compound correlation structure. The time points are assumed to be more correlated when there are less time differences.

The results of the bias of the compound symmetry correlation structure visualized in Figure 6.2a are consistent with the theory for the missing data mechanisms presented in Section 4.2. The methods of cLDA and ANCOVA can produce unbiased results when the missing data mechanism is MCAR or MAR, but bias may be introduced when the missing data mechanism is MNAR. The results of the power is as expected: The power is stable independent of missing data assumption, but lower than for no missing data.

8.3 Longitudinal data with different missing percentages

The Figure 6.3 and the Table 6.4 shows the results when longitudinal data are simulated, and the effect of the bias and power when the missing data percentage and the missing data mechanisms differ. The power and bias varies when the percentage of missing data increases. As Figures 6.3a and 6.3b shows, when the percentage of missing data is 0.05 the bias is low and the power high. This amount of missing data have almost no effect on the results. Thus, missing data ≤ 0.5 does not need to be taken into account, no matter what type of missing data mechanism there are. When the percentage of missing data increases, the results of the power decreases. The results of the bias are low when the missing percentage is $\leq 0.15\%$, but with higher values of missing percentage the results of the bias are significantly higher. Schulz and Grimes (2002) and Fielding et al. (2012) argued that when the amount of missing data was between 5% and 20%, the effect of the missing data should be discussed. This is confirmed in the simulations presented in Table 6.4. If the missing percentage is equal or lower than 5%, the bias is minimal and the power is reasonable high. As the missing percentage increases, the bias is varying and the power is decreasing. Schulz and Grimes (2002) states that trials with more than 20% missing data is refused by some journals, while Altman and Bland (2007) argues that more than 20% missing data is problematic. The large increased risk of bias when there are 20% missing data, visualized in Figure 6.3a confirms this. This is independent of missing data mechanism. In the report by Zhang et al. (2017) was the average missing data percentage found to be 10%, thus missing data should be a part of the analysis when conducting RCTs.
The missing data mechanism is said to have impact on the results of a trial. When the data are MAR, using LMM should result in unbiased estimates. This is confirmed by the simulations. The Figures 6.3a and 6.3b shows that the method of cLDA and ANCOVA with MI results in almost identical results with respect to power and bias. The method of longitudinal ANCOVA without MI results in a smaller amount of power and more variation in the results of the bias. Figure 6.3b shows that the methods and the missing data mechanisms varies more when the percentage of missing data increases. Figure 6.3c visualizes the bias when the missing data mechanism vary. One would expect that the missing data mechanism MNAR would produce more biased results than the missing data mechanisms MAR and MCAR but the results differ between the methods used. However, the bias are in general low (≤ 0.015). Figure 6.3a shows that a high percentage of missing data is a more severe problem than the missing data mechanism in these simulations.

8.4 Conclusion

When there are missing data, it is necessary to take the information about the missing data into account. The missing data mechanism should be handled with right type of method to analyze the data. The methods of change score analysis and follow-up score should be used when the missing data mechanism is MCAR, while the method of ANCOVA and cLDA can in addition be used when the missing data mechanism is MAR. The missing data mechanism affects the bias of the results, while the power is stable no matter what type of missing data there is. However, the analysis of the missing data is also affected by the percentage of missing data, which also affects the results of the different missing data mechanisms. Analysis with more than 20% missing data should be avoided. In this case the change in the bias is significantly higher, and the power is low. The percentage of missing data is also affecting the choice of method and the effect of the missing data mechanism. When the percentage of missing data is lower than 5%, the effect on the results is low, and it is not necessary to take the missing data into account.
For analyzing RCTs with two time points and missing data several methods are possible to use, and the choice of method depends on the correlation between the time point and missing data mechanism. The methods of follow-up score and change score analysis are shown to have unbiased results when the missing data mechanism is MCAR, while the methods of ANCOVA and cLDA can also be used in the case of MAR data. The method of follow-up is comparable with the methods of ANCOVA and cLDA when there are missing data, and the correlation is low ($\rho < 0.4$). In this case the method of change score should be avoided, because of low results of the power. When the correlation is high ($\rho > 0.8$) the method of change score is comparable with the methods of ANCOVA and cLDA. In this case the method of follow-up should be avoided. The methods of ANCOVA and cLDA are in most cases the methods of choice, and superior to the methods of follow-up and change score.
When there are missing data, the method of ANCOVA has a drawback compared to the method of cLDA: Using the method of ANCOVA the baseline values can not be missing. In this case the entire unit is removed from the analysis. This does not happen to the method of cLDA, thus more information are available when using the method of cLDA. The method of cLDA is thus the method of choice. However, applying MI to the missing baseline values have shown to have good effect on the results. The missing baseline data can be imputed, and using the method of ANCOVA with MI, no units have to be removed. Thus, the drawback using the method of ANCOVA is removed. The methods of cLDA and ANCOVA with MI are comparable, and produces equal results.

Bibliography

Altman, D. G., Bland, J. M., 2007. Statistical notes: Missing data. BMJ (Clinical research ed.) 334, 424.
URL https://10.1136/bmj.38977.682025.2C

Attia, A., 2005. Bias in RCTs: Confounders, selection bias and allocation concealment. Middle East Fertil Soc J 10, 258–260.

Bland, J. M., Altman, D. G., 1994. Some examples of regression towards the mean. BMJ 309, 780.

Carpender, J. R., Kenward, M. G., 2007. Missing data in randomised controlled trials: a practical guide. Health Technology Assessment Methodology Programme, Birmingham, 199.
URL http://researchonline.lshtm.ac.uk/id/eprint/4018500

Coffman, C. J., Edelman, D., Woolson, R. F., 2016. To condition or not condition? analysing 'change' in longitudinal randomised controlled trials. BMJ Open 6.
URL https://10.1136/bmjopen-2016-013096

Collins, L., L. Schafer, J., Kam, C.-M., 2002. A comparison of restrictive strategies in modern missing data procedures. Psychological methods 6, 330–51.
URL https://doi.org/10.1037/1082-989X.6.4.330

Diggle, P. J., Heagerty, P., Liang, K. Y., Zeger, S. L., 2001. Analysis of Longitudinal Data. Oxford University Press.

Dinh, P., Yang, P., 2011. Handling baselines in repeated measures analyses with missing data at random. Journal of Biopharmaceutical Statistics 21 (2), 326–341.
URL https://doi.org/10.1080/10543406.2011.550113

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Regression. Models, Methods and Applications. Springer - Verlag Berlin Heidelberg.

Fielding, S., Fayers, P., Ramsay, C. R., 2012. Analysing randomised controlled trials with missing data: Choice of approach affects conclusions. Elsevier 33, 461–469.
URL https://doi.org/10.1016/j.cct.2011.12.002

Frison, L., Pocock, S. J., 1992. Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. Statistics in medicine 11, 1685 – 1704.

Hollis, S., Campbell, F., 1999. What is meant by intention to treat analysis? survey of published randomised controlled trials. BMJ 319 (7211), 670–674.
URL https://www.bmj.com/content/319/7211/670

Ingul, C. B., Dias, K. A., Tjonna, A. E., Follestad, T., Hosseini, M. S., Timilsina, A. S., Hollekim-Strand, S. M., Ro, T. B., Davies, P. S., Cain, P. A., Leong, G. M., Coombes, J. S., 2018. Effect of high intensity interval training on cardiac function in children with obesity: A randomised controlled trial. Progress in Cardiovascular Diseases 61 (2), 214–221.
URL http://www.sciencedirect.com/science/article/pii/S0033062018300379

Liang, K.-Y., L. Zeger, S., 01 2000. Longitudinal data analysis of continuous and discrete responses for pre-post designs. Sankhy: The Indian Journal of Statistics, Series B (1960-2002) 62, 134–148.
URL https://doi.org/10.2307/25053123

Little, R. J. A., Rubin, D. B., 2002. Statistical analysis with missing data, 2nd Edition. John Wiley & Sons, Inc., Hoboken, New Jersey.

Liu, G. F., Lu, K., Mogg, R., Mallick, M., Mehrotra, D. V., 2009. Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? Statistics in Medicine 28 (20), 2509–2530.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3639

Lord, F. M., 1967. A paradox in the interpretation of group comparisons. Psychological Bulletin 68, 304–305.

Marshall, A., Altman, D. G., Holder, R. L., Royston, P., 2009. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Medical Research Methodology 9 (1), 57.
URL https://doi.org/10.1186/1471-2288-9-57

Matthews, J. N. S., 2006. Introduction to Randomized Controlled Clinical Trials, 2nd Edition. Chapman Hall/CRC.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team, 2018. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131.1.
URL https://CRAN.R-project.org/package=nlme

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/

Rombach, I., Rivero-Arias, O., Gray, A. M., Jenkinson, C., Burke, Ó., 2016. The current practice of handling and reporting missing outcome data in eight widely used proms in rct publications: a review of the current literature. Quality of Life Research 25 (7), 1613–1623.
URL https://doi.org/10.1007/s11136-015-1206-1

Rubin, D. B., 1987. Multiple imputation for nonresponse in surveys. John Wiley & Sons.

Schulz, K. F., Altman, D. G., Moher, D., 2010. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 340.
URL https://www.bmj.com/content/340/bmj.c332

Schulz, K. F., Grimes, D. A., 2002. Sample size slippages in randomised trials: exclusion and the lost and wayward. Lancet, 781–785.
URL https://doi.org/10.1016/S0140-6736(02)07882-0

Sedgwick, P., 2013. What is per protocol analysis? BMJ 346.
URL https://www.bmj.com/content/346/bmj.f3748

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., Carpenter, J. R., 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 338.
URL https://www.bmj.com/content/338/bmj.b2393

Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., Thabane, M., Giangregorio, L., Dennis, B., Kosa, D., Debono, V. B., Dillenburg, R., Fruci, V., Bawor, M., Lee, J., Wells, G., Goldsmith, C. H., 2013. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. BMC Medical Research Methodology 13 (1), 92.
URL https://doi.org/10.1186/1471-2288-13-92

van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: Multivariate Imputation by Chained Equations in R.
URL https://www.jstatsoft.org/v45/i03/

Veieroed, M. B., Lydersen, S., Laake, P., 2012. Medical statistics in clinical and epidemiological research. Gyldendal.

Vickers, A. J., 2001. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Medical Research Methodology 1 (1), 1.
URL https://doi.org/10.1186/1471-2288-1-6

Vickers, A. J., Altman, D. G., 2001. Analyzing controlled trials with baseline and follow up measurements. BMJ 323, 1123–1124.

Warnes, G. R., Bolker, B., Lumley, T., from Randall C. Johnson are Copyright SAIC-Frederick, R. C. J. C., by the Intramural Research Program, I. F., of the NIH, Institute, N. C., for Cancer Research under NCI Contract NO1-CO-12400., C., 2015. gmodels: Various R Programming Tools for Model Fitting. R package version 2.16.2.
URL https://CRAN.R-project.org/package=gmodels

Wei, W. W. S., 2006. Time Series Analysis, Univariate and multivariate Methods, 2nd Edition. Pearson.

Zhang, Y., D., F. I., Lozano, L. E. C., et al., 2017. A systematic survey on reporting and methods for handling missing participant data for continuous outcomes in randomized controlled trials. Journal of Clinical Epidemiology 88, 57–66.
URL https://doi.org/10.1016/j.jclinepi.2017.05.017

Appendix

R-code

Simulations

```
#Simulations with two time points and two groups,
#with different correlation. Methods: Follow-up score,
#change score analysis, ANCOVA and cLDA.

library(MASS)
library(gmodels)
library(lme4)
library(nlme)
library(lattice)
library(mice)
library(Matrix)
library(metap)

#Global values
mis.per = 0.1       #Percentage of missing data
n = 100             #Total number of participants
m0 = c(0,0)         #Expected values, control group
m1 = c(0,1)         #Expected values, treatment group
var.sim = 4         #Variance for both groups
alpha = 0.05        #Type 1-error
m = 5000            #Number of simulations
th = m1[2]-m0[2]    #True parameter of interest
V = 2               #Number of time points

#Group vector
gr.1 = c(rep(0,n/2), rep(1,n/2))

#Pattern used in ampute - missing only at the y's
pat.1 = matrix(data = c(1,1,0,1,1,0), nrow = 2, ncol = 3)

#Weights MAR data. Missing dependent of group
w.mar = matrix(data = c(1,1,0,0,0,0), nrow = 2, ncol = 3)

#Weight MNAR data.
w.mnar = matrix(data = c(0,0,1,0,0,1), nrow = 2, ncol = 3)
```

```
#Function generating missing data mechanisms
miss.func2 <- function(cor = 0.6){

  sig = matrix(data = c(var.sim, var.sim*cor, var.sim*cor, var.sim),
               nrow=2, ncol=2)  #Sigma

  #Simulations of missing data
  gr.a1 <- mvrnorm(n/2, m0, Sigma = sig)
  gr.b1 <- mvrnorm(n/2, m1, Sigma = sig)
  y1 = matrix(0,n,(V+1))
  y1[(1:(n/2)),(2:(V+1))] = gr.a1
  y1[((n/2+1):n),(2:(V+1))] = gr.b1
  y1[,1] = gr.1
  y1.res = c(y1[,2], y1[,3])

  #Missing data assumptions:
  #MCAR
  mcar1 = ampute(y1, prop = mis.per*((V-1)/V), mech = "MCAR",
                 bycases = FALSE, patterns = pat.1)
  y.mcar1 = c(mcar1$amp[,2], mcar1$amp[,3])

  #MAR - missing depends on groups. More missing in control group
  mar1 = ampute(y1, prop = mis.per*((V-1)/V), mech = "MAR",
                bycases = FALSE, patterns = pat.1,
                type = c("LEFT", "LEFT"), weights = w.mar)
  y.mar1 = c(mar1$amp[,2], mar1$amp[,3])

  #MNAR
  mnar1 = ampute(y1, prop = mis.per*((V-1)/V), mech = "MNAR",
                 bycases = FALSE, patterns = pat.1,
                 weights = w.mnar, type = c("LEFT", "LEFT"))
  y.mnar1 = c(mnar1$amp[,2], mnar1$amp[,3])

  y.tot1 = data.frame(id1 = rep(seq(1, n), 2), gr1 = rep(gr.1, 2),
                      td1 = c(rep(0, n), rep(1, n)), y1.res,4,
                      y.mcar1, y.mar1, y.mnar1)
  names(y.tot1) = c("ID", "Group", "Time", "y", "MCAR", "MAR", "MNAR")

  return(y.tot1)
}

#ANCOVA data frame for the analysis
Y1 = c(rep(0, n))
Y0 = c(rep(0, n))
dta.ancova = data.frame(Y1, Y0, gr.1)
dta.ancova$Y1 = as.numeric(dta.ancova$Y1)
dta.ancova$Y0 = as.numeric(dta.ancova$Y0)
dta.ancova$gr.1 = as.factor(dta.ancova$gr.1)
cmat = matrix(0, 1, 3)
```

```
cmat[1, ] = c(0, 0, 1)
colnames(cmat) = c("(Intercept)", "Y0", "group11")

#cLDA data frame for use for each simulation
id = rep(seq(1, n), 2)
Y = c(rep(0, 2*n))
gr.2 = c(rep(gr.1,2))
time = c(rep(0, n), rep(1, n))
t1gr1 = gr.2*time
dta = data.frame(id, Y, time, t1gr1)
dta$id = as.factor(dta$id)
dta$time = as.numeric(dta$time)
dta$tVgr1 = as.factor(dta$tVgr1)
cmat2 = matrix(0, 1, 3)
cmat2[1, ] = c(0, 0, 1)
colnames(cmat2) = c("(Intercept)", "time", "t1gr1")

correlation = seq(0.1,0.9,0.1)

for(j in 1:(length(correlation))){

  corr = correlation[j]

  #Data frame with results
  res <- data.frame(matrix(0, nrow = 15, ncol = 4))
  names(res) = c("no missing", "MCAR", "MAR", "MNAR")
  row.names(res) = c("Bias ANCOVA", "Power ANCOVA", "CI ANCOVA",
    "Bias cLDA", "Power cLDA", "CI cLDA","Bias change", "Power change",
    "CI change", "Bias follow-up", "Power follow-up", "CI follow-up",
    "Bias ANCOVA MI", "Power ANCOVA MI", "CI ANCOVA MI")

  #Simulations
  for(k in 1:m){

    df <- miss.func2(cor = corr)

    for(i in 1:4){

      #cLDA
      dta[,2] = df[,3+i]

      model = lme(Y ~ time + t1gr1, random = ~1|id, data = dta,
                  na.action=na.exclude)
      est2 = estimable(model, cmat2, conf.int = (1-alpha))

      res[4,i] = res[4,i] + (est2$Estimate - th)
      if(est2$`Pr(>|t|)` > alpha){
        res[5,i] = res[5,i] + 1
      }
      if(th >= est2$Lower.CI & th <= est2$Upper.CI){
```

```
  res[6,i] = res[6,i] + 1
}

#ANCOVA
dta.ancova$Y0 = df[(1:n),3+i]
dta.ancova$Y1 = df[((n+1):(2*n)),3+i]

fit <- lm(Y1 ~ Y0 + gr.1, data = dta.ancova, na.action = na.exclude)
est <- estimable(obj = fit, cm = cmat, conf.int = (1-alpha))

#ANCOVA without MI
res[1,i] = res[1,i] + (est$Estimate - th)
if(est$`Pr(>|t|)` > alpha){
  res[2,i] = res[2,i] + 1
}
if(th >= est$Lower.CI & th <= est$Upper.CI){
  res[3,i] = res[3,i] + 1
}

#ANCOVA with MI
if(i!=1){
  ancova.mice = is.na(dta.ancova)
  ancova.mice[,1] = FALSE
  MI = mice(data = dta.ancova, m=20, where = ancova.mice,
            print = FALSE)
  MI2 = summary(pool(with(data = MI,
        expr = lm(Y1 ~Y0+gr.1, na.action = na.exclude))))
  est = MI2[3,]

  #Bias
  res[13,i] = res[13,i] + (est$estimate - th)

  #Power
  if(est$p.value > alpha){
    res[14,i] = res[14,i] + 1
  }

  #CI (Rubins rule)
  if(th >= (est$estimate -
            (qt((1-alpha/2),est$df))*(est$std.error)) & th <=
     (est$estimate + (qt((1-alpha/2),est$df))*(est$std.error))){
    res[15,i] = res[15,i] + 1
  }
}

#Change score
test.c <- t.test(x =
    (df[((n+n/2+1):(2*n)),3+i] - df[((n/2+1):n),3+i]),
    y = (df[((n+1):(n+n/2)),3+i] - df[(1:(n/2)),3+i]),
    var.equal = TRUE, na.action=na.exclude)
```

```
      res[7,i] = res[7,i]+((test.c$estimate[1] - test.c$estimate[2]) - th)
      if(test.c$p.value > alpha){
        res[8,i] = res[8,i]+1
      }
      if(th >= test.c$conf.int[1] & th <= test.c$conf.int[2]){
        res[9,i] = res[9,i]+1
      }

      if(j==1){
        #Follow-up
        test.f <- t.test(x = df[(n+n/2+1):(2*n),3+i],
                         y = df[(n+1):(n+n/2),3+i], var.equal = TRUE)
        res[10,i] = res[10,i] +
          ((test.f$estimate[1] - test.f$estimate[2]) - th)
        if(test.f$p.value > alpha){
          res[11,i] = res[11,i] + 1
        }
        #CI
        if(th >= test.f$conf.int[1] & th <= test.f$conf.int[2]){
          res[12,i] = res[12,i]+1
        }
      }
    }
  }

  #Results
  for(l in 1:4){
    for(p in 1:15){
      if(p==2|p==5|p==8|p==11|p==14){
        res[p,l] = 1-res[p,l]/m  #Power
      }else{
        res[p,l] = res[p,l]/m    #CI/Bias
      }
    }
  }
  print(corr)
  print(res)
}


#Longitudinal data



#Global values
alpha = 0.05   #Type I error
m = 5000       #Number of simulations
n = 100        #Number of individuals in the trial
V = 4          #Number of time points
```

```
mu0 = 0          #Expected value at baseline for both groups
mu1 = 1          #Expected value at last follow-up for treatment group
th = mu1 - mu0     #True difference
var.sim = 4        #Variance. Equal for both group and time.
M = 20             #Number of multiple imputations
M0 = c(rep(mu0, V))                    #Expected values group 0
M1 = c(seq(mu0, mu1, by= 1/(V-1)))     #Expected values group 1
phi = 0.8      #Exponential decay rate (higher values, faster toward zero)
Vm = matrix(rep(0,V*V), nrow = V)

#Decaying exponential correlation matrix
for(i in 1:(V)){
  for(j in 1:(V)){
    Vm[i,j] = exp(-phi*abs(i-j))
  }
}
#Compound correlation structure
#cor = 0.6
#Vm = matrix(data = c(rep(c(1, rep(cor,V)),(V-1)),1), nrow=V, ncol=V)

sig = var.sim*Vm     #Covariance matrix

#Patterns and weights for missing data assumptions in the function ampute
long.pat = matrix(data = c(1,1,1,1, 0,1,1,1, 1,0,1,1,
          1,1,0,1, 1,1,1,0),  nrow = 4, ncol = 5)
w.mar.long = matrix(data = c(1,1,1,1, 0,0,0,0,
          0,0,0,0, 0,0,0,0, 0,0,0,0), nrow = 4, ncol = 5)
w.mnar.long = matrix(data = c(0,0,0,0, 1,0,0,0,
          0,1,0,0, 0,0,1,0, 0,0,0,1), nrow = 4, ncol = 5)

gr.short = c(rep(0,n/2), rep(1,n/2))

#cLDA
gr = rep(c(rep(0,n/2), rep(1,n/2)),V)
time = rep(0:(V-1), each = n)
id = rep(seq(1, n), V)

#ANCOVA
id2 = rep(seq(1, n), (V-1))
time2 = rep(1:(V-1), each = n)
timegroup = rep(1:((2*V)-2), each = (n/2))

#ANCOVA MI
id3 = seq(1,n)

miss.func.long.2 <- function(mis.per = 0.10){

  #Simulations
  gr.a <- mvrnorm(n/2, M0, Sigma = sig)
  gr.b <- mvrnorm(n/2, M1, Sigma = sig)
```

```
  y = matrix(0,n,(V+1))
  y[(1:(n/2)),(2:(V+1))] = gr.a
  y[((n/2+1):n),(2:(V+1))] = gr.b
  y[,1] = gr.short

  y.res = as.vector(y[,2:(V+1)])

  #Missing data mechanisms
  #MCAR
  mcar = ampute(y, prop = mis.per*((V-1)/V), mech = "MCAR",
                bycases = FALSE, patterns = long.pat)
  y.mcar = c(mcar$amp[,2], mcar$amp[,3], mcar$amp[,4], mcar$amp[,5])

  #MAR
  mar = ampute(y, prop = mis.per*((V-1)/V), mech = "MAR",
               bycases = FALSE, patterns = long.pat,
               type = c("LEFT", "LEFT", "LEFT", "LEFT"),
               weights = w.mar.long)
  y.mar = c(mar$amp[,2], mar$amp[,3], mar$amp[,4], mar$amp[,5])

  mnar = ampute(y, prop = mis.per*((V-1)/V), mech = "MNAR",
                bycases = FALSE, patterns = long.pat,
                weights = w.mnar.long)
  y.mnar = c(mnar$amp[,2], mnar$amp[,3], mnar$amp[,4], mnar$amp[,5])

  y.tot = data.frame(id, gr, time, y.res, y.mcar, y.mar, y.mnar)
  names(y.tot) = c("ID", "Group", "Time", "y", "MCAR", "MAR", "MNAR")

  return(y.tot)
}

#Contrast matrix for cLDA
cmat2 <- matrix(0, 3, 7)
cmat2[1,] = c(0,-1,1,0,0,0,0)
cmat2[2,] = c(0,0,0,-1,1,0,0)
cmat2[3,] = c(0,0,0,0,0,-1,1)
colnames(cmat2) <- c("(Intercept)", "tVgr12", "tVgr13",
              "tVgr14", "tVgr15", "tVgr16", "tVgr17")

#Contrast matrix for ANCOVA
cmat = matrix(0, 3, 9)
cmat[1,] = c(0,0,0,0,1,0,0,0,0)
cmat[2,] = c(0,0,0,0,0,-1,1,0,0)
cmat[3,] = c(0,0,0,0,0,0,0,-1,1)
colnames(cmat) = c("(Intercept)", "Y0", "Y0It2", "Y0It3",
                   "timegroup2", "timegroup3", "timegroup4",
                   "timegroup5", "timegroup6")
rownames(cmat) = c("Theta1", "Theta2", "Theta3")
```

```
run.long <- function(miss.perc = 0.1){

  #cLDA data frame for each simulation
  Y = c(rep(0, V*n))
  tVgr1 = c(rep(1,n), rep(2,n/2), rep(3,n/2), rep(4,n/2),
            rep(5,n/2), rep(6,n/2), rep(7,n/2))
  dta = data.frame(id, Y, time, tVgr1)
  dta$Y = as.numeric(dta$Y)
  dta$id = as.factor(dta$id)
  dta$time = as.factor(dta$time)
  dta$tVgr1 = as.factor(dta$tVgr1)

  #ANCOVA data frame for each simulation
  Y0 = c(rep(0,(V-1)*n))
  Y1 = c(rep(0,(V-1)*n))
  Y0It2 = c(rep(0,(V-1)*n))
  Y0It3 = c(rep(0,(V-1)*n))
  dta.ancova = data.frame(id2, Y0, Y1, timegroup,
                          time2, Y0It2, Y0It3)
  dta.ancova$id2 = as.factor(dta.ancova$id2)
  dta.ancova$Y0 = as.numeric(dta.ancova$Y0)
  dta.ancova$Y1 = as.numeric(dta.ancova$Y1)
  dta.ancova$timegroup = as.factor(dta.ancova$timegroup)
  dta.ancova$time2 = as.factor(dta.ancova$time2)

  #ANCOVA MI data frame
  Y0.m = c(rep(0,n))
  Y1.m = c(rep(0,n))
  Y2.m = c(rep(0,n))
  Y3.m = c(rep(0,n))
  dta.mice = data.frame(id3, gr.short, Y0.m, Y1.m, Y2.m, Y3.m)

  #Data frame with results
  res.long = data.frame(matrix(0, nrow = 3*(V-1), ncol = 11))
  names(res.long) = c("no missing cLDA", "MCAR cLDA", "MAR cLDA",
        "MNAR cLDA", "no missing ANCOVA", "MCAR ANCOVA",
        "MAR ANCOVA", "MNAR ANCOVA", "MCAR ANCOVA MI",
        "MAR ANCOVA MI", "MNAR ANCOVA MI")
  row.names(res.long) = c("Bias theta1", "CI theta1","Power theta1",
        "Bias theta2", "CI theta2","Power theta2","Bias theta3",
        "CI theta3","Power theta3")

  for(k in 1:m){

    set.seed(sample(1:(2^20),1))
    df <- miss.func.long.2(mis.per = miss.perc)

    #Analysis of different missing assumptions
    for(i in 1:4){
```

```
#cLDA
dta$Y = df[,(3+i)]
model = lme(Y ~ tVgr1, random = ~1|id, data = dta,
    correlation = corExp((1/phi), form = ~ 1|id),
    na.action = na.exclude)
#Compound correlation structure
#model = lme(Y ~ tVgr1, random = ~1|id, data = dta,
    #na.action = na.exclude)

est2 = estimable(model, cmat2, conf.int = (1-alpha))

for(l in 1:(V-1)){
  #Bias
  res.long[(3*l-2),i] = res.long[(3*l-2),i] +
    (est2$Estimate[l]-Ml[(l+1)])

  #Power
  if(est2$`Pr(>|t|)`[l] > alpha){
    res.long[(3*l),i] = res.long[(3*l),i] + 1
  }

  #CI
  if(Ml[(1+l)] >= est2$Lower.CI[l] &
     Ml[(1+l)] <= est2$Upper.CI[l]){
    res.long[(3*l-1),i] = res.long[(3*l-1),i] + 1
  }
}

#ANCOVA
dta.ancova$Y0 = rep(df[(1:n),3+i], (V-1))
dta.ancova$Y0It2[101:200] = df[(1:n),3+i]
dta.ancova$Y0It3[201:300] = df[(1:n),3+i]
dta.ancova$Y1 = df[((n+1):(V*n)),3+i]

#ANCOVA without MI
model2 = lme(Y1 ~ Y0 + Y0It2 + Y0It3 + timegroup,
       random = ~1|id2, data = dta.ancova,
       correlation = corExp((1/phi), form = ~ 1|id2),
       na.action = na.exclude)
#Compound correlation structure
#model2 = lme(Y1 ~ Y0 + Y0It2 + Y0It3 + timegroup,
           #random = ~1|id2,
           #data = dta.ancova, na.action = na.exclude)
est = estimable(obj = model2, cm = cmat,
                conf.int = (1-alpha))

for(l in 1:(V-1)){
  #Bias
  res.long[(3*l-2),(4+i)] = res.long[(3*l-2),(4+i)] +
    (est$Estimate[l] - Ml[(l+1)])
```

```
    #Power
    if(est$`Pr(>|t|)`[1] > alpha){
      res.long[(3*l),(4+i)] = res.long[(3*l),(4+i)] + 1
    }

    #CI
    if(Ml[(1+l)] >= est$Lower.CI[1] & Ml[(1+l)] <=
       est$Upper.CI[1]){
      res.long[(3*l-1),(4+i)] = res.long[(3*l-1),(4+i)] + 1
    }
  }
  if(i!=1){
    dta.mice$Y0.m = df[1:n, 3+i]
    dta.mice$Y1.m = df[((n+1):(2*n)), 3+i]
    dta.mice$Y2.m = df[((2*n+1):(3*n)), 3+i]
    dta.mice$Y3.m = df[((3*n+1):(4*n)), 3+i]

    #MI
    imp.dta = is.na(dta.mice)
    imp.dta[,4:6] = FALSE
    MI = mice(data = dta.mice, where = imp.dta, m=M,
              print = FALSE)
    MI2 = complete(MI, "all")

    tote = matrix(0,20,3)
    totvlc = matrix(0,3,3)

    for(j in 1:M){
      temp <- as.data.frame(MI2[j])
      dta.ancova$Y0 = rep(temp[,3],3)
      dta.ancova$Y0It2[((n+1):(2*n))] = temp[,3]
      dta.ancova$Y0It3[((2*n+1):(3*n))] = temp[,3]

      model2 = lme(Y1 ~ Y0 + Y0It2 + Y0It3 + timegroup,
                   random = ~1|id2,
                   data = dta.ancova, correlation = corExp((1/phi),
                   form = ~ 1|id2), na.action = na.exclude)
      #Compound correlation structure
      #model2 = lme(Y1 ~ Y0 + Y0It2 + Y0It3 + timegroup,
      #random = ~1|id2,
      #data = dta.ancova, na.action = na.exclude)

      estlc = cmat%*%fixef(model2)  #Estimates
      vlc = cmat%*%vcov(model2)%*%t(cmat) #Variance-covariance matrix

      tote[j,] = estlc
      totvlc = totvlc + vlc
    }
```

```
        #Rubin's rule
        meanMI = c(0,0,0)
        betvar = c(0,0,0)
        dfMI = c(0,0,0)
        pval = c(0,0,0)
        withvar = (1/M)*totvlc
        totvar = withvar

        for(l in 1:(V-1)){
          #Estimates
          meanMI[l] = mean(tote[,l])
          #Between variance
          betvar[l] = (sd(tote[,l]))^2
          #Total variance
          totvar[l,l] = totvar[l,l] + (1+1/M)*(betvar[l])
          #Degrees of freedom
          dfMI[l] = (M-1)*(1 +
                    withvar[l,l]/(((1+M^(-1))*betvar[l])))^2
          #P-values
          pval[l] = 1-pf(((meanMI[l])^2)/(totvar[l,l]),
                         df1 = 1, df2 = dfMI[l])

          res.long[(3*l-2),(7+i)] = res.long[(3*l-2),(7+i)] +
                    (meanMI[l] - (Ml[l+1]))

          if(Ml[(1+l)] >=
              (meanMI[l]-qt((1-alpha/2),dfMI[l])*sqrt(totvar[l,l])) &
              Ml[(1+l)] <= (meanMI[l]+qt((1-alpha/2),
                                dfMI[l])*sqrt(totvar[l,l]))){
            res.long[(3*l-1),(7+i)] = res.long[(3*l-1),(7+i)] + 1
          }
          if(pval[l] > alpha){
            res.long[(3*l),(7+i)] = res.long[(3*l),(7+i)] + 1
          }
        }
      }
    }
  }


  #Results
  for(q in 1:11){
    for(z in 1:((V-1)*3)){
      if(z==3|z==6|z==9){
        res.long[z,q] = 1- res.long[z,q]/m  #Power
      }else{
        res.long[z,q] = res.long[z,q]/m      #CI/Bias
      }
    }
  }
```

```
  print(miss.perc)
  print(res.long)
}

mp = c(0.05,0.1,0.15,0.20)

for(p in 1:length(mp)){
  run.long(miss.perc = mp[p])
}
```

Example

```
#Example RCT: Obese children with different types of exercises
library(foreign)
library(nortest)
library(nlme)

#Data from file
datadir = "~/Documents/NTNU/Master/Real_example"
dfile = "/15.11.2017-Childrenstudy-NTNU-mainfileLONG_echo-HE.sav"

#From SPSS to R
alldata = read.spss(paste(datadir, dfile, sep = ""),
                    use.value.labels = FALSE, to.data.frame = T)

#GLS variable

#Data with ID, Y0, Y1, Y2 and Group
look.data = cbind((alldata[grep("1", alldata$Time),])$SubjectID,
    (alldata[grep("1", alldata$Time),])$Rest_LV_GLS,
    (alldata[grep("2", alldata$Time),])$Rest_LV_GLS,
    (alldata[grep("3", alldata$Time),])$Rest_LV_GLS,
    (alldata[grep("1", alldata$Time),])$Group)
colnames(look.data) = c("ID", "Y0", "Y1", "Y2", "Group")
ld = as.data.frame(look.data)

#Diet group
dt = ld[grep("3", ld$Group),]
ndt = dim(dt)[1]
mdt = c(mean(dt$Y0,na.rm = TRUE), mean(dt$Y1,na.rm = TRUE),
        mean(dt$Y2,na.rm = TRUE))
vdt = c(var(dt$Y0,na.rm = TRUE), var(dt$Y1,na.rm = TRUE),
        var(dt$Y2,na.rm = TRUE))
cdt = cor(dt[,2:4],use= "na.or.complete")
nadt = c(sum(is.na(dt$Y0)), sum(is.na(dt$Y1)), sum(is.na(dt$Y2)))
dres = data.frame(nadt, ndt, mdt, vdt, cdt)
print(dres)

#MICT group
```

```
mc = ld[grep("2", ld$Group),]
nmc = dim(mc)[1]
mmc = c(mean(mc$Y0,na.rm = TRUE), mean(mc$Y1,na.rm = TRUE),
        mean(mc$Y2,na.rm = TRUE))
vmc = c(var(mc$Y0,na.rm = TRUE), var(mc$Y1,na.rm = TRUE),
        var(mc$Y2,na.rm = TRUE))
cmc = cor(mc[,2:4],use= "na.or.complete")
namc = c(sum(is.na(mc$Y0)), sum(is.na(mc$Y1)), sum(is.na(mc$Y2)))
mcres = data.frame(namc, nmc, mmc, vmc, cmc)
print(mcres)

#HIIT group
ht = ld[grep("1", ld$Group),]
nht = dim(ht)[1]
mht = c(mean(ht$Y0,na.rm = TRUE), mean(ht$Y1,na.rm = TRUE),
        mean(ht$Y2,na.rm = TRUE))
vht = c(var(ht$Y0,na.rm = TRUE), var(ht$Y1,na.rm = TRUE),
        var(ht$Y2,na.rm = TRUE))
cht = cor(ht[,2:4],use= "na.or.complete")
naht = c(sum(is.na(ht$Y0)), sum(is.na(ht$Y1)), sum(is.na(ht$Y2)))
htres = data.frame(naht, nht, mht, vht, cht)
print(htres)

#Checking for normality in the data - only baseline values
bl = (alldata[grep("1", alldata$Time),])$Rest_LV_GLS
qqnorm(bl, col = "slateblue",xlab = "", ylab = "", pch = 16,
       main = "")
qqline(bl, col = "slateblue")
hist(bl, main = "", xlab = "GLS", ylab = "Freq",
     col = "slateblue")

#Plot diet group
Tplot = c(0, 1, 2)
plot(Tplot, c(dt$Y0[1], dt$Y1[1], dt$Y2[1]), type = "b",
     col = "slateblue", main = "Diet group", ylim = c(-25, -11),
     xlim = c(0, 2), ylab = "GLS", xlab = " ", xaxt = 'n',
     pch = 16)
for(i in 2:(ndt)){
  lines(Tplot, c(dt$Y0[i], dt$Y1[i], dt$Y2[i]),
        type = "b", col = "slateblue", pch = 16)
}
axis(1, at = 0:2, labels = c("Baseline", "Follow-up 1",
                             "Follow-up 2"))

#Plot MICT group
Tplot = c(0, 1, 2)
plot(Tplot, c(mc$Y0[1], mc$Y1[1], mc$Y2[1]), type = "b",
     col = "slateblue", main = "MICT group", ylim = c(-25, -11),
     xlim = c(0, 2), ylab = "GLS", xlab = " ", xaxt = 'n',
     pch = 16)
```

```
for(i in 2:(nmc)){
  lines(Tplot, c(mc$Y0[i], mc$Y1[i], mc$Y2[i]), type = "b",
        col = "slateblue", pch = 16)
}
axis(1, at = 0:2, labels = c("Baseline", "Follow-up 1",
                             "Follow-up 2"))

#Plot HIIT group
Tplot = c(0, 1, 2)
plot(Tplot, c(ht$Y0[1], ht$Y1[1], ht$Y2[1]), type = "b",
     col = "slateblue", main = "HIIT group",
     ylim = c(-25, -11),
     xlim = c(0, 2), ylab = "GLS", xlab = " ", xaxt = 'n',
     pch = 16)
for(i in 2:(nht)){
  lines(Tplot, c(ht$Y0[i], ht$Y1[i], ht$Y2[i]), type = "b",
        col = "slateblue", pch = 16)
}
axis(1, at = 0:2, labels = c("Baseline", "Follow-up 1",
                             "Follow-up 2"))

#Analysis of the data
#cLDA
Y1 = as.numeric(alldata$Rest_LV_GLS)
Time = as.factor(alldata$Time)
Group = as.factor(alldata$Group)
ID = as.factor(alldata$SubjectID)
TVGR = c(rep(0,length(Y1)))

for(i in 1:length(TVGR)){
  if(Time[i] == 1){
    TVGR[i] = 1
  }else if(Time[i] == 2 & Group[i] == 3){
    TVGR[i] = 2
  }else if(Time[i] == 2 & Group[i] == 2){
    TVGR[i] = 4
  }else if(Time[i] == 2 & Group[i] == 1){
    TVGR[i] = 6
  }else if(Time[i] == 3 & Group[i] == 3){
    TVGR[i] = 3
  }else if(Time[i] == 3 & Group[i] == 2){
    TVGR[i] = 5
  }else{
    TVGR[i] = 7
  }
}
TVGR = as.factor(TVGR)
clda = data.frame(Y1, ID, Time, Group, TVGR)  #cLDA data frame

model.clda = lme(Y1 ~ TVGR, random = ~1|ID, data = clda,
```

```
        correlation = corCompSymm(), na.action = na.exclude)

#Contrast matrix for cLDA
CMAT2 <- matrix(0, 6, 7)
CMAT2[1,] = c(0,0,-1,0,0,0,1)
CMAT2[2,] = c(0,-1,0,0,0,1,0)
CMAT2[3,] = c(0,0,0,0,-1,0,1)
CMAT2[4,] = c(0,0,0,-1,0,1,0)
CMAT2[5,] = c(0,0,-1,0,1,0,0)
CMAT2[6,] = c(0,-1,0,1,0,0,0)
colnames(CMAT2) <- c("(Intercept)", "TVGR2", "TVGR3", "TVGR4",
                     "TVGR5", "TVGR6", "TVGR7")

alpha = 0.05
EST <- estimable(model.clda, CMAT2, conf.int = (1 - alpha))

#Results cLDA
clda.res = matrix(0,6,4)
row.names(clda.res) = c("HvsD_T2", "HvsD_T1", "HvsM_T2", "HvsM_T1",
                        "MvsD_T2", "MvsD_T1")
colnames(clda.res) = c("Estimate", "lower_CI", "Upper_CI", "p-value")

clda.res[,1] = EST$Estimate
clda.res[,2] = EST$Lower.CI
clda.res[,3] = EST$Upper.CI
clda.res[,4] = EST$`Pr(>|t|)`

print(clda.res)

#ANCOVA
Y = c(ld$Y1, ld$Y2)
Y0 = rep(ld$Y0, 2)
Y0IT2 = c(rep(0, length(Y0)/2), ld$Y0)
ID2 = rep(ld$ID, 2)
GROUP = rep(ld$Group, 2)
TIME = c(rep(2,length(Y)/2),rep(3,length(Y)/2))
TMGR = c(rep(0,length(Y)))

for(i in 1:length(TMGR)){
  if(TIME[i] == 2 & GROUP[i] == 3){
    TMGR[i] = 1
  }else if(TIME[i] == 2 & GROUP[i] == 2){
    TMGR[i] = 3
  }else if(TIME[i] == 2 & GROUP[i] == 1){
    TMGR[i] = 5
  }else if(TIME[i] == 3 & GROUP[i] == 3){
    TMGR[i] = 2
  }else if(TIME[i] == 3 & GROUP[i] == 2){
    TMGR[i] = 4
  }else{
```

```
    TMGR[i] = 6
  }
}

Y = as.numeric(Y)
Y0 = as.numeric(Y0)
Y0IT2 = as.numeric(Y0IT2)
ID2 = as.factor(ID2)
TMGR = as.factor(TMGR)

ancova = data.frame(Y, Y0, Y0IT2, ID2, GROUP, TIME, TMGR)

model.ancova = lme(Y ~ Y0 + Y0IT2 + TMGR, random = ~1|ID2,
      data = ancova, correlation = corCompSymm(),
      na.action = na.exclude)

#Contrast matrix for ancova
CMAT <- matrix(0, 6, 8)
CMAT[1,] = c(0,0,0,-1,0,0,0,1)
CMAT[2,] = c(0,0,0,0,0,0,1,0)
CMAT[3,] = c(0,0,0,0,0,-1,0,1)
CMAT[4,] = c(0,0,0,0,-1,0,1,0)
CMAT[5,] = c(0,0,0,-1,0,1,0,0)
CMAT[6,] = c(0,0,0,0,1,0,0,0)
colnames(CMAT) <- c("(Intercept)", "Y0", "Y0IT2", "TMGR2",
                    "TMGR3", "TMGR4", "TMGR5", "TMGR6")

EST2 <- estimable(model.ancova, CMAT,
                  conf.int = (1 - alpha))

#Results ANCOVA
ancova.res = matrix(0,6,4)
row.names(ancova.res) = c("HvsD T2", "HvsD T1", "HvsM T2",
                          "HvsM T1", "MvsD T2", "MvsD T1")
colnames(ancova.res) = c("Estimate", "lower CI",
                         "Upper CI", "p-value")

ancova.res[,1] = EST2$Estimate
ancova.res[,2] = EST2$Lower.CI
ancova.res[,3] = EST2$Upper.CI
ancova.res[,4] = EST2$`Pr(>|t|)`

print(ancova.res)


#LVS

#Logtransformation of data
alldata$Rest_LV_s_mean = log(alldata$Rest_LV_s_mean)
```

```
#Data with ID, Y0, Y1, Y2 and Group
look.data = cbind((alldata[grep("1", alldata$Time),])$SubjectID,
            (alldata[grep("1", alldata$Time),])$Rest_LV_s_mean,
            (alldata[grep("2", alldata$Time),])$Rest_LV_s_mean,
            (alldata[grep("3", alldata$Time),])$Rest_LV_s_mean,
            (alldata[grep("1", alldata$Time),])$Group)
colnames(look.data) = c("ID", "Y0", "Y1", "Y2", "Group")
ld = as.data.frame(look.data)

#Correlation between time - kanskje ikke n dvendig    ha med?
cor.all.data = cor(look.data[,2:4], use = "na.or.complete")
print(cor.all.data)

#Diet group
dt = ld[grep("3", ld$Group),]
ndt = dim(dt)[1]
mdt = c(mean(dt$Y0,na.rm = TRUE), mean(dt$Y1,na.rm = TRUE),
        mean(dt$Y2,na.rm = TRUE))
vdt = c(var(dt$Y0,na.rm = TRUE), var(dt$Y1,na.rm = TRUE),
        var(dt$Y2,na.rm = TRUE))
cdt = cor(dt[,2:4],use= "na.or.complete")
nadt = c(sum(is.na(dt$Y0)), sum(is.na(dt$Y1)),
         sum(is.na(dt$Y2)))
dres = data.frame(nadt, ndt, mdt, vdt, cdt)
print(dres)

#MICT group
mc = ld[grep("2", ld$Group),]
nmc = dim(mc)[1]
mmc = c(mean(mc$Y0,na.rm = TRUE), mean(mc$Y1,na.rm = TRUE),
        mean(mc$Y2,na.rm = TRUE))
vmc = c(var(mc$Y0,na.rm = TRUE), var(mc$Y1,na.rm = TRUE),
        var(mc$Y2,na.rm = TRUE))
cmc = cor(mc[,2:4],use= "na.or.complete")
namc = c(sum(is.na(mc$Y0)), sum(is.na(mc$Y1)),
         sum(is.na(mc$Y2)))
mcres = data.frame(namc, nmc, mmc, vmc, cmc)
print(mcres)

#HIIT group
ht = ld[grep("1", ld$Group),]
nht = dim(ht)[1]
mht = c(mean(ht$Y0,na.rm = TRUE), mean(ht$Y1,na.rm = TRUE),
        mean(ht$Y2,na.rm = TRUE))
vht = c(var(ht$Y0,na.rm = TRUE), var(ht$Y1,na.rm = TRUE),
        var(ht$Y2,na.rm = TRUE))
cht = cor(ht[,2:4],use= "na.or.complete")
naht = c(sum(is.na(ht$Y0)), sum(is.na(ht$Y1)),
         sum(is.na(ht$Y2)))
htres = data.frame(naht, nht, mht, vht, cht)
```

```
print(htres)

#Checking for normality in the data - only baseline values
bl = (alldata[grep("1", alldata$Time),])$Rest_LV_s_mean
qqnorm(bl, col = "slateblue", xlab = "", ylab = "",
       pch = 16, main = "")
qqline(bl, col = "slateblue")
hist(bl, main = "", xlab = "", ylab = "", col = "slateblue3")

#Plot diet group
Tplot = c(0, 1, 2)
plot(Tplot, c(dt$Y0[1], dt$Y1[1], dt$Y2[1]), type = "b",
     col = "slateblue",
     main = "Diet group", ylim = c(1.7, 2.7),
     xlim = c(0, 2), ylab = "log(LVS)", xlab = "",
     xaxt = 'n', pch = 16)
for(i in 2:(ndt)){
  lines(Tplot, c(dt$Y0[i], dt$Y1[i], dt$Y2[i]), type = "b",
        col = "slateblue", pch = 16)
}
axis(1, at = 0:2, labels = c("Baseline",
            "Follow-up 1", "Follow-up 2"))

#Plot MICT group
Tplot = c(0, 1, 2)
plot(Tplot, c(mc$Y0[1], mc$Y1[1], mc$Y2[1]), type = "b",
     col = "slateblue", main = "MICT group", ylim = c(1.7, 2.7),
     xlim = c(0, 2), ylab = "log(LVS)", xlab = "",
     xaxt = 'n', pch = 16)
for(i in 2:(nmc)){
  lines(Tplot, c(mc$Y0[i], mc$Y1[i], mc$Y2[i]), type = "b",
        col = "slateblue", pch = 16)
}
axis(1, at = 0:2, labels = c("Baseline",
                             "Follow-up 1", "Follow-up 2"))

#Plot HIIT group
Tplot = c(0, 1, 2)
plot(Tplot, c(ht$Y0[1], ht$Y1[1], ht$Y2[1]), type = "b",
     col = "slateblue",
     main = "HIIT group", ylim = c(1.7, 2.7),
     xlim = c(0, 2), ylab = "log(LVS)", xlab = "",
     xaxt = 'n', pch = 16)
for(i in 2:(nht)){
  lines(Tplot, c(ht$Y0[i], ht$Y1[i], ht$Y2[i]), type = "b",
        col = "slateblue", pch = 16)
}
axis(1, at = 0:2, labels = c("Baseline",
            "Follow-up 1", "Follow-up 2"))
```

```
#cLDA
Y1 = as.numeric(alldata$Rest_LV_s_mean)
Time = as.factor(alldata$Time)
Group = as.factor(alldata$Group)
ID = as.factor(alldata$SubjectID)
TVGR = c(rep(0,length(Y1)))

for(i in 1:length(TVGR)){
  if(Time[i] == 1){
    TVGR[i] = 1
  }else if(Time[i] == 2 & Group[i] == 3){
    TVGR[i] = 2
  }else if(Time[i] == 2 & Group[i] == 2){
    TVGR[i] = 4
  }else if(Time[i] == 2 & Group[i] == 1){
    TVGR[i] = 6
  }else if(Time[i] == 3 & Group[i] == 3){
    TVGR[i] = 3
  }else if(Time[i] == 3 & Group[i] == 2){
    TVGR[i] = 5
  }else{
    TVGR[i] = 7
  }
}
TVGR = as.factor(TVGR)
clda = data.frame(Y1, ID, Time, Group, TVGR)  #cLDA data frame

model.clda = lme(Y1 ~ TVGR, random = ~1|ID, data = clda,
                 correlation = corCompSymm(),
                 na.action = na.exclude)

#Contrast matrix for cLDA
CMAT2 <- matrix(0, 6, 7)
CMAT2[1,] = c(0,0,-1,0,0,0,1)
CMAT2[2,] = c(0,-1,0,0,0,1,0)
CMAT2[3,] = c(0,0,0,0,-1,0,1)
CMAT2[4,] = c(0,0,0,-1,0,1,0)
CMAT2[5,] = c(0,0,-1,0,1,0,0)
CMAT2[6,] = c(0,-1,0,1,0,0,0)
colnames(CMAT2) <- c("(Intercept)", "TVGR2", "TVGR3",
                     "TVGR4", "TVGR5", "TVGR6", "TVGR7")

alpha = 0.05
EST <- estimable(model.clda, CMAT2, conf.int = (1 - alpha))

#Results cLDA
clda.res = matrix(0,6,4)
row.names(clda.res) = c("HvsD_T2", "HvsD_T1", "HvsM_T2",
                        "HvsM_T1", "MvsD_T2", "MvsD_T1")
colnames(clda.res) = c("Estimate", "lower_CI",
```

```
                        "Upper␣CI", "p-value")

clda.res[,1] = EST$Estimate
clda.res[,2] = EST$Lower.CI
clda.res[,3] = EST$Upper.CI
clda.res[,4] = EST$`Pr(>|t|)`

print(clda.res)

#ANCOVA
Y = c(ld$Y1, ld$Y2)
Y0 = rep(ld$Y0, 2)
Y0IT2 = c(rep(0, length(Y0)/2), ld$Y0)
ID2 = rep(ld$ID, 2)
GROUP = rep(ld$Group, 2)
TIME = c(rep(2,length(Y)/2),rep(3,length(Y)/2))
TMGR = c(rep(0,length(Y)))

for(i in 1:length(TMGR)){
  if(TIME[i] == 2 & GROUP[i] == 3){
    TMGR[i] = 1
  }else if(TIME[i] == 2 & GROUP[i] == 2){
    TMGR[i] = 3
  }else if(TIME[i] == 2 & GROUP[i] == 1){
    TMGR[i] = 5
  }else if(TIME[i] == 3 & GROUP[i] == 3){
    TMGR[i] = 2
  }else if(TIME[i] == 3 & GROUP[i] == 2){
    TMGR[i] = 4
  }else{
    TMGR[i] = 6
  }
}

Y = as.numeric(Y)
Y0 = as.numeric(Y0)
Y0IT2 = as.numeric(Y0IT2)
ID2 = as.factor(ID2)
TMGR = as.factor(TMGR)

ancova = data.frame(Y, Y0, Y0IT2, ID2, GROUP, TIME, TMGR)

model.ancova = lme(Y ~ Y0 + Y0IT2 + TMGR, random = ~1|ID2,
             data = ancova, correlation = corCompSymm(),
              na.action = na.exclude)

#Contrast matrix for ancova
CMAT <- matrix(0, 6, 8)
CMAT[1,] = c(0,0,0,-1,0,0,0,1)
CMAT[2,] = c(0,0,0,0,0,0,1,0)
```

```
CMAT[3,] = c(0,0,0,0,0,-1,0,1)
CMAT[4,] = c(0,0,0,0,-1,0,1,0)
CMAT[5,] = c(0,0,0,-1,0,1,0,0)
CMAT[6,] = c(0,0,0,0,1,0,0,0)
colnames(CMAT) <- c("(Intercept)", "Y0", "Y0IT2", "TMGR2",
                    "TMGR3", "TMGR4", "TMGR5", "TMGR6")

EST2 <- estimable(model.ancova, CMAT, conf.int = (1 - alpha))

#Results ANCOVA
ancova.res = matrix(0,6,4)
row.names(ancova.res) = c("HvsD_T2", "HvsD_T1", "HvsM_T2",
                          "HvsM_T1", "MvsD_T2", "MvsD_T1")
colnames(ancova.res) = c("Estimate", "lower_CI",
                         "Upper_CI", "p-value")

ancova.res[,1] = EST2$Estimate
ancova.res[,2] = EST2$Lower.CI
ancova.res[,3] = EST2$Upper.CI
ancova.res[,4] = EST2$`Pr(>|t|)`

print(ancova.res)
```

Heidi Elisabeth Sando

Statistical methods for the analysis of randomized controlled trials with repeated measurements and missing data

NTNU
Norwegian University of
Science and Technology