NTNU

Norwegian University of
Science and Technology

Analysis of Longitudinal Data with Missing Values.

Methods and Applications in Medical Statistics.

Ingrid Garli Dragset

Master of Science in Physics and Mathematics

Submission date: June 2009

Supervisor: Jo Eidsvik, MATH

Co-supervisor: Stian Lydersen, IKM

Norwegian University of Science and Technology
Department of Mathematical Sciences

Problem Description

In medical research, data sets are seldom complete. That is, some of the values that should have been recorded, are missing for some of the persons in the study.

Traditionally, "complete case" analysis has been used in such situations. That is, only cases (persons) with complete data are included in the analysis. In addition reducing the sample size, this method typically gives biased results. Better methods include multiple imputation (MI), full maximum likelihood (ML), mixed models, and generalized estimating equations (GEE).

The candidate shall give a short description of alternative methods for handling missing values. The main part of the thesis consists of analyses of longitudinal data from a research project at the Faculty of Medicine, using alternative methods.

Assignment given: 21. January 2009
Supervisor: Jo Eidsvik, MATH

Preface

This paper is the result of my master thesis in statistics (TMA4905), the final assignment of my five year study in mathematics and physics at the Norwegian University of Science and Technology (NTNU). The subject of this thesis is missing values in longitudinal data, and methods to analyze datasets with missing values. The thesis is a continuation of the project study in the autumn semester 2008 with same topic, and the specialization in Biometrics that I started during my exchange study period at KU Leuven in Belgium.

Working with this master thesis has broaden my knowledge about methods to handle missing data, particularly in longitudinal data. It has also broaden my view of statistical methods and models, and given me more insight to the statistical diversity and the amounts of situations where the standard analysis methods cannot be utilized. As I have continued my studies I have realized how much I still don't know about statistics. This makes me curious and eager to learn more.

First I want to thank Stian Lydersen for introducing me to the subject of missing data, and giving me the opportunity to work with applied statistical methods on real datasets. His inspiring and enthusiastic being and always accessible help has been a great support in writing this thesis. Kari Hanne Gjeilo has letting me analyze one of her datasets, and made this assignment possible. I want to thank Eirik Skogvoll for the assistance in the statistical software R, and Jo Eidsvik for accepting to be my administrative supervisor at the department of mathematical science. The Unit of Applied Clinical Research has provided me with considerable amounts of refreshing coffee and an inspiring working environment.

Collaboration with my fellow students has been important to me during the years of studies. This last semester with the individual thesis has therefore been more lonesome and demanding, especially when problems arise. I want to thank my boyfriend Audun Torp for suggesting to meet on a weekly basis and discuss our progress and problems concerning the thesis. He has also been a great support when the thesis has appeared too demanding or the days and nights of studying have made my mental health slightly unstable. His loving care and splendid sence of humour is of great importance to me.

I want to thank Sara Ghaderi for introducing me to some excellent literature on the subject of mixed models. I also want to thank my younger brother Øystein Garli Dragset for his everlasting care and brotherliness which have been helpful after late evenings at the office. The study resulting in this master thesis has occupied most of my time this last semester. I want to thank my family and friends for being patient and understanding, and given me the possibility to loose myself in this assignment.

Ingrid Garli Dragset
Trondheim, June 2009

Summary

Missing data is a concept used to describe the values that are, for some reason, not observed in datasets. Most standard analysis methods are not feasible for datasets with missing values. The methods handling missing data may result in biased and/or imprecise estimates if methods are not appropriate. It is therefore important to employ suitable methods when analyzing such data.

Cardiac surgery is a procedure suitable for patients suffering from different types of heart diseases. It is a physical and psychical demanding surgical operation for the patients, although the mortality rate is low. Health-related quality of life (HRQOL) is a popular and widespread measurement tool to monitor the overall situation of patients undergoing cardiac surgery, especially in elderly patients with naturally limited life expectancies [Gjeilo, 2009].

There has been a growing attention to possible differences between men and women with respect to HRQOL after cardiac surgery. The literature is not consistent regarding this topic. Gjeilo et al. [2008] studied HRQOL in patients before and after cardiac surgery with emphasis on differences between men and women. In the period from September 2004 to September 2005, 534 patients undergoing cardiac surgery at St Olavs Hospital were included in the study. HRQOL were measured by the self-reported questionnaires Short-Form 36 (SF-36) and the Brief Pain Inventory (BPI) before surgery and at six and twelve months follow-up. The SF-36 reflects health-related quality of life measuring eight conceptual domains of health [Loge and Kaasa, 1998]. Some of the patients have not responded to all questions, and there are missing values in the records for about 41% of the patients. Women have more missing values than men at all time points.

The statistical analyses performed in Gjeilo et al. [2008] employ the complete-case method, which is the most common method to handle missing data until recent years. The complete-case method discards all subjects with unobserved data prior to the analyses. It makes standard statistical analyses accessible and is the default method to handle missing data in several statistical software packages. The complete-case method gives correct estimates only if data are missing completely at random without any relation to other observed or unobserved measurements. This assumption is seldom met, and violations can result in incorrect estimates and decreased efficiency.

The focus of this paper is on improved methods to handle missing values in longitudinal data, that is observations of the same subjects at multiple occasions. Multiple imputation and imputation by expectation maximization are general methods that can be applied with many standard analysis methods and several missing data situations. Regression models can also give correct estimates and are available for longitudinal data. In this paper we present the theory of these approaches and application to the dataset introduced above. The results are compared to the complete-case analyses published in Gjeilo et al. [2008], and the methods are discussed with respect to their properties of handling missing values in this setting.

The data of patients undergoing cardiac surgery are analyzed in Gjeilo et al. [2008] with respect to gender differences at each of the measurement occasions; Presurgery, six months, and twelve months after the operation. This is done by a two-sample Student's *t*-test assuming unequal variances. All patients observed at the relevant occasion is included in the analyses. Repeated measures ANOVA are used to determine gender differences in the evolution of the

HRQOL-variables. Only patients with fully observed measurements at all three occasions are included in the ANOVA.

The methods of expectation maximization (EM) and multiple imputation (MI) are used to obtain plausible complete datasets including all patients. EM gives a single imputed dataset that can be analyzed similar to the complete-case analysis. MI gives multiple imputed datasets where all dataset must be analyzed separately and their estimates combined according to a technique called Rubin's rules. Results of both Student's *t*-tests and repeated measures ANOVA can be performed by these imputation methods.

The repeated measures ANOVA can be expressed as a regression equation that describes the HRQOL-score improvement in time and the variation between subjects. The mixed regression models (MRM) are known to model longitudinal data with non-responses, and can further be extended from the repeated measures ANOVA to fit data more sufficiently. Several MRM are fitted to the data of cardiac surgery patients to display their properties and advantages over ANOVA. These models are alternatives to the imputation analyses when the aim is to determine gender differences in improvement of HRQOL after surgery.

The imputation methods and mixed regression models are assumed to handle missing data in an adequate way, and gives similar analysis results for all methods. These results differ from the complete-case method results for some of the HRQOL-variables when examining the gender differences in improvement of HRQOL after surgery.

Contents

1 Introduction

The aim of this thesis is to study methods that handle missing values in longitudinal data. Longitudinal data come from repeated observations of a sample of units over multiple occasions, and are also denoted panel data, hierarchical or multilevel data, Such data have a correlated structure of observations measured on the same subjects that has to be taken into account when handling the missing values.

The focus of my project thesis last semester [Dragset et al., 2008] was methods to handle missing data, with an application of the method of multiple imputation (MI) on a cross-sectional dataset. Cross-sectional data are measurements observed once on each unit without relation to the time difference. The data were previously analyzed in Klepstad et al. [2003], with focus on scores from the Brief Pain Inventory (BPI). The variables of this questionnaires were scaled to intervals, which resulted in challenges when imputing on these variables. This is explained in Dragset et al. [2008]. The method of multiple imputation is further examined in this thesis, with emphasize on application for longitudinal data.

The longitudinal data studied in this paper are previously analyzed in Gjeilo et al. [2008], and consists of 534 patients undergoing cardiac surgery at St Olavs Hospital in Trondheim. Health-related quality of life (HRQOL) are measured by the self-reported Short-Form 36 (SF-36) and the Brief Pain Inventory (BPI) questionnaires. The patients are intended to be observed at baseline, that is before surgery, and then repeatedly at six and twelve months after surgery. These repeated measurements make us able to monitor the gender differences at each time point and the gender difference in evolution in time after surgery. The two statistical analyses employed to examine these differences are the Student's *t*-test for the differences at each time point and the repeated measures ANOVA to examine the differences in improvement pattern. Both these methods require balanced data, that is all subjects are measured at all time points. The opposite is denoted *unbalanced design*, and is characterized by an unequal number of measurements for the units in the study.

As much as 40% of the subjects included in the study have missing values for one or more measurement occasions. In Gjeilo et al. [2008] the complete-case analysis is used to obtain balanced and complete data, that is all units that have one or more missing values in their records are excluded from the analysis. This leads to deletion of considerable amounts of information

in the data. Further the results can be *biased*, that is skewed expectatins and erroneous standard deviations, and the statistical power are decreased.

Comprehensive research is performed on the field of missing data, and several methods are developed during the last decades. The alternatives of methods for analyses of longitudinal data are numerous, including imputation methods of expectation maximization and multiple imputation, mixed regression models, and marginal models estimating the correlation in data externally.

The focus on methods of multiple imputation introduced by Rubin [1976, 1987] and the expectation maximization algorithm by Dempster et al. [1977] are growing within most research communities exposed to missing data nowadays. Common statistical software implement these methods and make new features accessible. Especially the iterative algorithm of multiple imputation, denoted *MICE*, has been extensively used, and was first introduced by Schafer [1997]. Royston [2004, 2005] have implemented multiple imputation by MICE in Stata [2007].

Mixed regression models (MRM) are described under a variety of names, and some of the first to be published were variance components models [Dempster et al., 1981], random-effects models [Laird and Ware, 1982] and random regression models [Bock, 1983]. The correlation structure of the data are fitted by random effects, for example the random intercepts and slopes effects. These models are similar to the covariance pattern models (CPM) introduced by Jennrich and Schluchter [1986]. CPM reflect the covariance in data through estimation of a specified correlation structure. A third type of regression models for longitudinal data were introduced during the 80's, denoted generalized estimating equations (GEE) [Liang and Zeger, 1986, Zeger and Liang, 1986, Zeger et al., 1988]. GEE are a quasi-likelihood estimating processes that are able to give unbiased results under some missing data assumptions, but are not as general as imputation by EM and MI, or the mixed regression models and covariance pattern models. The attractiveness of this method is rather focused on its ability to analyze both discrete and continuous outcome variables.

Recent work on the regression models introduced above are found in Hedeker and Gibbons [2006], Fitzmaurice et al. [2009] and Diggle et al. [2002], and a more applied setting for Stata users are given in Rabe-Hesketh and Skrondal [2008]. We observe that there are a set of commands specially designed to analyze longitudinal datasets, and these are labeled with the prefix `xt-`. Some of these commands will be described through this paper.

The structure of this paper is based on the methods described above. Chapter 2 introduces concepts and definitions of missing data and examples of situations where missing values typically occur. The complete-case method and other deficient methods are presented in Section 2.2, followed by the imputation methods and more complex methods. An overview of methods to handle longitudinal data are given in Chapter 3, including a description of repeated measures ANOVA, mixed regression models, covariance pattern models and methods for discrete outcome variables. The dataset analyzed in Gjeilo et al. [2008] are described in Chapter 4 with emphasize on the missing data structure and the observed correlation structure. The application of the complete-case method, imputation methods and the regression models are explained in Chapters 5, 6 and 7, including results of the respective methods. Discussion of methods and results are given in Chapter 8, followed by a conclusion that summarizes the most important results and features of the methods.

2 Theory

To describe methods that handle missing values in dataset and to evaluate the properties of these methods we introduce some definitions and concepts that are helpful throughout the report.

2.1 Missing data

Day [1999] defines missing data as follows

> ”Missing data refers to a data value that should have been recorded but, for some reason, was not”.

Missing data are also referred to as non-response or unobserved data, and occur in most types of studies. Missing values can occur due to failure of measurement, data loss, out-of-range data and data loading issues, units fail to answer all questions, loss of follow-up or other plausible reasons. It is important to handle the problem of missing data in a proper way to obtain unbiased results that can be used in research. We can characterize the missing values based on occurrence in time, for example do intermittent missing values correspond to unobserved values at some time points where the subject is measured at occasions both before and after the missing occasion. This may be due to panel waves (subjects missing the whole occasion) or subjects omitting the specific item at that time point. Another type of missing values is dropout that occur when subjects terminates all further observation through the study. This may happen when subjects relocate, decides to end the study, die or just due to random reasons.

The issue of missing data is widespread within the field of clinical trials, especially when the study is of longitudinal structure. The objects that are measured are denoted units or subjects. These subjects can be humans, which is often the case in clinical studies, but may as well be animals, plants or groups of individuals, to mention some examples.

2.1.1 Notation

The notation of missing data were introduced by Rubin [1976] and is still in use as the common notation. The data that are planned to be observed are denoted X_i for the covariate matrix and

Y_i for the dependent variable vector, both for subject i and intentionally observed for n time points. Further we can partition the dependent variable vector Y_i in the observed variable vector Y_i^O and the missing variable vector Y_i^M, thus we get $Y_i = (Y_i^O, Y_i^M)$.

R is the missing data distribution that describes the occurrence of missing data in the dataset. R_{ij} is a dichotomous variable (takes two levels of values), equal to one if the value of the dependent variable is observed for subject i at time point j, and zero if Y_{ij} is missing. This missing data distribution is important to describe the reasons for missing data, the *missing data mechanism*. There are three main missing data mechanisms described in the literature, these are denoted MCAR, MAR and MNAR. These concepts are explained slightly different by various authors, the notation described here are based on Hedeker and Gibbons [2006].

2.1.2 MCAR

Missing completely at random (MCAR) is the assumption that missing data occur totally random without any relation to the other observed or unobserved data. This is the most basic missing data mechanism and assumes missing data to occur for completely random reasons. The distribution of missing values R are thus assumed to be independent of both covariates and the dependent variable as

$$\mathrm{P}(R|Y, X) = \mathrm{P}(R). \tag{2.1}$$

Figure 2.1 displays graphically the principle of MCAR.

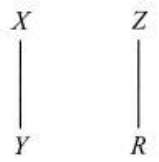

Figure 2.1: Graphical display of the missing data mechanism MCAR. Y represents the dependent variable vector (Y^O,Y^M), X are observed covariate variables, Z are variables representing reasons for missing data and R is the missing data distribution [Schafer and Graham, 2002].

A less stringent case of MCAR is the *covariate-dependent MCAR*. This missing data mechanism is dependent on the fully observed covariates X_i, expressed as

$$\mathrm{P}(R|Y, X) = \mathrm{P}(R|X). \tag{2.2}$$

This special case of MCAR allows the fraction of missing data to vary across variables. An example of a situation where the covariate-dependent MCAR is more appropriate than the strict MCAR mechanism is when the fraction of missing data varies in time. It is important to include predictors for the missing data in the given analysis. In the above example the time variable must be included in the covariate matrix of the analysis to yield unbiased results.

Another example of a covariate-dependent MCAR mechanism is found in a study on patients aged 18 to 70, where all the patients are asked to answer a questionnaire. The elder patients may

have higher proportions of missing data. This can be related to their age, not necessarily by the values of the unobserved data. Again, it is important to include the variable age as a covariate since the missing data may be explained by this variable.

A graphical presentation of covariate-dependent MCAR is given in Figure 2.2.

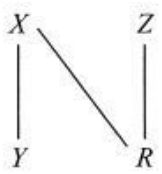

Figure 2.2: Graphical display of the missing data mechanism covariate-dependent MCAR. Y represents the dependent variable vector (Y^O,Y^M), X are observed covariate variables, Z are variables representing reasons for missing data and R is the missing data distribution [Schafer and Graham, 2002].

2.1.3 MAR

MAR is the short form for *missing at random* and describes how the missing data distribution R is dependent on both observed covariates X_i and the observed dependent variable vector Y_i^O, but not on the unobserved variable vector Y_i^M. This is expressed as

$$\mathrm{P}(R|Y, X) = \mathrm{P}(R|Y^O, X). \tag{2.3}$$

We illustrate the missing data mechanism by a simple example. Missing values due to subject dropout is a common issue in longitudinal data, and occur when subjects that have entered a study do not respond at a given time or any subsequent time points. This kind of missing values may be related to the observations measured prior to dropout. Thus the missing data distribution is assumed to be related to the observed dependent variable vector Y_i^O of the dropout subjects. This can be found for example in a study of mental health, where the dropout subjects have lower scores than the remaining subjects.

The observed values of the dependent variable are related to the missing values through the correlation structure of the data. In the previous section we described covariate-dependent MCAR, and emphasized the importance of including covariates in the analysis model that explain possible reasons of missing values, and this is equally important for the MAR mechanism. In addition it is essential to specify the correct variance-covariance matrix of the distribution of the data. If these factors are not proper implemented, the analyses may not necessarily be consistent with the underlying MAR mechanism and we can get biased estimates although the method are expected to give unbiased results.

An expression often used in the literature of missing data is *ignorable missing*. Ignorable missing is a collective term that requires the missing data mechanisms MCAR and MAR, and in addition the parameters of the data model and for the missing data mechanism must be distinct. This means that data can be analyzed without respect to the missing data distribution. The

opposite situation of *non-ignorable missing* refers to the missing data mechanism explained below, and must be handled with caution.

2.1.4 MNAR

Missing not at random (MNAR) is the missing data mechanism where missing values are assumed to be related to the unobserved dependent variable vector Y_i^M in addition to the remaining observed values. This is expressed as

$$\mathrm{P}(R|Y,X) = \mathrm{P}(R|Y^M, Y^O, X). \tag{2.4}$$

An example of the MNAR mechanism is when special values of a dependent variable lead to non-response. When investigating smoking habits we may experience that missing data rely on the actual value of the unobserved measurements, the subjects that are heavy smokers often omit the questions about smoking while non-smokers gladly fulfill these questions.

There are no ways to confirm or reject MAR versus MNAR, since the unobserved dependent variable vector Y_i^M is involved. There are developed methods that can handle data with missing data due to the unobserved values and some of these are presented in Section 2.4, and MNAR missing data is a topic of intense and emerging research [Hedeker and Gibbons, 2006].

2.1.5 Missing data patterns

In the literature different missing-data patterns are presented, and these are important for the types of methods that are recommended used for each specific dataset. Some of the single imputation methods explained in Schafer and Graham [2002] rely on missing data to occur only as dropout, while the method of expectation maximization described in Section 2.3.1 handle complex missing data patterns. Some frequently used examples of missing-data patterns are the univariate, monotone and arbitrary patterns displayed in Figure 2.3. The univariate missing data pattern can be described as one single variable containing unobserved values due to dropout, while the monotone non-response pattern includes several variables with missing values. These variables can be sorted in a way that consecutive variables have equal or more missing values. In a longitudinal study with units lost to follow-up, a monotone missing-data pattern may occur. Arbitrary missing data patterns are more complex allowing missing values in all variables and at all time points.

2.1.6 Criteria for methods to handle missing data

Appropriate methods to handle an analysis of data with missing values are characterized by three criteria to yield valid inferences about the data. The first two criteria are given by Rässler et al. [2008], while the criterion of consistency (3) are presented in Carpenter and Kenward [2007].

1. Estimates of coefficients of the missing data analysis must be approximately equal to the population estimates (that includes for example means and variance estimates).

2. The confidence intervals for estimates (with a significance level α) should have the property of covering the true population mean at least (1-α)$\times$100% of the time.

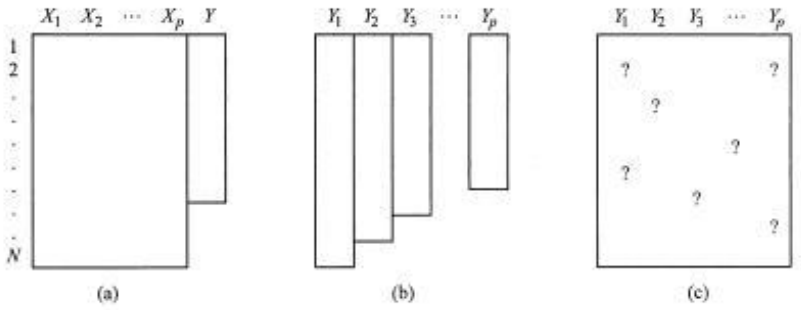

Figure 2.3: Graphical display of missing data patterns, (a) univariate pattern, (b) monotone pattern and (c) arbitrary pattern. Rows correspond to units and columns correspond to variables, X's are completely observed variables and Y's are partially observed variables [Schafer and Graham, 2002].

3. The methods must lead to consistency, which means that the confidence intervals are expected to be narrower when the number of units increases in the data set.

The methods described in Section 2.2 do not meet the criteria above, but, if data are assumed to be ignorable, the methods in the subsequent Sections 2.3, 3.4 and 3.5 are stated as proper methods to handle missing data analyses.

2.2 Simple but deficient methods

There are two main categories of deficient methods explained in Schafer and Graham [2002]. The first is those excluding units with missing values, denoted *complete-case analysis* and explained in the following section. The second category is those replacing missing values with an alternative best guess. The latter is denoted single imputation methods and are described in Section 2.2.2. Note: There exists a single imputation method that meets the criteria of methods to handle ignorable missing data outlined in Section 2.1.6. This method will be explained in Section 2.3.1.

2.2.1 Complete-case analysis

One of the more widespread and straight forward ways to handle missing data is the *complete-case analysis*. Other names for this method are *listwise deletion* and *case deletion*. It excludes all units in the dataset with one or more unobserved values. This leads to a complete dataset that consists of units with completely observed variables. Many of the statistical software packages (for example SPSS and Stata) employ this as default, which, in addition to being simple makes this method well-known and often applied.

The complete-case analysis requires data to be strict MCAR to be valid, that is the missing values are independent of all the covariates and observations at other measurement occasions.

This is a rather strong assumption, especially for longitudinal data, and are seldom met. Violation of this assumption can typically lead to selection bias. For example, consider a group of patients that participates to a questionnaire prior to a medical treatment. The same group of patients are requested to answer the questionnaire again after this treatment. Suppose that only the patients that still fell ill return the questionnaire, while the cured patients don't. The results will be too pessimistic for these patients due to selection bias. The opposite situation is also possible, and will give more uplifting results than realistic. The degree of bias depends on the extent of violation of the assumption of MCAR, the proportion of excluded patients and the analysis being implemented [Rässler et al., 2008].

Even if the data can satisfy the MCAR assumptions, there are several problems arising. A small percentage of missing values can lead to a considerable proportion of excluded units, even when data are MCAR. An example is found in a dataset where a high number of variables are measured, and many of the units have a few missing values. The percentage of missing values will be low, but a considerable number of units with non-observed variables may be excluded. This decreases the amount of data to be analyzed and also the power of the statistical results. The number of units is limited in most studies due to economic, ethic or other reasons, and with the complete-case analyses these samples are further confined. The complete-case method might be inefficient, even in case of MCAR, since the number of units decrease. In multivariate cases this gets even worse. The degree of inefficiency depends strongly on the fraction of excluded units.

In studies where data can be assumed MCAR and only a small proportion of units are excluded, this method can be a sensible choice. It is easy to implement and makes the way forward to the real analysis short. But it should be handled with caution, if the small group of deleted units keeps a large proportion of one feature, the analysis can still be biased.

A special case of the complete case method is the *available case analysis*, sometimes referred to as *pair-wise deletion* or *pair-wise inclusion*. This method deals with one analysis at the time, and includes all units where all the variables are observed. For a longitudinal study where each of the measurement occasions are examined separately, only those units with unobserved values at that single time point are excluded from the analysis. As long as no comparison of the analyses at the different measurement occasions are performed this is a legal analysis method, and leads to a higher number of included units in each analysis. As stated above, the degree of bias and efficiency depends on the proportion of excluded units, thus this method can reduce some of the bias and lead to higher efficiency. However the issue of bias and inefficiency is still present, and an additional disadvantage is that different groups of units contribute to the analysis at different time points depending on the missing-data pattern.

2.2.2 Single imputation

In a statistical setting the expression *imputation* is applied with replacement of missing values by values estimated from covariates or other observed values [Rosner, 2006]. The estimation function can be based on available information from the observed data of the unit itself and from other units with (similar) observed values. No units are excluded from the analysis, thus the original number of included units is maintained at all time points. There are many ways to estimate these imputations and some of these are described in Schafer and Graham [2002], including hot deck imputation, marginal mean imputation, conditional mean imputation and

conditional distribution imputation. A special case of single imputation is the hierarchical scales imputation, outlined in Fayers and Machin [2007].

The extent of the bias of results are different for the above mentioned single imputation methods. Some of these imputation methods may give approximately unbiased estimates. The underestimation of variance in the estimates is common for all the single imputation methods, and is also the reason why these methods are unsuitable for data that are not missing completely at random (MCAR). Expectation maximization algorithm and multiple imputation are both extensions of the single imputation methods, and are more appropriate for data with missing values since they model the variation in a better way. This is explained further in Section 2.3.

An imputation model is a useful tool to make imputations for unobserved values in a dataset. It describes the conditional distribution of the missing values dependent on other variables in the dataset, and also possibly on other variables important for the missing-data structure. The imputation model must not be confused with the analysis model, that is used to examine the dataset, for example to do inference about a sample. All variables with missing values and variables that can predict or explain the missing values are included in the imputation model, and can be thought of as multivariate responses. This way of structurizing the variables makes the imputation model a device to preserve important features of the joint distribution of non-observed variables. The analysis model on the other hand, makes distinctions between dependent and independent variables, and keeps up this relationship of variables throughout the analysis.

Last observation carried forward

The method of *last observation carried forward* imputes values for unobserved data based on former observed values, thus it is one of the single imputation methods described above. This method applies especially to longitudinal data where the units are observed at several occasions, and some units are lost-to-follow up or have intermittent missing values. It imputes values equal to the last observed response for the variable for each unit. This method gives potentially bias for all types of plausible missing-data mechanisms, and is almost never good fitted to the expected behavior. It should practically never be used according to Carpenter and Kenward [2007].

2.3 Imputation methods

In the section above we have described the missing data mechanisms MCAR (including dependent MCAR), MAR and MNAR, and the properties of these mechanisms. Both MCAR and MAR mechanisms are referred to as *ignorable missing*, and can be analyzed without loss of information. It must be emphasized that this relies on the use of analysis methods based on the full likelihood function [Diggle et al., 2002]. The expectation maximization (EM) algorithm and multiple imputation (MI) are imputation methods that provide unbiased parameter estimates and their belonging standard errors as required by the criteria in Section 2.1.6.

2.3.1 Expectation maximization

Expectation maximization (EM) is a method of obtaining estimates of coefficients from an analysis on base of Bayesian thinking, introduced by Dempster et al. [1977]. It applies maximum

likelihood estimation (ML) to draw population inferences based on observed values from a full likelihood function. Schafer and Graham [2002] describes the EM algorithm as follows:

"The key idea of EM is to solve a difficult incomplete-data estimation problem by iteratively solving an easier complete-data problem."

For monotone missing data patterns (results of dropout, see Figure 2.3) the full-distributional likelihood can be expressed in closed form, and the log-likelihood function may be expressed as a sum of functions, each function dependent of one parameter only. Maximum likelihood (ML) estimates are calculated by maximizing the log-likelihood function with respect to each parameter separately. For some models the full likelihood function exists, but the log-likelihood does not represent the parameters distinctly and thus maximizing the factors separately will not necessarily maximize the likelihood. Especially for complex missing data structures like the unstructured pattern in Figure 2.3 the direct calculation of ML estimates is not feasible. Iterative methods to approximate the full likelihood function have been developed for situations where explicit calculation of ML estimates are not available [Diggle et al., 2002].

The expectation maximization algorithm is a general iterative algorithm for ML estimation in incomplete-data problems. It applies the observed-data likelihood, also referred as the likelihood ignoring the missing data mechanism [Schafer and Graham, 2002].

The estimated parameter values $\hat{\theta}$ that maximizes the observed-data likelihood function have attractive properties. The estimated parameters tends to be approximately unbiased in large samples, and the estimated variances obtained are close to what is theoretically desirable. Thus the EM algorithm fulfill the criteria of methods handling data with missing values. The method assumes a large number of data so that the EM estimates can be approximately unbiased and normally distributed. In addition it assumes data to be ignorable, that is MCAR or MAR mechanism. The procedure of the EM algorithm builds on a relatively old *ad hoc* idea for handling missing data Diggle et al. [2002]. First we replace the missing values by initial values (for example estimated values). Then iteration over the following steps are carried out.

1. Estimate parameters of the full likelihood model by maximum likelihood estimation.
2. Re-estimate the "missing values" assuming the new estimated parameters from second step is correct.

Iteration must be continued until convergence, that is until values that are re-estimated by the second step approximate the previous estimated values. The notation "missing values" is used to separate the EM algorithm from the *ad hoc* idea. In EM the "missing data" are not the actual missing data Y_{mis}, rather the functions of Y_{mis} appearing in the complete-data likelihood. The second step above is called the *E step*, which calculates the conditional expectation of the "missing data" given the observed data and the estimated parameters $\hat{\theta}$. The first step is denoted the *M step* and estimates the new set of parameters from maximization of the observed complete-data log-likelihood. The algorithm can be quite easily implemented, and each iteration consists of one run of the E step and one of the M step. EM can be shown to converge reliably if the observed-data likelihood function is bounded [Diggle et al., 2002].

One drawback for the expectation maximization algorithm is that with large fractions of missing data the convergence of the iterations can be very slow. In some problems the compu-

tation of the M step can be difficult because no closed form of the likelihood function exists. Diggle et al. [2002] presents some extended algorithms of EM that gives solutions for the problems above. Some statistical analysis software available for EM estimation are described in Acock [2005] and Schafer and Graham [2002].

Another imputation method based upon the likelihood function is multiple imputation described in the following section.

2.3.2 Multiple imputation

The method of multiple imputation (MI) is a continuation of the method of single imputation from a conditional distribution [Schafer and Graham, 2002]. It retains much of the attractiveness of the single imputation method and gives unbiased results with respect to the estimated variance. MI produces m plausible datasets where each could have been the complete dataset if all values were observed. The completed datasets can be combined by easy arithmetic [Rubin, 1987] to obtain estimates and standard errors that reflect uncertainty in the missing-data and the finite-sample variation. This method makes use of complete-data techniques and software already possible which is very favorable.

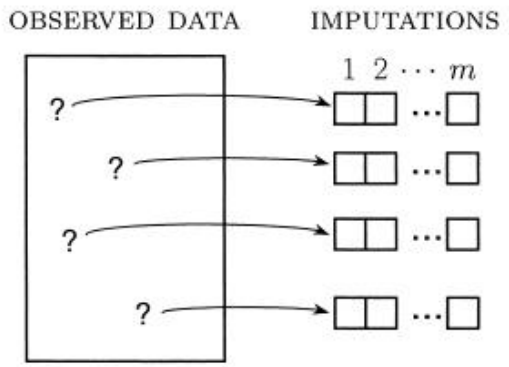

Figure 2.4: Schematic representation of multiple imputation with m imputed datasets [Schafer and Graham, 2002].

Multiple imputation is based on a Bayesian way of thinking, in the sense that the distribution in which the imputations are drawn from is a full-conditional posterior distribution. To perform MI successfully, the imputations need to be proper according to the criteria in Section 2.1.6. This means that the uncertainty about the parameters in the imputation model must be taken into account when imputing unobserved values. Both the missing data and parameters of the imputation model have distributions, and this feature is important in the MI setting.

2.3.3 Why use MI?

To perform proper multiple imputation we need a prior distribution for the parameters of the imputation model and a likelihood function for the variables that have unobserved values. If a large sample is obtained, the likelihood function can be approximated by the imputation model

as done by EM algorithms, and the prior distribution for the parameters can be uninformative. Further a posterior distribution $[\theta, y_{mis}|y_{obs}]$ must be obtained, where θ represents the sampled parameters of the imputation model drawn from the prior distribution. The sampled values of y_{mis} are used as imputations in the dataset. This sampling is done *m* times to obtain *m* imputed datasets. In this procedure the parameters of the imputation model are drawn for each imputation, so these parameters are unique for each imputation. This way to treat the model parameters as stochastic variables with uncertainty and not as fixed values (as in prediction) is the important property of MI that ensures the variance estimates to be unbiased. The total variance, that consists of within-imputation and between-imputation variance is therefore approximately unbiased for data assumed MCAR or MAR. The uncertainty is represented in both the parameters of the imputation model and the random sampling from the imputation model to draw imputations for the missing values. The EM algorithm on the other hand, estimates the parameters of the analysis model only once, thus they are considered as fixed parameters.

2.3.4 Imputation model

As for single imputation from a conditional distribution we must create an imputation model to obtain the imputations. The imputation model must be *at least as rich* as the analysis models. This is to ensure that all variables that may contain information about other variables in the analysis model are represented. In addition some additional variables may be included in the imputation model, according to Buuren et al. [1999]. These variables are divided in two groups, variables that are known to have influenced the missing-data pattern, called U-variables, and variables that explain much of the variance in the unobserved variables, denoted V-variables. Last step in the determination of the imputation model is to exclude those of the U- and V-variables that have too many unobserved values within the group of incomplete cases.

2.3.5 Properties

In Schafer and Graham [2002] an example of MI is presented, with a bivariate normal distribution as base for the full likelihood function. MI does not require all the variables to be normally distributed, but this assumption is used in many publications about MI. This is because the multivariate normal distribution is one of the easiest approach to the method. MI shares some of its properties with the EM algorithm, for example do both rely on large-sample approximations (though this assumption is stronger for the EM algorithms). MI and EM both use the observed-data likelihood function to approximate the likelihood function. MI implies uncertainty in the imputed values through both the parameters of the imputation model and the drawn imputation values themselves. This stands in contrast to EM that computes the joint likelihood function with fixed parameters, and imply uncertainty only in the imputed draws from this imputation model.

Multiple imputation gives unbiased estimates and variance under MCAR and MAR, and satisfies the criteria for methods to handle missing data. Even if data are MNAR the method of MI is assumed to be a better method than the above described simple methods [Rässler et al., 2008]. Software available to perform MI is listed in Schafer and Graham [2002] and Acock [2005].

Rubin's rules

Rubin's rules Rubin [1987] is a method to combine the results from *m* imputed datasets for a scalar parameter. For notation here we use Q as the parameter quantity (for example a regression coefficient), $\hat{Q}$ is the estimated value for Q and $\sqrt{U}$ denotes the estimated standard error if the original data were complete. This method assumes a large enough sample to ensure that $\hat{Q}$ is approximately normal distributed with mean Q and variance U. There are one estimated $\hat{Q}$ and U for each of the *m* datasets, thus we have the notation [$Q^{(j)}$,$U^{(j)}$], where $j = 1, 2, \cdots, m$.

The overall estimate is expressed as

$$\bar{Q} = m^{-1} \sum_{j=1}^{m} \hat{Q}^{(j)} \tag{2.5}$$

The uncertainty in $\hat{Q}$ has two parts, within-imputation variance

$$\bar{U} = m^{-1} \sum_{j=1}^{m} U^{(j)}$$

and between-imputation variance

$$B = (m-1)^{-1} \sum_{j=1}^{m} \left[\hat{Q}^{(j)} - \bar{Q}\right]^2$$

The total variance is a sum of the two parts of variance

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B$$

Now we have an estimated combined parameter $\bar{Q}$, with appurtenant standard error $\sqrt{T}$ which is the unbiased estimator for the dataset if no values were missing. To obtain p-values, Rubin recommended using an approximation to the Student's *t* distribution with ν degrees of freedom

$$\frac{\bar{Q} - Q}{\sqrt{T}} \sim t_\nu$$

$$\nu = (m-1) \left[1 + \frac{1}{\left(1 + \frac{\bar{U}}{m}\right) B}\right]^2$$

When the degrees of freedom ν is large, total variance is well estimated and thus the number of imputed datasets *m* is large enough.

The number of imputed datasets required varies from dataset to dataset, dependent on the quality of the imputation model and the structure in the data. Quantities like 5, 10, up to 20 and 50 imputations are advised, and this issue is well discussed in the literature [Schafer and Graham, 2002]. In recent years the recommended number of imputed datasets *m* has increased

due to new methods to estimate the error in results because of a finite number of imputations. As a rule of thumb the number of imputations should be raised if p-values are to be calculated. The time to make 10 versus 20 imputed datasets is seldom of large consequence compared to the effort and time it takes to establish the imputation model.

2.4 Full model-based methods assuming MNAR

All of the methods described above assume the missing data mechanism to be MCAR or MAR, but in certain situations the assumption that data are MNAR may be plausible. To handle this the missing data distribution R must be taken into account when unobserved values are to be imputed. Especially is MNAR potentially the issue in clinical studies, where the patients may have a high fraction of unobserved values, or possibly drop-out, closely related to the values that are not observed. Schafer and Graham [2002] presents two main types of model-based methods that assumes data to be MNAR, that is the selection models and the pattern-mixture models, both described below. Hedeker and Gibbons [2006] state that several researchers warn against reliance of a single MNAR model, because the assumptions about the missing data are impossible to access with the observed data. Thus we should use MNAR models with caution, and perhaps examine several models to conduct a sensitivity analysis of the missing data.

2.4.1 Selection models

A selection model consists of a distribution for the complete data, and a distribution of the missing-data given the data itself. This means that a joint distribution of the complete data *Y* and the missing-data distribution *R* as

$$\mathrm{P}(Y, R|\theta, \xi) = \mathrm{P}(Y|\theta) \cdot \mathrm{P}(R|Y, \xi)$$

θ is the unknown parameters of the complete-data population, while ξ is the parameters of the conditional distribution of the missing-data distribution given the complete data. The likelihood function may be calculated by integrating the equation above over all values of the missing-data Y_{mis}. The maximum of the likelihood is not always possible to find analytically, so iterative methods are needed to approximate $\hat{\theta}$'s. Another issue with the selection models is that they are extremely sensitive to the distributional shape that is chosen for the population.

2.4.2 Pattern-mixture models

The other type of model-based models presented in Schafer and Graham [2002] is the pattern-mixture models, which groups the whole sample on basis of the missing-data distribution. A model of this form may be written as

$$\mathrm{P}(Y, R|\theta, \xi) = \mathrm{P}(R|\eta) \cdot \mathrm{P}(Y|R, \nu)$$

θ and ξ have the same interpretation as above, η represents the proportions of the population that end up in each of the missing-data groups and ν is the parameters of the complete-data

distribution given the missing-data group. Pattern-mixture models are suffering from the fact that it is difficult to estimate ν for the groups containing unobserved values, so heavy restrictions or unverifiable assumptions must be made. On the other hand, these models are not as sensitive to the distribution of the population as the above described selection models. Development and incorporation of full model-based methods into suitable software are necessary to be able to apply these methods for scientific research. Schafer and Graham [2002] recommend the use of MI and EM algorithms for handling missing-data problems until the appropriate software is accessible.

3 Longitudinal data analysis

Longitudinal studies have both advantages and disadvantages over cross-sectional studies. First, a longitudinal study can give more powerful results with the same number of subjects compared with a cross-sectional study. Another way to formulate this is that longitudinal studies need fewer subjects included in the study to obtain the same statistical power as a cross-sectional study. Second, in a longitudinal study each subject is measured more than once, and can serve as his/her own control. The within-subject variability is often smaller than the variance between subjects, and the between-subject variability can be separated from the measurement error. This results in more efficient estimators of treatment-related effects compared with cross-sectional studies. Third, a longitudinal study can measure individual trends, or evolution in time, of dependent variables. This determines growth and changes at individual level, which is not possible in cross-sectional studies. Finally, these kind of studies allow us to separate the effect of changes over time within subjects from the differences between subjects at baseline.

Observations from a repeated measures study are assumed clustered dependent since more observations of the same variable are measured on each subject. This leads to a need of more sophisticated statistical methods than for ordinary cross-sectional studies. Very common issues of longitudinal studies are dropout or intermittent missing values at one or more occasions during the study. Simple statistical analysis methods do not handle this problem without inducing possibly biased results. One solution of this is to impute the missing values prior to statistical analysis. Other methods like mixed regression models and covariance pattern models use all available data in the analyses, which increases the statistical power and prevents biased estimates.

3.1 Analysis considerations

When modeling longitudinal data there are several properties that must be considered. Quite simple statistical analyses are available for continuous normally distributed outcome variables in data without missing values, for example the repeated measures ANOVA. To employ large sample theory to approximate non-normally distributed variables, the number of subjects m

allocated in each group should exceed 50. Also the number of observations per subject n_i should be considered. If each subject is measured twice, a simple change score may be calculated. An example of a change score is to calculate the difference in observations for each subject. This makes analysis methods for cross-sectional data suitable, such as the analysis of covariance (ANCOVA). This approach relies on balanced design of data and is not very suitable for data with missing values. For datasets where the number of observations varies between subjects, more general methods are required, for example mixed regression models (MRM).

For one-sample analyses, the only terms needed in the analysis model is the random factors. No between-subject covariates is required since all subjects are assumed to be randomly sampled from the same population. For two-sample situations the variables that describe the differences between groups must be included in the model. The terms used above are explained in the following sections. The last thing to consider for longitudinal data modeling is the variance/covariance structure of the panel data. This is further explained in the following sections.

3.2 General approaches

A wide range of methods to model longitudinal data are available. The following methods are discussed in Hedeker and Gibbons [2006]. The simplest method, also denoted as the *derived variable* approach, implies combination of the repeated measurements into one summary variable. This causes a longitudinal study to be simplified to a cross-sectional study since each subject obtain one single combined measurement. A derived variable may for example be an 'averaged over time'-variable, 'changed score'-variable, 'area under a curve'-score or 'linear trend across time'-variable. A problem with this methods is the strong dependency of balanced design, which makes it unsuitable for data with missing values.

The second approach to longitudinal data analyses are the univariate repeated measures analysis of variance (ANOVA). This method requires several assumptions to be met and is therefore limited in its application but is relatively easy to compute and implement. The most critical assumption for longitudinal data, besides the assumption of balanced data, is sphericity. Sphericity is defined in Rabe-Hesketh and Skrondal [2008] as the assumption that all pairwise differences between responses have the same variance. This is rarely met when analyzing longitudinal data, and violations can lead to skewed F-distributions. ANOVA takes into account that subjects can have individual baseline observations but no subject-specific evolution in time.

Multivariate ANOVA (MANOVA) is an approach to analyze longitudinal datasets but holds similar restrictions as the univariate ANOVA described above. It does not handle missing values in data and in addition it assumes the variables to be measured at the same occasions. It is therefore not suitable for longitudinal datasets with non-responses.

The fourth approach is the mixed regression models (MRM) that can be used for both categorical and continuous, and non-normally and normally distributed outcome variables. MRM give unbiased results if data are assumed ignorable (MCAR and MAR), and allow the measurement occasions to vary among the units. The method handle both time-invariant and time-varying variables and is therefore a suitable method to analyze longitudinal data with non-responses. In this report the term *mixed regression models* refers to the linear mixed regression models that assumes dependent variables to be continuous, if not specified otherwise. The

generalized linear mixed regression models are presented in Section 3.6.1.

The method of generalized estimating equations (GEE) is an alternative to the generalized mixed regression models in the sense that both methods handle some missing values, take time-varying covariates and different types of outcome variables. The drawback for this method is the more strict requirement for missing values to be ignorable, missing data are assumed to be explained by covariates in the model and not by observed values at other measurement occasions.

3.3 Repeated measures ANOVA

In longitudinal datasets some observations are dependent of other observations since more than one observation may be measured on the same subject. This must be taken into account when analyzing such types of data. Repeated measures ANOVA is a method to examine possible differences in means, that is between two or more groups, and from one, two or several samples. The method models the correlation structure of observations within subjects by the inclusion of a random subjects effect in the regression formulation. This random subject effect allows subjects to have individual baseline observations, thus we can partition the variance in explained variance by the random effect and unexplained variance, also known as the residual error.

3.3.1 Assumptions of repeated measures ANOVA

As stated above the method of repeated measures ANOVA have many assumptions that must be met to obtain unbiased estimates from the analysis. All subjects must be measured at the same fixed time points. The data are assumed complete and balanced in terms of time points n, but the size of the groups h may be unequal. Thus data with missing values must be handled by the complete-case method or one of the imputation methods prior to the analysis. The dependent variables are assumed multivariate normally distributed, and the variance-covariance matrix of the data with respect to the occasions is assumed equal to the compound symmetry structure. The latter consists of two layers, that is the assumption of homoskedasticity and sphericity. Homoskedasticity is defined as the variance of the dependent variable to be equal at all measurement occasions, while sphericity refers to the assumption that all pairwise differences between responses have the same variance, as stated above. This implies that consecutive observations are equally correlated as observations more distant in time, and that the variance at all occasions are equal. The assumption of compound symmetry is seldom met for longitudinal data. Rabe-Hesketh and Skrondal [2008] write that a less strict assumption of sphericity alone is sufficient for the repeated measures ANOVA to be valid.

Repeated measures ANOVA claims to be fairly robust to violations of normality and homoskedasticity. If sphericity cannot be assumed, the Greenhouse-Geisser or Huynh-Feldt correction methods can be applied to adjust the degrees of freedom for the F-test (explained in Section 3.3.2). The idea of these correction methods is to decrease the degree of freedom because more parameters are necessary to model the covariance matrix.

In the next section some notation is introduced to further explain the method of repeated measures ANOVA. This notation is also used with mixed regression models, covariance pattern models and generalized estimating equations in subsequent sections.

Notation

All sample members in the longitudinal study are denoted subjects. These subjects are the units measured during the study, and can be patients treated at a hospital, football players on a team or plants on a field as examples. The dependent variable measured repeatedly on each subject is called the within-subjects factor. Examples of such within-subjects effects are outcome of quality of life-questionnaires, type of treatment for patients or quantity of irrigation for plants. Variables measured on independent groups of sample members are called between-subject variables, and examples are gender of subjects, the football team, the area of which the plant grows and so on.

When performing repeated measures ANOVA we have to calculate sum of squares for both fixed and random terms in our analysis model. First we look at the analysis model expressed as

$$y_{hij} = \beta_0 + \beta_1 t_j + \beta_2 x_h + \beta_3 x_h t_j + v_{0i} + \epsilon_{hij}, \tag{3.1}$$

where h is the number of groups, $h = 1, 2, \cdots, g$, i is the number of subjects in group h, $i = 1, 2, \cdots, N_h$, and j is the number of measurement occasions, $j = 1, 2, \cdots, n$. An indicator variable must be created for each of the groups except the chosen baseline category. There are $\mathrm{N} = \sum_{h=1}^{g} N_h$ subjects included in the dataset. The measurement occasions are assumed equal for all subjects, and we also assume balanced design. y_{hij} is the outcome variable, assumed normally distributed and continuous. The rest of the variables can be interpreted as follows:

$$\begin{aligned}
\beta_0 &= \text{grand mean},\\
\beta_1 &= \text{effect of time } t_j,\\
t_j &= \text{time variable},\\
\beta_2 &= \text{effect of group } x_h,\\
x_h &= \text{group variable},\\
\beta_3 &= \text{interaction coefficient for the group by time interaction},\\
v_{0i} &= \text{individual difference component for subject } i,\\
\epsilon_{hij} &= \text{measurement error for subject } i \text{ in group } h \text{ at time } j.
\end{aligned} \tag{3.2}$$

Equation (3.1) may be expressed in matrix form as

$$\mathbf{y}_i = \mathbf{X}_i \underline{\beta} + \mathbf{Z}_i \underline{v}_i + \underline{\epsilon}_i, \tag{3.3}$$

where $\mathbf{y}_i$ is the response vector for subject i, $\mathbf{X}_i$ is the covariate matrix for subject i, $\underline{\beta}$ is the vector of fixed regression parameters, $\mathbf{Z}_i$ is the random effects design matrix for subject i, $\underline{v}_i$ denotes the vector of random subjects effects and $\underline{\epsilon}_i$ is the error vector. Note: It is only one random term in a repeated measures ANOVA, the random subjects effect. In Section 3.4 we explore mixed regression models with more than one random factor, therefore the matrix representation for the random effects are introduced here. For the analysis model in Equation (3.1) the covariate matrix $\mathbf{X}_i$ and random effects design matrix $\mathbf{Z}$ are represented as

$$\mathbf{X}_i = \begin{bmatrix} 1 & t_{i1} & x_i & x_i t_{i1} \\ 1 & t_{i2} & x_i & x_i t_{i2} \\ \vdots & & & \\ 1 & t_{in_i} & x_i & x_i t_{in_i} \end{bmatrix} \qquad \mathbf{Z}_i = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad . \tag{3.4}$$

The random subjects variables v_{0i} are assumed normally distributed with mean zero and variation $\sigma^2_{v_0}$, and ϵ_{hij} are similarly assumed normally distributed with mean zero and variation σ^2.

We express the association between random terms in model (3.1) as

$$\begin{aligned} \mathrm{E}(y_{hij}) &= \beta_0 + \beta_1 t_j + \beta_2 x_h + \beta_3 x_h t_j \\ \mathrm{Var}(y_{hij}) &= Var(v_{0i} + \epsilon_{hij}) = \sigma^2_{v_0} + \sigma^2_{\epsilon}, \\ \mathrm{Cov}(y_{hij}, y_{hi'j}) &= 0 \qquad \text{for } i \neq i', \\ \mathrm{Cov}(y_{hij}, y_{hij'}) &= \sigma^2_{v_0} \quad \text{for } j \neq j'. \end{aligned} \tag{3.5}$$

The expectation of y_{hij} is only related to the fixed terms of the model, that is the intercept, the covariate gender, the time variable and the gender by time interaction. In a simple linear regression model all the variance between observations (assumed independent) are due to the measurement error ϵ, and all observations are assumed to have equal variance. This variance is now partitioned into two parts, the subject-specific variance and the residual variance. The unexplained variance is decreased because some variance is explained by the variability in the population of subjects.

The first covariance statement in Equation (3.5) indicates that measurements from different subjects are independent of each other, while the second covariance statement tells us that measurements of the same subject are correlated equally for any pair of measurements observed on the same subject. This correlation is denoted the intra-class correlation, and is given as

$$\mathrm{Corr}(y_{hij}, y_{hij'}) = \frac{\sigma^2_{v_0}}{\sigma^2_{v_0} + \sigma^2_{\epsilon}}.$$

The covariance matrix of the repeated measures ANOVA model in Equation (3.1) is presented as

$$\begin{bmatrix} \sigma^2_{v_0} + \sigma^2_{\epsilon} & & & & \\ \sigma^2_{v_0} & \sigma^2_{v_0} + \sigma^2_{\epsilon} & & & \\ \sigma^2_{v_0} & \sigma^2_{v_0} & \sigma^2_{v_0} + \sigma^2_{\epsilon} & & \\ \vdots & & & \ddots & \\ \sigma^2_{v_0} & & \cdots & & \sigma^2_{v_0} + \sigma^2_{\epsilon} \end{bmatrix} \tag{3.6}$$

and we observe that it is equal to the above described compound symmetry structure. This is seldom realistic for longitudinal data because the variance tends to differ between groups and in time. In addition are consecutive measurements often more correlated than observations more distant in time.

3.3.2 Repeated measures ANOVA table

When analyzing a repeated measures ANOVA, we introduce the *dot notation* to represent the mean of groups, time points, and subjects. This notation is presented as follows:

$\bar{y}_{...} =$ average across groups, time points and subjects,
$\bar{y}_{hi.} =$ average for subject *i* in group *h* across time points,
$\bar{y}_{h.j} =$ average for group *h* at time point *j* across all subjects in the group,
$\bar{y}_{h..} =$ average for group *h* across time points and subjects,
$\bar{y}_{..j} =$ average for time point *j* across groups and subjects.

Further, *SS* is the sum of squares, calculated as listed in Table 3.1. *MS* is the Mean of Squares and is calculated by dividing the sum of squares by the corresponding degrees of freedom. The sum of squares can be found for all terms in the analysis model in Equation (3.1), and are used in the calculation of F statistics as described below.

Table 3.1: Repeated measures ANOVA table for the regression model in Equation (3.1). SS is the sum of squares, g is the number of groups, n is the number of measurement occasions and N_h is the number of subjects in group h.

Source	SS	df
Group	$\text{SS}_G = n\sum_{h=1}^{g} N_h(\bar{y}_{h..} - \bar{y_{...}})^2$	g-1
Time	$\text{SS}_T = N\sum_{j=1}^{n}(\bar{y}_{..j} - \bar{y_{...}})^2$	n-1
Group×Time	$\text{SS}_{GT} = \sum_{h=1}^{g}\sum_{j=1}^{n} N_h(\bar{y}_{h.j} - \bar{y}_{h..} - \bar{y}_{..j} + \bar{y_{...}})^2$	(g-1)×(n-1)
Subjects in groups	$\text{SS}_{S(G)} = n\sum_{h=1}^{g}\sum_{i=1}^{N_h}(\bar{y}_{hi.} - \bar{y_{h..}})^2$	N-g
Residual	$\text{SS}_R = \sum_{h=1}^{g}\sum_{i=1}^{N_h}\sum_{j=1}^{n}(\bar{y}_{hij} - \bar{y}_{h.j} - \bar{y}_{hi.} + \bar{y_{h..}})^2$	(N-g)×(n-1)
Total	$\text{SS}_T = \sum_{h=1}^{g}\sum_{i=1}^{N_h}\sum_{j=1}^{n}(\bar{y}_{hij} - \bar{y_{...}})^2$	(Nn-1)×(n-1)

Group by time interaction

The first term in model (3.1) that we want to investigate is the group by time interaction. The null hypothesis is formulated as

$$\mathrm{H}_{GT} : \beta_3 = 0 \quad \mathrm{F}_{GT} = \frac{MS_{GT}}{MS_R} = \frac{SS_{GT}/(g-1)(n-1)}{SS_R/(N-g)(n-1)}.$$

F is the statistic that is calculated and compared to the F-distribution, as given in Kvaløy and Tjelmeland [2000]. If this hypothesis is rejected there are three conclusions that can be drawn:

1. The between-group differences vary across time,
2. The between-group curves are not parallel across time,
3. The group and time effect are confounded by the interaction, and separate group effects or time effects cannot be estimated.

The latter describes the importance of testing the interaction prior to the main effects. Tests of the main effects group and time can be carried out if the group by time interaction is insignificant. This is performed in a similar way as the above gender by time interaction.

3.4 Mixed regression models

A repeated measures ANOVA can be thought of as one of the simplest mixed regression models. A repeated measures ANOVA consists of both random subjects effects and covariates, and in a similar way a mixed regression model consists of both fixed and random effects (hence the name *mixed model*). Fixed factors are the covariates of the model that we want to estimate in the analysis. Examples of fixed factors are *age* and *gender*, and the interaction of fixed factors can be included in the same way as in a simple regression model. Random factors are variables in which we are not explicitly interested in estimating coefficients. The reason for inclusion is to control for the variance related to the random variables. Examples of such variables are subjects in a repeated measures ANOVA model, schools and children in a multi center study of children and reading and mothers and children in a study of birth weight. If we can control for these between-subjects factors we are rewarded with increasing statistical power, which makes us able to discover correlations among variables that otherwise are hidden.

The following sections are based on the theory in Hedeker and Gibbons [2006], Rabe-Hesketh and Skrondal [2008] and Fitzmaurice et al. [2009].

3.4.1 Introduction

We have previously explored the univariate and multivariate ANOVA models for repeated measures. The univariate ANOVA models assumes sphericity which is not a common feature of longitudinal datasets. Neither of the methods handle missing values in the data, thus we have to omit subjects with non-responses from the data prior to the analysis and the results may be biased. The time variable is restricted in these models, all subjects must be measured at the same fixed time points, and in a MANOVA the distances in time between all consecuting measuring

occasions must be equal. Both univariate and multivariate ANOVA models have possibilities to examine group trends across time, but subject-specific trends are not accessible.

Mixed regression models (also denoted MRMs) are more general models to analyze longitudinal data in a flexible and correct way with respect to bias induced by missing values. First, the variance structure can be specified more generally than the compound symmetry, but less costly than the alternative unstructured covariance matrix where all variances and covariances are allowed to be unequal. The time variable can be continuous, and subjects can be measured at individual time points. The mixed models can handle ignorable non-responses in longitudinal datasets without inducing bias to the estimates, since all subjects with observed values are included in the study. The sample is assumed representative for the population of subjects. Compared to complete-case data the mixed models analysis has increased power, which is an important additional advantage in such analyses. Both time-variant and time-invariant covariates can be included in mixed models.

Variants of mixed models are known as 'random-effects models', 'multilevel models', 'hierarchical linear models', 'variance component models', 'two-stage models' and 'random regression models', to mention some. The common feature for all these models is the inclusion of random subjects effects into the regression models. The random subjects effects describe each units deviation from the population mean intercept, and take part in the explanation of the correlation structure in the longitudinal data. Alternatively, they describe the degree of variation between subjects in the population of subjects.

Simple linear regression models

A simple linear regression model with time variable t_{ij} and dichotomous grouping variable x_h can be written as

$$y_{hij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_h + \beta_3 x_h t_{ij} + \xi_{hij}. \tag{3.7}$$

y_{hij} is the response for subject *i*, $i = 1, 2, \cdots, N_h$ at occasion t_{ij}, where the time variable can be continuous and individual for each subject. The model above assumes all observations to be independent. This can be expressed by the variance component ξ, that is assumed to be independent normally distributed with mean zero and common variance σ^2.

3.4.2 Random intercepts mixed regression models

In longitudinal data the assumption of independence is violated since we cannot assume observations on the same subject to be uncorrelated. To control for this correlation we simply add a random subjects effect to the regression model, thus the variance due to subject-specific features is separated from the residual variance. A random intercepts mixed regression model for a two-sample analysis is presented as

$$y_{hij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_h + \beta_3 x_h t_{ij} + v_{0i} + \epsilon_{hij}. \tag{3.8}$$

Interpretation of the parameters are equal to the parameters in the repeated measures ANOVA model as given in Equation (3.3.1). The variance ξ_{hij} in Equation (3.7) is partitioned into two

variance parts, the random subjects effect and the residual error, expressed as

$$\xi_{hij} = v_{0i} + \epsilon_{hij}.$$

The random subjects effects v_{0i} have mean zero and common variance $\sigma^2_{v_0}$. The variation due to subject-specific features is thus separated from the residual and leads to a decrease in the unexplained variance. The variances $\sigma^2_{v_0}$ are assumed identical and independent between subjects, and the residuals ϵ_{hij} are assumed identical and independent of each other and the $\sigma^2_{v_0}$ conditionally on y_{hij}. This leads to a compound symmetry structure, which is one of the "strict" assumptions of repeated measures ANOVA. Thus the analysis by a random intercepts mixed regression model and the repeated measures ANOVA assuming compound symmetry gives similar results.

The compound symmetry can be expressed by the following covariance matrix, here with three measure occasions, t=0, 1, 2:

$$\begin{bmatrix} \sigma^2 + \sigma^2_{v_0} & & \\ \sigma^2_{v_0} & \sigma^2 + \sigma^2_{v_0} & \\ \sigma^2_{v_0} & \sigma^2_{v_0} & \sigma^2 + \sigma^2_{v_0} \end{bmatrix} \quad . \tag{3.9}$$

(3.10)

An alternative way to represent the model in Equation (3.8) is by a hierarchical, or multilevel structure. The equation is partitioned into a within-subjects model

$$y_{hij} = b_{0i} + b_{1i}t_{ij} + \epsilon_{hij} \tag{3.11}$$

and a between-subjects model

$$\begin{aligned} b_{0i} &= \beta_0 + \beta_2 x_h + v_{0i}, \\ b_{1i} &= \beta_1 + \beta_3 x_h. \end{aligned}$$

The random subjects effects v_{0i} can be interpreted as each subjects deviation from the group intercept, thus it represents the individual intercepts and are denoted random subjects or random intercepts effects. A random intercepts MRM assumes the slopes of all subjects to be equal, so all lines on a profile plot are parallel for all group means. A profile plot displays the group means of the dependent variable over time, with lines connecting the time points.

3.4.3 Random slopes mixed regression model

Now we have explored the random intercepts mixed regression model, where random subjects effects are included in the linear regression model. In many situations this model may be too simplistic. By using this model we assume that all women have the same progress for the response variable over time, and all men the same. The strict variance assumption of compound symmetry is often violated in longitudinal data. The next step in developing a MRM is to examine the extent of heterogeneity of subjects within groups with respect to their slopes. In

other words, we are examining whether a random slopes effect is statistically significant. A random slopes MRM, also denoted random-coefficient MRM, is expressed as

$$y_{hij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_h + \beta_3 x_h t_{ij} + v_{0i} + v_{1i} t_{ij} + \epsilon_{hij}. \tag{3.12}$$

We now want to express Equation (3.12) as a two-level model similar to the two-level model in Section 3.4.2. The within-subjects model in Equation (3.11) is unchanged, but the between-subjects model is augmented as

$$b_{0i} = \beta_0 + \beta_2 x_h + v_{0i},$$
$$b_{1i} = \beta_1 + \beta_3 x_h + v_{1i}.$$

As before, b_{0i} is the intercept parameter for each subject *i* and b_{1i} is the subject-specific slope parameter indicating the development in response variable over time for subject *i*. The new term v_{1i} is the individual slope deviation from the population slope for subject *i*.

We now have a model with subject-specific intercepts and time trends (v_{0i} and v_{1i}), in addition to the population intercept and change in time (β_0 and β_1). The fixed effects of the random intercepts MRM are unaltered, but the variance-components of the random slopes MRM must be investigated further. The error ϵ_{hij} is independent conditional on v_{0i} and v_{0i}, and normally distributed with mean zero and variance σ^2. The random subjects and slopes effects are assumed to be bivariate normal with mean vector zero and covariance matrix given as

$$\Sigma_v = \begin{bmatrix} \sigma^2_{v_0} & \sigma_{v_0 v_1} \\ \sigma_{v_0 v_1} & \sigma^2_{v_1} \end{bmatrix}.$$

The interpretation of $\sigma^2_{v_0}$ and $\sigma^2_{v_1}$ is the heterogeneity of subject intercepts and slopes, respectively. The notation $\sigma_{v_0 v_1}$ denotes the covariance of v_0 and v_1, which may be interpreted as the relation between a subjects intercept and the subjects individual slope parameter. So, if $\sigma_{v_0 v_1}$ is positive we would interpret this as follows: A subject with an intercept above the population mean (higher initial values) have steeper slopes than subjects with smaller initial values.

One of the important advantages of the random slopes mixed regression model compared with the random intercepts model is the more slack variance assumptions, compound symmetry is no longer required. A total of four variance components are estimated in the analysis, and model a more flexible variance-covariance relation of the data, see covariance matrix in Equation (3.13). This correlation structure is well explained in Rabe-Hesketh and Skrondal [2008]. In longitudinal data it is natural to expect observations that are measured subsequently to be more correlated than observations more distant in time. This model include this in the model assumptions, and is therefore more flexible and in many settings more realistic than the compound symmetry.

Random coefficient structure are displayed in Equation (3.13) for data with three measure occasions, t=0,1,2:

$$\begin{bmatrix} \sigma^2_{v_0} + \sigma^2 & & \\ \sigma^2_{v_0} + \sigma_{v_0 v_1} & \sigma^2_{v_0} + 2\sigma_{v_0 v_1} + \sigma^2_{v_1} + \sigma^2 & \\ \sigma^2_{v_0} + 2\sigma_{v_0 v_1} & \sigma^2_{v_0} + 3\sigma_{v_0 v_1} + 2\sigma^2_{v_1} & \sigma^2_{v_0} + 4\sigma_{v_0 v_1} + 4\sigma^2_{v_1} + \sigma^2 \end{bmatrix}. \tag{3.13}$$

Decomposing the time effect

Till now we have assumed the time trend to be linear, but this may be a strict assumption that models the data poorly. Particularly this gets critical for studies with many measurement occasions. The next step in modeling longitudinal data can therefore be examination of quadratic (and maybe cubic) time trends. One alternative to introduce a quadratic trend is to simply include a quadratic time term in the regression model as

$$y_{hij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_h + \beta_3 x_h t_{ij} + \beta_4 t_{ij}^2 + \beta_5 x_h t_{ij}^2 + v_{0i} + v_{1i} + \epsilon_{hij}. \tag{3.14}$$

An interaction group by time squared may also be included.

An alternative approach to introduce polynomial time trends is to create a dichotomous variable for each measuring occasion. This requires equal fixed time points for all subjects, and makes comparison of progress in time between gender less informative. Several parameters must be estimated, thus we loose degrees of freedom. The overall progress of the dependent variable over time is not described.

For data with measures at time $t = 0, 1, 2$ the quadratic time term is $t^2 = 0, 1, 4$, which is nearly collinear to *t*. To avoid this we can express time in centered form, for example $t'_j = (t_{ij} - \bar{t})$. The interpretation of the main intercept must be adjusted for the shift in the time variable, and must be interpreted as the mean of all observations at the midpoint of time t'_{ij}.

The centered time t_{ij} denotes the original time notation, while t'_{ij} corresponds to the centered time. β_0 is the original intercept coefficient and β'_0 is the centered time intercept coefficient displayed as

$$\beta_0 + \beta_1 t_{ij} = \beta_{0'} + \beta_1 t_{ij'} = \beta_{0'} + \beta_1 (t_{ij} - \bar{t}). \tag{3.15}$$

The group by time interaction is also affected by the centering of time. Results for fixed effects are not altered by the centering of time, neither are the other effects in a random intercepts regression model (group effect, random subject effects and residual variance). Note: Models with polynomial time trends are still denoted *linear regression models* since they are linear in terms of regression covariates other than time [Hedeker and Gibbons, 2006].

When the number of measuring occasions increases, the method of centering time points gets rather complicated. A more general method to examine polynomial time effects is by using orthogonal polynomials. Comparisons of time-related effects such as change relative to baseline, consecutive time comparisons or contrasting each time point to the mean of subsequent time points are also possible to examine through orthogonal polynomials.

3.4.4 Random slopes mixed regression model, quadratic time trend

We can introduce a squared time variable to model the response variable as a quadratic curve along the time axis. By including t_{ij}^2 we obtain the model expressed as

$$y_{hij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_h + \beta_3 x_h t_{ij} + \beta_4 t_{ij}^2 + \beta_5 x_h t_{ij}^2 + v_{0i} + v_{1i} t_{ij} + \epsilon_{hij}.$$

Again this model can be represented as a two-level model. The within-subjects model is given as

$$y_{hij} = b_{0i} + b_{1i}t_{ij} + b_{2i}t_{ij}^2 + \epsilon_{hij} \tag{3.16}$$

and the between-subjects model is presented as

$$\begin{aligned} b_{0i} &= \beta_0 + \beta_2 x_h + v_{0i}, \\ b_{1i} &= \beta_1 + \beta_3 x_h + v_{1i}, \\ b_{2i} &= \beta_4 + \beta_5 x_h. \end{aligned}$$

The squared time parameter β_4 represents the curvation of the group mean, while β_5 refers to the deviation in curvation for the group h when $x_h \neq 0$. The interpretation of the remaining parameters in this model are not modified compared to the simpler random slopes model. The mixed model can be extended even further by including higher-degree time variables or more random effects, but the principle of analyzing these models are basically the same as for the models described above.

3.4.5 Curvilinear mixed regression models

The random slopes MRM with quadratic time trend induces a possible extension of the random effects model, that is a random squared time effect. This effect can be described as the subject-specific deviation in curvation from the population mean. This mixed regression model is denoted *curvilinear MRM*, and is displayed as

$$y_{hij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_h + \beta_3 x_h t_{ij} + \beta_4 t_{ij}^2 + \beta_5 x_h t_{ij}^2 + v_{0i} + v_{1i} t_j + v_{2i} t_j^2 + \epsilon_{hij}. \tag{3.17}$$

The random curvilinear effect v_{2i} is multiplied by the squared time variable. In the same way as for the previous MRMs, the curvilinear model can be given as a two-level model, with within-subjects model equal to the random slopes MRM with quadratic time term in (3.16) and between-subjects model given as

$$\begin{aligned} b_{0i} &= \beta_0 + \beta_2 x_h + v_{0i}, \\ b_{1i} &= \beta_1 + \beta_3 x_h + v_{1i}, \\ b_{2i} &= \beta_4 + \beta_5 x_h + v_{2i}. \end{aligned}$$

The fixed effects of the random slopes and curvilinear MRM are equal, and the random effects are further partioned into subject-specific intercept, slope and curvilinear variance and measurements error (unexplained variance). The subject-specific variances are correlated within subjects in a similar matter as in matrix (3.4.3), which leads to a total of six estimated variance and covariance parameters. The error variances are independent normally distributed with mean zero and variance σ^2, and conditionally independent on the other variance terms as before.

The curvilinear MRM allows each subject to have an individual intercept, slope and curve parameter, and is the most general model accessible for longitudinal data restricted to three fixed measurement occasions.

3.4.6 Comparison of models

The fixed effects in a mixed regression model can be tested by the Wald's test, which compare the test statistics to a standard normal frequency table. The null hypothesis is that the parameters are zero. For the variance and covariance parameters this test is not suitable since the variance parameters are bounded to positive values [Hedeker and Gibbons, 2006]. This is the reason why no p-values are calculated for the random terms in mixed models. Likelihood ratio tests are used to compare nested models, thus we can use this technique to examine if a random effect is necessary to model the data.

Simple linear regression models and random intercepts and slopes models are hierarchically nested in each other, and we apply the likelihood ratio test to compare these models. This test uses the difference in deviance values for the models and compares them with a chi-square distribution given as

$$-2\ln\left(\frac{L_0}{L_1}\right) \sim \chi^2_\nu. \tag{3.18}$$

The degree of freedom ν is determined by the difference in number of estimated parameters in the two models. L_0 is the maximum value of the likelihood for the data from the more specified model (for example the random intercepts model), while L_1 is the maximum value of the likelihood for the data from the general model (for example the random slopes model). For these two models there are estimated two extra parameters in the random slopes model, that is the slope variance and intercept-slope covariance. Thus the degrees of freedom ν for the likelihood ratio test equals two. Because the variance terms are restricted to be positive, the likelihood ratio test for random effects are too conservative. Hedeker and Gibbons [2006] refers to Berkhof and Snijders [2001] to state that more correct p-values are obtained by dividing the test values by two. This is also described in Rabe-Hesketh and Skrondal [2008].

Regression parameters are calculated by maximum likelihood (ML) estimates or restricted maximum likelihood (REML) estimates. REML estimates are often preferred over ML estimates because REML adjusts the likelihood for the number of covariates in a model. This is also why these estimates cannot be applied when models are compared with a likelihood ratio test. The parameters of all models must be estimated by maximum likelihood according to the above mentioned [Rabe-Hesketh and Skrondal, 2008].

The null hypothesis of likelihood ratio tests is formulated in the following matter; the more restricted model is sufficient to model the data with less parameters compared to the more general model. If the likelihood ratio test is statistically significant, the null hypothesis is rejected and the general model with more parameters is preferred.

3.5 Covariance pattern models

In Section 3.4 we found that mixed regression models can be thought of as an extension of the univariate repeated measures ANOVA models, and similarly we can imagine the covariance pattern models (CPM) as an extension of the multivariate ANOVA models for repeated measures. CPM are formulated as regression models in the same way as MANOVA, but assumes

the variance-covariance matrix to be of a certain form. No random effects are included in the regression model, so these models do not distinguish the variance term in within-subjects and between-subjects variance. CPM treat time as a categorical variable, so the time is considered fixed for all subjects. Unlike MANOVA the CPM allow unbalanced design. The regression model for CPM on matrix form can be written as

$$\mathbf{y}_i = \mathbf{X}_i\underline{\beta} + \mathbf{e}_i. \tag{3.19}$$

The $\mathrm{n}_i \times 1$ vector $\mathbf{y}_i$ contain the responses for subject *i* with $j = 1, 2, \cdots, n_i$ observations, $i = 1, 2, \cdots, N$. $\mathbf{X}_i$ is the covariance matrix for subject *i*, $\underline{\beta}$ corresponds to the fixed effects parameter vector and $\mathbf{e}_i$ is the $\mathrm{n}_i \times 1$ error vector.

$\mathbf{e}_i$ is the model specification that separates CPM from MANOVA models, where the error vector is assumed normally distributed with mean zero and variance $\sigma^2\mathbf{I}_{n_i \times n_i}$. We assume that the error vector satisfies the assumptions

$$\mathbf{e}_i \sim \mathrm{N}(0, \Sigma_i). \tag{3.20}$$

The variance-covariance matrix Σ_i is estimated for the n_i time points in which subject *i* is measured. Note that CPM estimates the regression parameters jointly, that is based on all covariates and observed values of the dependent variable simultaneously. Thus these models handle missing data structures of both MCAR and MAR mechanisms. Equation (3.20) leads to the model assumptions $\mathbf{y}_i \sim \mathrm{N}(\mathbf{X}_i\underline{\beta}, \Sigma_i)$ and $\mathrm{Var}(\mathbf{y}_i|\mathbf{X}_i) = \Sigma_i$.

3.5.1 Covariance patterns

Different covariance patterns are feasible for CPM, and some of these are explained below.

Independent covariance structure

The independent covariance structure implies independent measurements and equal variance σ^2 at each time point, that is $\sigma^2\mathbf{I}_{n_i \times n_i}$. This structure can be found in studies where different subjects are measured at the different time points, as a simple example, but this occurs rarely with longitudinal data.

Exchangeable covariance structure

The exchangeable covariance structure is previously referred to as compound symmetry, and is induced by the random intercepts MRM and repeated measures ANOVA models assuming homoskedasticity and sphericity. It is defined by two parameters, σ^2 and σ_1^2, where the first is the residual variance and the latter being the estimated random intercepts variance. The covariance pattern is displayed for a repeated measures ANOVA in Equation (3.6), and expressed by the newly introduced parameters as

$$\begin{bmatrix} \sigma_1^2+\sigma^2 & & & & \\ \sigma_1^2 & \sigma_1^2+\sigma^2 & & & \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2+\sigma^2 & & \\ \vdots & & \ddots & \ddots & \\ \sigma_1^2 & \cdots & & \sigma_1^2 & \sigma_1^2+\sigma^2 \end{bmatrix}. \tag{3.21}$$

First-order autoregressive structure, AR(1)

This covariance pattern is often used with time series and requires estimation of the measurement error σ^2 and the autoregressive coefficient ρ. The covariance is assumed to diminish by *lags*, so consecutive measurements in time are more correlated than measurements more distant in time. Lags describe the distance in time between two fixed (equally distanced) measuring occasions, where lag-1 corresponds to two consecutive occasions, lag-2 corresponds to two fixed time intervals between measurement occasions, and so on. Estimated covariance between time points j and j' are calculated as $\sigma_{jj'} = \sigma^2\rho^{|j-j'|}$.

This covariance structure assumes at least two consecutive observations to be observed for each subject to be able to estimate the autoregressive parameter. It also assumes all intervals between measuring occasions to be equal, thus no gaps between measures are tolerated. These assumptions force constraints on the missing data structure, and subjects that do not fulfill the requirements are either omitted or should be imputed prior to analyses.

The first-order autoregressive structure can be written as

$$\sigma^2 \begin{bmatrix} 1 & & & & \\ \rho & 1 & & & \\ \rho^2 & \rho & \ddots & & \\ \vdots & & \ddots & & \\ \rho^{n-1} & \rho^{n-2} & \cdots & \rho & 1 \end{bmatrix}. \tag{3.22}$$

Toeplitz structure

The toeplitz structure, also referred to as the banded structure, holds a covariance parameter for each lag. This leads to a total of n estimated parameters, that is one for each time point. The assumptions implied by this covariance structure are equal variance at all time points and equal covariance for all measurements with equal lags. θ_1 corresponds to the variance of measurements, θ_2 is the lag1 covariance and so on. The measuring occasions are restricted to be fixed and with equal distances between all consecutive measurements. The first-order autoregressive structure is a special case of the toeplitz structure, the latter given as

$$\begin{bmatrix} \theta_1 & & & & \\ \theta_2 & \theta_1 & & & \\ \theta_3 & \theta_2 & \theta_1 & & \\ \vdots & & \ddots & \ddots & \\ \theta_n & \cdots & & \theta_2 & \theta_1 \end{bmatrix}. \tag{3.23}$$

Unstructured form

The above covariance structures AR(1) and toeplitz hold restrictions for the time point intervals to be equally distanced and equal within lags. A more general structure, denoted the unstructured form, allows all variances and covariances to be different. This covariance structure allows all types of missing data structure among subjects, and is therefore often the preferred for longitudinal datasets with few measuring occasions. The total number of parameters that must be estimated is $n(n+1)/2$, where n is the number of time points. For studies with many observational time points this structure demands estimation of a high number of parameters which may lead to imprecise estimates of variances and covariances. The unstructured correlation form is displayed as

$$\begin{bmatrix} \theta_{11} & & & & \\ \theta_{21} & \theta_{22} & & & \\ \theta_{31} & \theta_{32} & \theta_{33} & & \\ \vdots & & \ddots & \ddots & \\ \theta_{n1} & \cdots & & \theta_{n(n-1)} & \theta_{nn} \end{bmatrix} \tag{3.24}$$

Covariance pattern models are dependent of a correct fixed regression model and a suitable covariance pattern to obtain unbiased results. When selecting the model, the set of fixed covariates must include all possible variables that may affect the response variable, and this set of covariates must stay unchanged through the selection process of the covariance pattern. Once the covariance structure is determined, the fixed regression model selection can be performed as in a standard regression analysis. To select the most appropriate covariance structure Σ a likelihood ratio test is employed as explained in Section 3.4.6.

3.6 Analyses of discrete outcome variables

We have now described methods of linear models to analyze continuous responses in longitudinal datasets. Both mixed regression models and covariance pattern models handle ignorable missing, and have possibilities to adjust the covariance matrix to the data. When the response variable is discrete (for example binary, ordinal or a count), these linear models do not model the changes in the mean response due to covariates in an adequate way. This section will give an introduction to modeling of ordinal variables in a longitudinal dataset, when the data contain unobserved values.

Generalized linear models (GLM) is applied for analyzing univariate discrete outcome variables, via known variances and link functions. For longitudinal data the GLM is not sufficient

to model discrete responses because of the dependency between observations within subjects. There are two ways to extend GLMs to longitudinal data, either by generalized linear mixed regression models (GLMM) as described in Section 3.6.1, or by generalized estimating equations (GEE) described in Section 3.6.2, according to Hedeker and Gibbons [2006]. These sections focus on analysis with an ordinal outcome variable, but other discrete outcome variables can be analyzed in a similar way.

An ordinal variable is a categorical variable where the categories are ordered. Examples of ordinal variables are agreement ratings with categories 'disagree', 'undecided' and 'agree', rating of movies with categories one to six, and items of the SF-36 questionnaire with categories one to five where higher scores refers to higher quality of health for the given item.

3.6.1 Generalized linear mixed regression models

Proportional odds models (POM) are a common choice for analysis of ordinal data, thus many of the mixed models for ordinal data are generalizations of this model. POM is based on logistic regression formulation and characterizes the ordinal responses in C categories in terms of $C-1$ cumulative category comparisons. The notation $P_{hijc} = P(Y_{hij} \leq c) = \sum_{k=1}^{c} p_{hijk}$ is introduced, where p_{hijk} is the probability of response in category k for subject i in group h at time j. POM can be extended to a random intercepts mixed effects proportional odds model as follows

$$\log\left(\frac{P_{hijc}}{1-P_{hijc}}\right) = \gamma_c - (\beta_1 t_{ij} + \beta_2 x_h + \beta_3 x_h t_{ij} + v_{0i} + \epsilon_{hij}), \tag{3.25}$$

where $c = 1, 2, \cdots, C-1$. γ_c are category-specific parameters denoted *model thresholds* and are assumed to be strictly increasing in c. The interpretation of the random intercepts MRM parameters is equivalent to the model described in Section 3.3.1. The mixed effects POM is essentially a proportional odds model with random effects in the linear regression equation [Rabe-Hesketh and Skrondal, 2008].

The model in Equation (3.25) is a cumulative model for ordinal responses in terms of the linear regression model linked to a cumulative probability, not the mean of response as in a GLM. POM is a multiplicative model. To illustrate this, let us look at a continuous covariate that increase by 5 units from x_i to x_i'. The odds ratio is $exp(\beta)^5$ if β is the parameter of variable x_i.

An alternative modelrepresentation of Equation (3.25) is expressed by the latent-response formulation, also called the threshold model, in Rabe-Hesketh and Skrondal [2008] as

$$y_i^* = \beta_1 t_{ij} + \beta_2 x_h + \beta_3 x_h t_{ij} + v_{0i} + \epsilon_{hij}.$$

The terms $\epsilon_{hij}|t_{ij}, x_h, v_{0i}$ have logistic distributions and are assumed to be independent across subjects and measurement occasions. The continuous latent variable y_i^* are related to the ordinal variable via the threshold model $y_i = s$ if $\gamma_{s-1} < y_i^* \leq \gamma_s$. Further, more complex mixed effects POM can be fitted in a similar way.

An alternative approach to model ordinal variables is by probit regression formulation. Hedeker and Gibbons [2006] describes both approaches in detail, and Rabe-Hesketh and Skrondal [2008] explain the modeling of such methods in Stata.

3.6.2 Generalized estimating equations

As a generalization of GLM, the generalized estimating equations (GEE) support many different types of dependent variables. The method was developed for non-continuous variables as dichotomous responses and counts, and is rarely used for continuous data [Rabe-Hesketh and Skrondal, 2008]. Marginal distribution of Y_{ij} at each time point need to be specified (denoted quasi-likelihood models), in contrast to full likelihood-models where the joint distribution of a subject's response vector y_i must be specified. The variance-covariance structure is treated as a nuisance, and this is an important feature of GEE. Since the unobserved variables are dependent only on the covariates, the missing data structure implied is the covariate-dependent MCAR. Thus the method of GEE does not handle ignorable missing data assuming MAR, which is a major disadvantage when working with missing values. GEE assumes all subjects to be measured at the same fixed occasions, but it assumes no restrictions on the missing data pattern (dependent on the chosen correlation structure).

Fitzmaurice et al. [2009] introduced the term *marginal models* which refers to models for longitudinal data without random effects. This includes among others the covariance pattern models described in Section 3.5 and GEE. Both models assume the number of measurement occasions to be fixed, but missing values among the time points are allowed. The correlation structures induced by the CPM and GEE are similar, but there are some basic differences that separate the properties of the methods. CPM specify the joint distribution and likelihood of the dependent variable vector Y_i and apply only for continuous normal distributed outcome. GEE, on the other hand, specifies the marginal distribution and likelihood of Y_i for each time j, but can be applied for many types of outcome variables.

Specification of a GEE is similar to a GLM, with a linear predictor, a link function and variance described as a function of the mean. An additional feature of GEE is the "working" correlation structure R, a $n \times n$ correlation matrix common for all subjects. If subject i are observed at n_i occasions he or she achieve a $n_i \times n_i$ correlation matrix with the appropriate rows and columns for the observed time points from R. The choice of "working" correlation matrix should be consistent with the observed correlation matrix, and is often selected from the structures displayed in Section 3.5.1. A common set of association parameters **a** are estimated, where the size of the vector **a** dependent on the "working" correlation structure chosen.

Selection of the correlation structure for the repeated measurements is not as critical for GEE as for mixed regression models or covariance pattern models. This is because GEE provides estimated parameters and standard errors that are robust to misclassification of the variance-covariance structure, as long as the univariate analysis models at each time point are specified correctly. The statistical power decreases with misclassification of the correlation structure, but the loss of power is small when the number of subjects increases. GEE should be applied when the research interest is focused on estimates and inference of the regression parameters, but is not suitable when modeling variance-covariance structures of longitudinal data.

Solving the GEE involves iterating between the quasi-likelihood solutions for estimating β's and a robust method for estimating a set of correlation parameters **a** as a function of β. Hedeker and Gibbons [2006] describes the iterating process as follows

1. Given estimates of $R_i(a)$, calculate estimates of β using iteratively reweighted least squares.
2. Given estimates of β, calculate Pearson residuals and use these residuals to consistently estimate **a**.

Iteration over the two steps continues until convergence, that is when the updated estimates approximates the rejected estimates. Hedeker and Gibbons [2006] describes in more detail the process of GEE, and examples where the method is applied.

There are two general approaches for handling data assuming the MAR mechanism within the framework of GEE. The first approach is to analyze multiple imputed data by generalized estimating equations. The technique of GEE described above can be performed without further adjustment. The properties of multiple imputation leads to unbiased analysis results if the GEE is correctly specified.

The second approach is the weighted estimating equations. The idea of weighted estimating equations is to account for subjects with missing responses by giving extra weight to the subjects with observed measurements and similar observed covariates and the same history of responses. The weights are calculated as the inverse probability of being observed. This approach is suitable when the missing data pattern is monotone, typically as a result of dropout, and is discussed in particular by Fitzmaurice et al. [2009]. In many longitudinal studies the subjects are missing at an arbitrary missing data pattern, as explained in Section 2.1.5. For such settings approximative methods have been proposed for handling missing data that are assumed to be MAR. Further, if data are assumed MNAR, more complex methods must be applied.

4

The data of patients undergoing cardiac surgery

Missing values or non-response occur frequently in datasets with measurements of persons involved, especially when the observations are self-reported and repeated at several timepoints. A dataset of this type is analyzed in the article 'The role of sex in health-related quality of life after cardiac surgery: a prospective study', written by Gjeilo et al. [2008]. Of all patients undergoing cardiac surgery at the Cardiothoracic Surgery at St Olavs Hospital, Norway, in the period from September 2004 to September 2005, a total of 534 patients were included in the study. Of these were 413 men and 121 women. Data were prospectively collected using the Norwegian verson of the Short-Form Health Survey (SF-36) version 1.2 [Loge and Kaasa, 1998], among others. The variables with main focus in the article by Gjeilo et al. [2008] are age, gender, marital status and health-related quality of life variables (denoted HRQOL and measured by SF-36 and health transition).

We have selected four of the HRQOL-variables, that is general health, bodily pain, social functioning and role emotional based on fraction of missing data and results from the complete-case analyses. Men and women were found to improve different for the variables role emotional and bodily pain, while the differences between genders presurgery and follow-up occasions differed between general health and social functioning. This selection is performed to decrease the amount of results and the size of tables in the report, and thus help the reader to focus on the important results. These four variables are ment to represent the features that exist in the data.

Age and gender are variables that are measured presurgery, and do not change during the study (age increases by one year during the study, this increase is equal for all patients). Marital status is measured at all three time points, and have nine options of status; married, couple living together, widow/widower, single, separated/divorced and combinations of these. The research questions involve the status of living together or alone, thus the variable is implemented as a dichotomous variable where zero refers to living alone and one is living together. This variable may change during the study.

4.1 SF-36 questionnaire

The SF-36 [Loge and Kaasa, 1998] is a short-form health survey with 36 questions/items developed to access HRQOL. The 36 items yield eight scales of functional, physical and mental health, that is general health (GH), physical function (PF), bodily pain (BP), mental health (MH), role limitations owing to physical problems (role physical, RP), role limitations owing to emotional problems (role emotional, RE), vitality (VT) and social functioning (SF). These scales are made up from different number of items, as seen in Table 4.1.

The SF-36 has been extensively applied in several countries, including Norway, and found satisfactory for evaluating HRQOL in cardiac surgery. In addition to the eight scales from SF-36 the item health transition (HT) is recorded. All nine variables are transformed into intervals from zero to 100, where higher scores reflect better health.

Table 4.1: SF-36 scales and health transition (HT). Each scale is made up from several items, which leads to different number of possible responses for each scale.

Scales	Items	Levels
PF	10	21
RP	4	5
BP	2	11
GH	5	21
VT	4	21
SF	2	9
RE	3	4
MH	5	26
HT	-	5

The compound scales presented in Table 4.1 are computed as the mean of the items that form the scale. Such scales are sensitive to missing data, if one item is missing then the whole scale will be missing. Questionnaires often have guidelines to handle missing values to ensure that most of the composite scores are obtained. For SF-36 the mean of the available items replaces the missing items if at least half of the items of a score are observed [Ware et al., 2000].

4.2 Missing-data structure

A total of 311 patients (or 58%) of the 534 patients included in the study responded to all of the variables at the three time points. 470 of the 534 patients, that is 88%, have fully observed variables before surgery. The middle columns in Table 4.2 displays the missing data structure of the dataset for four of the HRQOL-variables general health, bodily pain, social functioning and role emotional, in addition to the variables age, gender and marital status. A similar table for all HRQOL-variables are found in Appendix, Table A.1. Each subject are given an identification number (denoted *patkey*), displayed in the first row of the table. Patkey, age, gender and the presurgery measurement of marital status have no missing values. All the four HRQOL-variables

have less missing data presurgery compared to the follow-up time points. Role emotional has most missing values, with more than 20% non-response at both follow-up occasions.

In studies where the same subjects are observed repeatedly, the phenomena of missing forms may appear. When a subject included in the study have missing values for all variables at one time-point, we call this a missing form. In this study none of the subjects have missing forms presurgery, but at six months a total of 72 subjects have completely missing responses. At six months there are 69 missing forms, and 43 of these are found at both follow-up time points. The reason for missing forms can be many, for example because of forgetfulness, the patient is very ill, or so healthy that he or she sees no point in responding, patient moved or did not recieve the questionnaire due to other reasons, or death.

When examining the missing-data structure after omitting subjects with missing forms, the pattern of missing data are more similar at all three timepoints, see the two right columns in Table 4.2. None of the variables at any of the timepoints have more than 9% missing values. The missing data due to missing items only seems to be similar for the subjects without missing forms. The variable with most missing values is role emotional.

521 patients were alive after 12 months from operation time, and approximately 90% of them responded at least one of the two observational timepoints after surgery. The patients that did not survive until 12 months after surgery are a special case of missing data in the dataset, the reason for missing is obvious but it may be difficult to interpret in the analyses. One solution to this is to exclude all patients that die during the first 12 months after surgery, and base the analyses on the patients that survive their first year postoperation. In this way the research question is altered, we only want to look at the surviving patients after heart surgery. Another approach is to evaluate the reason of death, and separate the patients on base of this. If the reason for death is a consequence of the heart surgery or in any other way connected to the operation or their heart condition, the missing values should be handled by care. In this situation the missing data may be cathegorized as either MAR or MNAR. In comparison, if the reason for death has no connection to the surgery, the missing values are arbitrary and thus may be treated as MCAR. This is discussed further in Section 8.

4.3 Observed covariance and correlation matrices

Further we want to look at the relation between measurements within subjects to examine the correlation induced by repeated measures on the same subjects. The covariance matrix of the observed data for role emotional is displayed to the left in Table 4.3. The diagonal elements equals the variances at each time point, covariances below the diagonal are based upon listwise deletion and covariances above the diagonal are computed from pairwise observed data. The right matrix in Table 4.3 displays the correlation matrix of the same variable.

As we can see the variances are not equal at the three measurement occasions, most variability is found presurgery while the variability is more similar and lower after the surgery. Covariances between the presurgery measurements and the follow-up measurements are approximately equal, but less than the covariance between six and twelve months follow-up after surgery. This covariance matrix indicates that the assumptions of homoskedasticity and sphericity may be violated. The correlation matrix for role emotional can be interpretted the same way. Only mi-

Table 4.2: Missing-data structure for the variables patkey, age, gender, marital status, general health (GH), bodily pain (BP), social functioning (SF) and role emotional (RE). The second and third columns are calculated based on the original dataset, while the two last columns describe the missing data structure in the data after exclusion of missing forms. The percentages are calculated based on all SF-36 items and HT at the three time points.

Variable	**No. of missing**	**% missing**	**No. of missing**	**% missing**
Patkey (id)	0	0.0	0	0.0
Gender	0	0.0	0	0.0
Age	0	0.0	0	0.0
Marital status				
Before surgery	0	0.0	1	0.2
6 months follow-up	72	13.5	0	0.0
12 months follow-up	70	13.1	1	0.2
General Health				
Before surgery	40	7.5	33	7.6
6 months follow-up	94	17.6	17	3.9
12 months follow-up	98	18.4	26	6.0
Bodily Pain				
Before surgery	13	2.4	11	2.5
6 months follow-up	86	16.1	13	3.0
12 months follow-up	84	15.7	14	3.2
Social Functioning				
Before surgery	14	2.6	12	2.8
6 months follow-up	77	14.4	4	0.9
12 months follow-up	80	15.0	10	2.3
Role Emotional				
Before surgery	41	7.7	35	8.0
6 months follow-up	115	21.5	38	8.7
12 months follow-up	108	20.2	37	8.5

nor differences are found when comparing the complete-case and available-case samples with respect to correlations.

Table 4.3: Variance-covariance matrix (left) and correlation matrix (right) for the variable role emotional (RE) at the three measurement occasions presurgery (RE1), six months follow-up (RE2) and 12 months follow-up (RE3). Variable quantities below diagonal are based on listwise deletion, and quantities above on pairwise deletion.

	RE1	RE2	RE3
RE1	1887.91	*595.492*	*547.338*
RE2	550.642	1513.9	*875.153*
RE3	535.713	890.004	1504.48

	RE1	RE2	RE3
RE1	1.000	*0.357*	*0.324*
RE2	0.335	1.000	*0.608*
RE3	0.332	0.594	1.000

5 Complete-case analyses

In the article of Gjeilo et al. [2008] the hypotheses of possible differences between genders at each measuring occasion and the impact of gender on the improvement of HRQOL-variables over time are explored. The first research question is examined by a Student's *t*-test assuming unequal variances and the latter by repeated measures ANOVA. Both of these analyses require completely observed data for all subjects, and complete-case analyses are performed to handle the missing values in the dataset, that is available-case analyses for the *t*-test and listwise deletion for the repeated measures ANOVA. The following sections describe the methods and results published by Gjeilo et al. [2008] by a redo of the analyses.

5.1 Student's *t*-test

The first main goal of the study is to examine possible differences between men and women prior to surgery and during the restitution time when evaluating Quality-of-Life variables. The statistical analysis to examine this question is a two-sample Student's *t*-test, which test if the mean of men are significantly different from the mean of women.

Normality is assumed by Gjeilo et al. [2008], and can be examined graphically by normal Q-Q plots, or by normality tests like the Kolmogorov-Smirnov test and the Shapiro-Wilks test. For interval-restricted variables we might expect some floor and ceiling effects, but according to Sullivan and Dagostino [1992] such violations of the normality assumption do not necessarily lead to biased results of the *t*-test. At each of the measuring points baseline, six and twelve months after surgery the subjects are assumed independent. A test of homogeneous variance (F-test) in the two groups male and female indicates that the variance are unequal for some of the variables at some occasions. Thus the two-sample Student's *t*-test for continuous variables assuming unequal variances (also known as *Welch-Satterthwaite t-test*) is suitable to examine differences in means for men and women at each time point.

5.1.1 Implementation in SPSS and Stata

The statistical calculations that are published in the article of Gjeilo et al. [2008] are performed using SPSS for Windows version 13.0 (SPSS Inc., Chicago, Illinois, USA). The redo of the complete-case analyses are performed using SPSS version 16.0 for Windows [2007] and Stata/SE version 10.1 for Windows [2007]. In SPSS the results are obtained by the command `Compare Means -> Independent-Sample T Test`, and p-values are found in the row *GH-scale, Equal variances not assumed* of the output. In Stata the script in Listing 5.1 provides results for the three measuring occasions for the variable general health.

Listing 5.1: Student's *t*-test in Stata for general health (GH)

```
1 ttest gh1 , by(kjønn) unequal
2 ttest gh2 , by(kjønn) unequal
3 ttest gh3 , by(kjønn) unequal
```

gh1 corresponds to the presurgery measurements, *gh2* is the six months follow-up measurements and *gh3* is the observations at twelve months follow-up. The *by()* option specifies the grouping on gender, and *unequal* implies that the variances are assumed unequal in the groups. Note: The data are assumed to be stored in wide format. This will be explained in Section 5.2.2.

5.1.2 Results, Student's *t*-test

The results for variables general health, bodily pain, social functioning and role emotional are displayed in Table 5.1. A table including all the variables analyzed in Gjeilo et al. [2008] can be found in Appendix Table A.2.

All estimates for both men and women increase during the study, thus the HRQOL-variables seems to be positively affected by the heart surgery. The first impression of the differences between men and women is the high proportion of significant p-values that means there are differences in means for the genders. Men have higher estimated means for all variables at all time points.

From Table 5.1 we see that general health is considered equal between men and women at baseline, but after surgery this variable turns out to be different among the genders. Men score significantly higher than women at six and twelve months after surgery. This may suggest that men have more effect of the operation and that they are in better shape after the surgery than women. The same pattern is found for the variables role emotional and bodily pain. For social functioning the pattern is different, the genders score different at baseline and six months after surgery, but at twelve months after surgery the difference is insignificant. It must be noticed that the p-values for role emotional at baseline of 0.054 and for social functioning at twelve months follow-up of 0.067 are slightly larger than the significance level of $\alpha = 0.05$, and should be handled by care. Such p-values indicate a weak difference between genders at the given occasions.

Table 5.1: Results from a two-sample Student's t-test assuming unequal variances for the variables general health (GH), bodily pain (BP), social functioning (SF) and role emotional (RE), grouping on gender. Significant p-values are marked as boldface.

	Baseline			Six months			Twelve months		
SF-36	Mean	SD	P-value	Mean	SD	P-value	Mean	SD	P-value
GH									
Male	64.9	19.7	0.490	72.0	22.1	**0.011**	71.9	21.6	**0.004**
Female	63.3	20.7		65.2	21.9		64.7	20.9	
BP									
Male	56.5	27.2	0.137	75.7	25.7	**0.004**	78.7	25.2	**0.002**
Female	52.3	26.4		66.4	27.1		68.9	26.9	
SF									
Male	73.2	24.8	**0.042**	84.7	22.6	**0.013**	86.3	20.5	0.067
Female	67.7	25.4		77.3	25.8		81.4	24.1	
RE									
Male	58.7	42.6	0.054	74.6	37.0	**0.006**	75.6	36.7	**0.001**
Female	49.1	45.9		59.8	44.1		58.0	43.1	

5.2 Repeated measures ANOVA

The second research question in the article of Gjeilo et al. [2008] is to examine a possible different improvement of mental and physical health during and after a heart surgery. The analyses of differences between genders at each measuring timepoint are already examined above, and the same grouping in gender is examined with respect to improvement of the HRQOL-variables over time. This improvement are analyzed by repeated measures ANOVA as explained in Section 3.3.

The first step when performing a repeated measures ANOVA is to set up the analysis model. This must include main effects gender and time, and the gender by time interaction, in addition to the subject identification to ensure the correlation structure is maintained in the analysis. The analysis model is given as

$$y_{hij} = \beta_0 + \beta_1 t_j + \beta_2 x_h + \beta_3 x_h t_j + v_{0i(h)} + \epsilon_{hij}, \tag{5.1}$$

where the parameters are interpreted as described in Section 3.3.1. Specially the parameter β_3 that explains the interaction effect is of interest in these analyses.

Additional variables may be added as explanatory variables. In the article of Gjeilo et al. [2008] the variables *age* and *marital status* are explored as potential covariates, but are not considered significant and thus omitted in the final model.

5.2.1 Data layout

The model above leads to a structure of the data as displayed in Table 5.2. The subjects are nested within groups and crossed with the time factor. Each subject belong to one of the genders, and is observed once at each time point.

Table 5.2: Design of data for the repeated measurements ANOVA, subjects grouped by gender and crossed with time points.

		Time point		
Gender	Subject	Presurgery	Six months follow-up	Twelve months follow-up
1	1	y_{111}	y_{112}	y_{113}
1	2	y_{121}	y_{122}	y_{123}
⋮	⋮	⋮	⋮	⋮
1	N_1	y_{1N_11}	y_{1N_12}	y_{1N_13}
2	1	y_{211}	y_{212}	y_{213}
2	2	y_{221}	y_{222}	y_{223}
⋮	⋮	⋮	⋮	⋮
2	N_2	y_{2N_21}	y_{2N_22}	y_{2N_23}

We are now familiar with the analysis model. Listwise deletion are performed to obtain a balanced design, that is all subjects have observations at the same measurement occasions. This leads to the same time intervals between measurements for all subjects. The assumption of multivariate normally distributed variables can be tested in Stata by the command `omninorm` [Baum and Cox, 2007]. Repeated measures ANOVA is robust to violations of the assumption of multivariate normally distributed variables [The University of Texas at Austin Statistical Services, 1997]. The assumption of compound symmetry can be examined by a F-test. As described in Section 3.3.1 the repeated measures ANOVA is robust to violations of homoskedasticity, thus we can carry out the analyses although the F-test reveals heteroskedasticity of the data. Sphericity can be examined by Mauchly's test [Mauchly, 1940] implemented in SPSS, but not in Stata.

The method of repeated measures ANOVA requires listwise deletion prior to the analysis. This excludes a higher fraction of the subjects than the *t*-test, since it disregards all information from patients who have one or two observed values, while the *t*-test applies available-case analyses that includes observations at the relevant measurement occasion independent of the response at the other time points.

5.2.2 Implementation in SPSS and Stata

The repeated measures ANOVA requires data to be stored in *long format*, that means each variable is represented in one column each, and a variable *time* keeps track of the measurement occasions. This results in *n* rows for each subject in the data matrix, and each row for a subject corresponds to a measurement occasion. The corresponding *wide format* has one row for each subject, and the variables are kept in *n* columns, one for each time point. The dataset

in Gjeilo et al. [2008] is structured in wide format, thus we have to transform it to long format prior to the repeated measures ANOVA. This is done in SPSS by the command `Data -> Restructure`. In Stata this is performed by the command `reshape` presented in Listing 5.2 below. The variable *time* starts counting at one, thus the time will be coded as $t = 1, 2, 3$. If we want the baseline time point to be zero to ease interpretation this can be done as displayed by the `replace` command on line two.

Listing 5.2: Reshape of the data from wide to long format

```
1  reshape wide mh vt bp gh sf pf rp re ht, i(patkey) j(time)
2  replace time = time-1
```

The command `General Linear Models -> Repeated measures` is used to perform the complete-case repeated measures ANOVA in SPSS. The variable *patkey* holds the identification of each subject and ensures the correlation structure of the data to be maintained. The output of this command gives many interesting results about the current variable, for example Levene's test of equality of error variances, Mauchly's test of sphericity and Box's test of equality of covariance matrices, in addition to the test results for between- and within-subjects effects. These results indicate that sphericity cannot be assumed, thus we have to use methods available to adjust the degrees of freedom for the tests of within-subjects effects. In SPSS the Greenhouse-Geisser and Huynh-Feldt adjustments are given. In Gjeilo et al. [2008] the Greenhouse-Geisser correction is preferred when testing the within-subject effects.

The same analyses are performed in Stata. Two commands are available for repeated measures ANOVA, that is `anova` and `wsanova`. This two gives the same results, but are different in structure and the generality of the commands. `anova` is the applied command here because it is most general. The script to obtain results from the repeated measures ANOVA in Stata is given in Listing 5.3 for the variable general health. First line performs listwise deletion and second line transforms the data from wide to long format. The command `anova` is given on the last line.

Listing 5.3: Repeated measures ANOVA for general health (GH)

```
1  drop if re1==. | re2==. | re3==.
2  reshape wide mh vt bp gh sf pf rp re ht, i(patkey) j(time)
3  anova gh time kjønn patkey time*kjønn, repeated(time) bse(patkey)
```

The option *repeated()* tells the program that we have correlated data and takes as input the within-subjects variable, here represented by the variable time. *bse()* takes the between-subjects effects as input, here as the identification number *patkey*. It automatically omits subjects without any measurements for the relevant HRQOL-variable when adjusting for lack of sphericity (Greenhouse-Geisser and Huynh-Feldt). The `anova`-command returns both original and adjusted p-values as default. Note: To perform valid analyses with the `anova`-command in terms of the assumptions given in Section 3.3.1, listwise deletion must be performed prior to analyses.

5.2.3 Results, repeated measures ANOVA

The `anova` command presented above gives exactly the same results with respect to p-values for the gender by time interaction as found in SPSS and the published article. Table 5.2.3 presents the results of the repeated measures ANOVA for the variables general health, bodily pain, social functioning and role emotional. The results for all SF-36 variables and health transition are given in Appendix Table A.

Table 5.3: Repeated measures ANOVA with Greenhouse-Geisser adjustment for non-sphericity for the SF-36-variables general health (GH), bodily pain (BP), social functioning (SF) and role emotional (RE). Analyses are based on complete-case subjects and grouped on gender. Significant p-values are marked as boldface.

SF-36	Number of fully observed patients	Repeated measures ANOVA P-value
GH		
gender by time	374	0.179
BP		
gender by time	401	**0.046**
SF		
gender by time	414	0.224
RE		
gender by time	348	**0.025**

The gender by time interaction for bodily pain and role emotional are found statistically significant, which means that the improvement of these scores are indicated to be different for men and women. The intercepts for the two remaining variables general health and social functioning are insignificant, thus there are no proven difference in improvement between genders for these variables.

5.3 Graphical presentation of subject samples

The complete-case subjects and available-case subjects at each time point form different samples of subjects from the original data. These samples are displayed as profile plots and tables of means, standard deviations and sample sizes for men and women separately. Profile plots display the means for men and women at each time point, with drawn lines between the occasions. These plots are informative when examining improvement pattern for the genders, and to visualize the differences in genders at each time point.

Student's *t*-tests base the analyses on available-case samples which correspond to all observed values in the profile plots below. The repeated measures ANOVA base upon complete-case samples displayed in the left profile plots. The omitted subjects are also displayed in profile plots for the four variables. This is to give information about the excluded information in

complete-case analyses. Note: The profile plots for all observations and omitted subjects do not base the estimation of means at each time point upon the same set of subjects, since some of the subjects included only have one or two observations.

Table 5.4: Mean of variable general health (GH) for men and women at the three measurement occasions presurgery, six and twelve months follow-up. 'CC' refers to the completely observed subjects, 'All' is all subjects included in the study and 'Omit' denotes the subjects omitted by the complete-case method.

GH	Time point	Mean			Standard deviation			Number of subjects		
		CC	All	Omit	CC	All	Omit	CC	All	Omit
Men	Pre	66.076	64.882	60.479	19.223	19.746	21.106	306	389	83
	6 mths	72.260	72.024	70.553	22.175	22.115	21.904	306	355	49
	12 mths	71.867	71.89	72.068	21.71	21.638	21.34	306	345	39
Women	Pre	65.238	63.322	59.802	20.633	20.679	20.576	68	105	37
	6 mths	66.684	65.182	59.176	21.851	21.916	21.781	68	85	17
	12 mths	67.765	64.697	55.627	19.519	20.874	22.52	68	91	23

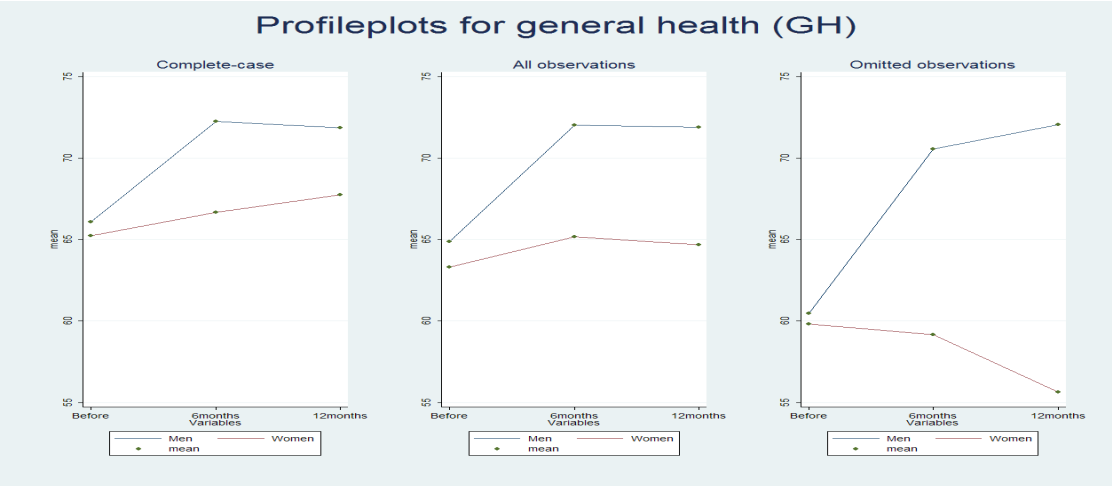

Figure 5.1: Plot of mean for GH for men and women at the three time points before, 6months and 12 months after surgery. The left plot is based on complete-case subjects, the plot in the middle is based on all observations of all subjects and the right plot is based on excluded subjects from complete-case methods.

Table 5.5: Mean of variable bodily pain (BP) for men and women at the three measurement occasions presurgery, six and twelve months follow-up. 'CC' refers to the completely observed subjects, 'All' is all subjects included in the study and 'Omit' denotes the subjects omitted by the complete-case method.

BP	Time point	Mean			Standard deviation			Number of subjects		
		CC	All	Omit	CC	All	Omit	CC	All	Omit
Men	Pre	55.938	56.473	58.617	26.595	27.196	29.565	325	406	81
	6 mths	75.985	75.659	72.455	25.235	25.668	29.846	325	358	33
	12 mths	78.717	78.732	78.897	24.94	25.21	28.541	325	354	29
Women	Pre	52.632	52.278	51.59	28.079	26.389	23.072	76	115	39
	6 mths	66.368	66.444	66.857	25.906	27.059	33.773	76	90	14
	12 mths	67.105	68.896	75.7	27.839	26.936	22.513	76	96	20

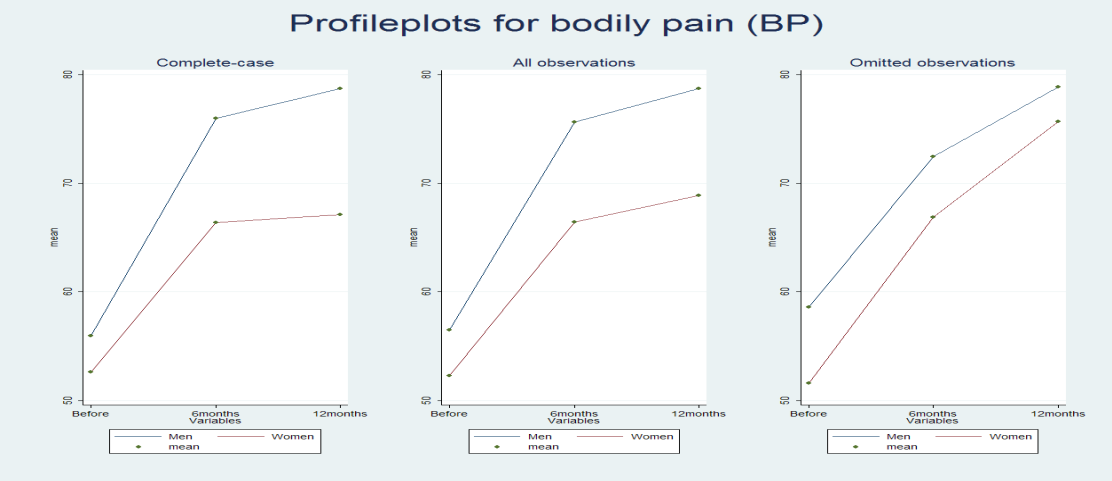

Figure 5.2: Plot of mean for BP for men and women at the three time points before, 6months and 12 months after surgery. The left plot is based on complete-case subjects, the plot in the middle is based on all observations of all subjects and the right plot is based on excluded subjects from complete-case methods.

Table 5.6: Mean of variable social functioning (SF) for men and women at the three measurement occasions presurgery, six and twelve months follow-up. 'CC' refers to the completely observed subjects, 'All' is all subjects included in the study and 'Omit' denotes the subjects omitted by the complete-case method.

SF	Time point	Mean			Standard deviation			Number of subjects		
		CC	All	Omit	CC	All	Omit	CC	All	Omit
Men	Pre	73.69	73.179	70.775	24.539	24.832	26.215	334	405	71
	6 mths	85.442	84.692	76.613	22.217	22.559	24.946	334	365	31
	12 mths	86.789	86.348	80.208	20.434	20.498	20.824	334	358	24
Women	Pre	69.375	67.717	63.929	25.924	25.38	24.02	80	115	35
	6 mths	75.781	77.31	87.5	26.336	25.806	19.943	80	92	12
	12 mths	81.094	81.38	82.813	25.47	24.129	16.378	80	96	16

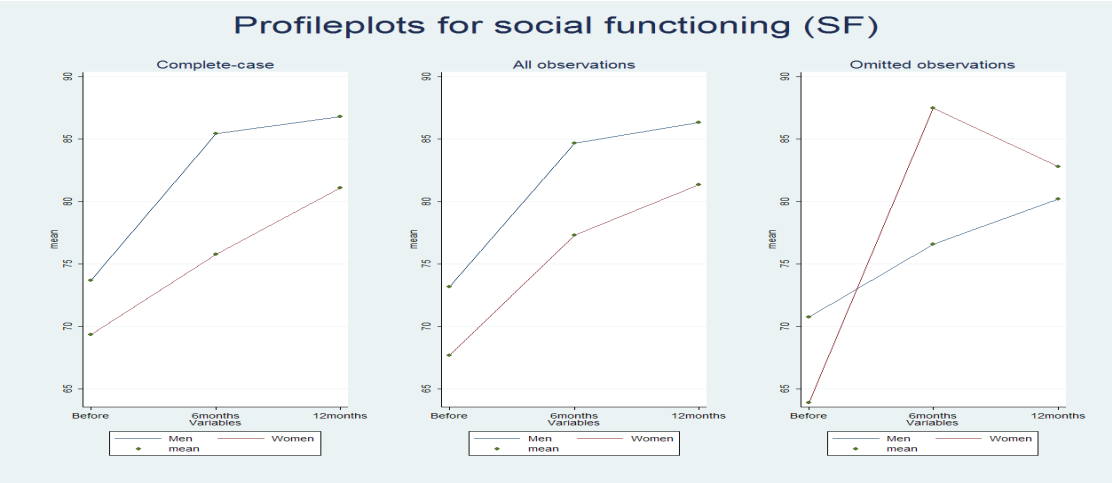

Figure 5.3: Plot of mean for SF for men and women at the three time points before, 6months and 12 months after surgery. The left plot is based on complete-case subjects, the plot in the middle is based on all observations of all subjects and the right plot is based on excluded subjects from complete-case methods.

Table 5.7: Mean of variable role emotional (RE) for men and women at the three measurement occasions presurgery, six and twelve months follow-up. 'CC' refers to the completely observed subjects, 'All' is all subjects included in the study and 'Omit' denotes the subjects omitted by the complete-case method.

RE	Time point	Mean			Standard deviation			Number of subjects		
		CC	All	Omit	CC	All	Omit	CC	All	Omit
Men	Pre	60.781	58.679	52.667	41.738	42.558	44.49	286	386	100
	6 mths	75.408	74.603	70.0	36.586	37.009	39.412	286	336	50
	12 mths	76.224	75.645	72.333	36.112	36.736	40.351	286	336	50
Women	Pre	58.065	49.065	36.667	46.237	45.949	43.023	62	107	45
	6 mths	58.602	59.839	63.492	45.04	44.099	42.038	62	83	21
	12 mths	58.602	57.963	56.548	42.864	43.074	44.291	62	90	28

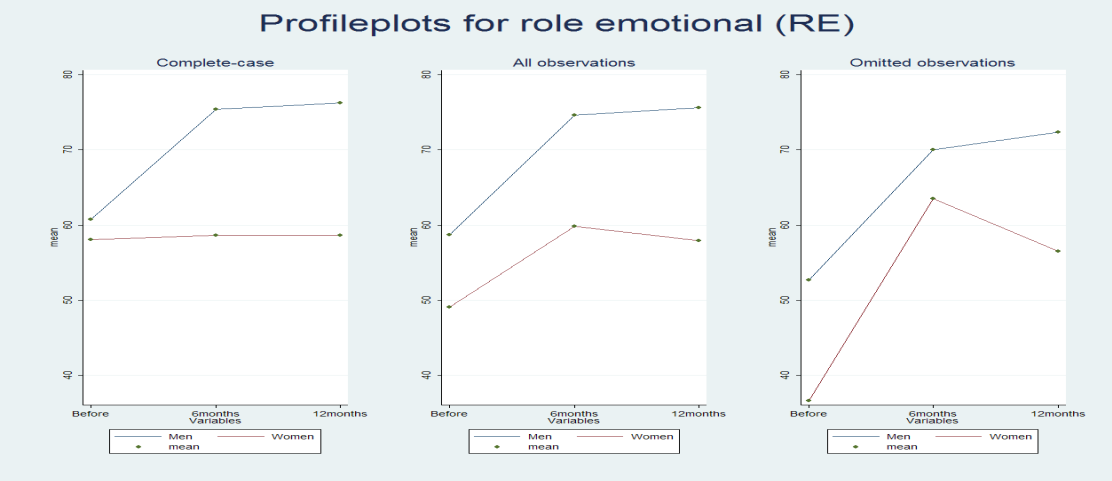

Figure 5.4: Plot of mean for RE for men and women at the three time points before, 6months and 12 months after surgery. The left plot is based on complete-case subjects, the plot in the middle is based on all observations of all subjects and the right plot is based on excluded subjects from complete-case methods.

5.4 Interpretation of profile plots

For variable general health the mean for both men and women presurgery are considerably affected by the listwise deletion of subjects, the proportion of omitted subjects is large for both groups and the mean of omitted subjects are smaller than for complete-case subjects. Thus we induce a higher intercept mean for both genders by applying listwise deletion. The mean for the three subject samples for men at 6 months and 12 months follow-up are approximately equal for the samples, but for women these means still remain much smaller for omitted females than for complete-case females. The number of women with non-responses is not of the same size as for baseline, thus the influence of the means calculated on base of all subjects are not of equal magnitude as for the presurgery occasion. The standard deviations for both genders at all three time points are similar in magnitude among the three samples of subjects.

We have seen that the means for general health are altered when grouping the subjects in samples of complete-case and omitted subjects. The plots for the different samples are not as different for the variable bodily pain. The mean for men is slightly higher for omitted subjects compared to complete-case subjects presurgery, while the mean for women at 12 months follow-up is higher for the omitted subjects than for the complete-case subjects. The standard deviations are in general higher for omitted male subjects compared with complete-case subjects. For women this is not always the case, prior to surgery and at 12 months follow-up the omitted subjects give less variance than the complete-case subjects.

The profile plots of social functioning in Figure 5.3 reveal that the omitted female subjects with observations at 6 months follow-up have an extremely high mean value compared with the complete-case subjects. The standard deviation for these subjects are also smaller than for the complete-case female subjects, but it is important to notice the size of this group, only 12 of the omitted subjects have observed values at this time point. The means for male subjects that are omitted are much smaller than the complete-case subjects at 6 and 12 months follow-up. The number of subjects observed at these occasions are very small, thus the means for all subjects are only slightly altered compared with the means for complete-case subjects.

The last variable role emotional is the variable with most significant gender by time interaction from the complete-case repeated measures ANOVA, but with insignificant intercept effect from the imputation analyses and the analysis by curvilinear MRM. This variable is therefore of special interest when looking at observations of the omitted subjects and their affection on the analysis results based on all data. The plots in Figure 5.4 draws the attention to the female subjects, where the presurgery mean of omitted subjects are much smaller than for complete-case subjects. Among the 59 omitted subjects as many as 45 subjects have observed responses at baseline. Compared to the sample of complete-case women that consists of 62 subjects we observe that a considerable portion of the information of female patients are excluded in the complete-case analysis. The mean for all observed women presurgery is remarkably altered by the inclusion of omitted subjects, from 58.1 to 49.1. This is also commented in the article by Gjeilo et al. [2008]. The means at six and twelve months follow-up are more similar for the samples for both men and women.

6 Imputation analyses

We have now presented the published results obtained by complete-case analyses. As proposed in Section 2.3 the methods of expectation maximization (EM) and multiple imputation (MI) are able to handle data with non-response if the missing data can be assumed ignorable. The method of EM is performed in SPSS [2007] and R [2008], while MI is performed mainly in Stata [2007].

6.1 Expectation Maximization

Expectation maximization (EM) is an general algorithm that is used for multiple purposes and is relatively easy to implement, and is therefore incorporated in most statistical software (for example SPSS, R, NORM and SAS). To demonstrate the use of this approach we explain how the analyses performed in Gjeilo et al. [2008] are obtained by EM in both SPSS and R. First, we must settle a joint multivariate model including all variables in the analyses. This model consists of an identification key for each subject, gender and the nine HRQOL-variables.

The missing values analysis (MVA) routine in SPSS is found in the Analyze menu, and we choose EM as estimation method. The number of maximum iterations must be increased from the default number of 25 iterations, since the number of variables with missing values is large. von Hippel [2004] was critical to the MVA module in SPSS on the basis of an incorrect implementation of the residual variation for each imputed value. This random disturbance term should be included in imputation to reflect uncertainty associated with the imputation. The absence of the random error term results in a deterministic imputation process, so imputed datasets based on the same imputation model for the dataset are identical. This violates the assumptions of imputation by the EM algorithm.

The methods of conditional single imputation are described as deficient in Schafer and Graham [2002], and is partitioned in conditional *mean* imputation and conditional *distributional* imputation. The difference of these two methods is the inclusion of residual variation in the latter. It must be emphasized that single imputation using these two methods can give biased standard errors of estimates under the MAR assumption, while the EM algorithm provides theoretically unbiased estimates if implemented properly.

Another software package that supports single imputation by the EM algorithm is R. The `norm` package [Novo and Schafer, 2006] is developed for this purpose, and yield many helpful tools when working with the EM algorithm. Prior to the analyses we must examine the data with respect to the assumption of normal distributed variables. This may be done in Stata without too much effort (explained in Section 6.2.2), before exporting the datafile to R. As we will see in Section 6.2.3 the arc sine transformation [Box et al., 2005] is suitable to achieve a more normally distribution of variables, additional to the feature of restricting imputed values to the relevant intervals. This transformation of all HRQOL-variables are performed before the imputation process, and both original and imputed values are back-transformed after the imputation (the latter performed in R).

6.1.1 Implementation of EM in R

An extract of the script to perform EM imputation in R is displayed in Listing 6.1.

Listing 6.1: R script to perform imputation by the EM algorithm

```
1  library(norm)
2  mat <- read.table('originalfile.txt', header=TRUE, sep=",", na.strings=".")
3  s <- prelim.norm(mat)
4  thetahat <- em.norm(s)
5  getparam.norm(thetahats,thetahat,corr=TRUE)
6  rngseed(7654321)
7  ximp <- imp.norm(s,thetahat,mat)
```

The first line specifies that the norm package must be read (only the basic packages of R are stored in memory when starting the software, and additional packages must be specified by the function `library()`). Further the datafile exported from Stata is read by `read.table()`, where observations are separated by a comma, and missing values are given as periods. The `prelim.norm()` function performs preliminary manipulations of the data matrix and is the input object of the functions on the consecutive lines. The EM algorithm itself is executed by calling `em.norm()`, and provides a vector of maximum likelihood-estimates of the normal parameters on a transformed scale and in packed storage. To achieve these estimates the function `getparam.norm()` is helpful. Before imputation we must initialize a random number generator seed (function `rngseed()`), and imputation is performed by the `imp.norm()` function on the last line of the script. For further explanation of the functions in the `norm` package we recommend the help file of R [Novo and Schafer, 2006].

After imputation we have obtained a balanced dataset including all the 534 patients with values of all variables at all occasions. The HRQOL-variables must be transformed to the original 0 to 100 interval to ease the interpretation of the results and to make comparisons with other analysis methods possible. The analyses described in Chapter 5 can now be repeated, but without listwise deletion

6.1.2 Results, EM analyses

The results of Student's *t*-test assuming unequal variances and the repeated measures ANOVA for the dataset containing observed and imputed values after EM are computed using the func-

tions `t.test()` and `aov()`, and the script for these analyses are found in Appendix Section B. The results are displayed in Table 6.1.

Table 6.1: Estimates of means and p-values for the two-sample Student's t-test assuming unequal variances and p-values for the gender by time interaction for the repeated measures ANOVA after imputing missing values by the EM algorithm. Results displayed for the variables general health (GH), bodily pain (BP), social functioning (SF) and role emotional (RE), grouping on gender. Significant p-values are marked as boldface.

	Baseline		Six months		Twelve months		ANOVA
SF-36	Mean	P-value	Mean	P-value	Mean	P-value	P-value
GH							
Male	65.12	0.171	71.06	**0.002**	70.50	**0.005**	0.071
Female	62.20		63.76		64.26		
BP							
Male	56.38	0.114	75.20	**0.005**	78.12	**<0.001**	0.096
Female	52.02		67.34		67.79		
SF							
Male	73.00	0.055	84.01	**0.004**	85.70	**0.049**	0.432
Female	67.96		76.52		81.15		
RE							
Male	58.61	**0.029**	71.31	**<0.001**	73.04	**<0.001**	0.503
Female	48.47		57.27		57.91		

The function `t.test()` in R gives no estimated standard deviations for the means, thus we must base out results on the estimated mean values and the corresponding p-values for these analyses.

The first impression is that most the estimates are approximately equal to the original complete-case estimates. The changes of largest magnitude are found in estimates for men and women at six months after surgery for role emotional, and for men at twelve months follow-up for the same variable. The means decrease when analyzing the data after EM imputation. The corresponding p-values for the differences between men and women at these two occasions are highly significant. The p-values for the baseline differences of social functioning and role emotional are altered from the complete-case p-values, but these changes are relatively unessential when interpreting the results. The same yields for the difference at twelve months follow-up for social functioning. Changes in p-values can be due to different estimates for the genders, and the fact that the sample sizes are increased by including all subjects. By imputing missing values we ensure that the analyses are based on more of the information from the dataset, thus the statistical power increases and p-values decreases.

We now want to focus on the analysis of improvement for men and women represented by the gender by time interactions in repeated measures ANOVA. EM imputation leads to insignificant p-values for the interactions regarding all the variables. Especially the estimate for role emotional is altered from quite significant in the complete-case analysis (p-value of 0.025) to

clearly insignificant in the EM analysis (p-value of 0.503). The variable general health on the other hand obtains a decreased p-value compared to the complete-case analysis. Also the p-value for the interaction for bodily pain is in the grey area what significance concerns, modified from significant in the complete-case analysis. A p-value less than 0.10 should be handled by care, it is not necessarily one single way to interpret such p-values.

Profile plots of the variables general health and role emotional are found in Figures 6.1 and 6.2 respectively. We see that the improvement patterns of role emotional for men and women are more similar in the right plot after EM imputation than for the complete-case subjects in the left plot. These profile plots are in accordance with the results found in Table 6.1. The profile plots for general health seems to be approximately equal, but the change in p-value for the interaction may be due to increased statistical power induced by the EM imputation.

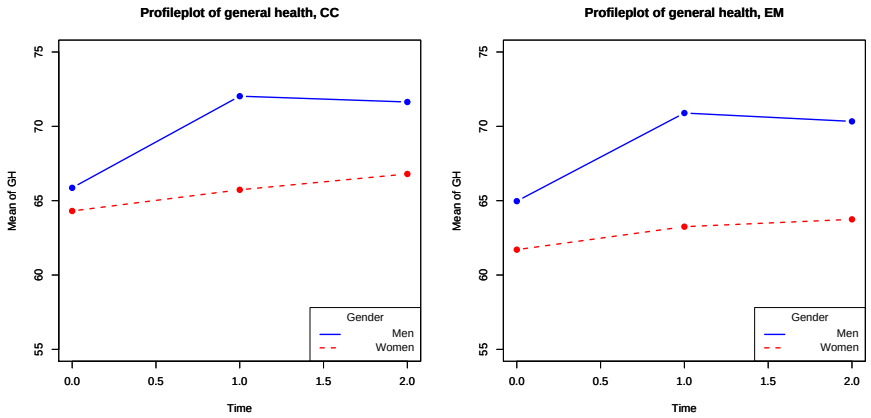

Figure 6.1: Profile plots for the variable general health (GH) based on the complete-case subjects (denoted CC) and expectation maximization imputed datasets (denoted EM).

6.2 Multiple imputation

The theory of multiple imputation (MI) is presented in Chapter 2.3.2, and will be described in a more practical setting in this section. MI is performed on the dataset of Gjeilo et al. [2008]. MI makes use of all information available in the observed data, and meets the criteria for methods to analyze data with missing values assumed MCAR or MAR described in Section 2.1.6.

A multiple imputation analysis is performed in two steps,

1. Imputation of the missing values in the original dataset that leads to a total of m imputed datasets. Each dataset is a plausible version of the original dataset if all values were observed.

2. Analyze of each of the datasets and combination by Rubin's rules to obtain analysis results.

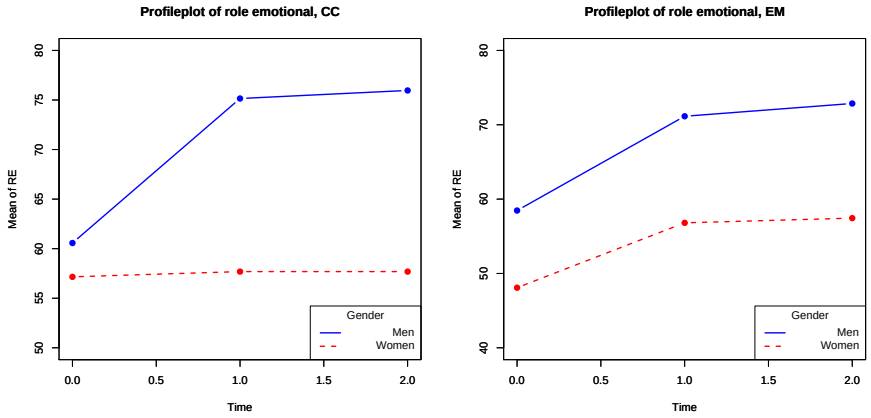

Figure 6.2: Profile plots for the variable role emotional (RE) based on the complete-case subjects (denoted CC) and expectation maximization imputed datasets (denoted EM).

The first step of multiple imputation of the dataset is described in the Sections 6.2.1 to 6.2.8. The analysis step is presented in Sections 6.2.5 to 6.2.10.

6.2.1 Imputation by MICE

Multiple imputation is performed in Stata [2007] by the command `ice`. This command creates a user-specified number of datasets where the missing values are replaced with imputed values [Royston, 2005]. It performs multiple imputation by the MICE procedure, where MICE stands for Multiple Imputation by Chained Equations. The missing values are imputed by the method of switching regression, an iterative multivariate regression technique introduced by Buuren and Oudshoorn [2000] for S-plus, and `ice` implements MICE for Stata.

In the Theory Section 2.3.2 of MI we find that imputations are obtained as draws from a joint multivariate distribution including all the variables in the imputation model and the parameters of the imputation model. It is seldom possible to obtain this joint distribution in closed form, and in these situations iterative algorithms like MICE are useful. The switching regression makes use of Gibbs sampler iteration algorithm to perform the imputation.

MICE assumes existence of a joint multivariate distribution, from which conditional distributions for each variable with missing values can be derived. Thus the complex multivariate problem can be parted into easier univariate problems. Let $X_1, X_2, ..., X_k$ be the set of variables in the imputation model where some of the X's have missing values. The 'chained equations' method for imputing data described in Carpenter and Kenward [2007] are similar to the the technique of MICE. First, initialize missing values, for example by the mean of the observed values or randomly drawn values of the observations. Then draw imputations

- X_1^t from f($X_1^t|X_2^{t-1}, X_3^{t-1}, ..., X_k^{t-1}$)
- X_2^t from f($X_2^t|X_1^t, X_3^{t-1}, ..., X_k^{t-1}$)
- ...
- X_k^t from f($X_k^t|X_1^t, X_2^t, ..., X_{k-1}^t$)

Repeat the above as a loop until convergence, that is typically ten to twenty times. Then repeat the loop a further *m* times to obtain the imputed datasets.

Modern Markov Chain simulation methods often need up to thousands of iterations to obtain a sample of $m = 20$ imputed datasets, while the 'switching regressions' need less. The reason is that 'switching regressions' draws independent samples for each variable, while Markov Chain methods generates dependent samples and therefore need to iterate and reject many samples to achieve approximately statistical independent imputation draws. The number of iterations for 'switching regressions' increases with the amount of unobserved data.

van Buuren et al. [2006] write the following about their implementation of MICE

"The theoretical weakness of this approach is that the specified conditional densities can be incompatible, and therefore the stationary distribution may not exist. It appears that, despite the theoretical weakness, the actual performance of conditional model specification of multivariate imputation can be quite good, and therefore deserves further study."

Further the authors advise to perform MI with an extra caution to the imputations obtained. The method does not guarantee that all imputations are drawn from the joint multivariate normal distribution, rather from other distributions that the method may converge to in case the joint distribution don't exists. A technique to examine the outlier datasets are explained in Section 6.2.10. Note: Carpenter and Kenward [2007] did not recommend the method presented above, based on the uncertainty in the convergence to the joint multivariate normal distribution.

6.2.2 The imputation model

To perform multiple imputation we need to make a selection of variables as a base for the imputation model. This model includes variables with missing values, variables that can predict the missing values and variables that can explain the reason for non-response, in addition to the variables in the analysis models. First we look at these variables contained in the analysis models. All variables except the subject identification number patkey, gender and age have missing values, distributed at all three measurements occasions.

MI assumes all variables in the imputation model to be multivariate normally distributed, and we want to examine whether the variables in the dataset meet this assumption. Graphical presentation of the variables can be useful to examine the behavior of the variables. Histograms of general health, bodily pain, social functioning and role emotional are displayed in Figures 6.3 to 6.6.

General health has 21 levels of theoretical possible values, uniformly distributed on the interval 0 to 100. Histogram of the variable are displayed in Figure 6.3 for the three occasions. At baseline a Gaussian shape can be seen, and also at six and twelve months after surgery this bell shape is present, though more right-skewed than for the first measuring point.

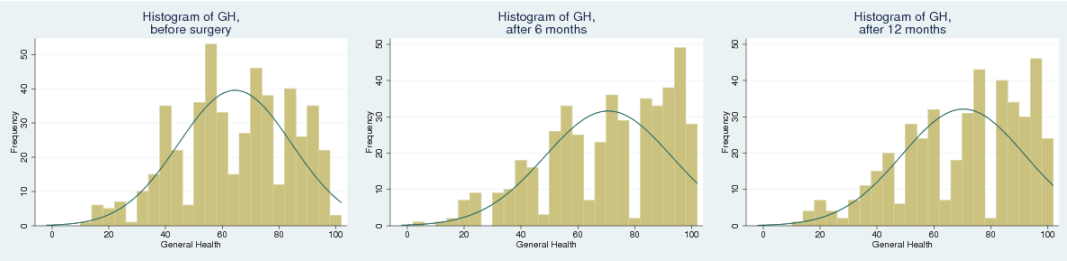

Figure 6.3: Histograms of the observed values of the variable general health (GH) before surgery and at six and twelve months follow-up.

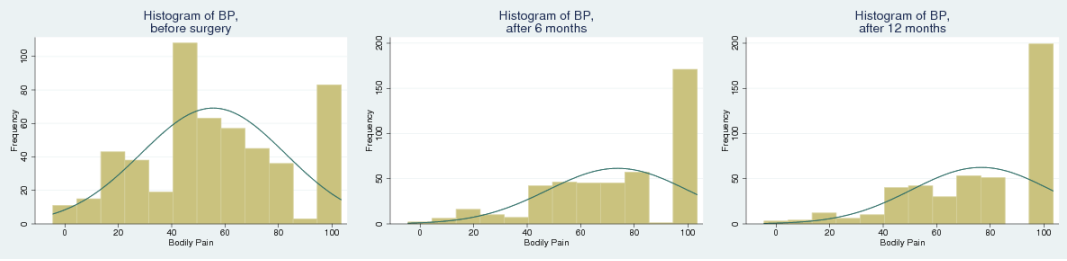

Figure 6.4: Histograms of the observed values of the variable bodily pain (BP) before surgery and at six and twelve months follow-up.

Histograms for the variable bodily pain is displayed in Figure 6.4. This variable has eleven levels of theoretical possible values on the interval 0 to 100. Before surgery the data looks normally distributed with a peak at the middle of the interval and a considerable ceiling-effect. At six and twelve months after surgery this ceiling-effect is still present and dominate the histogram, while the bell shape of an Gaussian variable is suppressed.

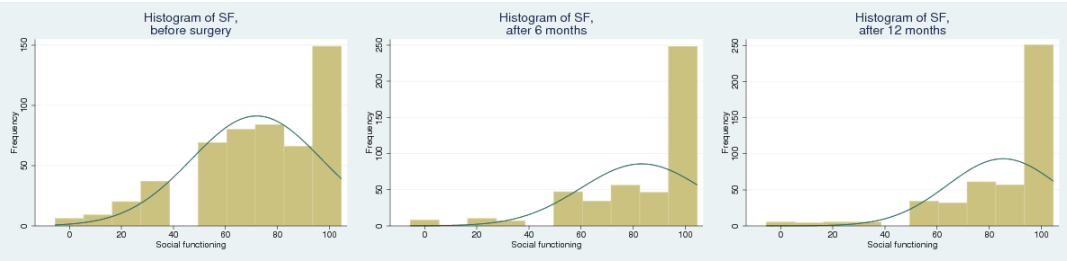

Figure 6.5: Histograms of the observed values of the variable social functioning (SF) before surgery and at six and twelve months follow-up.

Social functioning has nine ordered levels of values. By looking at the histograms in Figure 6.5 we see that the variable possibly violates the assumption of multivariate normally distributed data by the clear ceiling effect at all time points.

Role emotional is distributed on four values, and does not seem convincingly bell shaped

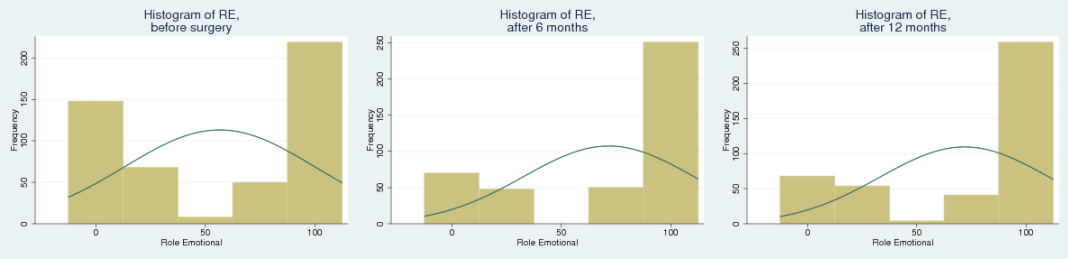

Figure 6.6: Histograms of the observed values of the variable role emotional (RE) before surgery and at six and twelve months follow-up.

from the histograms in Figure 6.6. For the histogram of the presurgery values and at twelve months follow-up a total of five levels of the variable can be found. This is due to the mean value imputation described in Section 4.1, and this feature is also found for the other variables.

6.2.3 Constraints of imputed values

As we can see from the histograms above the variables are not convincingly normally distributed, thus the assumption of multivariate normally distributed variables is possibly violated. There are several ways to handle this in such a way that we still can apply the method of multiple imputation.

Another aspect of the imputation process that must be considered is the possible imputation values for each variable. All SF-36 variables and health transition are restrained to intervals and we expect the unobserved data to take values on the same intervals. Dragset et al. [2008] examine three possible alternatives to restrain the imputed values to their intervals, and if possible influence the variables to behave more normally distributed. These methods can be summarized as follows

- *Truncation,* the method of restraining imputations after the imputation process,
- *Ordinal logistic regression,* which leads to restraints during the imputation process, and
- *Transformation,* the method of modifying data before the imputation process.

Truncation is a method of replacing the imputed values that exceeds the interval of each variable. This means that the imputed values are manually altered after imputation which may induce biased analysis results. It may also lead to artificial enhanced floor and ceiling effects which is not desirable. Dragset et al. [2008] concluded that this method is not preferable. Variables with a handful possible levels may be treated as ordinal (or nominal) variables and with that avoid the assumption of normality. The last approach to restrict the imputed values is transformation prior to the imputation process. With a suitable transformation we may obtain data that behave more normally distributed, and at the same time ensures the imputed values to take values on the intervals.

Ordinal logistic regression

The variable social functioning takes nine levels of possible values. The histograms in Figure 6.5 does not reveal a convincing bell shaped distribution for the variable. This may violate the assumption of multivariate normality of MI, thus imputation of this variable must be handled with caution. Variables that take only a few levels of values can be treated as nominal or ordinal variables without loosing too much statistical power. This affects the imputations to take values found among the observed values, which means that all imputed values are restricted to the respective intervals.

This also applies for other variables with few levels of observations, that is role emotional, role physical and health transition. The implementation of ordinal logistic regression for the specified variables are explained in Section 6.2.4.

Transformation of variables

Transformation of variables prior to multiple imputation is a technique to reduce violations of the normality assumption of the data. We can transform the data to obtain a distribution of the variables that are more Gaussian shaped, impute the missing values and then back-transform the variables after the imputation process. The back-transform is necessary to obtain comparable estimates for the analyses.

The desired transformation formula should influence the data to behave approximately normal distributed. Arcsine transformation [Box et al., 2005] is a variance stabilizing transformation for binomial proportions of the form y/n, where y is the original variable and n is the upper limit of the interval of y. The transformation formula is given as

$$x = \arcsin\left(\sqrt{\frac{y}{n}}\right), \tag{6.1}$$

where x is the transformed of the original variable y and that holds the characteristic of a more bell shaped distribution than y.

The arcsine transformation is preferable when the variables take values close to the limits of the intervals. Applying the arc sine of the variables leads to stretching the tails of the distribution, thus we can transform the variables by the arcsine transformation to imply a more Gaussian shape of the distributions than for the original variables.

An additional advantage of the arcsine transformation when applied to data that are limited to intervals is that it constrains imputed values to these interval. Thus we avoid the problem of imputing values outside the original intervals. Consider for example the variable general health that holds original values on the interval 0 to 100. Higher scores indicate better general health, and a score of zero is interpreted as death. If we get an imputed value below 0 it is not obvious how to interpret this value.

Implementation of the transformation in `ice` can be summarized as follows

- generate transformed values of the variables prior to imputation,
- include only the transformed version of the variables in the imputation process, and

- back-transform the imputed variables to obtain imputed values for the original format variables.

This approach is presented in Section 6.2.4 below. An alternative way to implement transformation with `ice` is to back-transform the variables directly in the imputation process:

- generate transformed values of the variables prior to imputation,
- include both original and transformed variables in the imputation process, and
- specify imputation equations for all variables that are to be imputed. Note: The original continuous variables (not transformed) should not be imputed by the imputation process, nor included in any of the imputation equations.
- Back-transform the imputed variables before finishing the imputation process.

These two approaches to impute transformed variables by multiple imputation generates the same imputation process, but differs in the implementation in Stata. The effort and time spent may vary for different analyses, and the easiest of the approaches are recommended.

Histograms of the transformed variables general health and bodily pain are displayed in Figures 6.7 and 6.8 respectively. We observe that the distributions of general health look more Gaussian shaped, and the skewed trend of the follow-up occasions are not as evident as for the original variable.

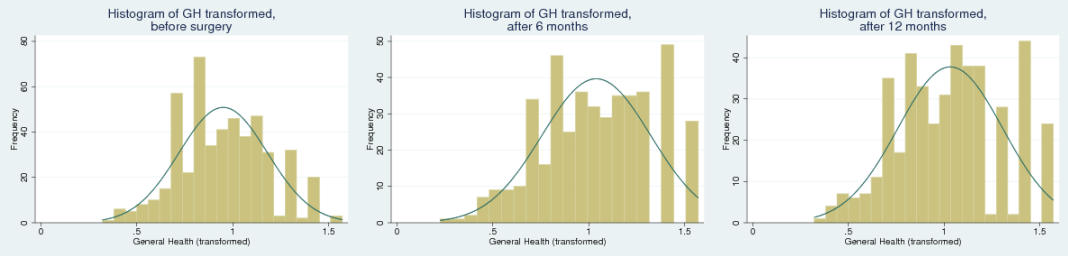

Figure 6.7: Histograms of the transformed values of the variable General Health (GH) before surgery and at six and twelve months follow-up.

The histograms for bodily pain are also ameliorated with respect to the bell shape, but the ceiling effect is still overwhelming for the follow-up measurement occasions.

6.2.4 Implementation of `ice` in Stata

Multiple imputation is performed in Stata by the command `ice`, as stated above. This command imputes *m* datasets that are stored in a new file *Impfile* together with the original dataset. A variable *_mi* keeps track of the subjects in each dataset, and another variable *_mj* holds the information to separate the datasets. The script displayed in Listing 6.2 below is based on the imputation model including each patients identification number, gender and age at the time of

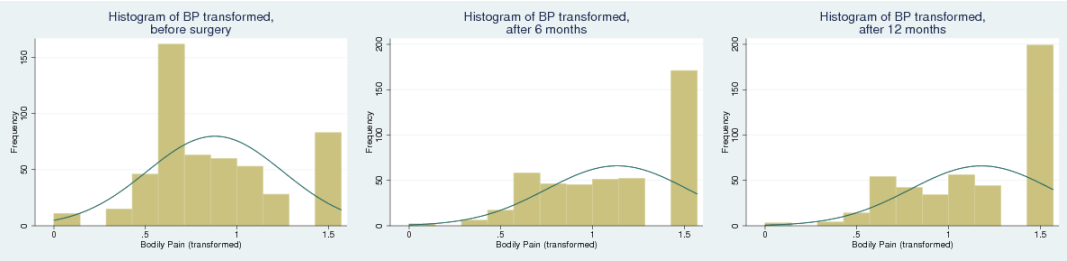

Figure 6.8: Histograms of the transformed values of the variable Bodily Pain (BP) before surgery and at six and twelve months follow-up.

Listing 6.2: Multiple imputation of dataset from Gjeilo et al. [2008]

```
1  ge gh1t = asin(sqrt(gh1/100))
2  ge gh2t = asin(sqrt(gh2/100))
3  ge gh3t = asin(sqrt(gh3/100))
4
5  ice patkey gh1t gh2t gh3t pf1t pf2t pf3t sf1 sf2 sf3 rp1 rp2 rp3 re1 re2 re3 mh1t
       mh2t mh3t vt1t vt2t vt3t bp1t bp2t bp3t ht1 ht2 ht3 ALDER kjønn sivilstand1
       sivilstand2 sivilstand3 using Impfile, m(20) genmiss(mis) cmd(sf1 sf2 sf3 rp1
       rp2 rp3 re1 re2 re3 ht1 ht2 ht3: ologit)
6
7  replace gh1=100*(sin(gh1t))^2 if misgh1== 1
8  replace gh2=100*(sin(gh2t))^2 if misgh2== 1
9  replace gh3=100*(sin(gh3t))^2 if misgh3== 1
```

surgery, and all HRQOL-variables, health transition and marital status at the three measurement occasions.

Data must be arranged in wide format prior to imputation. The variables social function, role physical, role emotional and health transition are imputed by ordinal logistic regression specified by the option *cmd()*. Marital status is a dichotomous variable and thus handled by logistic regression as default. The remaining of the HRQOL-variables are transformed by the arcsine transformation formula prior to the imputation as shown for general health on line one to three. These transformed variables are included in the variable list instead of the original variables as explained in Section 6.2.3, together with the subject identification number patkey, gender, marital status and the four ordinal variables in original format.

There are several more options specified in the script above. *genmiss()* creates an indicator variable for each variable with non-response in the imputation model, where one refers to an imputed value and zero is an observed value. *m()* declares the desired number of imputed datasets and *dryrun* displays the imputation equations without performing imputation (left out in this script).

After the imputation process is finished, the back-transformation of the continuous variables are performed in the newly created *Impfile*, the datafile in which the original and imputed datasets are stored. This file are now ready for analyses and combination by Rubin's rules as described in Section 2.3.5.

An alternative procedure to obtain imputed datasets for the transformed variables is by in-

cluding both original and transformed HRQOL-variables in the variable list, and back-transforming the imputed values directly in `ice`. This is done by specifying two extra options *passive()* and *eq()*. *passive()* allows a variable to be computed from both complete and partially imputed variables, and ensures that this variable is not imputed. One must now specify the imputation equations so that the imputation process is based on the transformed HRQOL-variables and not the original (or both). This is done by the option *eq()*. This option can be comprehensive if many variables are included in the imputation model. The two alternatives to perform multiple imputation yield the same imputation process, thus we are free to choose the method that is easiest to implement.

6.2.5 T-test after multiple imputation

We have seen how multiple imputation is performed for the dataset of Gjeilo et al. [2008]. The next step is to analyze these imputed datasets with respect to the differences in mean for men and women at each measuring occasion. The Stata command `mim` is designed to combine such datasets by Rubin's rules for several analysis methods, but does not support Student's *t*-test assuming unequal variances. If the variances were assumed equal a simple linear regression analysis could have been applied.

Student's *t*-test assuming equal variance

The two-sample Student's *t*-test with equal variance gives the same results as a linear regression with the HRQOL-variable as dependent variable y_i and gender as covariate x_i, given as

$$y_i = \beta_0 + \beta_1 x_i.$$

Linear regression is one of the supported analyses in `mim`, thus it is easy and time-saving to implement compared to implementing the analysis by Rubin's rules ourselves. These analyses serve for several purposes. First, we can examine the number of imputations to decide if *m* is large enough to achieve accurate estimates. Second, we can get an impression of the fraction of missing information due to the unobserved values in the dataset. Third, the approximate analysis can give an impression of the results prior to the correct and more time-demanding implementation of the *t*-test assuming unequal variances. The first two features are further described in Section 6.2.10.

The two groups men and women have different variances at the different measuring occasions, and this violates the assumption of linear regression about equal variances in the groups. To compute the *t*-test for unequal variances we have to combine estimates of the test statistic in each dataset and combine these by Rubin's rules "manually". This is achieved by exporting the imputed datasets file Impfile.dta to a .txt-file and compute the desirable estimates in the statistical software R.

Student's *t*-test assuming unequal variances

To perform a Student's *t*-test assuming unequal variances by Rubin's rules, some notation must be introduced. First, we assume the outcome variables to be normally distributed, displayed as

$$X_M \sim \mathrm{N}(\mu_M, \sigma_M^2),$$
$$X_W \sim \mathrm{N}(\mu_W, \sigma_W^2),$$

where X_M and X_W are the relevant HRQOL-variable for men and women, respectively. Estimators for μ_M and σ_M are given as

$$\hat{\mu}_M = \bar{X}_M = \frac{1}{n_M}\sum_{i=1}^{n_M} X_{Mi},$$
$$\hat{\sigma}_{X_M} = S^2_{X_M} = \frac{1}{n_M}\sum_{i=1}^{n_M}(X_{Mi} - \bar{X}_M)^2.$$

The estimators for women are identical to the above, but based on the values for women. n_M is the number of men included in the study, and n_W is the number of included women. Note: These sizes are not necessarily equal for MI and complete-case analyses, since subjects with missing values are omitted in the latter.

Further, we want to test the null hypothesis H_0: $\bar{X}_M - \bar{X}_W = 0$, that is the hypothesis of no difference between the genders. These differences in means for men and women are the estimates that we must combine with Rubin's rules. We write the difference in each imputed dataset as

$$\hat{Q}_i = \bar{X}_{Mi} - \bar{X}_{Wi},$$

where the estimator $\hat{Q}_i$ for imputed datasets $i = 1, 2, \cdots, m$ is the parameter quantity that we want to examine. We apply Rubin's rule for the parameter quantity in Equation (2.5)) to obtain

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i.$$

To estimate the variance of $\bar{Q}$ we must find the within-imputation variance and between-imputation variance and combine these as given in Equation (6.2). In the calculations below we assume independent observations for men and women. This is an approximation for the imputed values, since all values for all subjects (both men and women) are used as predictors for the imputed values in the imputation process. Thus some dependency between imputed values for women and values for men might have been induced, and vice versa. The imputation process is complicated and the covariance structure is difficult to model (if possible). As an approximation we assume values at each time point to be independent.

The variance of $\hat{Q}_i$ in each imputed dataset is calculated as

$$U_i = \frac{1}{m}\sum_{i=1}^{m}(\sigma^2_{\bar{X}_M} + \sigma^2_{\bar{X}_W}) = \frac{1}{m}\sum_{i=1}^{m}(\frac{\sigma^2_{X_M}}{n_M} + \frac{\sigma^2_{X_W}}{n_W}).$$

Within-imputation variance is given as

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} U_i,$$

and between-imputation variance is given as

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})^2.$$

The total variance for $\bar{Q}$ is expressed as

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B. \tag{6.2}$$

From Equation (6.2) we see that the total variance for $\bar{Q}$ is computed as a sum of the within-imputation and between-imputation variance, where the latter is weighted by the number of imputed datasets *m*. For the dataset of Gjeilo et al. [2008] we find the within-imputation variance $\bar{U}$ to be large compared to the between-imputation variance B, thus the variance of our parameter quantity $\bar{Q}$ is highly determined by the within-imputation variance. This indicates that most of the variance in the multiple imputed datasets are due to variance in the data, not because of uncertainty in the imputed values. Thus the uncertainty about the imputed values is relatively low.

The degrees of freedom ν is calculated as expressed as

$$\nu = (m-1)\left(1 + \frac{\bar{U}}{\left(1 + \frac{1}{m}\right) B}\right)^2.$$

We observe that the calculation of the degrees of freedom ν includes the term $\bar{U}/B$. Since the within-imputation variance is huge compared to the between-imputation variance, the expression above leads to a degree of freedom of large magnitude for all variables at all occasions. When the degrees of freedom increases towards ∞ (typically more than 100), a good approximation of the *t*-distribution is the standard normal distribution [Kvaløy and Tjelmeland, 2000]. Note: The degrees of freedom for complete-case analyses are also generally large, but not as huge as for multiple imputation analyses.

The test statistic for the standard normal distribution is

$$Z = \frac{\bar{Q}}{\sqrt{T}},$$

and the computed test statistics are compared with the standard normal distribution $\mathrm{N}(0, 1)$ to yield estimated p-values.

R is a free software environment for statistical computing and graphics [R Development Core Team, 2008]. The estimates of the difference of means for men and women and the corresponding total variances are calculated in R as displayed in Appendix Section B. The command `dnorm()` computes p-values for the z-test.

6.2.6 Results, two-sample *t*-test assuming unequal variances

Results for the two-sample Student's *t*-test after multiple imputation (MI) and combination by Rubin's rules in R are found in Table 6.2.

Table 6.2: Estimates for the two-sample Student's t-test after multiple imputation (MI) and combination by Rubin's rules. Results are displayed for general health (GH), bodily pain (BP), social functioning (SF) and role emotional (RE), grouping on gender. Significant p-values are marked as boldface. $\tilde{\theta}$ is the estimated difference in mean for men and women combined for the 20 datasets, and $\sqrt{\tilde{T}}$ is the estimated standard deviation for $\tilde{\theta}$.

	Baseline			Six months			Twelve months		
SF-36	Mean	$\tilde{\theta}$ ($\sqrt{\tilde{T}}$)	P-value	Mean	$\tilde{\theta}$ ($\sqrt{\tilde{T}}$)	P-value	Mean	$\tilde{\theta}$ ($\sqrt{\tilde{T}}$)	P-value
GH									
Male	65.1	2.4 (5.0)	0.222	71.3	6.4 (6.8)	**0.020**	70.8	5.6 (5.4)	**0.022**
Female	62.7			64.9			65.2		
BP									
Male	56.6	4.4 (8.0)	0.115	75.4	7.8 (9.1)	**0.013**	78.1	8.6 (8.8)	**0.006**
Female	52.2			67.6			69.5		
SF									
Male	73.2	5.4 (7.0)	0.051	84.4	6.5 (7.5)	**0.023**	85.6	3.6 (6.0)	0.132
Female	67.8			77.9			82.0		
RE									
Male	58.6	10.5 (23.2)	**0.036**	72.4	12.4 (24.4)	**0.017**	74.4	16.6 (21.6)	**<0.001**
Female	48.1			60.0			57.8		

The estimates in Table 6.2 are similar to those found for complete-case analyses in Gjeilo et al. [2008], displayed in Table 5.1. The estimates of means for men and women are of the same magnitude in both complete-case and multiple imputation analyses. Standard deviations for the observations are not accessible due to the design of the multiple imputation analyses, and the variance estimates `T` are not comparable to the estimated standard deviations for complete-case analyses.

P-values follows the same trend in both analyses, but are slightly altered in the multiple imputation analyses compared to complete-case analyses. For social functioning at baseline, the multiple imputation method changes the p-value from 0.042 to 0.051, and is therefore not marked as boldface in Table 6.2. It must be emphasized that this change in the p-value is not of importance, the estimated means for men and women are not particularly affected, and the p-values are both low. For role emotional on the other hand, the p-value at baseline is altered from insignificant to significant when applying multiple imputation analysis. The p-value decreases from 0.054 to 0.036, but the estimates are also here approximately unaltered. The interpretation of these results are analogous to the complete-case results.

6.2.7 Repeated measurements ANOVA after MI

Repeated measurements ANOVA is performed in Gjeilo et al. [2008] to explored a possible difference in improvement between the genders in time. This analysis requires complete-case data which leads to a comprehensive loss of patients since all patients with missing value of a variable at one, two or all of the measurement occasions are omitted. In the following Section we want to analyze the imputed datasets with respect to the gender by time interaction without listwise deletion of subjects with missing values.

The command `mim` is one of the Stata commands that carries out the requested analysis with multiple imputed datasets [Royston, 2004, 2005]. It implements the methods by Rubin [1987] of combining estimates from the imputed datasets. `mim` is a prefix that supports several analysis commands, for a complete list we refer to the help-file of Stata [2007]. To apply `mim` for the dataset, it has to be stacked with indexes _mj (dataset numbering) and _mi (subject number in each dataset). This is done automatically for multiple imputed datasets in `ice`. The output of the command includes estimates, standard deviations, Student's *t*-quintiles, p-values, confidence intervals and FMI for the parameters of the selected analysis.

FMI is the estimated fraction of missing information, or relative increase in variance due to missing values [Royston et al., Submitted for publication, 2008], given as

$$\text{FMI} = \frac{1}{(r+1)}\left(r + \frac{2}{\text{d.f.}+3}\right). \tag{6.3}$$

r is the relative increase in variance due to non-response. This increase is estimated by

$$r \approx \left(1 + \frac{1}{m}\right)\frac{\text{B}}{\bar{\text{U}}},$$

where $\bar{\text{U}}$ and B are the within- and between-imputation variance. The degrees of freedom, denoted d.f. in Equation (6.3), is determined to be positive, thus FMI is a quantity between zero and one. A large FMI (close to one) refers to a great estimated loss of information due to missing data, and hence loss of prediction. Carpenter and Kenward [2007] indicates a direct connection between the FMI and the number of patients with complete observed records as a measure of how precise the imputations of MI are.

We now want to perform repeated measurements ANOVA on the imputed datasets. It is no supported command for this analysis in `mim`, thus we have two alternatives to produce this analysis. Either we may implement the analysis ourselves as done for Student's *t*-test assuming unequal variances, or we can apply an analysis that approximates the repeated measures ANOVA with respect to the gender by time interaction.

We observe that the repeated measurements ANOVA assuming compound symmetry is similar to the analysis by a random intercepts mixed regression model as described in Section 3.4.2. A random intercepts MRM is analyzed in Stata by the command `xtmixed`, which is a supported command in `mim`. Thus we can analyze the imputed datasets by the above mentioned mixed regression model and obtain approximate results for repeated measurements ANOVA after multiple imputation. This approximation corresponds to assuming homoskedasticity and sphericity for the data when performing repeated measurements ANOVA. We can compare the

results from the combined mixed regression model after MI with the results from a complete-case repeated measurements ANOVA without correction of the degrees of freedom (such as Greenhouse-Geisser or Huynh-Feldt).

6.2.8 Implementation of `mim` in Stata

The analysis of the random intercepts MRM must be carried out in the file created by `ice`, here denoted Impfile. In this file the original dataset and the *m* imputed datasets are stored. Prior to the analysis we must ensure that the data are arranged in long format, and generate the necessary time and interaction variables.

The script to analyze the variable general health in Stata is given in Listing 6.3 below.

Listing 6.3: Approximation to repeated measurements ANOVA after multiple imputation for general health (GH)

```
1 reshape long mh vt bp gh sf pf rp re ht, i(patkey _mj) j(time)
2 egen timemean = mean(time)
3 ge timeshift = time - timemean
4 ge timeshift_kjønn = timeshift*kjønn
5 ge timeshift2 = timeshift^2
6
7 mim, noisily: xtmixed gh kjønn timeshift timeshift_kjønn timeshift2  || patkey:, mle
8 mim, mcerror
```

The first line transform the data to long format. Further the time is centered by generating the variable *timeshift* on the second and third line, and the gender by time interaction and quadratic time variable is generated on the two consecutive lines. The time variable is centered to ensure comparable results with subsequent analyses, and is explained in Section 7.1.3.

The actual multiple imputation analysis is performed by the command on line eight, where the prefix `mim` specifies analysis and combination of multiple imputed datasets, the option *noisily* ensures that the analysis output from each dataset are displayed, and the command `xtmixed` performs the random intercepts MRM on each dataset. The `xtmixed` command is explained in more detail in Chapter 7.

The last command given in Listing 6.3 is the replay option *mcerror*, that leads to an output of the last analysis results including the estimated Monte Carlo errors. This feature is also explained in the consecutive Chapter 7.

6.2.9 Results, random intercepts MRM after MI

The results of the random intercepts mixed regression model with quadratic time trend for general health, bodily pain, social functioning and role emotional are represented in Table 6.2.9.

A first glance at the results in Table 6.2.9 above reveals that the gender by time interactions are insignificant for all the four variables. For general health the p-values are approximately equal for the complete-case and multiple imputation analyses. The interaction estimates for bodily pain and role emotional are significant in the repeated measures ANOVA. After imputing

Table 6.3: Estimates, standard deviations and p-values of the gender by time interaction from the random intercepts mixed regression model with quadratic time trend after multiple imputation. Results displayed for general health (GH), bodily pain (BP), social functioning (SF) and role emotional (RE), grouping on gender. The results from the repeated measures ANOVA assuming compound symmetry for the gender by time interaction are also displayed. Significant p-values are marked as boldface.

	Random intercepts quadr. MRM				Rep Meas ANOVA
SF-36	Estimate	SD	P-value	FMI	P-value
GH					
Gender by time	-1.68	1.25	0.180	0.318	0.175
mcerror	0.15	0.05	0.045	0.060	
BP					
Gender by time	-2.50	1.52	0.101	0.136	**0.038**
mcerror	0.12	0.03	0.017	0.035	
SF					
Gender by time	0.73	1.37	0.595	0.185	0.160
mcerror	0.13	0.03	0.069	0.044	
RE					
Gender by time	-3.82	2.52	0.130	0.155	**0.021**
mcerror	0.22	0.06	0.021	0.039	

the missing values by multiple imputation these interactions are no longer significant, the p-values are far from an α-level of 0.05. General health and social functioning have insignificant interaction estimates for both analysis.

Profile plots for both complete-case sample and multiple imputed sample of the variable bodily pain are presented in Figure 6.9. Left plot in Figure 6.9 indicates a different profile for men and women after listwise deletion, which is the sample that the complete-case repeated measures ANOVA is based upon. After imputing missing values by multiple imputation these improvement profiles are more parallel for the genders, as found in the right plot in Figure 6.9). This confirms the increased p-value of 0.081 for the gender by time intercept found in Table 6.2.9. The documented difference in progress between genders in Gjeilo et al. [2008] are found to be weaker when imputing the missing values prior to the analysis.

Similar profile plots for role emotional are displayed in Figure 6.10. The p-value for gender by time intercept is 0.021 in the repeated measures ANOVA assuming equal variances, based on fully observed subjects only. The left plot in Figure 6.10 reflects this result. Men seems to improve considerably after heart surgery, while no progress is observed within the female group. The right plot in Figure 6.10 displays the profile plot for men and women of the role emotional scale for the data after multiple imputation. By imputing the missing values we see that the genders are more equal in improvement, and this confirms the increased p-value for gender by

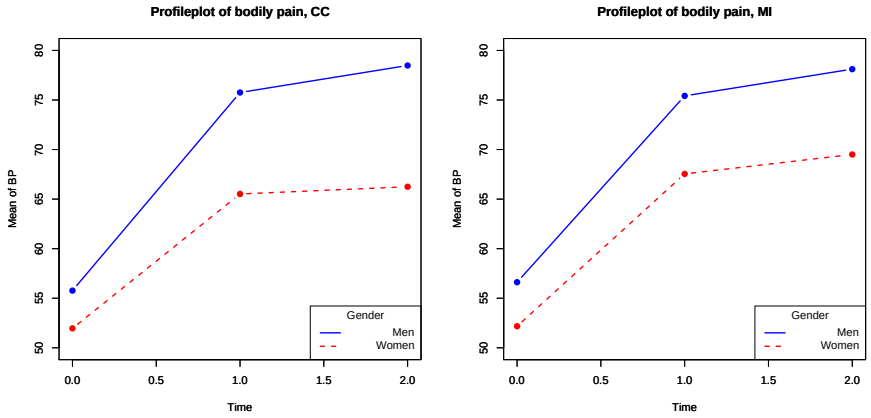

Figure 6.9: Profile plots for the variable bodily pain (BP) based on the complete-case subjects (denoted CC) and multiple imputed datasets (denoted MI).

time interaction found in Table 6.2.9 (p-value of 0.143).

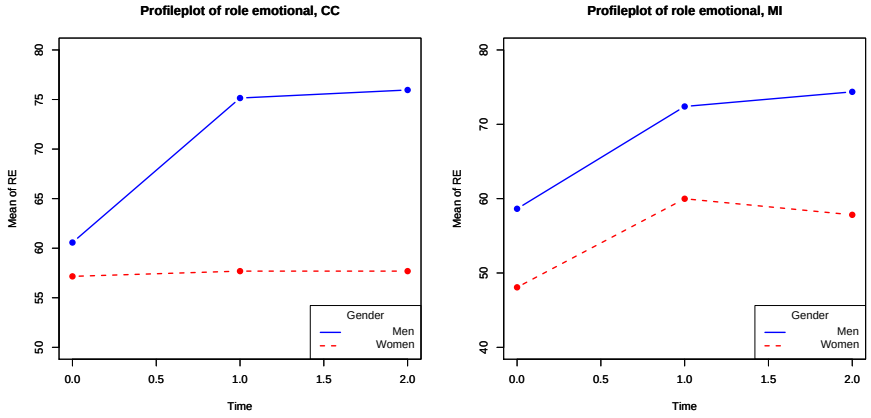

Figure 6.10: Profile plots for the variable role emotional (RE) based on the complete-case subjects (denoted CC) and multiple imputed datasets (denoted MI).

6.2.10 Features of `mim`

The estimates from analysis by multiple imputation depend on the number of imputed datasets m. If we have obtained ∞ imputed datasets from multiple imputation the estimates are assumed accurate, when also assuming the imputation model is correct and sufficiently comprehensive. Instead we have a finite number of imputations, here $m = 20$, and some uncertainty about the estimates are induced. The option *mcerror* in `mim` returns the Monte Carlo standard errors for each estimated value. This is an estimate of the error in combined estimates and standard errors

caused by analyzing a finite number of imputed datasets. These estimates give indication of the uncertainty of the parameters estimated by MI, and thus how reliable the results should be considered.

The standard errors of the estimates $\hat{\beta}$ are computed using the jackknife approach [Royston et al., Submitted for publication, 2008], and these errors behave similar to the Monte Carlo errors. So when *m* increases the standard error of the estimates will decrease, and the Monte Carlo errors follow the same trend. *mcerror* varies with a factor $\sqrt{m}$ so by increasing the number of datasets from 20 to 100 the estimated uncertainty in the results decreases by a factor $\sqrt{5}$. This estimate can be used to examine the necessary number of imputed datasets to obtain precise results.

Estimated FMI are given in Table 6.2.9 for the four variables presented. General health have a FMI of 0.318, which is rather high. This indicates that the subjects with missing values may keep a special feature of the population, and this part is missing. For the remaining variables bodily pain, social functioning and role emotional the FMI lies below 0.2. This indicates that most of the information is found in the data that are observed.

Atypical datasets are imputed datasets where imputations are drawn from conditional posterior distributions not derived from the joint multivariate normal distribution. This may occur when the joint multivariate normal distribution does not exist. It is important to detect such datasets to prevent biased estimates and abnormal high variance estimates due to the values from atypical datasets. The option *mcerror* can be used to detect such atypical datasets by analyzing subsets of the imputed datasets separately and compare the estimates. Another option that can be more useful is *noicily*, that displays the analysis output for each dataset during computation of the combined results. To detect an atypical dataset we can graph dot-plots of the estimates obtained in all datasets to search for outliers. If the estimates seems to behave randomly and without outliers, the datasets have probably converged. Examples of dot plots for estimated parameters and standard deviations of the gender by time interaction for general health are displayed in Figure 6.11.

All the twenty imputed datasets for general health seems to behave random and no outliers are detected. This is the case for all variables and we conclude that the process have converged to the joint multivariate normal distribution for the 20 imputations. Thus we can rely on the estimates to be unbiased and that the estimated variances are not affected by outliers in some of the datasets.

Note: The random intercepts quadratic MRM is used with `mim` to approximate a repeated measurements ANOVA assuming compound symmetry after multiple imputation. A random slopes quadratic MRM is not a reasonable alternative to further develop the modeling of the data here since the aim of these analyses is to compare the complete-case method to multiple imputation.

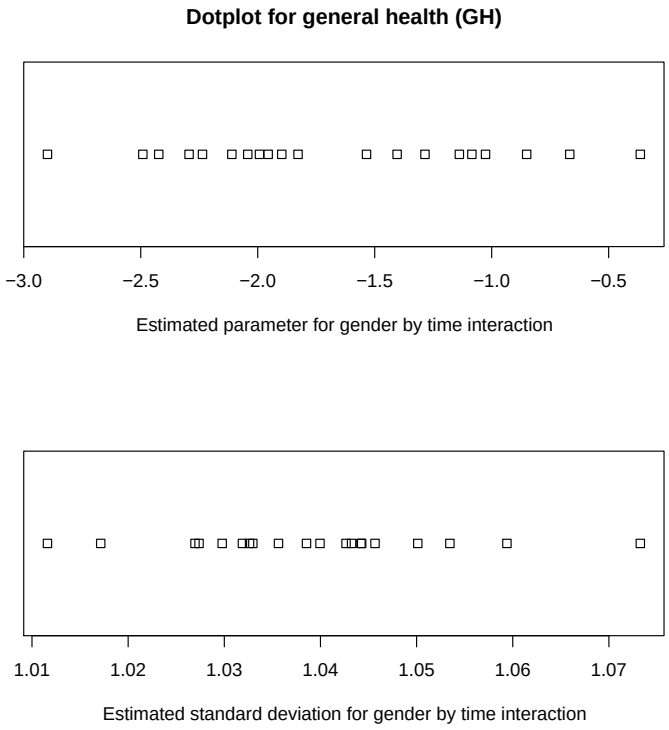

Figure 6.11: Estimated parameters and standard deviations for gender by time interaction from each of the 20 imputed datasets, for the variable general health (GH).

7 Analyses by regression models

The imputation methods of expectation maximization and multiple imputation are two ways to apply a full likelihood distribution to handle missing values in datasets. Analyses by regression models are alternative ways to examine the improvement patterns in HRQOL-variables in data with non-response. The mixed models described in Section 3.4 and the covariance pattern models in Section 3.5 employ a full likelihood function to estimate coefficients and the corresponding standard deviations, and thus have possibilities to yield unbiased results. The generalized estimating equations described in Section 3.6.2 are exceptions of the regression models in the Theory Chapter, in the sense that these models assume data to be covariate-dependent MCAR while the remaining regression models handle ignorable missing data assuming MAR. The use of the latter regression models can be combined with multiple imputation, and this is explained in Section 7.3.

7.1 Mixed regression models in Stata

We are now about to employ mixed regression models to analyze the dataset of Gjeilo et al. [2008] to look for gender by time interactions of importance for the HRQOL-variables. The regression approach allows subjects with unbalanced data to take part in the analyses, thus no listwise deletion are necessary. This is one of the major advantages of mixed regression models over repeated measures ANOVA. All analyses are performed in Stata/SE 10.0 for Windows [2007] with the command `xtmixed`. Data must be arranged in long format prior to all consecutive analyses.

7.1.1 Random intercepts MRM

We will first examine a random intercepts mixed regression model (RI MRM) with fixed effects gender, time and gender by time interaction as displayed in Equation (3.8) in Section 3.4.2. The script to analyze the variable general health is displayed in Listing 7.1.

This command is similar to the command applied with `mim` in Listing 6.3. First the dataset is transformed to long format and the centered time and interaction variables are generated. This

Listing 7.1: Random intercepts linear MRM for general health (GH)

```
1 reshape long mh vt bp gh sf pf rp re ht, i(patkey) j(time)
2 egen timemean = mean(time)
3 ge timeshift = time - timemean
4 ge timeshift_kjønn = timeshift*kjønn
5 ge timeshift2 = timeshift^2
6
7 xtmixed gh kjønn timeshift timeshift_kjønn || patkey:, mle
8 estimates store rit
```

is done to ensure comparable results with mixed regression models including a quadratic time trend in consecutive analyses. The mixed model analysis are specified on the seventh line, where the command `xtmixed` are followed by the dependent variable and the fixed effects equation. The random effects are given after the two vertical lines, and the random intercepts effect are especially marked with the punctuation mark ':'. We want to compare hierarchical models by likelihood ratio tests, thus the parameters must be estimated by maximum likelihood estimators. This is achieved by specifying the option *mle*. The estimates of the analysis is stored in memory as *rit* by the last line.

From the output of the `xtmixed` command we obtain estimated parameters, standard errors, p-values and confidence intervals for the fixed effects derived from Maximum likelihood estimation and Wald's test. Further we achieve estimated standard deviations and their standard errors and confidence intervals for the random factors including the error term ϵ_{hij}. We also get the log likelihood value for the model, and test result for the likelihood ratio test of the MRM model versus linear regression. Descriptive information about the data included in the analysis, for example the number of subjects and observations, and the average number of observations per subject are given at the upper right hand side of the output.

To detect whether the fitted random intercepts MRM from Listing 7.1 describes the data better than a more strict hierarchical model (here: a simple linear regression model), we look at the output of the likelihood ratio test. The test gives a significant p-value even when dividing by two, thus we conclude that the random intercepts term is necessary to analyze the data appropriately. The p-values from the likelihood ratio test must be halved as described in Rabe-Hesketh and Skrondal [2008]. We continue to model the data with the random intercepts MRM as the strict hierarchical model for comparison with more general models. It is important that the models compared by the likelihood ratio test are nested (hierarchical) in each other and that parameters are estimated by maximum likelihood estimators.

7.1.2 Random slopes MRM

To examine whether a random slopes effect is necessary to model the data we employ a random slopes mixed regression model with the same centered time variable and fixed effects as above. A random time term is introduced to the regression model as expressed in Equation (3.12). This random factor is included in `xtmixed` in the random factors equation after the two vertical lines as displayed in Listing 7.2.
The error term ϵ_{hij} is independent of the other variance parameters, but the random intercepts

Listing 7.2: Random intercepts linear MRM for general health (GH)

```
1 xtmixed gh kjønn timeshift timeshift_kjønn || patkey: timeshift, mle cov(unstructured)
2 estimates store rst
3
4 lrtest rit rst
```

and random slopes effects are not necessarily independent. The option *cov(unstructured)* allows the variance estimates of the two random factors to be different and correlated as expressed in Equation (3.5.1). The output of the analysis is stored as *rst*.

We find the likelihood ratio test result for the random slopes MRM compared with the simple linear regression model from the output of `xtmixed`. This test is no longer of interest since we detected the random subjects effect to be significant, and therefore want to compare the random intercepts and random slopes MRM instead. The command `lrtest` performs the likelihood ratio test for the specified models as displayed on the fourth line in Listing 7.2. The random slopes mixed regression model is found to model the data better than the random intercepts MRM, thus we continue the modeling based on the more general model.

7.1.3 Quadratic time trend MRM's

We have now examined mixed regression models assuming a linear time trend. Profile plots in Section 5.3 indicate a non-linear improvement of the HRQOL-variables over time, thus we want to explore the MRM including the fixed quadratic time trend. One way to achieve this is by including a squared-time term in the regression model as expressed in Equation (3.4.4). As explained in Section 3.4.3 we transform the time variable to centered form to ensure the linear and quadratic time variables are not collinear. After centering the time the linear time term takes the values -1, 0 and 1, while the quadratic time effect takes the values 1, 0 and 1. A gender by the squared time interaction is included in the regression models, but found highly insignificant for all variables and therefore omitted in the following analyses.

The centering of time is performed as explained in Listing 7.1. Random intercepts and random slopes MRM with quadratic time trend are obtained by the script in Listing 7.3.

Listing 7.3: Random intercepts and random slopes quadratic MRM for general health (GH)

```
1 xtmixed gh kjønn timeshift timeshift_kjønn timeshift2 || patkey:, mle
2 estimates store rit2
3
4 xtmixed gh kjønn timeshift timeshift_kjønn timeshift2  || patkey: timeshift , mle cov(
      unstructured)
5 estimates store rst2
6
7 lrtest rit2 rst2
```

The quadratic time trend is found highly significant for all the four variables general health, bodily pain, social functioning and role emotional. Further the random intercepts random effect is found necessary to model the data in the output from the first command above. From the

likelihood ratio test on the last line we conclude that the random slopes effect leads to a better fit of the data, thus the random slopes mixed regression model with a quadratic time trend models the data most satisfactory of the above examined MRM's.

7.1.4 Curvilinear trend MRM's

When introducing the quadratic time trend in Section 7.1.3 we achieve a third possible random effect in addition to the random intercepts and slopes effects. This is denoted the random quadratic time trend and is explained in Section 3.4.5. The model specification of the curvilinear MRM is found in Equation (3.17) the same section. The implementation of this model is given in Listing 7.4.

Listing 7.4: Curvilinear trend MRM for general health (GH)

```
1 xtmixed gh kjønn timeshift timeshift_kjønn timeshift2 || patkey: timeshift timeshift2,
      mle cov(unstructured)
2 estimates store rs2t2
3
4 lrtest rst2 rs2t2
```

The included random term v_{2i} in the random factors equation represents the individual deviation from the quadratic trend components for subject *i*. The random slopes quadratic MRM is nested in the curvilinear MRM, thus we apply the likelihood ratio test on script line four to determine if the random quadratic time trends are of importance when analyzing the data. This test is highly significant and we conclude that the curvilinear mixed regression model yields the best fit of the data in Gjeilo et al. [2008].

7.2 Results of analyses by MRM

We have achieved results for all of the mixed regression models described in Section 7.1 above. These results include parameter estimates, standard deviations and p-values for each fixed effect, in addition to estimated variances and covariances for the random effects. All estimates, standard deviations and p-values for the fixed effects are presented in Appendix Tables A.4, A.5, A.6 and A.7 for general health, bodily pain, social functioning and role emotional respectively. As mentioned in Section 3.4.6 no p-values are obtained for the random effects. The results from likelihood ratio tests of the hierarchical models are found in Appendix Table A.8.

The likelihood ratio tests reveal that the most general mixed regression model is necessary to model the data satisfactory, as found in Section 7.1.4. This curvilinear MRM, also denoted the mixed model from now on are examined for the four selected HRQOL-variables, and estimates for the gender by time interaction are given in Table 7.1. We also present results from the complete-case repeated measures ANOVA to ease the comparison of the results.

Table 7.1: Estimates of parameters, standard deviations and p-values for the gender by time interaction from the repeated measures ANOVA and from the curvilinear mixed regression model (MRM) with a quadratic time trend based on the original and the complete-case dataset. Results displayed for the variables general health (GH), bodily pain (BP), social functioning (SF) and role emotional (RE). Significant p-values are marked as boldface.

	Repeated measures		
SF-36	ANOVA	Curvilinear MRM	
Gender by time	Complete-case	All subjects	Complete-case
GH			
Estimate		-1.33	-1.00
SD		1.18	1.36
P-values	0.179	0.260	0.460
BP			
Estimate		-2.82	-3.64
SD		1.61	1.82
P-values	**0.046**	0.081	**0.046**
SF			
Estimate		0.77	0.19
SD		1.36	1.50
P-values	0.164	0.571	0.898
RE			
Estimate		-3.86	-5.96
SD		2.63	3.06
P-values	**0.025**	0.143	0.052

The p-values of the interactions for all the four selected variables alter when applying the curvilinear model to all data compared to the complete-case repeated measurements ANOVA. More of the information contained in the data are included in the analyses and the sample sizes of men and women increases. This should give more significant p-values if the data are missing completely at random, as assumed by the complete-case analyses. Instead we observe that the p-values increases and this indicates that some information is lost due to complete-case analyses, and the repeated measures ANOVA yields biased estimates. In addition does the observed correlation structure of the data violate the assumption of sphericity from repeated measures ANOVA. Approximations to make the repeated measures ANOVA achievable are performed by Greenhouse-Geisser corrections, and this leads to approximations of the assumed correlation matrix of the data. The curvilinear MRM have possibilities to model the correlation structure more generally, and are therefore assumed to give a better fit of the data than the approximative method of Greenhouse-Geisser and repeated measures ANOVA.

For the variable role emotional the analysis by mixed model alters the interpretation of the fixed gender by time interaction. With complete-case ANOVA this variable is significant with a p-value of 0.025. This p-value is increased by a factor of six when analyzing the data with the mixed model based on all subjects (p-value of 0.143). For the variable bodily pain the interaction p-value is also altered from significant (0.046) to insignificant (0.081) when comparing complete-case repeated measurements ANOVA with the mixed model. It must be emphasized that this change in p-values does not necessarily alter the interpretation of the effect. The gender by time interaction is insignificant for both analyses for the remaining variables general health and social functioning.

Further it is desirable to compare the results from the curvilinear mixed regression model performed on both the complete-case selection of patients and the dataset containing all subjects. These analyses make us able to compare the impact of complete-case analyses directly, since the assumptions are equal for the analyses of both subject samples. Results of the analyses of the curvilinear MRM performed on the complete-case data are displayed in the right column of Table 7.1.

For the variable role emotional we discover that the parameter estimate for the gender by time interaction is altered from -3.85 for all subjects to -5.19 for complete-case subjects, that is a decrease of approximately size 1.3. The standard deviance is only slightly increased by a number 0.3, from 2.63 to 2.92. The p-value for the interaction when analyzing the complete-case subjects is almost half of the p-value for the same analysis performed on all subjects. For bodily pain the estimate of the interaction is decreased from -2.82 to -3.64 when analyzing the subjects remaining after listwise deletion by the mixed model, and the p-value is decreased below the α-level 0.05. For the remaining variables general health and social functioning the estimated interactions are close to zero and small compared with the estimated standard deviations, thus the p-values yield insignificant results.

The difference between men and women at each measurement occasions seems to be of greater magnitude when describing all subjects than found for the complete-case subjects, but the improvement profile for men and women seems to be more similar. This is reflected in the results from the curvilinear MRM where the gender by time interaction is increased (see Table A.4).

The plot of all subjects in Figure 5.2 displays that by including the omitted subjects the improvement profile of bodily pain become more similar for men and women than for complete-case subjects. This is analogous with the results found in Table 7.1. Gender by time interaction for social functioning is far from significant for any of the analyses, and by including all subjects with observed measurements the improvement profiles become even more parallel than for complete-case analyses.

The overall change in results for role emotional when including omitted subjects is that the mean for women at baseline becomes smaller, thus the profile for women is altered and the improvement profiles for men and women are more similar. The gender by time interaction is found insignificant when analyzing all subjects by the curvilinear MRM, and this is reflected in the middle profile plot in Figure 5.4.

7.3 Analyses by generalized estimating equations

Generalized estimating equations (GEE) are described in Section 3.6.2 as a general analysis tool that handles several types of dependent variables. But, since it assumes covariate-dependent MCAR, it is not the first choice of analysis methods to handle data with missing values. We have conducted an analysis by GEE with both the original data including subjects with missing values and with the multiple imputed datasets to present an application of this method. As described in the previous section the unstructured form seems to be the most suitable correlation structure for the data in Gjeilo et al. [2008]. There are no restrictions of the missing data pattern when applying this working correlation matrix, thus no subjects are omitted.

The Stata script to perform the GEE method is found in Listing 7.5. The data are reshaped to long format and the centered time and intercept variables are generated as displayed in Listing 6.3. The `xtgee` command on first line performs the regression analysis, with the correct between-subjects and time variables specified by the options *i()* and *j()*. The unstructured correlation form is given by *corr()* and the standard deviations are estimated by a robust method as given in the last option *vce()*. Finally, if we want to look at the estimated working variance-covariance matrix produced by the analysis this is achieved by the second line.

Listing 7.5: Generalized estimating equations for the variable role emotional (RE) with unstructured correlation

```
1  xtgee ret kjønn timeshift timeshift_kjønn timeshift2, i(patkey) t(timeshift) corr(uns
       ) vce(robust)
2  matrix list e(R)
```

The results of the analysis of role emotional by GEE are found in Table 7.3. The command `xtgee` is one of the supported commands of `mim`, thus it is easy to apply GEE to the multiple imputed datasets to compare the results obtained from original data. These results are also listed in Table 7.3, including the fraction of missing information (FMI) and mcerrors.

The gender by time interaction is the focus of these analyses, analogous to the repeated measures ANOVA and the remaining regression models explained previously. We observe that all four variables obtain insignificant p-values for the interaction effect with both original data

and the combined results for the multiple imputed datasets. Bodily pain are close to the α-level of 0.05 for original data analysis, which indicates a possible difference in improvement for men and women. The multiple imputed datasets gives a p-value that is far from significant, and because of the features of MI we have reasons to believe in this value rather than the p-value from the original data analysis.

Role emotional is the other variable with significant interaction by complete-case repeated measures ANOVA. It is not found significant in either of the GEE analyses displayed in Table 7.3 above. We obtain the same interpretation of results from the GEE analyses after MI as for the curvilinear mixed regression models and the imputation methods of EM and MI.

Table 7.2: Generalized estimating equations conducted on the original dataset and the multiple imputed datasets (MI datasets) for the variables general health (GH), bodily pain (BP), social functioning (SF) and role emotional (RE), grouping on gender. Significant p-values are marked as boldface.

SF-36 Gender by time interaction	Original dataset P-value	MI datasets P-value	MI datasets FMI
GH			
Estimate	-1.00	-0.57	0.267
mcerror		0.14	0.066
SD	1.18	1.29	
mcerror		0.06	
P-value	0.399	0.656	
mcerror		0.081	
BP			
Estimate	-2.93	-2.12	1.169
mcerror		0.14	0.043
SD	1.51	1.60	
mcerror		0.04	
P-value	0.052	0.185	
mcerror		0.033	
SF			
Estimate	0.42	0.93	0.155
mcerror		0.13	0.040
SD	1.43	1.47	
mcerror		0.04	
P-value	0.770	0.529	
mcerror		0.061	
RE			
Estimate	-4.28	-3.79	0.170
mcerror		0.24	0.043
SD	2.63	2.70	
mcerror		0.07	
P-value	0.103	0.162	
mcerror		0.025	

8 Discussion

Missing data arise within numerous of research fields and is clearly an issue that we must handle by care. As mentioned in Section 4.2 there are different reasons for missing values. A special type of missing data occur when the reason for non-response is that patients die during a longitudinal study. The question is whether or not these unobserved values should be treated as missing data, or rather as censored values. If the reason for death can be related to the research question under evaluation, the missing values should be regarded as missing at random (MAR), or even missing not at random (MNAR). On the other hand, if missing values due to mortality not related to the hypothesis of the study, the missing values can be regarded as completely random, and some of the deficient methods described in Schafer and Graham [2002] can be applied without too much loss of efficiency and give unbiased estimates.

Normality of data are assumed in Gjeilo et al. [2008], which is approximately correct for many of the variables that are analyzed, but not for all. Role emotional is a variable that takes only five levels of values, and with a clear roof and ceiling effect that violates the normality assumption. This can affect the results and lead to biased estimates and/or decreased statistical power. This concerns for example the Student's *t*-test. Some methods claim to be robust to violations of normality, for example the repeated measures ANOVA. In Section 6.2.3 we state that the arc sine transformation formula modifies variables to be more Gaussian shaped, so we have performed all analyses with transformed variables as the dependent variable. This is done to examine the sensitivity of methods to the normality assumption of variables. The results of the improvement hypothesis gives slightly decreased p-values for the gender by time interaction for bodily pain, but leads to equal interpretation. The remaining variables obtain similar results of p-values for the interaction for both original and transformed variables. The analyses on the original variables gives interpretable parameter estimates, and are the only results reported in this paper.

Chapter 4 describes the dataset and the missing data structure under study in this paper. We find that the proportion of missing values are considerably higher for female patients than for male. This indicates that the probability of missing values are related to the gender of patients, that is data are covariate-dependent missing completely at random (MCAR). Complete-case analysis leads to a selection of patients that have all values observed, thus a higher fraction of

the women are excluded from the analyses. This again leads to possible selection bias. Based on these observations we advise more sophisticated analysis methods are used for this dataset.

The effect of missing data on the samples of subjects from complete-case analyses are examined descriptively in Section 5.3. We remember that the Student's *t*-tests are performed on the available subjects at each time point, and are displayed in the middle plots denoted 'All observations' in Figures 5.1 to 5.4. The samples of which the analyses are performed are not necessarily equal for the different time points. As long as no direct comparisons of these results are performed, this available-case method is preferred over complete-case analyses since a higher proportion of subjects are included in analyses. Note that this analysis method may still give biased results.

The graphical display of the complete-case and available-case samples for role emotional in Figure 5.4 reveal that many of the omitted women hold exceptionally low observations at baseline. It is natural to assume subjects with low intercept value to improve during the study, but remain lower than the subjects with higher baseline score. Since these observations are missing we expect the estimated means to be too optimistic for the available-case analyses of the *t*-tests. Thus we image the real differences between men and women to be larger than the published results for these two time points. The analysis results of the same *t*-test after multiple imputation (in Table 6.2) and expectation maximization (in Table 6.1) are highly significant, which confirms this assumption. There seems to be a relation of the missing values at six and twelve months follow-up and the observed baseline values for the omitted women. Thus we have reason to expect the missing data mechanism in this data are not MCAR (we can not determine whether the mechanism is MAR or MNAR).

The extent of exclusion of subjects are more serious for the repeated measures ANOVA, which base the analyses on the complete-case samples of subjects. The latter are displayed in the left plots in the same figures as described above (Figures 5.1 to 5.4). If we compare the profile plots of role emotional in Figure 5.4 we see that the omitted women (right plot) that are observed at baseline have a mean score of 36.7, more than twenty points less than the complete-case women. These women are not observed at one or both of the consecutive measurement occasions. It is rather easy to spot that the complete-case method leads to selection bias, and we examine the extent of the bias for these analysis results in subsequent paragraphs.

The correlation matrix of the observed dataset is given in Matrix 4.3, and indicates that the assumption of compound symmetry structure is violated. Repeated measures ANOVA assumes normality, homoskedasticity and sphericity, but claims to be robust to violations of the first two assumptions. The latter is violated in the data of Gjeilo et al. [2008], thus we employ Greenhouse-Geisser adjustment of degrees of freedom as an approximation to be able to perform the analysis nevertheless. This means that the repeated measures ANOVA is not very suitable for the specified dataset, and we should apply more appropriate methods to the data with less strict assumptions about the correlation structure of the method. The more general mixed regression models, and the covariance pattern models are examples of such alternative methods that are more suitable to fit the correlation structure of the patients undergoing cardiac surgery. This concerns both complete and incomplete longitudinal data.

Gjeilo et al. [2008] analyzed a longitudinal dataset, where each subject was intended to be measured at three occasions. This longitudinal structure gives potentially more information

about the missing values and the reasons for missing, which strengthens the imputation model when performing expectation maximization algorithm or multiple imputation. A missing value at one time point may have been observed at some of the other measurement occasions for the subject, which is a strong predictor for the values to be imputed.

Expectation maximization (EM) is a full likelihood-based method that gives unbiased estimates assuming ignorable missing data in longitudinal datasets. It gives one plausible version of how the dataset could have been if there were no unobserved values. This is a weak point of EM compared to multiple imputation (MI), that gives multiple plausible versions of the original dataset. This makes MI able to reflect about the accuracy and reliability of results. In addition, the parameters of MI are regarded as stochastic variables, drawn from a posterior distribution for each imputed dataset. This gives different parameters for the imputation model for all the imputed datasets. An extra layer of uncertainty about the imputed values are induced in addition to the residual error of the drawn imputations them self. EM computes a fixed set of parameters for the imputation model, and induces uncertainty only in the imputed values.

The methods of EM and MI predict imputations for the missing values based on the values of all subjects. The gender variable is included in the imputation model to ensure that women are more equal each other, and equally for the male patients. The imputation process is complex and the correlation structure is difficult (if possible) to examine. There may be induced some correlation between imputed values and observed data for both men and women after imputation. The implementation of Rubin's rules for multiple imputed datasets explained in Section 6.2.5 does not calculate any correlation since the two-sample Student's t-test assumes independent samples, and the same goes for EM. The results of the t-test after EM and MI should be regarded as approximations, based on the above.

We have explained the imputation model for multiple imputation where the variables with few levels of values are assumed to be ordinal variables. The imputation by EM applies arc sine transformation to all variables prior to imputation, and assumes all variables to be continuous and normally distributed. We have performed an imputation by MI where all variables are transformed as an inspection of the results from the two imputation models used. The parameter estimates are similar for the two analyses with MI and random intercepts MRM, and give the same interpretation of results. The results from MI with transformed variables are not reported here.

The method of random intercepts mixed regression models (MRM) are stated as an approximative method to the repeated measures ANOVA when sphericity is assumed. The analysis of random intercepts MRM must be regarded as an approximation to the repeated measures ANOVA without adjustment methods to compare the estimates of gender by time interactions for MI versus complete-case.

The mixed regression models with more than one random factor employ the unstructured correlation form. This means that the variances and correlations of these effects are allowed to be different. Unstructured form is chosen based on the structure of the observed variance-covariance matrix, which is unlike the more strict covariance patterns. Since we have three fixed measurement occasions this leads to a total of six random variance and covariance parameters to be estimated. We have a large sample of both men and women, thus the accuracy of the estimates of model parameters will not be seriously affected.

The gender by time interaction is found insignificant for all four variables under study. If we remove the interaction from the regression equations we can examine the main effect of gender and linear and quadratic time trends. All three parameters are significant for the variables, therefore we conclude that gender is an important factor for HRQOL scores at baseline, the improvement is considerable at the first months after surgery for both genders, and seems to stabilize after the first half year of recovery. The results of these analyses are not given in this report, since the research question is directed to the gender by time interaction.

The theory of covariance pattern models (CPM) is described in Section 3.5. The use of CPM to analyze longitudinal data is just barely explained in Rabe-Hesketh and Skrondal [2008] by the command `xtgls`, which is a panel data command to perform the method of generalized least squares. This command is able to model the independent and first-order auto-regressive (AR(1)) covariance patterns explained in Section 3.5.1, and allows heteroskedasticity in the data. Heteroskedasticity is the opposite of homoskedasticity, that is variance parameters are allowed to be different for the random effects. Obviously, the independent covariance matrix is not suitable for longitudinal data with correlations within subjects, but the AR(1) structure could be an appropriate alternative for many situations. AR(1) requires subjects to have at least two consecutive observations to be able to estimate the auto-regressive coefficient ρ, and no intermittent missing values are allowed. The option *force* specifies that the analysis should be carried through including subjects with intermittent non-response, but the subjects with less than two observations are still omitted.

The deletion of subjects with less than two observations may lead to selection bias, as explained for the complete-case method. For the data in Gjeilo et al. [2008] a total of 62 patients are omitted when analyzing the variable role emotional, that is 12% of all subjects included in the study. This fraction is small compared with complete-case method where 35% of the patients are excluded. When analyzing role emotional by the command `xtgls` and the option *force* we obtain an estimated auto-regressive coefficient of 0.285. This correspond to the lag1 correlation. The lag2 correlation is further computed as $\rho^2 = 0.081$, which is far from the observed lag2 correlation of 0.332. Although the AR(1) structure seems unappropriate for the data of Gjeilo et al. [2008] it can be suitable for studies with many measurement occasions where most of the subjects are observed at several occasions. The unstructured correlation form is sensible for studies with limited number of fixed occasions.

The observed correlation structure displayed in Table 4.3 indicates a more complex covariance structure than the exchangeable or banded structures, and with only three fixed measurement occasions the unstructured correlation form seems to be most suitable. Unfortunately this structure is not available with the `xtgls` command. The generalized estimating equations (GEE) command `xtgee` handles most of the correlation structures except AR(1), and Rabe-Hesketh and Skrondal [2008] suggest to use this command to fit the unstructured form. A drawback for this method is the requirement of the missing data distribution to rely only on covariates, not on values observed at other occasions, also known as covariate-dependent MCAR. The use of GEE with the original data is performed under the assumption that the missing data distribution is covariate-dependent MCAR although we have observed indication of MAR from the descriptive analysis of role emotional in Section 5.3. This is done as an sensitivity analysis to examine the features of the method for the data in Gjeilo et al. [2008]. In addition we

have applied the method with the imputed datasets from MI, which make us able to inspect the estimates based on original data.

Role emotional takes five levels of values and can be regarded as an ordinal variable as described in Section 6.2.3. The methods of mixed regression models and covariance pattern models assume the dependent variable to be normally distributed and continuous, and role emotional violates possibly both assumptions. This may lead to biased estimates and/or decreased statistical power. Section 3.6 describes some methods to handle discrete outcome variables with missing values. Generalized mixed regression models can be performed in Stata [2007] by the command `gllamm` and corresponding options *link(ologit)* or *link(oprobit)*, according to Rabe-Hesketh and Skrondal [2008]. Unfortunately, there was not enough time for this during this thesis. The GEE command `xtgee` does not have an option to handle ordinal variables according to Rabe-Hesketh and Skrondal [2008], and is rather not performed for the dataset. Weighted estimating equations, or any other weighting methods for handling missing values have not been examined in this paper, but can be appropriate for both continuous and discrete outcome variables in longitudinal datasets with missing data.

Full likelihood-based methods make assumptions about the joint distribution which rely on unobserved information. These methods provide approximative results for the datasets with missing values. When analyzing data with some missing information such methods give better results than the simpler methods of complete-case or single imputation. More complex methods for missing data analyses are developed that make assumptions about the missing data distribution and are able to model data that are missing not at random. The selection models and pattern-mixture models described in Schafer and Graham [2002] can give unbiased estimates for data with missing values when the assumed missing data mechanism is MNAR. Comprehensive research are performed on these methods, and will be continued in the future.

The methods for data missing not at random (MNAR) make assumptions about the missing data distribution, that is the probability that a value is observed. These assumptions are difficult to examine since they are based on the unobserved data. If the assumptions of the missing data distribution is faulty, the methods provide biased results. As discussed in Schafer and Graham [2002], one or two models assuming to handle MNAR data can be performed as a sensitivity analysis for the full likelihood-based methods treated in this report.

9 Conclusion

The overall results of the gender differences in HRQOL scores at each time point in Gjeilo et al. [2008] remains when the data are analyzed by propere imputation methods. For the four HRQOL variables general health, bodily pain, social functioning and role emotional there are a proven difference in favor of men at six and twelve months after surgery. Some results are even more significant when employing imputation techniques because the methods use all the accessible information in the data and thus yield higher statistical power. The differences between the genders are not as convincingly presurgery, only role emotional gives a significant difference in score for men and women.

The improvement in time after a cardiac surgery is found equal for men and women for six of the eight scales of SF-36 when employing complete-case analysis. For the remaining variables bodily pain and role emotional the analysis indicate that female patients improve less than the male patients during the recovery period. When we analyze the evolution in time by proper imputation methods or the curvilinear mixed regression model this difference disappear for role emotional, and diminish considerable for bodily pain. We conclude that there is no statistical proven gender differences in improvement of HRQOL scores after cardiac surgery.

The method of expectation maximization and multiple imputation makes us able to employ the standard statistical analysis methods used in Gjeilo et al. [2008]. The mixed regression models and covariance pattern models represent alternative analysis methods to examine the gender difference in improvement patterns. The imputation methods and mixed regression models are assumed to handle missing data in an adequate way, and gives similar analysis results for all methods. These results differ from the complete-case method results for two of the HRQOL-variables when examining the gender differences in improvement of HRQOL after surgery. The method of generalized estimating equations are able to give unbiased estimates only if the missing data mechanism are assumed covariate-dependent MCAR.

From the curvilinear MRM without interaction we find that gender is an important factor for HRQOL scores at baseline, the improvement is considerable and equal for both genders at the first months after surgery, and the evolution seems to stabilize after the first half year of recovery.

References

A.C. Acock. Working with missing values. *Journal of Marriage and the Family*, 67(4):1012–1028, NOV 2005.

C.F. Baum and N.J. Cox. Omninorm: Stata module to calculate omnibus test for univariate/multivariate normality. `http://ideas.repec.org/c/boc/bocode/s417501.html`, July 2007.

Johannes Berkhof and Tom A. B. Snijders. Variance Component Testing in Multilevel Models. *Journal of Educational and Behavioral Statistics*, 26(2):133–152, 2001.

R.D. Bock. The discrete bayesian. *In H. Wainer & S. Messick (Eds.), Modern advances in psychometric research*, pages 103–115, 1983.

George E. P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley-Interscience, 2nd edition, 2005.

S. Van Buuren and C.G.M. Oudshoorn. Multivariate imputation by chained equations: Mice v1. 0 users's manual. Technical report, TNO Prevention and Health, 2000.

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

J.R. Carpenter and M.G. Kenward. Missing data in randomised controlled trials, a practical guide. http://www.lshtm.ac.uk/msu/missingdata/, 2007.

Simon Day. *Dictionary for Clinical Trials*. J. Wiley and sons, 1999.

A.P. Dempster, M.N. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 1977.

A.P. Dempster, Donald B. Rubin, and R.K. Tsutakawa. Estimation in covariance components models. *Journal of the American Statistical Association*, 76(374):341–353, 1981.

Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott Zeger. *Analysis of longitudinal data*. Oxford University Press, 2002.

Ingrid G. Dragset, Stian Lydersen, and Pål Klepstad. Missing data in quality-of-life measurements: Methods and an application to a questionnaire of pain evaluation. Project thesis, 2008.

Peter M. Fayers and David Machin. *Quality of life: the assessment, analysis and interpretation of patient-reported outcomes*. J. Wiley and Sons, 2nd edition, 2007.

Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, and Geert Molenberghs. *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*. Chapman & Hall/CRC, 2009.

Kari Hanne Gjeilo. *Health-related quality of life and chronic pain in patients undergoing cardiac surgery*. PhD thesis, Norwegian University of Science and Technology, June 2009.

Kari Hanne Gjeilo, Alexander Wahba, Pal Klepstad, Stian Lydersen, and Roar Stenseth. The role of sex in health-related quality of life after cardiac surgery: a prospective study. *European Journal of Cardiovascular Prevention and Rehabilitation*, 15(4):448–452, AUG 2008.

Donald R. Hedeker and Robert D. Gibbons. *Longitudinal data analysis*. J. Wiley and Sons, 2006.

R. Jennrich and M. Schluchter. Unbalanced repeated measures models with structured covariance matrices. *Biometrics*, 42:805–820, 1986.

P Klepstad, PC Borchgrevink, O Dale, K Zahlsen, T Aamo, P Fayers, B Fougner, and S Kaasa. Routine drug monitoring of serum concentrations of morphine, morphine-3-glucuronide and morphine-6-glucuronide do not predict clinical observations in cancer patients. *Pallative Medicine*, 17(8):679–687, 2003.

Jan Terje Kvaløy and Håkon Tjelmeland. *Tabeller og formler i statistikk*. Tapir akademisk forlag, 2000.

N.M. Laird and J.H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

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

J.H. Loge and S. Kaasa. Short Form 36 (SF-36) health survey: normative data from the general Norwegian population. *Scandinavian Journal of Social Medicine*, 26(4):250–258, DEC 1998.

John W. Mauchly. Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, 11(2):204–209, 1940.

Alvaro A. Novo and Joseph L. Schafer. The norm package. http://www.stat.psu.edu/ jls/misoftwa.html#aut, 2006.

The University of Texas at Austin Statistical Services. Repeated measures anova using sas proc glm. `http://www.ats.ucla.edu/stat/sas/library/repeated_ut.htm`, July 1997.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Sophia Rabe-Hesketh and Anders Skrondal. *Multilevel and Longitudinal Modeling Using Stata*. Stata Press, 2nd edition, 2008.

Bernard Rosner. *Fundamentals of biostatistics*. Thomson, 6th edition, 2006.

P. Royston. Multiple imputation of missing values. *Stata Journal*, 4:227–241, 2004.

P. Royston. Multiple imputation of missing values: update. *Stata Journal*, 5(2):188–201, 2005.

P. Royston, J.B. Carlin, and I.R. White. Multiple imputation of missing values; new features for mim. *The Stata Journal*, Submitted for publication, 2008.

Susanne Rässler, Donald B. Rubin, and Elizabeth R. Zell. *Incomplete Data in Epidemiology and Medical Statistics*, volume 27 of *Handbook of Statistics*. Elsevier, 2008.

Donald B. Rubin. Inference and Missing Data. *Biometrica*, 63(3):581–590, 1976.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. J. Wiley and Sons, New York, 1987.

Joseph L. Schafer. *Analysis of incomplete multivariate data*. Chapman and Hall, London, 1997.

Joseph L. Schafer and John W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, JUN 2002.

SPSS 16.0 for Windows. SPSS Inc., Chicago, Illinois, USA, Nov 2007.

Stata/SE 10.1 for Windows. StataCorp LP, 4905 Lakeway Drive, College Station, TX 77845 USA, Aug 2007.

L.M. Sullivan and R.B. Dagostino. Robustness of the t-test applied to data distorted from normality by floor effects. *Journal of Dental Research*, 71(12):1938–1943, Dec 1992.

S. van Buuren, J.P.L. Brand, C.G.M. Groothuis-Oudshoorn, and Donald B. Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064, Dec 2006.

Paul T. von Hippel. Biases in spss 12.0 missing value analysis. *The American Statistician*, 58 (2), May 2004.

J.E. Ware, M. Kosinski, and B. Gandek. *SF-36 Health Survey: Manual and interpretation guide*. Quality Metric Inc., Lincoln, Rhode Island, 2nd edition, 2000.

S. Zeger and L. Liang. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

S.L. Zeger, K.Y. Liang, and P.S. Albert. Models for longitudinal data - a generalized estimating equation approach. *Biometrics*, 44(4):1049–1060, Dec 1988.

A

Additional tables

Table A.1: Missing data structure for identification number patkey, gender, age, the SF-36 scores, health transition (HT) and marital status, from the data of Gjeilo et al. [2008]. The second and third columns describes the original dataset, while the two last columns describes missing data structure in data after missing forms have been excluded. Missing form is a totally unobserved occasion for a subject.

Variable	**No. of missing**	**% missing**	**No. of missing**	**% missing**
Patkey (id)	0	0.0	0	0.0
Gender	0	0.0	0	0.0
Age	0	0.0	0	0.0
General health (GH)				
Before surgery	40	7.5	33	7.6
6 months follow-up	94	17.6	17	3.9
12 months follow-up	98	18.4	26	6.0
Physical functioning (PF)				
Before surgery	19	3.6	15	3.4
6 months follow-up	90	16.9	16	3.7
12 months follow-up	83	15.5	11	2.5
Social functioning (SF)				
Before surgery	14	2.6	12	2.8
6 months follow-up	77	14.4	4	0.9
12 months follow-up	80	15.0	10	2.3
Role physical (RP)				
Before surgery	24	4.5	20	4.6
6 months follow-up	105	19.7	30	6.9
12 months follow-up	101	18.9	30	6.9

Continued on next page

Variable	No. of missing	% missing	No. of missing	% missing
Role emotional(RE)				
Before surgery	41	7.7	35	8.0
6 months follow-up	115	21.5	38	8.7
12 months follow-up	108	20.2	37	8.5
Mental health (MH)				
Before surgery	32	6.0	27	6.2
6 months follow-up	93	17.4	15	3.4
12 months follow-up	95	17.8	24	5.5
Vitality (VT)				
Before surgery	27	5.1	23	5.3
6 months follow-up	91	17.0	15	3.4
12 months follow-up	93	17.4	22	5.0
Bodily pain (BP)				
Before surgery	13	2.4	11	2.5
6 months follow-up	86	16.1	13	3.0
12 months follow-up	84	15.7	14	3.2
Health transition (HT)				
Before surgery	6	1.1	5	1.1
6 months follow-up	78	14.6	5	1.1
12 months follow-up	76	14.2	6	1.4
Marital status				
Before surgery	0	0.0	1	0.2
6 months follow-up	72	13.5	0	0.0
12 months follow-up	70	13.1	1	0.2

Table A.2: Results from a two-sample Student's t-test assuming unequal variances for all SF-36-variables and health transition (HT). Analyses are based on complete-case subjects and grouped on gender. Significant p-values are marked as boldface.

	Baseline			Six months			Twelve months		
SF-36	Mean	SD	P-value	Mean	SD	P-value	Mean	SD	P-value
GH									
Male	64.9	19.7	0.490	72.0	22.1	**0.011**	71.9	21.6	**0.004**
Female	63.3	20.7		65.2	21.9		64.7	20.9	
PF									
Male	60.2	26.0	**0.000**	80.3	22.2	**0.000**	80.3	22.2	**0.000**
Female	43.9	26.3		66.8	26.8		65.1	26.7	
SF									
Male	73.2	24.8	**0.042**	84.7	22.6	**0.013**	86.3	20.5	0.067
Female	67.7	25.4		77.3	25.8		81.4	24.1	
RP									
Male	23.4	36.7	**0.002**	58.3	42.8	**0.010**	58.1	42.8	**0.000**
Female	13.2	27.6		44.4	44.1		40.0	42.4	
RE									
Male	58.7	42.6	0.054	74.6	37.0	**0.006**	75.6	36.7	**0.001**
Female	49.1	45.9		59.8	44.1		58.0	43.1	
MH									
Male	77.5	17.0	**0.000**	82.0	16.5	**0.032**	82.3	15.4	**0.001**
Female	68.9	18.6		77.9	15.7		75.6	16.5	
VT									
Male	50.7	22.9	**0.001**	62.4	22.1	**0.010**	61.5	22.2	**0.004**
Female	42.5	22.0		55.4	22.5		54.1	21.2	
BP									
Male	56.5	27.2	0.137	75.7	25.7	**0.004**	78.7	25.2	**0.002**
Female	52.3	26.4		66.4	27.1		68.9	26.9	
HT									
Male	32.0	24.4	0.195	82.0	24.6	0.791	86.8	20.6	0.566
Female	28.7	24.5		82.8	26.0		85.3	22.6	

Table A.3: Repeated measures ANOVA with Greenhouse-Geisser adjustment for non-sphericity for all SF-36-variables and health transition (HT). Analyses are based on complete-case subjects and grouped on gender. Significant p-values are marked as boldface.

SF-36 Gender by time interaction	Number of fully observed patients	Repeated measures ANOVA P-value
GH	374	0.179
PF	402	0.812
SF	414	0.224
RP	373	0.083
RE	348	**0.025**
MH	382	0.275
VT	386	0.797
BP	401	**0.046**
HT	420	0.493

Table A.4: Estimates for parameters, standard deviations and p-values from the mixed regression models for the variable general health (GH). MRM is short form for mixed regression models. Random intercepts linear MRM corresponds to a random intercepts mixed regression model with linear time trend (3.8), random slopes linear MRM includes a random slopes effect (3.12), random slopes quadratic MRM includes a random slopes effect and a quadratic time trend (3.14) and the curvilinear MRM includes random linear and quadratic time effects and a quadratic time trend (3.17). The last column corresponds to a curvilinear MRM performed on complete-case data.

GH	**Rep. Meas. ANOVA**	**Random intercepts linear MRM**	**Random slopes linear MRM**	**Random slopes quadratic MRM**	**Curvilinear MRM**	**Curvilinear MRM CC**
Gender						
Estimate		-5.526717	-5.418016	-5.30598	-4.718743	-2.75168
SD		1.959242	1.975778	1.974303	1.940176	2.328532
P-value	<0.001	0.005	0.006	0.007	0.015	0.237
Time						
Estimate		3.012206	3.006856	2.945783	2.892607	2.781373
SD		0.5174225	0.5577031	0.557216	0.5494746	0.5995567
P-value	0.003	<0.001	<0.001	<0.001	<0.001	<0.001
Gender$\times$ Time						
Estimate		-1.841928	-1.849577	-1.795574	-1.329212	-1.003433
SD		1.145846	1.234114	1.23255	1.179587	1.356995
P-value	0.179	0.108	0.134	0.145	0.260	0.460
Time^2						
Estimate				-2.939087	-2.815899	-2.723262
SD				0.7293293	0.7209759	0.7755459
P-value				<0.001	<0.001	<0.001

Table A.5: Estimates for parameters, standard deviations and p-values from the mixed regression models for the variable bodily pain (BP). MRM is short form for mixed regression models. Random intercepts linear MRM corresponds to a random intercepts mixed regression model with linear time trend (3.8), random slopes linear MRM includes a random slopes effect (3.12), random slopes quadratic MRM includes a random slopes effect and a quadratic time trend (3.14) and the curvilinear MRM includes random linear and quadratic time effects and a quadratic time trend (3.17). The last column corresponds to a curvilinear MRM performed on complete-case data.

BP	**Rep. meas. ANOVA**	**Random intercepts linear MRM**	**Random slopes linear MRM**	**Random slopes quadratic MRM**	**Curvilinear MRM**	**Curvilinear MRM CC**
Gender						
Estimate		-8.098923	-8.034678	-7.814501	-7.554636	-8.011724
SD		2.280957	2.287777	2.288619	2.278322	2.648945
P-value	0.006	<0.001	<0.001	0.001	0.001	0.002
Time						
Estimate		11.24742	11.2278	11.00562	10.89872	11.29177
SD		0.7163395	0.7762027	0.7784732	0.7655617	0.827171
P-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Gender×Time						
Estimate		-3.284021	-3.233921	-3.227943	-2.816709	-3.638176
SD		1.55517	1.681012	1.683951	1.612395	1.824467
P-value	**0.046**	**0.035**	0.054	0.055	0.081	**0.046**
Time^2						
Estimate				-7.751016	-7.817136	-8.24813
SD				0.9474952	0.932124	0.9687647
P-value				<0.001	<0.001	<0.001

Table A.6: Estimates for parameters, standard deviations and p-values from the mixed regression models for the variable social functioning (SF). MRM is short form for mixed regression models. Random intercepts linear MRM corresponds to a random intercepts mixed regression model with linear time trend (3.8), random slopes linear MRM includes a random slopes effect (3.12), random slopes quadratic MRM includes a random slopes effect and a quadratic time trend (3.14) and the curvilinear MRM includes random linear and quadratic time effects and a quadratic time trend (3.17). The last column corresponds to a curvilinear MRM performed on complete-case data.

SF	**Rep. Meas. ANOVA**	**Random intercepts linear MRM**	**Random slopes linear MRM**	**Random slopes quadratic MRM**	**Curvilinear MRM**	**Curvilinear MRM CC**
Gender						
Estimate		-5.927744	-5.903225	-5.781197	-5.510533	-6.285521
SD		2.040383	2.029119	2.02721	2.020287	2.33862
P-value	0.005	0.004	0.004	0.004	0.006	0.007
Time						
Estimate		6.452945	6.478433	6.393658	6.258906	6.378669
SD		0.6166398	0.6511259	0.6509613	0.6423493	0.6823583
P-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Gender$\times$ Time						
Estimate		0.1307206	0.1432712	0.1389689	0.7700118	0.1935107
SD		1.340295	1.411287	1.41028	1.360416	1.5023
P-value	0.164	0.922	0.919	0.922	0.571	0.898
Time^2						
Estimate				-4.308828	-4.289153	-4.302536
SD				0.8775461	0.8687816	0.9048887
P-value				<0.001	<0.001	<0.001

Table A.7: Estimates for parameters, standard deviations and p-values from the mixed regression models for the variable role emotional (RE). MRM is short form for mixed regression models. Random intercepts linear MRM corresponds to a random intercepts mixed regression model with linear time trend (3.8), random slopes linear MRM includes a random slopes effect (3.12), random slopes quadratic MRM includes a random slopes effect and a quadratic time trend (3.14) and the curvilinear MRM includes random linear and quadratic time effects and a quadratic time trend (3.17). The last column corresponds to a curvilinear MRM performed on complete-case data.

RE	**Rep. Meas. ANOVA**	**Random intercepts linear MRM**	**Random slopes linear MRM**	**Random slopes quadratic MRM**	**Curvilinear MRM**	**Curvilinear MRM CC**
Gender						
Estimate		-14.65376	-14.51358	-14.36809	-14.14039	-12.33781
SD		3.47494	3.465802	3.455653	3.462541	4.297665
P-value	<0.001	<0.001	<0.001	<0.001	<0.001	0.004
Time						
Estimate		8.303575	8.33102	8.196229	8.023801	7.455226
SD		1.177078	1.268306	1.269118	1.245756	1.354061
P-value	0.015	<0.001	<0.001	<0.001	<0.001	<0.001
Gender×Time						
Estimate		-4.873948	-4.60745	-4.575106	-3.858221	-5.958365
SD		2.572712	2.761353	2.761383	2.634411	3.060166
P-value	**0.025**	0.058	0.095	0.098	0.143	0.052
Time^2						
Estimate				-5.934938	-6.027423	-5.72318
SD				1.663914	1.636294	1.804549
P-value				<0.001	<0.001	0.002

Table A.8: Likelihood ratio test results from the mixed models for variables general health (GH), bodily pain (BP), social functioning (SF) and role emotional (RE).

Parameter	Nested model	Generalized model	Test statistic	P-value
GH	Linear regression	RI linear MRM	403.76	<0.001
	RI linear MRM	RS linear MRM	22.47	<0.001
	”Quadratic” linear regression	RI quadratic MRM	410.68	<0.001
	RI quadratic MRM	RS quadratic MRM	24.94	<0.001
	RS quadratic MRM	Curvilinear MRM	49.53	<0.001
BF	Linear regression	RI linear MRM	249.19	<0.001
	RI linear MRM	RS linear MRM	15.44	<0.001
	”Quadratic” linear regression	RI quadratic MRM	271.04	<0.001
	RI quadratic MRM	RS quadratic MRM	29.92	<0.001
	RS quadratic MRM	Curvilinear MRM	46.28	<0.001
SF	Linear regression	RI linear MRM	285.81	<0.001
	RI linear MRM	RS linear MRM	11.20	0.004
	”Quadratic” linear regression	RI quadratic MRM	294.93	<0.001
	RI quadratic MRM	RS quadratic MRM	14.14	<0.001
	RS quadratic MRM	Curvilinear MRM	40.85	<0.001
RE	Linear regression	RI linear MRM	177.16	<0.001
	RI linear MRM	RS linear MRM	15.17	<0.001
	”Quadratic” linear regression	RI quadratic MRM	178.65	<0.001
	RI quadratic MRM	RS quadratic MRM	17.02	<0.001
	RS quadratic MRM	Curvilinear MRM	42.72	<0.001

Stata and R scripts

Stata-script

```
/* Data in wide format */

/* MISSING DATA STRUCTURE */
/* keep the variables that are included in any of the analyses, */
/* to examine the missing values in the data structure */

preserve
keep patkey sivilstand1 sivilstand2 sivilstand3 ALDER kjønn mh1 mh2 mh3 vt1 vt2 vt3 gh1
     gh2 gh3 bp1 bp2 bp3 sf1 sf2 sf3 pf1 pf2 pf3 rp1 rp2 rp3 re1 re2 re3 ht1 ht2 ht3
misschk
restore

/* Examine dataset when complete forms are missing */
preserve
drop if sivilstand2==. & sivilstand3==. & mh2==. & mh3==. & vt2==. & vt3==. & bp2==. &
    bp3==. & gh2==. & gh3==. & sf2==. & sf3==. & pf2==. & pf3==. & rp2==. & rp3==. &
    re2==. & re3==. & ht2==. & ht3==. & sivilstand2==. & sivilstand3==.
drop if sivilstand2==. & mh2==. & vt2==. & bp2==. & gh2==. & sf2==. & pf2==. & rp2==. &
     re2==. & ht2==. & sivilstand2==.
drop if sivilstand3==. & mh3==. & vt3==. & bp3==. & gh3==. & sf3==. & pf3==. & rp3==. &
     re3==. & ht3==. & sivilstand3==.

keep patkey sivilstand1 sivilstand2 sivilstand3 ALDER kjønn mh1 mh2 mh3 vt1 vt2 vt3 gh1
     gh2 gh3 bp1 bp2 bp3 sf1 sf2 sf3 pf1 pf2 pf3 rp1 rp2 rp3 re1 re2 re3 ht1 ht2 ht3
misschk

restore

/* Examination of variables, histograms */
/* Displayed for general health, similar for all variables */

/* GH */
ta gh1
hist gh1 , norm start(-2) width(4) freq xtitle("General Health") title("Histogram of GH
    ," "before surgery")
graph save GHhist1a.gph, replace
graph export GHhist1a.eps, replace
```

```
hist gh2 , norm start(-2) width(4) freq xtitle("General Health") title("Histogram of GH
    ," "after 6 months")
graph save GHhist2a.gph, replace
graph export GHhist2a.eps, replace
hist gh3 , norm start(-2) width(4) freq xtitle("General Health") title("Histogram of GH
    ," "after 12 months")
graph save GHhist3a.gph, replace
graph export GHhist3a.eps, replace

graph combine GHhist1a.gph GHhist2a.gph GHhist3a.gph, rows(1) altshrink imargin(large)
    ycommon xcommon saving(GHhista.gph)
graph combine GHhist1a.gph GHhist2a.gph GHhist3a.gph, rows(1) altshrink imargin(large)
    ycommon xcommon saving(GHhista.eps)

/* Generate arc sine transformed variables */
/* GH */
ge gh1t = asin(sqrt(gh1/100))
ge gh2t = asin(sqrt(gh2/100))
ge gh3t = asin(sqrt(gh3/100))
hist gh1t , norm bin(21) freq xtitle("General Health (transformed)") title("Histogram
    of GH transformed," "before surgery")
graph save GHthist1a.gph, replace
graph export GHthist1a.eps, replace
hist gh2t , norm bin(21) freq xtitle("General Health (transformed)") title("Histogram
    of GH transformed," "after 6 months")
graph save GHthist2a.gph, replace
graph export GHthist2a.eps, replace
hist gh3t , norm bin(21) freq xtitle("General Health (transformed)") title("Histogram
    of GH transformed," "after 12 months")
graph save GHthist3a.gph, replace
graph export GHthist3a.eps, replace

graph combine GHthist1a.gph GHthist2a.gph GHthist3a.gph, rows(1) altshrink imargin(
    large) ycommon xcommon saving(GHthista.gph)
graph combine GHthist1a.gph GHthist2a.gph GHthist3a.gph, rows(1) altshrink imargin(
    large) ycommon xcommon saving(GHthista.eps)

/* Profile plot based on completecase, all observations and omitted subjects */
/* GH */

label define ggg 0 "Men" 1 "Women"
label values kjønn ggg

/* Complete-case subjects */
/* in a dataset */
drop if gh1==. | gh2==. | gh3==.
bysort kjønn: su gh1 gh2 gh3
bysort kjønn: sutex gh1 gh2 gh3
profileplot gh1 gh2 gh3, by(kjønn) xlabel(1 "Before" 2 "6months" 3 "12months") title("
    Complete-case") saving(GHprofileCC.gph)

/* All subjects */
/* In another dataset */
bysort kjønn: su gh1 gh2 gh3
bysort kjønn: sutex gh1 gh2 gh3
profileplot gh1 gh2 gh3, by(kjønn) xlabel(1 "Before" 2 "6months" 3 "12months") title("
    All observations") saving(GHprofileALL.gph)

/* Omitted subjects */
/* in a third dataset */
drop if gh1!=. & gh2!=. & gh3!=.
bysort kjønn: su gh1 gh2 gh3
```

```
bysort kjønn: sutex gh1 gh2 gh3
profileplot gh1 gh2 gh3, by(kjønn) xlabel(1 "Before" 2 "6months" 3 "12months") title("
    Omitted observations") saving(GHprofileOrig.gph)

/*Then combine the three plots */
graph combine GHprofileCC.gph GHprofileALL.gph GHprofileOrig.gph, rows(1) altshrink
    imargin(large)  ycommon xcommon saving(GHprofile3.gph)
graph combine GHprofileCC.gph GHprofileALL.gph GHprofileOrig.gph, rows(1) altshrink
    imargin(large)  ycommon xcommon title("Profileplots for general health (GH)")

/* Examining the residuals of the simple model, y ~ bx + e */
/* RE */

keep patkey re1 re2 re3 kjønn

reshape long re, i(patkey)
rename _j time

quietly regress re patkey kjønn time
predict res, residuals

/* Plot of the residuals */
quantile res
pnorm res

/* Observed variance-covariance matrix */
/* and correlation matrix */

keep patkey re time kjønn res
reshape wide re res, i(patkey) j(time)

tabstat re1 re2 re3, statistics(variance) format(%4.3f)
correlate re1 re2 re3, wrap
correlate re1 re2 re3, covariance

/* t-test */
/* Testing homogeneity of variance between genders */

sdtest re1, by(kjønn)
sdtest re2, by(kjønn)
sdtest re3, by(kjønn)

sdtest bp1, by(kjønn)
sdtest bp2, by(kjønn)
sdtest bp3, by(kjønn)

sdtest gh1, by(kjønn)
sdtest gh2, by(kjønn)
sdtest gh3, by(kjønn)

sdtest sf1, by(kjønn)
sdtest sf2, by(kjønn)
sdtest sf3, by(kjønn)

/* RE2, RE3 and SF3 are significantly different in variances. */

/* Two-sample student's t-test assuming unequal variances */

ttest gh1 , by(kjønn) unequal
ttest gh2 , by(kjønn) unequal
ttest gh3 , by(kjønn) unequal
```

```
ttest bp1 , by(kjønn) unequal
ttest bp2 , by(kjønn) unequal
ttest bp3 , by(kjønn) unequal

ttest sf1 , by(kjønn) unequal
ttest sf2 , by(kjønn) unequal
ttest sf3 , by(kjønn) unequal

ttest re1 , by(kjønn) unequal
ttest re2 , by(kjønn) unequal
ttest re3 , by(kjønn) unequal


/* Test of multivariate normality */
/* Data in long format */
reshape long mh vt bp gh sf pf rp re ht, i(patkey) j(time)
replace time = time-1

qnorm gh
pnorm gh
omninorm gh, by(kjønn)

/* Repeated-measures ANOVA */
/* start with the file in wide format */
/* delete patients with missing values */
/* reshape into long format */
/* same values with wsanova and anova, repeated() */

set matsize 2000
preserve

/*---------GH----------*/
drop if gh1==. | gh2==. | gh3==.
reshape long mh vt bp gh sf pf rp re ht, i(patkey) j(time)
wsanova gh time, id(patkey) between(kjønn) epsilon
anova gh time kjønn patkey time*kjønn, repeated(time) bse(patkey)
/*---------BP----------*/
drop if bp1==. | bp2==. | bp3==.
reshape long mh vt bp gh sf pf rp re ht, i(patkey) j(time)
wsanova bp time, id(patkey) between(kjønn) epsilon
anova bp time kjønn patkey time*kjønn, repeated(time) bse(patkey)
/*---------SF----------*/
drop if sf1==. | sf2==. | sf3==.
reshape long mh vt bp gh sf pf rp re ht, i(patkey) j(time)
wsanova sf time, id(patkey) between(kjønn) epsilon
anova sf time kjønn patkey time*kjønn, repeated(time) bse(patkey)
/*---------RE ----------*/
drop if re1==. | re2==. | re3==.
reshape long mh vt bp gh sf pf rp re ht, i(patkey) j(time)
wsanova re time, id(patkey) between(kjønn) epsilon
anova re time kjønn patkey time*kjønn, repeated(time) bse(patkey)

/* MI with ice, based on all 8 sf-36 scales, */
/* ht, age, marital status, patkey, gender. */
/* Re, sf, rp and ht as ordinal variables */
/* no back-transformation */

ice patkey gh1t gh2t gh3t pf1t pf2t pf3t sf1 sf2 sf3 rp1 rp2 rp3 re1 re2 re3 mh1t mh2t
    mh3t vt1t vt2t vt3t bp1t bp2t bp3t ht1 ht2 ht3 ALDER kjønn sivilstand1 sivilstand2
     sivilstand3 using ImpFull1a, m(20) genmiss(mis) cmd(sf1 sf2 sf3 rp1 rp2 rp3 re1
    re2 re3 ht1 ht2 ht3: ologit)
```

```
/* MI with ice, all variables arcsine-transformed */

ice patkey gh1t gh2t gh3t pf1t pf2t pf3t sf1t sf2t sf3t rp1t rp2t rp3t re1t re2t re3t
    mh1t mh2t mh3t vt1t vt2t vt3t bp1t bp2t bp3t ht1t ht2t ht3t ALDER kjønn
    sivilstand1 sivilstand2 sivilstand3 using ImpFull1trans, m(20) genmiss(mis) cmd(
    re1t re2t re3t ht1t ht2t ht3t: regress)

/* Back-transformation */
/* Perform in Impfile.dta */
/* displayed for general health */
/* equal for pf, mh, vt and bp */

replace gh1=100*(sin(gh1t))^2 if misgh1== 1
replace gh2=100*(sin(gh2t))^2 if misgh2== 1
replace gh3=100*(sin(gh3t))^2 if misgh3== 1

/*----------MIXED MODEL----------*/
/* GH variable, equal for BP, SF and RE */

/* data to long format */
reshape long mh vt bp gh sf pf rp re ht, i(patkey) j(time)

/* Generate the variables for time^2 and for timeshift. */
ge time_kjønn = time*kjønn
egen timemean = mean(time)
ge timeshift = time - timemean
ge timeshift_kjønn = timeshift*kjønn
ge timeshift2 = timeshift^2
ge timeshift2_kjønn = timeshift2*kjønn

/*-------Random intercepts models-------*/
/* Random intercept, timeshift */
/* gh = b0 + b1*t + b2*x + b3*t*x + zeta0 + epsilon */

xtmixed gh kjønn timeshift timeshift_kjønn || patkey:, mle
estimates store rit

/* Random intercept with squared time, timeshift */
/* gh = b0 + b1*t + b2*x + b3*t*x + b4*t^2 + zeta0 + epsilon */

xtmixed gh kjønn timeshift timeshift_kjønn timeshift2 || patkey:, mle
estimates store rit2

/* Random intercept with squared time and interaction squared time, timeshift */
/* gh = b0 + b1*t + b2*x + b3*t*x + b4*t^2 + b5*t^2*x + zeta0 + epsilon */

xtmixed gh kjønn timeshift timeshift_kjønn timeshift2 timeshift2_kjønn || patkey:, mle
estimates store rit2g
/* This term is NOT significant! */

/*-------Random slopes models-------*/
/* Random slopes, timeshift */
/* gh = b0 + b1*t + b2*x + b3*t*x + zeta0 + zeta1*t + epsilon */

xtmixed gh kjønn timeshift timeshift_kjønn || patkey: timeshift, mle cov(unstructured)
estimates store rst

/* To examine the covariance matrix of random effects */
estat recovariance

/* Random slopes with squared time , timeshift */
/* gh = b0 + b1*t + b2*x + b3*t*x + b4*t^2 + zeta0 + zeta1*t + epsilon */
```

```
xtmixed gh kjønn timeshift timeshift_kjønn timeshift2  || patkey: timeshift , mle cov(
    unstructured)
estimates store rst2

/* Random slopes with squared time and interactions, timeshift */
/* gh = b0 + b1*t + b2*x + b3*t*x + b4*t^2 + zeta0 + zeta1*t + epsilon */

xtmixed gh kjønn timeshift timeshift_kjønn timeshift2 timeshift2_kjønn || patkey:
    timeshift, mle cov(unstructured)
estimates store rst2g
/* obs interaction-term timeshift2_kjønn is not significant */

/*----------------Random quadratics models----------------*/

/* Random quadratics with squared time , timeshift */
/* gh = b0 + b1*t + b2*x + b3*t*x + b4*t^2 + zeta0 + zeta1*t + epsilon */

xtmixed gh kjønn timeshift timeshift_kjønn timeshift2  || patkey: timeshift timeshift2,
     mle cov(unstructured)
estimates store rs2t2

/*----Random quadratic model without interaction----*/
/* To examine the main effects gender, time and time squared */
xtmixed gh kjønn timeshift  timeshift2  || patkey: timeshift timeshift2, mle cov(
    unstructured)

/*----------Likelihood ratio tests----------*/
/* LRTEST1 - Find from output for random intercepts w/linear model */
/* LRTEST2 - Test of random slopes w/linear model */
lrtest rit rst
/* LRTEST3 - Test of random slopes w/quadratic model */
lrtest rit2 rst2
/* LRTEST4 - Test of random quadratics w/quadratic model */
lrtest rst2 rs2t2

/*----Covariance pattern models----*/
/* Data in long format */
/* Make panel-data */
xtset patkey timeshift

set matsize 2000
xtgls re timeshift kjønn timeshift_kjønn timeshift2,  corr(ar 1) force

/*----------GEE----------*/
/* xtgee command */
/* Data in long format */

/* AR(1) covmatr */
/* Panels with gaps or less than 2 observations are omitted */
xtgee re kjønn timeshift timeshift_kjønn, i(patkey) t(timeshift) corr(ar 1) vce(robust)
matrix list e(R)
/* covmatr not equal to the observed */

/* exchangeable covmatr (compound symmetry) */
/* all panels allowed */
xtgee re kjønn timeshift timeshift_kjønn, i(patkey) t(timeshift) corr(exc) vce(robust)
matrix list e(R)
/* covmatr not similar to the observed */

/* unstructured covmatr */
/* All panels allowed */
```

```
xtgee re kjønn timeshift timeshift_kjønn, i(patkey) t(timeshift) corr(uns) vce(robust)
estat wcorrelation, format(%4.3f)
matrix list e(R)
/* covmatr quite similar to the observed */
```

R-script for EM algorithm with consequtive t-test and repeated measures ANOVA

```
library(norm)
#Read the dataset with transformed variables
x <- read.table('EMtrans.txt', header=TRUE, sep=",", na.strings=".")
n <- NCOL(x)
mat <- cbind(x[,1:n])
mat <- as.matrix(mat)

#EM imputation
s <- prelim.norm(mat)
thetahat <- em.norm(s)
#getparam.norm(s,thetahat,corr=TRUE)
rngseed(7654321)
ximp <- imp.norm(s,thetahat,mat)

#Function to backtransform HRQOL-variables
backtrans <- function(ximp){
nrimp <- NROW(ximp)
ncimp <- NCOL(ximp)
for(i in 1:nrimp){
  for(j in 1:ncimp){
    ximp[i,j] <- 100*(sin(ximp[i,j]))^2
  }
}
return(ximp)
}

#Function that performs a t-test assuming unequal variances
ttest1 <- function(vec){
matrise <- matrix(nrow=length(vec),ncol=3)
for(f in 1:(length(vec))){
  num <- vec[f]
  men <- c()
  mi <- 1
  women <- c()
  wi <- 1
  for(i in 1:534){
    if(x[i,2]==0){
      men[mi] <- as.numeric(impmat[i,num])
      mi <- mi+1
    }else{
      women[wi] <- as.numeric(impmat[i,num])
      wi <- wi+1
    }
  }
  t <- t.test(men,women)
  matrise[f,1:2] <- t$estimate
  matrise[f,3] <- t$p.value
}
  return(matrise)
}

colnames(ximp)
impmat <- backtrans(ximp)
vektor <- c(12:14,9:11,21:23,27:29)
```

```
ttest1(vektor)

#----------Repeated measures ANOVA after EM----------

ximpa <- as.data.frame(cbind(x[,1:2],impmat))
attach(ximpa)
names(ximpa)

#General health
longdatagh <- reshape(ximpa, v.names=c("GH"), varying = c("gh1t","gh2t","gh3t"),
    timevar = "time", times = 1:3, direction = "long", idvar="patkey", ids=1:NROW(
    ximpa))
longdatagh.aov <- aov(GH ~ factor(time) * factor(kjønn) + Error(factor(patkey)), data =
     longdatagh)
summary(longdatagh.aov)

y2 <- cbind(ximpa$gh1t,ximpa$gh2t,ximpa$gh3t,ximpa$kjønn)
ny <- NROW(x)
men <- c(rep(1,3))
women <- c(rep(1,3))
for(i in 1:ny){
  if(y2[i,4]==0){
    men <- rbind(men,y2[i,1:3])
  }else{
    women <- rbind(women,y2[i,1:3])}
}
ymeanmatr <- c()
for(i in 1:3){
  ymeanmatr[i] <- mean(men[,i])
  ymeanmatr[i+3] <- mean(women[,i])
}

#Profile plot, general health (GH)
xvar <- c(0,1,2)
par(mfrow=c(1,1))
plot(xvar,ymeanmatr[1:3],type="b",pch=19,lwd=2,ylim=c(40,80),col="blue",xlab=("Time"),
    ylab=("Mean of GH"),main=("Profileplot of general health, EM"))
par(new=TRUE)
plot(xvar,ymeanmatr[4:6],type="b",pch=19,lty=2,lwd=2,ylim=c(40,80),col="red",xlab=("
    Time"),ylab=("Mean of GH"),main=("Profileplot of general health, EM"))

temp=legend("bottomright", legend = c(" ", " "),col=c("blue","red"),
                text.width = strwidth("Parametervei"),
                lty = c(1,2), lwd=2, xjust = 1, yjust = 1,
                title = "Gender")
text(temp$rect$left + temp$rect$w, temp$text$y,
     c("Men", "Women"), pos=2)
```

R-script for combination by Rubin's rules after MI with consequtive t-test

```
#----------Complete-case----------
#Use the dataset complete-case for the relevant variable
y <- read.table('ghcc.txt', header=TRUE, sep=",", na.strings=".")
y2 <- cbind(y$gh1,y$gh2,y$gh3,y$kjønn)
y <- read.table('bpcc.txt', header=TRUE, sep=",", na.strings=".")
y2 <- cbind(y$bp1,y$bp2,y$bp3,y$kjønn)
y <- read.table('sfcc.txt', header=TRUE, sep=",", na.strings=".")
y2 <- cbind(y$sf1,y$sf2,y$sf3,y$kjønn)
y <- read.table('recc.txt', header=TRUE, sep=",", na.strings=".")
y2 <- cbind(y$re1,y$re2,y$re3,y$kjønn)

ny <- length(y[,1])
```

```
men <- c(rep(1,3))
women <- c(rep(1,3))

for(i in 1:ny){
  if(y2[i,4]==1){
    men <- rbind(men,y2[i,1:3])
  }else{
    women <- rbind(women,y2[i,1:3])}
}
ymeanmatr <- c()
for(i in 1:3){
  ymeanmatr[i] <- mean(men[,i])
  ymeanmatr[i+3] <- mean(women[,i])
}

#Profile plot
xvar <- c(0,1,2)
par(mfrow=c(1,1))
plot(xvar,ymeanmatr[1:3],type="b",pch=19,lwd=2,ylim=c(65,90),col="blue",xlab=("Time"),
    ylab=("Mean of SF"),main=("Profileplot of social functioning, CC"))
par(new=TRUE)
plot(xvar,ymeanmatr[4:6],type="b",pch=19,lty=2,lwd=2,ylim=c(65,90),col="red",xlab=("
    Time"),ylab=("Mean of SF"),main=("Profileplot of social functioning, CC"))

temp=legend("bottomright", legend = c(" ", " "),col=c("blue","red"),
               text.width = strwidth("Parametervei"),
               lty = c(1,2), lwd=2, xjust = 1, yjust = 1,
               title = "Gender")
text(temp$rect$left + temp$rect$w, temp$text$y,
     c("Men", "Women"), pos=2)

#-----Plot of the estimates from m imputed datasets-----
#General health
x <- read.table('GHinteractionEstMIM.txt',header=FALSE)

pdf("dotplotGH.pdf")
par(mfrow=c(2,1))
stripchart(x[,1],xlab="Estimated parameter for gender by time interaction",main="
    Dotplot for general health (GH)")
stripchart(x[,2],xlab="Estimated standard deviation for gender by time interaction")
dev.off()

#----------Imputed datasets----------
#Read the dataset from Stata
x <- read.table('HRQOL_imp_long2.txt', header=TRUE, sep=",", na.strings=".")

#Function to partition the file in datasets
impnr <- function(y){
  y <- cbind(y[,1:12],y[,32:33],y[,59:62])
  y <- split(y,y$impnr)
  return(y)
}

#Function to partition the subjects with respect to gender
kjonn <- function(y,gender){
  ykjonn <- split(y,y$KJONN)
  if(gender==1){
    return(ykjonn$'1')
  }else{
    return(ykjonn$'2')
  }
}
```

```
#Function to partition each dataset in occasions
tid <- function(n,tid,kjonn){
  if(kjonn==1){
    string <- paste("mann",n,sep="")
  }else{
    string <- paste("kvinne",n,sep="")
  }
  y <- get(string)
  ytid <- split(y,y$time)
  if(tid==1){
    return(ytid$'1')
  }else if(tid==2){
    return(ytid$'2')
  }else{
    return(ytid$'3')
  }
}

#Function to generate the means for men and women
#Here: Role emotional (RE)
meanmatrix <- function(){
  matr <- matrix(ncol=40,nrow=6)
  for (i in 1:20){
    for (j in 1:3){
      midl1 <- tid(i,j,1)
      matr[j,i] <- mean(midl1$re)
      matr[j+3,i] <- var(midl1$re)
      midl2 <- tid(i,j,2)
      matr[j,i+20] <- mean(midl2$re)
      matr[j+3,i+20] <- var(midl2$re)
    }
  }
  return(matr)
}

#Applying the functions to the dataset
y <- impnr(x)

y0 <- y$'0'
y1 <- y$'1'
y2 <- y$'2'
#....
y20 <- y$'20'

mann0 <- kjonn(y0,1)
mann1 <- kjonn(y1,1)
mann2 <- kjonn(y2,1)
#...
mann20 <- kjonn(y20,1)

kvinne0 <- kjonn(y0,2)
kvinne1 <- kjonn(y1,2)
kvinne2 <- kjonn(y2,2)
#...
kvinne20 <- kjonn(y20,2)

matr <- meanmatrix()
nm <- length(mann0[,1])/3
nw <- length(kvinne0[,1])/3

#Establish all vectors and matrices
```

```
m <- 20
mannmh <- matr[,1:m]
kvinnemh <- matr[,(m+1):(2*m)]
diffmatr <- mannmh-kvinnemh
diffmean <- c()
diffvar <- c()
diffvarbetw <- c()
meanmatr <- c()
meanmatrvar <- c()
for(i in 1:3){
  meanmatr[i] <- mean(matr[i,1:20])
  meanmatrvar[i] <- mean(matr[i+3,1:20])
  meanmatr[i+3] <- mean(matr[i,21:40])
  meanmatrvar[i+3] <- mean(matr[i+3,21:40])
  diffmean[i] <- mean(diffmatr[i,])
  diffvar[i] <- (mean(mannmh[i+3,])/nm + mean(kvinnemh[i+3,])/nw)
  diffvarbetw[i] <- var(diffmatr[i,])
}

#Calculating T statistic
diffsd <- sqrt(diffvar + (1+1/m)*diffvarbetw)
#Calculating degrees of freedom
dof <- (20-1)*((1+diffvar/((1+1/20)*diffvarbetw))^2)

#This may be approximated to the standard normal distribution N(0,1)
N <- diffmean/diffsd

#Results
meanmatr
diffmean
diffsd^2
dnorm(N)
dt(N, dof)

#Profile plots
xvar <- c(0,1,2)
par(mfrow=c(1,1))
plot(xvar,meanmatr[1:3],type="b",pch=19,lwd=2,ylim=c(40,80),col="blue",xlab=("Time"),
     ylab=("Mean of RE"),main=("Profileplot of role emotional, MI"))
par(new=TRUE)
plot(xvar,meanmatr[4:6],type="b",pch=19,lty=2,lwd=2,ylim=c(40,80),col="red",xlab=("Time
     "),ylab=("Mean of RE"),main=("Profileplot of role emotional, MI"))

temp=legend("bottomright", legend = c(" ", " "),col=c("blue","red"),
                text.width = strwidth("Parametervei"),
                lty = c(1,2), lwd=2, xjust = 1, yjust = 1,
                title = "Gender")
text(temp$rect$left + temp$rect$w, temp$text$y,
     c("Men", "Women"), pos=2)
```

5	3	4	6	7	8	9	1	2
6	7	2	1	9	5	3	4	8
1	9	8	3	4	2	5	6	7
8	5	9	7	6	1	4	2	3
4	2	6	8	5	3	7	9	1
7	1	3	9	2	4	8	5	6
9	6	1	5	3	7	2	8	4
2	8	7	4	1	9	6	3	5
3	4	5	2	8	6	1	7	9

Figure B.1: Solution of the front side Sudoku